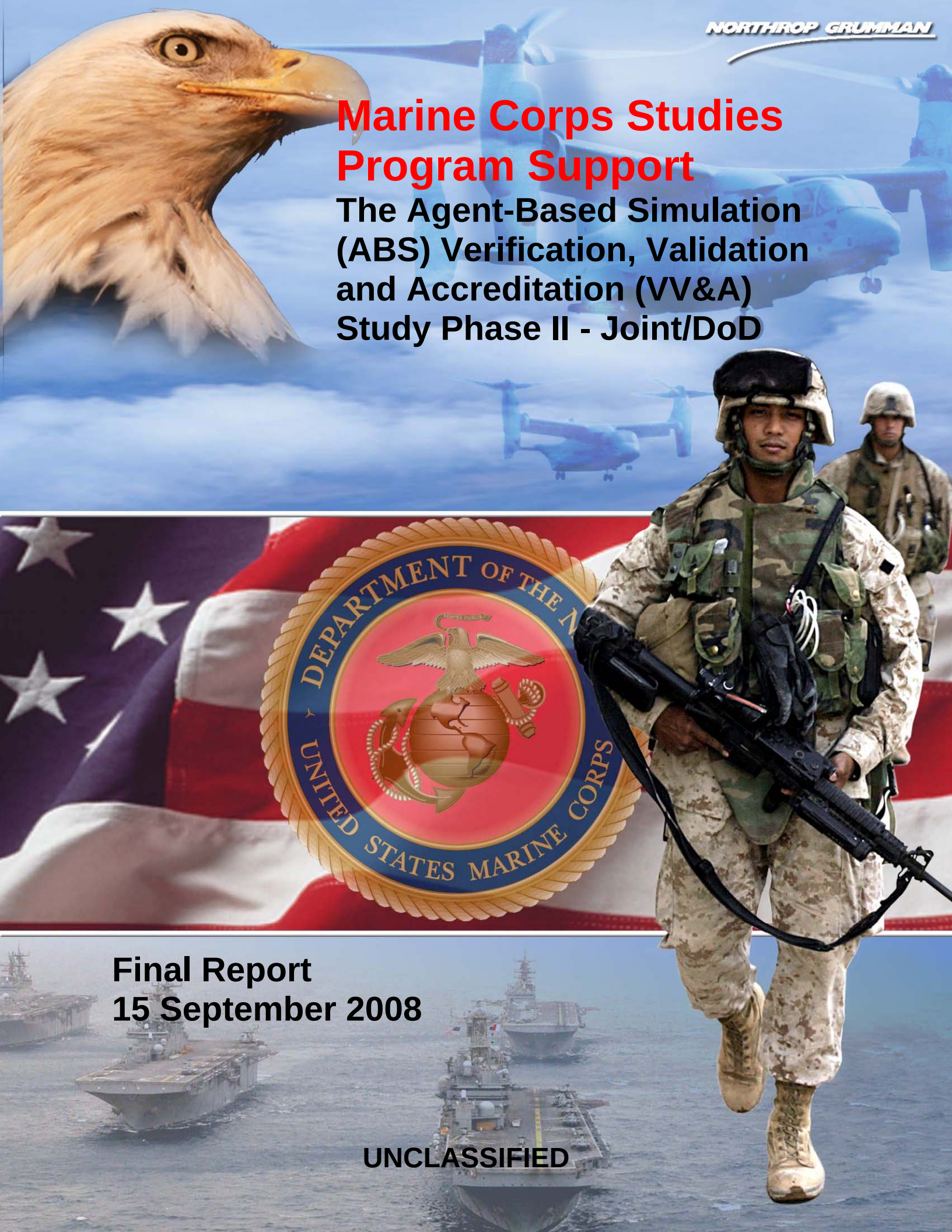

NORTHROP GRUMMAN
Marine Corps Studies
Program Support
The Agent-Based Simulation
(ABS) Verification, Validation
and Accreditation (VV&A)
Study Phase II - Joint/DoD
DEPARTMENT OF THE N
UNITED STATES MARINE CORPS
Final Report
15 September 2008
UNCLASSIFIED

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
14 MAR 2012	**Final**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Agent Based Model Validation -- Final report	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
Michael Bailey	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
USMC MCCDC OAD	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS
USMC Analysis Federation

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**UU**	**238**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

MARINE CORPS STUDIES PROGRAM SUPPORT

Final Report

THE AGENT-BASED SIMULATION (ABS) VERIFICATION, VALIDATION, AND ACCREDITATION (VV&A) STUDY PHASE II - JOINT/DoD

Contract #M00264-06-D-0001

Delivery Order #007

Prepared for:
Operations Analysis Division
Marine Corps Combat Development Command
3300 Russell Road
Quantico, VA 22134-5001

Prepared by:
Northrop Grumman Mission Systems,
Reston, VA 20191

15 September 2008

Points of Contact:

Mr. Mitch Youngs

Program Manager

Technical Leads

Dr. Eric Weisel

Ms. Lisa Jean Moya

ABSTRACT

This is the final report for the Agent Based Simulation (ABS) Verification, Validation, and Accreditation (VV&A) Framework Study Phase II. It reports on the tasks associated with this ABS VV&A (ABSVal) Framework Study project with an emphasis on insights gained in the application of the general ABSVal Framework developed in support of analysis applications.

The project consisted of four main tasks:

Task 1 – Create Measures of Validity

Task 2 – VV&A Process Application

Task 3 – Test Case Workshops

Task 4 – Publishable Documentation and Reports

The bulk of the effort was placed on Task 2. The development of measures of validity proceeded as part of the testing of the ABSVal Framework in specific application domains. The results of the overall effort were presented to the community at large through workshops and publications presented at various conference venues. Most of the material from this project resides at http://orsagouge.pbwiki.com/ABSVal.

The key insight gained by this project was that the validation of models in support of analysis resides within the analysis process itself. That is, validation cannot be decoupled from the analysis plan, process, and results. Validation in this intended use identifies the limitations and boundaries of the analysis itself, lending clarity to that process for the recipients of the simulation and analytical results.

This project resulted in sample validation reports that were independently critiqued, publications to the larger modeling and simulation community, and guidance for what ought to be included in a validation effort in support of analysis applications.

Table of Contents

List of Figures

EXECUTIVE SUMMARY

This Executive Summary provides a synopsis of the work performed for and results achieved in the U.S. Marine Corps (USMC) Agent-Based Simulation Verification, Validation, and Accreditation (VV&A) Study Phase II.

ES.0 OBJECTIVE

The Marine Corps Combat Development Command (MCCDC) Operations Analysis Division (OAD) intends to develop a credible and analytically valuable model of Irregular Warfare (IW). To achieve this goal, the model must demonstrate that it has been built on sound principles and that it can be used, at a minimum, to compare various courses of action (COAs), providing the relative merit of each. The ability to predict the possible outcomes of a COA would make the model that much more valuable. Therefore, MCCDC OAD needs a framework for verifying, validating, and accrediting models such as this.

The objective of this effort was to apply the framework developed in Phase I of this Study for VV&A of Agent Based Simulations (known as ABSVal) to at least two candidate model applications being considered for future entry into the USMC Irregular Warfare Analytic Baseline. The goals of Phase II of the Study included: 1) testing the viability and utility of ABSVal in a realistic institutional setting; 2) evaluating ABSVal in a seminar setting combining communities of ABS users and developers; 3) developing methodologies for applying ABSVal to future ABS development efforts; and 4) producing information products useful for the M&S community.

This effort concentrated on the validation of those portions of the selected ABS that are not physics-based. The goal was to maintain the analytic rigor of the traditional VV&A process, while expanding it to cover non-traditional topics (e.g., population dynamics and cultural shifts). The benefit is a demonstration of the use of ABSVal, an assessment of its utility and applicability to future ABS VV&A efforts, and an assessment of the validity of the two selected ABS applications. While the acronym VV&A covers three separate but highly interrelated processes, the primary thrust of this effort was on the validation process. The verification and accreditation processes were addressed with regard to their interdependencies with the validation process. The application of ABSVal provided a means for assessing the reliability, applicability, and feasibility of the ABS for its intended use.

The larger modeling and simulation community has gained access to these validation exercises and insights through three workshops hosted by Northrop Grumman and publications developed by the study team members and affiliated participants. Materials from this project and other related items are broadly accessible on a website hosted by the study sponsor, including publications, briefings, and reports: http://orsagouge.pbwiki.com/ABSVal.

ES.1 TASKS

The VV&A Phase II Task Order consisted of four main tasks:

Task 1 – Create Measures of Validity

Task 2 – VV&A Process Application

Task 3 – Test Case Workshops

Task 4 – Publishable Documentation and Reports

The bulk of the effort was placed on Task 2. The development of measures of validity proceeded as part of the testing of the ABSVal Framework in specific application domains. The results of the overall effort were presented to the community at large through workshops and publications presented at various conference venues.

ES.2 VALIDATION PROCESS

The process needed the following characteristics:

1) **Transparency** – to provide an understanding of the assumptions, decisions, and activities that went into VV&A
2) **Traceability** – to ensure the flow of activities and actions is logical and that appropriate referents for those activities can be located and consulted
3) **Reproducibility** – to provide for the event that the same model/data/users will be applied to a similar effort in the future
4) **Communicability** – to produce sufficient, understandable documentation so the effort can be independently duplicated, and so the consumer can make an informed, and perhaps qualified, decision

Other objectives included the ability of the process to do the following:

1) Describe the bounds of use for the specified purpose
2) Communicate the risk of use for the specified purpose

The key insight gained by this project was that the validation of models in support of analysis resides within the analysis process itself. That is, validation cannot be decoupled from the analysis plan, process, and results. Validation in this intended use identifies the limitations and boundaries of the analysis itself, lending clarity to that process for the recipients of the simulation and analytical results.

ES.3 TESTING THE FRAMEWORK

The main purpose of applying the ABSVal Framework to selected applications was primarily to test the framework developed in Phase I of this Study. This testing effort included three validation efforts. Each of these efforts also had an audit of the validation conducted to elicit further insights for the framework. The first was a validation of the Conceptual Model of the Pythagoras-COIN (P-COIN) model using assumptions testing. The second was an application of the full framework to an analysis application of the P-COIN model. The third was a validation of an Obstacle Clearing Model implemented in Pythagoras (P-OCM). The validation reports created in

these efforts not only provide the validation analysis, they also serve as a model for other future reports that may be based on this framework. The reports also provided insights into framework improvements and general ABS Validation methodology. Therefore, the purpose of these validation application efforts was to exercise the framework in order to gain insights and improve the validation process.

ES.4 CHALLENGES

There were several challenges in testing the ABSVal Framework. A major problem was finding ABS models to which the team could have access for the purposes of validating their use. Although much effort was expended to find a variety of ABS applications, the team was often unable to secure access to the model and the analysis use. Ultimately, the team applied the ABSVal Framework to two applications both developed in Pythagoras. Although the team had access to the developer in both instances, for differing reasons in each case, the validation exercise was limited by the available documentation. Further, neither of these applications had significant amounts of dynamic emergent behavior, which was initially desired for the Study. Lastly, both of these studies were low level of effort, small scope simulation applications, which inherently had an effect on the overall analysis supported by the chosen models.

A particular challenge in completing the validation effort for P-COIN was that the methodology and framework were in development concurrent with the P-COIN development and analysis. Additionally, significant insights into the framework occurred midway through the validation effort causing a significant redirection of effort and an alignment between two separate efforts: the P-COIN analysis and development and the P-COIN validation as a test of the ABS VV&A Framework. This concurrence and alignment, as well as the mid-course insight, has led to some areas having less reporting than others, as well as work completed early in the validation process not necessarily utilized as part of the final validation effort.

ES.5 AUDIT RECOMMENDATIONS

Following the validation activities, an independent audit was conducted of the work. The following summarizes recommendations found in the audit report:

(1) Validation reports should clearly describe the qualitative questions investigated and tests conducted. These descriptions should include the reasoning for the investigation and tests, as well as expected and desirable results. It should also state necessary conditions for “passing” the tests for the intended use (i.e., validation criteria) and conditions that would constitute a “failure.”
(2) The process of making qualitative assessments would be more objective with less reliance on the subjective opinion of the validator if the process were more structured, expectations were identified, and if the reasoning for the validator’s conclusions were made explicit.
(3) The significance of the results of all validity assessments (qualitative and quantitative) on the intended use and application of the model needs explicit explanation. That is, the effect of failing the test on the intended use needs to be

evident. Along these lines, the actual recommendation of whether or not to use the model needs to be stated clearly.

(4) The report should identify mitigation strategies: in use (e.g., sensitivity analysis), in corrective development, or in model improvement.

(5) The framework should include the minimum sets of information and materials sufficient to reach a final validity determination using the ABSVal Framework.

(6) Due to the brainstorming nature of this task, there should be two or more people involved in the extraction of the implicit assumption of the Conceptual Model.

(7) A practitioner's guide to validation resulting from the ABSVal Framework should include rules, stepwise approaches, branches, and criteria for determining necessary or sufficient information to support a given conclusion. Rules might include necessary information and model and simulation artifacts to embark on a validation investigation. The availability of material might influence the risk of use, the risk of use being stratified at different acceptable levels.

ES.6 DEVELOPER'S POINT OF VIEW

The following is a statement provided by Mr. Edmund Bitinas, developer of the two Pythagoras ABS models on which the ABSVal framework was applied during the Study as a critique of the results of applying the framework:

"None of the reports developed, audit or validation, helped the developer to understand necessary improvements to the model: in the Conceptual Model or the implementation, nor do the reports give the developer the sense of whether the model development is on the right track. From the reports, the developer cannot assess the salvageable elements of the model, if any, should a validator find deficiencies. Further, none of the reports accounted for the maturity of the application to which the models were intended. Requiring precision when there is none is as problematic as not achieving precision when it is required. This problem is explicitly identified in the VV&A Recommended Practices Guide (RPG) but is sometimes forgotten in a zeal for "accuracy with respect to the real world." Representative accuracy may be more applicable (e.g., equivalence) when the data available does not support a more definitive approach. Sensitivity analysis can mitigate data difficulties. Identifying representative accuracy requirements can mitigate the desire for data matching precision (e.g., magnitude and direction of change). Finally, the validation cannot be decoupled from the analysis. The VV&A RPG notes that without validation criteria (i.e., what must be met for the model to meet intended use) validation becomes merely an assessment of capabilities."

ES.7 FRAMEWORK

Phase I of the Study and early stages of Phase II looked at the fundamental theory that surrounds the validation of simulations. There are problems of tractability in generating a mathematical proof that shows that a simulation is valid. A key finding from this development of the underlying theory was that if one could not answer a posed analytical question using a perfect representation of the system, then one could not answer that question with any simulation representation. Thus, other techniques of validation must be employed.

Because of the inability to comprehensively validate a human system, the scientific method was employed in the framework as a method in which validation techniques are applied to the simulation in an attempt to disprove or “poke holes” in the integrity of the simulation with respect to the application. With rigorous validation techniques, the inability to invalidate the simulation subsequently provides evidence of its validity and usefulness toward an analysis application.

In the framework, the intended use of the simulation defined by the analytical application drives the abstraction process to determine what details and elements are critical for the simulation. Frequently, this entails establishing assumptions, both in the model and surrounding the analysis. The impact of these assumptions on results and conclusions must be assessed or tested. Results validation is often used to provide evidence of the validity of a simulation. Often for social systems, the referent is not solid; however results validation can still be employed (although much more subjectively) to assess if simulation output is sensible and useful. Intended use drives validation criteria in this assessment.

ES.8 CONCLUSIONS

Conveying the risk of using a model and its simulation results is the primary purpose of validation. However, the determination of risk and its communication to the users of models and the consumers of their simulation results is a critical area in the validation process that still requires additional work. Risk is a combination of the process used to evaluate validity, error in the model and its simulation, and the consequence of using the model. Errors in the core elements of the model or resulting from the absence of needed model interactions that cannot be otherwise mitigated may indicate a higher risk level. This study identified important characteristics of risk and utilized preliminary metrics in its evaluations. Additional research into the development of risk assessment methodologies in support of validation assessments is needed.

Critical to the validation of a model and its simulation results in an analysis application lies in understanding the decision context for the analysis. This includes the core elements for supporting the analytical decisions as well as the decision’s relationship to the influencing elements within the modeling environment. The validation process may uncover dynamic elements that need to be addressed more completely, mitigation techniques that ought to be taken or additional test cases that should be included within the analysis. Further, it is not enough to identify limitations in the model. These limitations must be explicitly linked to the analysis context and discuss the risk to using the model within that context.

1 STUDY INTRODUCTION

The Marine Corps Combat Development Command (MCCDC) Operations Analysis Division (OAD) intends to develop a credible and analytically valuable model of Irregular Warfare (IW). To achieve this goal, the model must be demonstrated to have been built on sound principles and that it can be used, at a minimum, to compare various courses of action (COAs), providing the relative merit of each. The ability to predict the possible outcomes of a COA would make the model that much more valuable. Therefore, MCCDC OAD needs a framework for verifying, validating, and accrediting models such as this.

Traditional physics and probability based models undergo a verification, validation, and accreditation (VV&A) process to obtain general acceptance within the user and modeling community. While there are formal definitions and procedures for each of these three functions, this study expanded those definitions and procedures to cover areas not previously addressed. Specifically, this effort addressed agent-based simulations (ABS) and IW. These subject areas are relatively new to the modeling community and involve phenomena that are somewhat difficult to apply within simulations (such as population dynamics and social and cultural factors). To expand the VV&A process to include ABS and IW features, the ABS VV&A Framework Study addressed three central questions:

- Do the model implementation and its associated data accurately represent the developer's conceptual description and specifications (verification)?
- What is the degree to which a model and its associated data accurately represent the real world from the perspective of the intended uses of the model (validation)?
- Is the scope of the model sufficiently applicable to one or more of the objectives of the study so that the model can be officially certified as acceptable for use for the specific purpose at hand (accreditation)?

The Department of Defense (DoD) procedures for accomplishing VV&A for traditional models are well known and documented. These procedures provided a starting point for this Study.

ABS, however, involves an additional modeling property in that they create software entities with the capability of choice, which may lead to unexpected or emergent behavior of the agents and of the system as a whole. Although many legacy simulations exhibit some capability for entities to make choices, ABSs are based on this capability. Moreover, the choices available to these entities may depend on many factors, such as an agent's perception of its operating environment, outside influences, and/or changing objectives. This capability sometimes results in unpredictable outcomes and, therefore, could cause users to view the model as unreliable. The challenge, therefore, was to devise a way to validate the unpredictable.

At its onset, this study envisioned focusing on the concept of validity, since ABS verification likely would be similar to the process used for legacy software; and accreditation is simply an agreement between analysts and the study sponsor that a particular model is useful for the problem being studied.

Phase I of the ABS Verification, Validation, and Accreditation Framework Study evaluated issues with validating ABS in the context of simulations representing to some degree social systems and human decision-making. The Study Team found that there are several characteristics inherent to social systems that make validation of these simulations challenging.

In the modeling of social systems, there may be little or no data against which to compare model results. Even when there are data, the available data may be biased or dated. The modeling of human decision making also has its share of difficulties:

1) Humans may have more information than Agents
2) Humans may include emotions and experience
3) Humans may think/plan ahead
4) Humans may anticipate the actions of others
5) A single human in precisely the same circumstances at two different points in time may make different decisions
6) Two humans, given the same information, may make different decisions

These and other factors make a heavy reliance on substantiating simulation results against a known referent and empirical data for validation difficult to apply and interpret. To combat these difficulties, the Study Team proposed applying the Modern Scientific Method to validation. In this approach, each validation experiment applied to the ABS would seek to invalidate the model within the context of its intended application. Then, each failure to invalidate the model provides supporting evidence of the model's validity for that particular application. The interpretation of these experimental results is unique to that specific application of the model. Although aspects of the effort may be applicable across numerous applications, each application of the model to a new situation must have its validation revisited. As people and society change and their circumstances change, so do the decisions they make.

An intrinsic element to this approach is that this process must be traceable, repeatable, transparent, and communicated to the consumer to allow the making of an appropriate decision on the use of the ABS. In order to apply the scientific method, validation experiments need to describe what would constitute an accurate result prior to their commencing. The users or consumers of the model's results must determine the degree of risk that they are willing to take in accepting the results of the ABS. The validation report must communicate some measure of risk to the user for using the model in the application areas tested. This degree of risk can be determined through the application of the validation experiments and their outcome.

Phase II of the study applied the framework developed in Phase I of the ABS VV&A Framework Study to actual ABS applications to both evaluate these applications in a

specified problem domain and to stress the framework to improve and test its usefulness to the consumers of ABS.

1.1 STUDY OBJECTIVE

The objective of this effort was to apply the Phase I developed framework for VV&A of Agent Based Simulations (known as ABSVal) to at least two candidate model applications being considered for future entry into the USMC Irregular Warfare Analytic Baseline. The goals of Phase II included: 1) testing the viability and utility of ABSVal in a realistic institutional setting; 2) evaluating ABSVal in a seminar setting combining communities of ABS users and developers; 3) developing methodologies for applying ABSVal to future ABS development efforts; and 4) producing information products useful for the Modeling & Simulation (M&S) community.

1.2 SCOPE OF STUDY

This effort concentrated on the validation of those portions of the selected ABS that are not physics-based. The goal was to maintain the analytic rigor of the traditional VV&A process, while expanding it to cover non-traditional topics (e.g., population dynamics and cultural shifts). The benefit would be a demonstration of the use of ABSVal, an assessment of its utility and applicability to future ABS VV&A efforts, and an assessment of the validity of the two selected ABS applications. While the acronym VV&A covers three separate but highly interrelated processes, the primary thrust of this effort was on the validation process. The verification and accreditation processes were addressed with regard to their interdependencies with the validation process. The application of ABSVal provided a means for assessing the reliability, applicability, and feasibility of the ABS for its intended use, in a quantifiable way.

Northrop Grumman assumed the following:

1) The Marine Corps mission, as prescribed in the National Security Act of 1947 (amended), will change to include more mission areas in Irregular Warfare environments. As such, analytic modeling and simulation of these missions and environments will increase in importance to the Marine Corps.
2) The Marine Corps Study Sponsor would facilitate access to models, data, information, and Points of Contact (POCs) of interest to this study.

1.3 TWO ASPECTS OF STUDY: ABS VALIDATION AND IW ANALYSIS

This effort had two distinct, but related, parts. One aspect of this effort was the issue of ABS validation; the second was the issue of a model's applicability to IW analysis.

Most ABS systems are not explicit models of anything in particular. Rather, they are toolboxes allowing the construction of scenarios that represent a particular situation. Pythagoras and the Map Aware Non-Uniform Automata (MANA) are two examples of ABS models that are currently in use. In both cases, the scenarios exercised in the model are constructed via data, and it is that data, including both behavioral and

physical attributes, that give meaning to the representation of the agents as being Marines, tanks, or other objects.

The other aspect of this effort was the domain of irregular warfare (IW). Like traditional warfare, IW analysis ranges from the tactical level (e.g., countering improvised explosive devices) through mission level (e.g., border security) to campaign level (e.g., nation building). It is envisioned that models will be identified or built that cover one or more levels of IW. It was an objective of the Study that this framework must be able to address the VV&A of the various elements of IW as well.

1.4 GENERAL STUDY APPROACH

This effort continued to involve close and frequent informal collaboration between OAD representatives and the Study Team (including members from industry, Government, and academia). The general approach for applying this framework was to first identify specific applications of ABS suitable for the framework and this level of effort; next, the framework was applied to the two selected applications. The effort concentrated on validation. An extensive code walk-through and other software V&V activities were not envisioned.

This Study approach also involved three Government hosted workshops. These workshops presented the framework and the results of the application of the framework to the community. The results included not only the results of the validation effort, but also a critique of the framework itself with recommendations for improvements.

1.4.1 ABS VV&A (ABSVal) Application Pair Selection

Following are several factors taken into consideration during the selection process to determine the suitable ABS for exercising the framework:

1.4.1.1 Sponsorship

Sponsorship of the Modeling Environment or the ABS itself was perceived to lend some credibility to the effort. It was thought that sponsorship would also facilitate the coordination of the necessary formalities of permission that the Team needed to have in order to exercise the ABSVal framework and publish the findings.

1.4.1.2 Analysis and Decision Making

If the ABS had been used for some real-world analysis and/or decision making, it was thought the results may have more tangible correlations to real-world operations. The fact that an ABS had been used for real-world analysis and/or decision making also indicated that the simulation was suitably robust and worthwhile to employ the VV&A resources on this task.

1.4.1.3 Artifact Availability and Access to the ABS Developer

Artifact availability and access to the ABS developer influences the level of rigor that can be achieved when applying the ABSVal Framework. Depending on the level of formality of the development of a particular ABS, certain development artifacts may or

may not be available to review during the validation process. Access to the ABS developer was seen as critical should artifact availability be scarce, to answer questions, and facilitate the process.

1.4.2 Study Tasks

The objectives of this effort were accomplished through the completion of the following four tasks, each discussed in its own section of this report.

Task	Section
Task 1 – Create Measures of Validity	3
Task 2 – VV&A Process Application	2
Task 3 – Test Case Workshops	4
Task 4 – Publishable Documentation and Reports	5

The main purpose of applying the ABSVal Framework to selected applications was primarily to test the framework developed in Phase I of this Study (Task 2). This testing included three validation efforts. Each of these efforts also had an audit of the validation conducted to elicit further insights for the framework. The validation reports created in these efforts not only provided the validation analysis, they also serve as examples for future reports that may be based on this framework. The reports also provided insights into framework improvements and general ABS Validation methodology.

The identification of the risk of using a simulation for a specified analysis problem was an integral part of communicating to the decision maker the limits of valid use for that simulation (Task 1). This included discussing the ability of the simulation to meet identified validation criteria, represent all necessary system elements at the needed levels of fidelity, and achieve accuracy and precision requirements. The Study Team approached identification of risk of use for a simulation in two ways. The first was theoretical, embodied by a review of the literature, which revealed the Validation Process Maturity Model. The second was practical, through the assessment of risk for the framework applications and assessments of the process maturity through the use of a survey in the Phase II Workshop #4.

As a research project, sharing and communicating with the broader analysis, validation, and ABS communities was a large part of this study. To support this goal, workshops have been held (Task 3). In addition, early in the project, an initial Publication/Media Plan was developed (Task 4).

The tasks for Phase II of the ABSVal Framework Study supported the larger goals for both the research project, in general, and to the larger validation community, in particular. Phase I of the project identified the following goals:

1) Establish processes for understanding how valid is “valid enough” for the intended application

2) Determine techniques for uncovering invalid models, since validation itself may not be universally possible due to the lack of a referent or the existence of a large body of referents that do not agree
3) Establish the boundaries of validation, which may limit applicability of the model/data to only a portion of the overall intended use
4) Establish a framework process that is not resource-intensive and can be accomplished with a small fraction of the resources devoted to the overall application

Further, the process needed the following characteristics:

1) **Transparent** – to provide an understanding of the assumptions, decisions, and activities that went into V&V
2) **Traceable** – to ensure the flow of activities and actions is logical and that appropriate referents for those activities can be located and consulted
3) **Reproducible** – to provide for the event that the same model/data/users will be applied to a similar effort in the future
4) **Communicable** – to produce sufficient, understandable documentation so the effort can be independently duplicated, and so the consumer can make an informed, and perhaps qualified, decision

Other objectives included the ability of the process to do the following:

1) Describe the bounds of use for the specified purpose
2) Communicate the risk of use for the specified purpose

Task 2 activities supported the assessment of the Framework's satisfaction of these goals.

2 VV&A PROCESS APPLICATION: TESTING THE ABSVAL FRAMEWORK (TASK 2)

The main purpose of applying the ABSVal Framework to selected applications was to test the framework developed in Phase I of this study. This testing included three validation efforts. Each of these efforts also had an audit of the validation conducted to elicit further insights for the framework. The first was a validation of the Conceptual Model of Pythagoras-COIN (P-COIN) using assumptions testing. The second was an application of the full framework to an analysis application of P-COIN. The third was a validation of an Obstacle Clearing Model implemented in Pythagoras (P-OCM). The validation reports created in these efforts not only provided the validation analysis, they also serve as examples for future reports that may be based on this framework. The reports also provided insights into framework improvements and general ABS Validation methodology.

2.1 VALIDATION APPLICATION EFFORTS OVERVIEW

This section gives an overview of the validation applications used to test the framework. Appendices to this final report contain the details for these validation activities in their respective validation reports. The website, http://orsagouge.pbwiki.com/ABSVal, contains briefings which provide an overview of the validation activities and conclusions.

2.1.1 Pythagoras Obstacle Clearing Model (P-OCM)

This validation effort assessed an Obstacle Clearance Model developed by Northrop Grumman Mission Systems in Pythagoras for the Marine Corps Warfighting Laboratory (MCWL) and the Office of Naval Research (ONR). The question being addressed was whether air-dropped bombs could be effective counter-obstacle devices in the surf and beach zones of an amphibious assault. The U.S. Air Force had conducted extensive static tests of alternative Mk 80-series bombs against mines and other obstacles in a large "pond" constructed at Eglin AFB. P-OCM was developed to apply the Air Force data to model and evaluate different tactics, techniques, and procedures for the use of guided and unguided bombs for obstacle clearance in the Surf Zone.

In P-OCM, everything was modeled as an agent (e.g., obstacles, bombers, aimpoints, hitpoints, and AAVs). The bombs themselves were not modeled explicitly; rather, the type of bomb, type of mine, and depth of water were used along with the Air Force damage data to determine the radius of the “lethal area” around the associated hitpoint. Mines that were outside the lethal area were assumed not to be destroyed and were subject to displacement. Mines that were inside the lethal area were assumed destroyed.

The basic concept of P-OCM was quite simple and was patterned after the results of the Air Force tests:

(1) Mines could be destroyed or displaced
(2) Hedgehogs could only be displaced
(3) Triple-Strand Concertina (TSC) would not be affected by bombs

(4) All displacements would be directly away from the point of detonation (the hitpoint)
(5) Magnitudes of displacement would be determined from an internal look-up table constructed from the actual Air Force displacement data (with necessary interpolations)

After a bombing run had destroyed or displaced some number of the obstacles, the AAVs would attempt to use the just-created assault lane to reach their Littoral Penetration Point (LPP). The AAVs would be subject to navigational errors and might have to maneuver off their intended course to avoid damaged or sunken AAVs or TSC. If they did have to maneuver, they would again set course directly for the LPP as soon as they were again clear. (Note that a succession of maneuvers could cause some AAVs to be outside their planned assault lane. The problem could be exacerbated by navigational errors).

2.1.1.1 P-OCM Study Questions

The study had four tactical questions:

1) What accuracy is best?
2) Do simultaneous or sequential detonations play a role in developing a better "lane?"
3) Is there significant difference between using precision bombs versus conventional bombs?
4) How many bombs should be dropped at each aim point?

2.1.1.2 P-OCM Measures of Effectiveness

Most counter-obstacle studies use as their measure of effectiveness the percentage of the lane that is cleared or the removal of a stated percentage of mines and obstacles (often stated as the confidence level of having cleared a threshold percentage of mines and obstacles). In this study, the measure of effectiveness used was the number of Amphibious Assault Vehicles (AAVs) surviving the attempted transit of the assault lane after bombs had been dropped to clear it (although all the reports actually were stated in terms of the numbers of AAVs killed). The study focused entirely on the Surf Zone.

2.1.1.3 P-OCM Conceptual Model Validation Findings

Four things stood out within the Conceptual Model as potential problems, with two of those as validity issues:

1) Choice of MOE. The selected MOE unnecessarily introduced too many new and difficult variables into the basic problem. The most difficult were AAV navigational errors and AAV maneuver doctrine, either of which could have a huge impact on reported results and could have suppressed otherwise useful information related to the assigned analytic questions.
2) Assumption of knowing the location of mines and obstacle belts. This assumption is highly suspect in the real world in the first place. It also injected a

great deal of risk into the analysis. Finally, it enabled the use of a counter-obstacle tactic that is extremely risky.

3) The counter-obstacle tactic used injected a high level of risk, specifically, attempting to place the "target boxes" directly on the mine/obstacle belts, leaving gaps between the belts. This tactic alone could be responsible for one set of counter-intuitive results.
4) The Lethal Area Radius approach used to determine damage to mines by bombs consistently over or under estimated the damage, even when done carefully.

None of the above, however, actually would falsify the hypothesis that the model was valid for this application. Therefore, none of the above serves to invalidate P-OCM.

2.1.1.4 P-OCM Results Validation Findings

Two extremely counter-intuitive sets of results were reported out of the P-OCM study. The first set indicated that less precision was preferable to greater precision with respect to the accuracy of the hitpoints on the aimpoints. The second set presented the even more unlikely proposition that fewer bombs can be more effective than more bombs.

Unfortunately, nothing in the assumption testing of the Conceptual Model gave even a hint of anything that could cause the results indicated above. The problem could have been with the dataset, or with the P-OCM instantiation, or with Pythagoras itself. But there was a problem, and it appeared to be serious.

Until the causes of the extremely counter-intuitive results are found and corrected, or at least are satisfactorily explained, P-OCM was considered invalid for analysis.

That finding was reinforced by an application of the notional risk assessment tool developed during the ABSVal study. Using that tool, it was assessed that the risk of using the P-OCM model in its current form for the study application as Very High to Extreme.

The "Conclusions and Recommendations" section below indicates a significant change that could be made to the P-OCM model and the MOE in any future study of the same problem. It is believed that those changes would yield a far more favorable risk rating.

2.1.1.5 P-OCM Validation Conclusions and Recommendations

P-OCM was considered invalid for analytic applications due to the near-impossibility of some of its results. It was strongly recommended for any future work on the problem addressed by P-OCM that the "AAVs surviving" MOE be discarded in favor of MOEs that directly measure the effectiveness of clearance TTPs on mine and obstacle density.

However, it was noted that there is the strong possibility that the obstacle clearance model within P-OCM may in fact be fully valid for this application (valid in the sense that it would not be falsified if allowed to operate separately). The problem that showed up in results validation could have been entirely due to the AAV lane transit model that had been imposed on top of the obstacle clearance model. Unfortunately, the data that

could have allowed examination of that possibility (the positions of destroyed and displaced mines and of displaced obstacles after each bomb detonation, and the positions of destroyed AAVs) were not collected. Stripping out the AAV lane transit model and changing the MOE as indicated above could yield a very useful obstacle clearance model.

2.1.2 Pythagoras Counter Insurgency (P-COIN) Conceptual Model Validation

The model was developed as a central component of a MCCDC Irregular Warfare (IW) Project. The scenario area of interest was the coastal Buenaventura region of Colombia, which struggles with the FARC insurgency and a large and growing drug trade. The external stressor in the scenario was a natural disaster, a tsunami that devastated the coast and displaced many of its residents.

The complete P-COIN Conceptual Model Validation Report is in Appendix I of this Report.

2.1.2.1 Objectives

There were three distinct objectives pursued simultaneously within the IW Project:

(1) “To make headway in developing a Counter-Insurgency model.” (MCCDC Study Sponsor)
(2) “To determine whether Pythagoras could model the Buenaventura Disaster Relief/Humanitarian Assistance scenario.” (NGMS)
(3) “To determine In the Buenaventura Disaster Relief/Humanitarian Assistance scenario whether it’s better to base the MAGTF ashore or afloat.” (MCCDC OAD)

2.1.2.2 Findings and Conclusions

The following gives the findings and conclusions of the P-COIN Conceptual Model validation, which primarily used assumptions testing.

2.1.2.2.1 Validity of Pythagoras COIN Model for Making Headway in Developing a Counter-Insurgency Model

The only assumption that had a validity problem with respect to this particular objective was that allegiance changes are the sole MOE. That assumption ignored the actual assigned missions of the MAGTF in favor of an implied mission.

The bottom line for the objective of “making headway” was that, with the one caveat noted, the theoretic sub-model of Pythagoras COIN appeared fully valid. It is actually believed that, more than just making headway; it could represent a real breakthrough in conflict modeling in general.

2.1.2.2.2 Validity of Pythagoras COIN Model for Modeling the Buenaventura Disaster Relief/Humanitarian Assistance Scenario

The validity of the Pythagoras COIN Model for the scenario at hand appeared to have been designed into the theoretic sub-model from the start. The clearest evidence of

that was the choice of affiliation changes as the sole MOE. That choice ignored the actual assigned missions of the MAGTF in favor of an implied mission. It also meant that the scenario and the analysis had become tightly circumscribed by the capabilities of the tool.

2.1.2.2.3 Validity of Pythagoras COIN Model for Answering the "Afloat vs. Ashore" Question

The theoretic sub-model of Pythagoras COIN to which the validation framework was applied had serious issues with respect to its validity for addressing the afloat/ashore question:

(1) Use of only a single MOE, with that one MOE addressing an implied MAGTF mission while ignoring the three assigned missions, all of which would undoubtedly be affected by the MAGTF basing decision.
(2) Use of the Markovian memoryless process model for a scenario in which human memory of past affiliations and events would be expected to play a significant role.
(3) Use of constant transition probabilities, thus effectively ignoring the inevitable react/adapt cycles of opposing forces.
(4) Use of a suspect algorithm to force single values from the three-dimensional E-P-A metric of semantic differential.
(5) Use of semantic differential word-scoring data developed in one context with words selected for a radically different context, and without experimental justification.
(6) Loss of context in general with the use of the semantic differential.

As a result, it must be said that the validity of Pythagoras COIN to answer "ashore versus afloat" was highly suspect. None of the above issues, however, proved it to be invalid for that particular purpose. Thus, strictly speaking, the assessment of the theoretic sub-model has failed to falsify the null hypothesis that it is valid.

That above fact of "failure to falsify," however, was at best a very weak endorsement of the validity of the system to answer the ashore versus afloat question. More research is needed into each of the areas noted above. It may be that the research will provide solid justification for some aspects, or cause some or all to change. Additional research could be particularly valuable in the case of semantic differential, which is seen as potentially the most important analytic tool emerging from this Pythagoras COIN study. It was advised to not report out any results from the model as actionable in any sense.

The most important determinant of the above recommendation was that the P-COIN model addressed only the mission implied by the selected MOE and effectively ignored the three missions specifically assigned to the MAGTF. If (and it's a very big "If") the MAGTF commander were to state that he is indifferent to shore versus sea-basing from the perspective of his primary missions, then the P-COIN model could reasonably be considered for use to inform the ashore/afloat decision. Using a notional risk assessment tool developed during the ABSVal study, the risk of using the P-COIN model under those circumstances would be High to Very High if the MAGTF

commander considered his implied mission very important and Minimal to Low risk if he considered it unimportant.

2.1.3 Pythagoras Counter Insurgency (P-COIN) Analysis Application Validation

The focus in P-COIN was on the changing distribution of insurgency sector orientation amongst population segments. This orientation changed over time due to the natural tendency of the population to be affiliated with the insurgency sector, the degree to which each population segment wanted to be like another, and the effect events had on the population segments. The Valle Del Cauca and Cauca provinces of Colombia were the population of interest developed for Pythagoras-COIN. The population modeled had eight general population segments based on ethnicity, political orientation, family history, socio-economic status, living location, and occupation: Catholic Church, Displaced Persons, Illicit Organizations, Military, Police, Old Money, Urban Middle Class, and Urban Poor. It was sub-divided further into five insurgency sectors along the spectrum between support for the government (GOVT) and for the insurgency, i.e., the Revolutionary Armed Forces of Colombia (FARC): FARC, Pro-FARC, Neutral, Pro-GOVT, and GOVT.

The population had a tendency to drift naturally between these insurgency sectors based on a natural tendency (vulnerability) and the influence that the various population segments had upon each other (salience). The arrival of the MAGTF, either ashore or afloat, was an event that further influenced these population segments in insurgency orientation for the duration of its stay. A Markov chain provided the base descriptive model and data for the vulnerability of the population segments toward the spectrum of insurgency. Salience factors modified this Markov chain to allow the population segments to influence each other. This allowed events to affect population segments not directly targeted or affected by an event. Additional factors were used to capture the influence of events in the simulation; in the case of this model application, the on-shore or afloat presence of the MAGTF. Figure 1 depicts this.

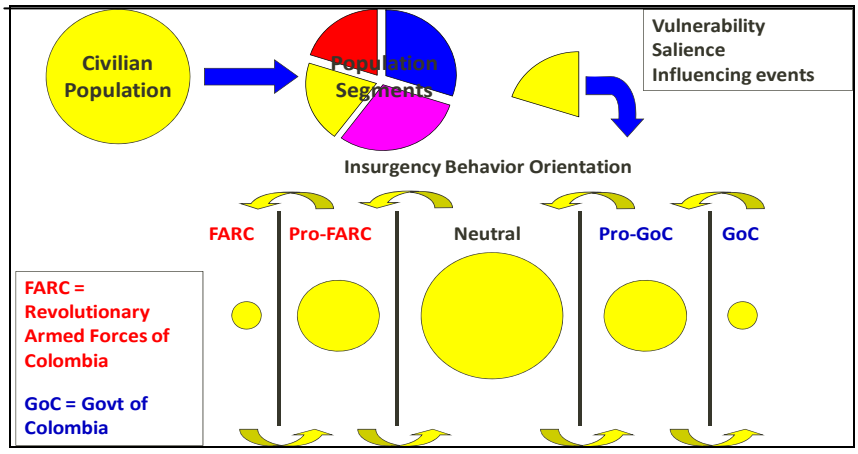


Figure 1 Conceptual Model of Civilian Population

2.1.3.1 Analysis Intended Use

While the primary intent for the development of P-COIN was to assess whether and how Pythagoras could be used to model population dynamics, the analysis question for this validation effort is

> In a Disaster Relief/Humanitarian Assistance mission for the above scenario, is it better to base the MAGTF ashore or afloat?

Answering this question was the intended use for the validation assessments.

2.1.3.2 Validation Conclusions

A summary of the conclusions of the validation assessments made to P-COIN were as follows:

1) The P-COIN simulation failed to capture the dynamic effects intended in the Conceptual Model of the insurgency in Colombia provided to the P-COIN developer. That is, P-COIN did not capture the secondary and tertiary effects of the natural drift of population segments between insurgency sectors or the salience between population segments resulting from the influencing event of the MAGTF.
2) The data supporting the P-COIN model was perishable and of low precision. Care should be taken when using the data beyond its origination date; in those cases, perhaps “warming-up” the Markov chains supporting the data used to build the P-COIN model may create more reasonable results. Further, the data cannot be deemed valid if an influencing event occurs that would cause the base data used in this simulation to change (e.g., events of July 2008).
3) The P-COIN model should not be used to evaluate long term effects on the population resulting from the influencing event of the MAGTF arrival.
4) This model and simulation cannot be deemed as predictive of the actual population distributions amongst insurgency sectors in the event that the scenario described in the scenario documentation actually occurs.
5) There is little risk in using the results of the analysis since the analysis does not advocate a change in current Marine Corps procedure. However, item 1 implies that P-COIN also provided little insight into the ashore or afloat question in its current implementation.

2.1.3.3 Desirable Additional Material

The validator would have found the following material useful in the validation assessment:

1) Better documentation on the P-COIN instantiation

2) Time series data

3) A descriptive walk-thru of results charts (meaning & implications)

4) Verification cases (isolated effects) to ensure dynamics have expected direction (first derivative) and order of magnitude with descriptions of what was believed to be correct

5) Better explanations of expected resulting effects from data values in the referent as most had to be inferred and order of magnitude differences unknown

6) Expected interaction effects

2.1.3.4 Validation Recommendations

Even though the results of the analysis reflected the expectations of the analysis based on the input data tables and the analysis did not recommend a change in course of action, it was difficult to trust the results without being able to trust the underlying dynamics of the model. The chosen instantiation of only applying attribute changers reflective of the start state of the agent population distribution should be thoroughly evaluated with the recommendation to apply these attribute changers more robustly to reflect how the population insurgency affiliation changes over time and in response to system events. The full dynamics intended by the use of salience and the natural drift aspects of the population segments ought to be included in the P-COIN model in order to allow for the emergence of secondary and tertiary effects of the influencing event of the MAGTF arrival. The developer should also develop robust test cases to evaluate the dynamic behavior in isolation to gain surety that the combined dynamic behavior should be trusted.

2.2 VALIDATION APPLICATION EFFORTS AUDIT OVERVIEW

The complete Audit Summary Report is at Appendix K of this Report.

To assess the utility of the ABSVal Framework, each validation effort underwent an independent audit. The task was to evaluate and comment on the content and application of the ABSVal process with a focus on its practical utility:

(1) How useful did the process appear to be in providing guidance to a Verification and Validation (V&V) team, and
(2) How useful are the products of an ABSVal guided V&V team to an accreditation agent or model manager with respect to the given test cases?

The auditors participated in the workshops held as part of this Study. While they attended the workshops, the auditors consciously avoided contributing to the content of ABSVal or attempting to steer the process of developing that content, in order to prevent a sense of ownership and compromise their position as independent observers and evaluators. The focus of the audit review was the content (what content is and is not present) and the impact on the process.

2.2.1 Audit Questions

The auditor used the following questions to bound the audit:

(1) Does the report clearly identify the application (set of study questions) for which the model is being validated and the model's role in addressing those questions?

It is important to bound the problem and document understanding (or reveal any disconnect) between the V&V team and the accreditor about what the model needs to do well.

(2) Does the report clearly describe the tests that were performed on the model, the possible outcomes for each test, and the criteria for passing?

A significant component of the discussion that underlies the inventing of ABSVal had to do with making the V&V process more science-based, and therefore objective and repeatable, than it has traditionally been. Specifically the notion that model validity is a falsifiable proposition subject to challenge by applying tests that, should the model be invalid, would have a reasonable chance to reveal that invalidity. Results were mixed, but this question assessed how far the V&V team was able to take that idea. Also, regardless of any connection to the scientific method, the accreditation agent would need an easily digested summary of the V&V team's investigations that led to the accreditation recommendation.

(3) For each test performed, is the result clearly presented in a way that relates directly to the specified acceptance criteria?

The auditors did not expect it in the test cases, but it is common that an investigator, perhaps in response to actual or perceived political pressure, did not seem able to confront the implications of facts uncovered in his or her investigation. This phenomenon typically manifests itself as a redesign of the scorecard after the game has been played, or an expression of the result in terms that do not relate directly any acceptance criteria, predetermined or otherwise. Obscuring the results diminishes the value of the work done to produce them, and this question is one that a reviewer must consider in any situation where it is appropriate.

(4) Does the report provide a recommended decision for the accreditation authority?

A recommendation with rationale is more useful to an accreditation authority than "here are some facts we found, make of them what you will," even if the rationale is not accepted by the accreditation authority.

(5) Does the report make a convincing argument that the tests conducted collectively provide a sufficient basis for the recommended accreditation decision?

In other words, does it tell the accreditation agent and authority why they should agree that the investigation was complete enough to make a recommendation that they and the accreditation agent can stand behind?

(6) To how broad an audience does the report make its findings accessible?

This question was included after having seen an early draft that the audit team found to have a somewhat esoteric style reminiscent of some academic journals. The

audit team wanted to encourage the V&V team to target its arguments to the broadest segment of people qualified to evaluate them.

(7) Are recommendations provided that are actionable by a model improvement program?

This question was added to get at the value to a model improvement program. Actionable recommendations are the vehicle for delivering that value. The same background information required by an accreditation agent would also provide rationale for these recommendations.

The audit team treated the two validation reports for P-COIN as a single effort. In general, the audit team found the reports easy to understand and follow. However, in some cases, the audit team had to infer completed tests and recommendations. Details are in the Audit Summary Report found in Appendix K. Recommendations from the audit are below.

2.2.2 Audit Recommendations

The following summarizes recommendations found in the audit report:

(1) Validation reports should clearly describe the qualitative questions investigated and tests conducted. These descriptions should include the reasoning for the investigation and tests, as well as expected and desirable results. It should also state necessary conditions for "passing" the tests for the intended use (i.e., validation criteria) and conditions that would constitute a "failure."
(2) The process of making qualitative assessments would be more objective with less reliance on the subjective opinion of the validator if the process were more structured, expectations were identified, and if the reasoning for the validators conclusions were made explicit.
(3) The significance of the results of all validity assessments (qualitative and quantitative) on the intended use and application of the model needs explicit explanation. The effect of failing the test on the intended use needs to be clear. Along these lines, the actual recommendation of whether or not to use the model needs to be stated clearly.
(4) The report should identify mitigation strategies: in use (e.g., sensitivity analysis), in corrective development, or in model improvement.
(5) The framework should include the minimum sets of information and materials sufficient to reach a final validity determination using the ABSVal Framework.
(6) Due to the brainstorming nature of this task, there should be two or more people involved in the extraction of the implicit assumption of the Conceptual Model.
(7) A practitioner's guide to validation resulting from the ABSVal Framework should include rules, stepwise approaches, branches, and criteria for determining necessary or sufficient information to support a given conclusion. Rules might include necessary information and model and simulation artifacts to embark on a validation investigation. The availability of material might influence the risk of use, the risk of use being stratified at different acceptable levels.

2.2.3 Post-Audit Activities

Work on the final audit report of the two ABSVal test cases continued after the final workshop, Workshop #4, Summary Report was submitted. Consequently, two important instances require clarification.

The most important instance regarded explicit risk assessments. The audit report made strong recommendations in Section 7 (see Appendix K) about the importance of a risk assessment as part of the ABSVal process. ABSVal Phase II Interim Report #2 had addressed three possible risk assessment tools that could be applied within the ABSVal process. Unfortunately, none of the ABSVal application report versions reviewed by the audit team had included such a risk assessment. However, both the final P-COIN Conceptual Model validity assessment report and the final P-OCM validity assessment report included an explicit assessment of the risk of using the models for their respective applications. The tool applied was very similar in concept but significantly different in detail to the three that had been addressed in Interim Report #2. This tool showed promise and could meet the intent of the audit team's recommendation. The P-COIN validation briefing presented at Workshop #4 (see Appendix E, Section E.5.2.2) also addressed risk.

The second instance concerned recommendations for improvement of the P-OCM model. The Audit Report noted in Section 5 that no model improvement recommendations had been made. In the final P-OCM report version, however, two strong improvement recommendations were made in the "Conclusions" section. These recommendations may obviate the need for that particular audit finding. Further, the P-COIN Workshop #4 briefing identified recommendations for improvement.

2.3 PYTHAGORAS COUNTER INSURGENCY (P-COIN) VALIDATION CHALLENGES

A particular challenge in completing the validation effort for P-COIN was that the methodology and framework were in development concurrently with the P-COIN development and analysis. Additionally, significant insights into the framework occurred midway through the validation effort causing a significant redirection of effort and an alignment between two separate efforts: the P-COIN analysis and development and the P-COIN validation as a test of the ABS VV&A Framework. This concurrence and alignment as well as the mid-course insight has led to some areas having less reporting than others as well as work completed early in the validation process not strictly utilized as part of the final validation effort.

2.3.1 Framework Study Insight

Insight gained through the application of the framework led to the notion that the focus on this validation effort and the ABSVal Framework needs to be on *analysis*. For example, the validation focus for P-COIN was on the questions that the P-COIN model could support and the degree to which this model could support those questions. In the validation process, there was a heavy reliance on the documentation of the analysis by the analyst. The results of the framework and the application of the validation experimental procedures provided a method for decisions makers to know the

boundaries of a model's use and the limits of the "believability" of their results. The development product was a procedure to evaluate whether a model or simulation could be used to support an analysis objective. Therefore, validation is, in part, a questioning or prodding of the analysis and cannot be decoupled from that analysis. In other words, it is an analysis of the analysis. In this validation process, the analyst assessed when answering the posed analytical question (intended use) the important elements for answering the question, the data required, and the tests (sensitivity, experiments, or excursions) needed so that the decision makers or recipients of the analysis could have their questions answered and confidence in the simulation results. That is, it was the role of the analyst in validating the use of the model for an analytical purpose to ensure the ability to assign a reason for all the elements in the model and the results.

In reviewing the DoD definition of validation,

> Validation is the process of determining the degree to which a model and its associated data are an accurate representation of the real world from the perspective of the intended uses of the model.

Three elements are critical: model, accuracy with respect to real world, and intended use.

Discovered in Phase II Workshop #3, a frequently overlooked aspect of this definition was the *intended use* of the model. While the accuracy requirements for a simulation in terms of precision, level of detail, and fidelity of the real world representation embedded in the model usually considered the intended use of a simulation, the specifics of *using a model in a specific application* was often overlooked. That is, validation efforts would frequently attack the Conceptual Model of a simulation without assessing whether the identified deficiencies were core elements to the application at hand, for instance, to answer the posed analysis question, or if these so-called problems with the model could be mitigated in some way, say, through sensitivity analysis. This was the key insight gained in Workshop #3: Validation in an analysis application is an analysis of the analysis. This insight required a mid-course correction of the P-COIN validation effort in the full application of the ABSVal Framework.

2.3.2 Access to Models

There were several challenges in testing the ABSVal Framework. A major problem was in finding ABS models to which the team could have access for the purposes of validating their use. Although much effort was expended to find a variety of ABS applications, the team was unable to secure access to the model and the analysis use. In the end, the team had to apply the ABSVal Framework to two applications both developed in Pythagoras. Although the team had access to the developer in both instances, for differing reasons in each case, the validation exercise was limited by the available documentation. Further, neither of these applications had significant amounts of dynamic emergent behavior, which was initially desired for the Study. Lastly, both of these studies were a low level of effort, small scope, simulation applications, which inherently has an effect on the overall analysis supported by the chosen models.

2.4 DEVELOPER'S POINT OF VIEW

The following is a statement provided by Mr. Edmund Bitinas, developer of the two Pythagoras ABS models on which the ABSVal framework was applied during the Study, as a critique of the results of applying the framework:

None of the reports developed, audit or validation, helped the developer to understand necessary improvements to the model: in the Conceptual Model or the implementation, nor did the reports give the developer the sense of whether the model development is on the right track. From the reports, the developer was not able to assess the salvageable elements of the model, if any, should a validator find deficiencies.

None of the reports accounted for the maturity of the application to which the models were intended. In each case, the analysts had little basis to begin their development (in contrast to models based in physics differential equations or known and accepted combat modeling). Thus, while the main insight was the need to more robustly include intended use throughout the validation process, the developer saw the need to further this aim in the process. That is, the validators did not pursue this insight far enough.

Not being able to prove a model is right is not the same as showing a model is wrong. That is, a validator has the obligation to demonstrate or identify the negative impact on a modeling or coding necessity. Merely identifying problems with substantiating a modeling choice within a referent is insufficient.

Requiring precision when there is none is as problematic as not achieving precision when it is required. This problem is explicitly identified in the VV&A Recommended Practices Guide (RPG) but is sometimes forgotten in a zeal for "accuracy with respect to the real world." Representative accuracy may be more applicable (e.g., equivalence) when the data available does not support a more definitive approach. Sensitivity analysis can mitigate data difficulties. Identifying representative accuracy requirements can mitigate the desire for data matching precision (e.g., magnitude and direction of change).

Finally, the validation cannot be decoupled from the analysis. Model development was designed to meet that analysis. The VV&A RPG notes that without validation criteria (i.e., what must be met for the model to meet intended use) validation becomes merely an assessment of capabilities. Thus, again meeting the insight of Workshop #3, the validation activities must assess the model and its simulation results on the terms in which it will be applied (e.g., the way its measures will be collected and used).

2.5 ABSVAL FRAMEWORK UPDATE

Traditional approaches to validation are embedded in math, science, and engineering applications. Empirical data is used to compare simulation results and is the basis against which accuracy measurements are measured. Accuracy is measured in the delta between the data and the simulation results. Higher levels of accuracy can often be achieved through improvements in the solver algorithms applied in numerical methods. In simulations of social systems, these mathematical representations are not

available. The approach for validating these models becomes an exercise in critically examining the Conceptual Model. Accuracy is the subjective assessment of one or more experts of the correctness of the Conceptual Model. Problems in the Conceptual Model descriptions can lead the validator to never evaluate the model instantiation into code or an evaluation of the simulation results.

Neither approach fully meets the needs of an analysis context as they rely on information that is either not available or ignores the use of the model altogether. The update made to the framework addresses both problems.

2.5.1 Role of Theory

Phase I of the study and early stages of Phase II looked at the fundamental theory that surrounds the validation of simulations. This theoretical investigation resulted in the development of the diagram shown in Figure 2, which shows the comparisons between the modeling and simulation elements available. The goal of developing a theoretical foundation for ABSVal was to tie the simulation to the real world in a mathematical way.

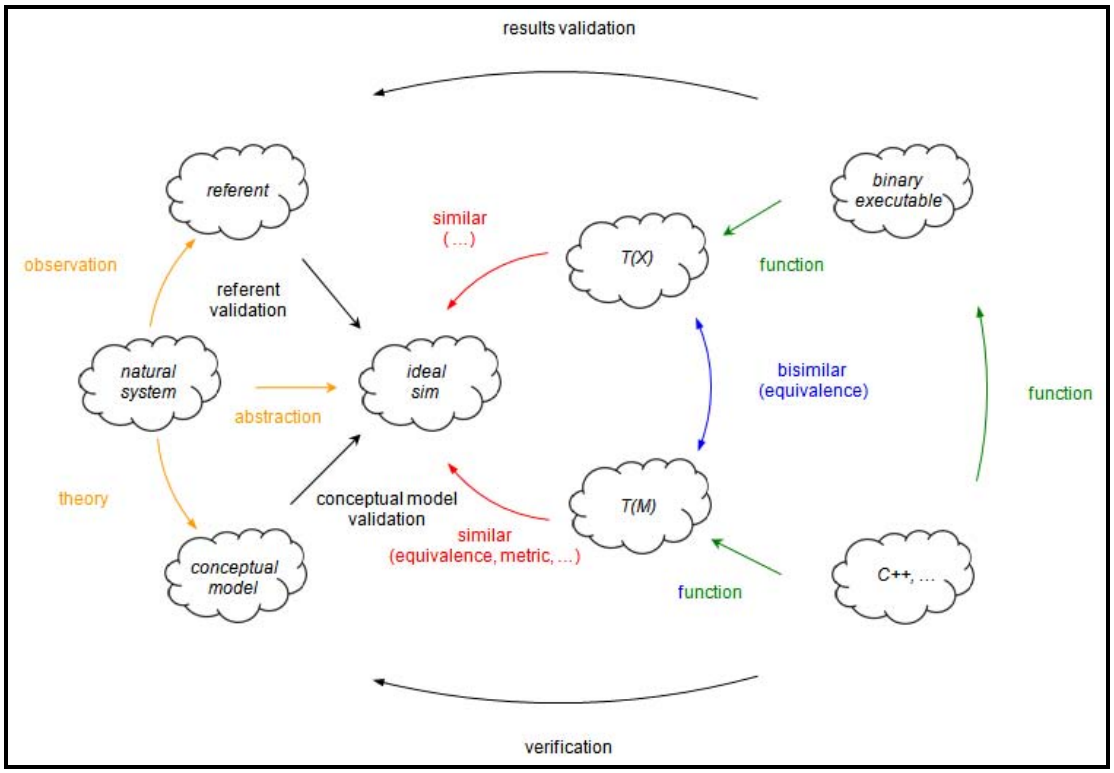


Figure 2 Validation Theory

The diagram places simulation elements on the right hand side and the simulated real-world elements on the left hand side. The outside nodes of the cloud diagram represent elements of practical validation (tangible), while the inside nodes generally represent elements of theoretical validation (theoretical / mathematical).

A digital computer running a programmed simulation creates a state transition system. Therefore, in the theoretical sense, validation is a comparison of transition systems, the simulated transition system compared to the transition system of the simulated system (when available).

2.5.1.1 Natural System and Ideal Simulation

The Natural System (aka the Target System or the System of Interest) is a subsystem of the universe in its entirety from the perspective of a perfect observer at a snapshot in time. Depicting the universe from this perspective as an infinite vector, the natural system is a notional "slice" of that vector.

Pragmatically, not all elements in this Natural System can be modeled. Abstraction focuses on elements of interest in the Natural System for the purposes of simulation. In mathematical terms, this abstraction amounts to the bounding of the state vector and the input vector. Further, abstraction creates non-deterministic transition systems, creating the Ideal Simulation as depicted in the cloud diagram. For the Ideal Simulation, the values of vectors match exactly with the corresponding values in the Natural System with no loss of accuracy (it extrapolates perfectly), but the elements of the Ideal Simulation are limited to what has been derived by abstraction.

2.5.1.2 Code

The Code element in the cloud diagram represents a transition system in the form of a binary executable program. The ultimate goal of the validation effort is to demonstrate that the transition system generated from the code simulates the Ideal Simulation (i.e., the coded simulation matches every transition made in the Ideal Simulation).

2.5.1.3 Provability of Validity

There are problems of tractability in generating a mathematical proof that shows that one transition system simulates the other, and there is no algorithm to evaluate these transition systems trajectory by trajectory. Further, a key finding from this development of the underlying theory was that if one could not answer a posed analytical question using the Ideal Simulation, then one could not answer that question with any simulation representation. Thus, for the inner loop (theoretical elements) of the cloud diagram validity cannot be proven (in an exclusively mathematical sense), and so other techniques of validation must be employed.

2.5.1.4 Results Validation

Results validation is often used to provide evidence of the validity of a simulation; the binary executable is run to acquire an array of trajectories, which can be compared to a referent, real-world empirical data from observations. Often for social systems, the referent is not solid; however results validation can still be employed (although much more subjectively) to assess if simulation output is sensible and useful. The referent data, in turn, requires validation as well, to determine accuracy and suitability to the application.

2.5.1.5 Intended Use

In the framework, the intended use of the simulation defined by the analytical application drives the abstraction process to determine what details and elements are critical for the simulation. Intended use also drives validation criteria (e.g., “Do the states in the simulation need to be within some tolerance compared to the Ideal Simulation?” or can more subjective methods be applied to determine the validity/usefulness of the simulation?).

2.5.1.6 Conceptual Model

The Conceptual Model arises out of the Natural System by theory – it is the “best that is known” about the Natural System from a theoretical sense. An aspect of validation may be to provide evidence that the elements in the Conceptual Model are true with respect to the Ideal Simulation. This often may occur with some degree of subjectivity, especially when simulating human systems. Because of the inability to comprehensively validate a human system, the scientific method was employed in the framework as a method in which validation techniques were applied to the simulation in an attempt to disprove or “poke holes” in the integrity of the simulation with respect to the application. With rigorous validation techniques, the inability to invalidate the simulation subsequently provided evidence of its validity and usefulness toward an analysis application.

2.5.2 Validation Supporting Analysis Applications

The ABSVal Framework evolved over time. Initially, the study team approached validation in a traditional way, such as that discussed in the VV&A Recommended Practices Guide (RPG). However, this guide has several limitations with respect to the needs of the ABSVal Framework Study. While the guidebook lists several techniques for V&V, they mostly focus on verification rather than validation. Further, it does not give methods for applying these techniques or appropriate times to use them. With respect to the challenges to agent based simulation (addressed as human behavior modeling), the VV&A RPG is woefully deficient in guidance. For instance, when discussing the validation techniques, the VV&A RPG states that face validation is to be avoided whenever possible and SMEs have particular difficulty in human behavior modeling, but then it goes on to say that it is likely the only technique available for validation of those types of models.

In recognizing the shortcomings of the VV&A RPG with respect to the simulations of interest and armed with the knowledge that not addressing the difficulties with validating these simulations was unacceptable, the study embarked on a journey to discover what could be determined about the validity of an agent-based simulation to communicate the risk of their use to decision-makers. The initial framework developed focused on validation of the Conceptual Model against the referent. In this context the Conceptual Model included any embedded mathematics or logic imposed on the system resulting from the need to code the simulation in a digital computer. In the agent-based simulation context, those mathematics and logic structures include any rule sets and agent decision-making structures. Results validation, while it could include empirical,

observed, or historical results, is essentially limited to expectation matching of behaviors. The early stages of the project never fully defined this concept, however. The structure provided was a recommended table of contents for a validation report to guide validators and validation report recipients to ensure all required information would at least be addressed during a validation assessment.

This approach still seemed to have shortcomings with respect to the analysis context. During Workshop #3, it became obvious that validation was an analysis process, intrinsically intertwined with the analysis for which the agent-based simulation was employed as a tool. This resulted in a significant redirection of the, then ongoing, validation applications to test the framework. This was challenging from a process/framework development perspective because often these analysis processes and techniques are difficult to describe in “handbook” form – there was an “art” involved driven by analyst intuition and subject matter expertise. Some aspects of this analysis included understanding parameters and their change mechanisms as well as understanding the surprise elements in the model and removing those surprise elements (i.e., uncontrolled emergent behavior).

Also evolving was the notion of what decision makers and model users needed to know about a model when using the simulation results to support analysis. In particular, the need to contextualize the results from an agent-based simulation and their potential credibility resulting from the information embedded within the model. This brought the concept of Constraints, Limitations, and Assumptions (CLAs) into sharper focus as the study progressed.

The original approach to the validation assessments prior to Workshop #3 focused on the general aspects of validation all but ignoring intended use. In these initial efforts, intended use was included to help the validators determine critical elements to include in the model and necessary accuracy requirements. Beyond this cursory inclusion, further thought to the actual use of the model to address the analysis questions was not apparent. It was determined that intended use is not just the field of use for a model (e.g., training, analysis, concept exploration, test and evaluation, or experimentation), it also includes the mechanisms of that use for the broad application, and these aspects need to be carefully considered during validation.

In an analysis context, the validation of a model’s use is not a pass/fail endeavor. It is rather a communication of the CLAs, the dynamic factors that influence the model, and the impact of CLAs on the model’s feasibility as a tool for answering the questions posed. The investigation process may lead to a reframing of the analysis questions posed. Validation is an “analysis” in and of itself, where validators are putting on an “analyst’s hat” and attempting to be as comprehensive and well documented as possible with the examinations and results.

The evolution of the framework with this emphasis on analysis introduced new potential validation questions such as follows:

- Under what conditions am I evaluating alternatives?

- What are the core elements of the simulation analysis?
- What is causing variability in test case results?
- How are dynamic influencers impacting the results?

An Assessment of Risk may be communicated using validation results that express the potential consequences of using a particular model for the analysis.

All this information needed to be contextualized for the decision maker, focusing on the intended use. Ultimately, the framework was a description of what a good analysis report for the question at hand would provide, recognizing that there are levels of rigor for validation that depend on the data available, the complexity of the model, and the validation resources available. The result from the validation effort in support of a simulation analysis application was a report (to decision makers, analysts, developers, etc.) that elucidates the limitations of the model/analysis. The report is intended to provide descriptions of the analysis tests completed and areas explored; the results of these tests; and the conclusions and recommendations drawn with supporting reasoning. This report should be written to the mid-level manager (O-3, O-4, or O-5) who will be familiar with the analytical need, may have some subject matter expertise, and will likely have to make a recommendation and defend it. This audience, however, typically will not be the decision maker or an analyst. However, appendices and supplementary documentation should be provided with the report to enable this mid-level manager to get a recommendation following a detailed evaluation of the validation analysis provided by a junior analyst that might review the report. The goals of the report are to communicate answers to the following questions:

1) Does the model represent what it is advertised to represent?
2) Are the limitations of the model explicit?
3) Does the analysis (using the model as a tool) answer the analytical question(s)?

Limitations of the simulation results to support answering the analysis questions need to be made explicit with these model limitations communicated to the end users. Recognizing the importance of intended use, it is critical that these limitations be traced back to the analysis questions and the impact of these limitations on the analysis be made explicit. That is, it is not enough to state a model's limitations, the effect of these limitations on addressing the analysis (use of the model) must be made clear. A failure to link a Conceptual Model validation to the analytical use was the reason for the initial difficulties in applying the framework prior to the Phase II Workshop #3. It must not be forgotten that a validation effort cannot exist independently from the actual use of the model and the decision context for the analysis. That is, validation does not exist in general. Finally, the validation report must contain the risk of using the model in the face of uncovered limitations.

In short, the validation of an agent-based simulation in support of an analysis application is rarely an attempt to draw a good/bad or valid/invalid Boolean type conclusion. Rather, it is an exercise to uncover and then communicate to the end user of the model an assessment of its capabilities and limitations. The assessment of limitations must inform the model user/decision maker of the impact, effect, and

criticality of that limitation. It could provide mitigation efforts that the analyst could use to reduce the risk of the limitations (e.g., robust statistical experimental designs, data farming, sensitivity analysis, or additional cases). The results of validation also may define subsequent, second tier validation questions that could further provide evidence of the model's usefulness that could not be answered due to technical or even resource constraints. These subsequent questions could be especially useful in attempting to determine if emergent behavior in the model is of analytical importance or a model anomaly.

2.5.3 Tools and Techniques

Tools available to the validator in making these assessments included not only those listed in the VV&A RPG but also any tools available to the analyst in general. The most important tool that was available to the validator in assessing a model's ability to provide simulation results supportive of answering an analysis question is the validator's ability to understand the analysis problem, the decision context, and the elements critical for inclusion. This includes the ability to describe the decision drivers that must be included in the model, the influencing dynamic elements that affect the answer, and the cases that must be included to understand the robustness of proposed decisions. It also includes the ability to understand the shape of the answer and presentation of the results in a way that helps illuminate the questions being asked and identify any "holes" that must be filled in stepping through the analysis or other collateral questions that should be addressed.

2.5.3.1 Results Testing

Whenever possible, simulation results should be compared to empirical or observed data. While a metric relation could be used to assess accuracy (i.e., the delta between values), other accuracy measurements are possible (e.g., comparisons of direction, slope, or relative magnitude). When this kind of data is not explicitly available, the validator still needs to assess whether the simulation output meets the analytical needs. In this case, results validation relies on robust test cases. Due to the inherent complexity of ABS and the general inability of any subject matter expert to understand the fully interacting effects of all the embedded elements with these models, these test cases may require a structured reduction of the complexity in the model. For instance, the test cases could evaluate pairings of the dynamic elements within the model. Subject matter experts could be used to provide the expected behavior or results in these pairings for specified data values. The simulation results are then compared to these expected behaviors to determine whether, at least on these "boundaries," the model performs as expected. This can be viewed as a type of Turing Test. The key to this evaluation is the a priori description of the expected results. This has the effect of filling out the referent for the system. If the model performs as expected in these test cases, then there is greater confidence that the interaction between multiple model elements would also perform correctly. When there are many elements that can take on a large number of values in the model, a robust statistical design might be used to determine which test cases are built. Expected behaviors might be expressed as trends, as specific data values, or anywhere in between. The validator should specify *a*

priori how achievement of those trends will be evaluated (e.g., goodness of fit or statistical tests).

2.5.3.2 Assumption Testing

Another important method for understanding the usability of a model's simulation results is assumption testing. As stated in the Logical Validity Assessment Process report (See Appendix M),

> Model assessments commonly address the hardware and software engineering aspects of a model, and particularly its *usability* characteristics (e.g., user interface, graphics, input data availability and formatting requirements, clarity and completeness of documentation, maintenance support). While those characteristics are unquestionably important, the *analytic* capabilities of any model are not determined by its hardware and software architecture, communications network, or user interface. Nor can they reliably be determined from the "capabilities" statements – often little more than advertising copy – that typically accompany a model. Rather, analytic capabilities are determined by a model's logic and control structures and their underlying assumptions, its computational algorithms and underlying mathematical assumptions, and its data manipulation and transformation algorithms – all of which are rarely seen by the end user. Moreover, a model's *bounds of validity* also are determined by its underlying assumptions, some of which may not be readily visible even to the model's developers.

As a practical matter, however, *explicitly* identifying every assumption in even simple models is impossible. However, there is no need to identify all of them. Only those assumptions that have significance to the *intended purpose* of the model and especially to the *analytic questions at hand* needed identification and testing. Part of the art, vice science, of assumption testing was to be able to recognize in at least broad terms which assumptions are likely to be significant, given only a description of the model, the context of the study, and the specific analytic questions at hand. Thus, validity assessment analysts generally have to cast a wider net than would be necessary if they had full knowledge going in as to which assumptions are significant. Assumptions having little or no apparent significance are set aside. Ones having apparent significance are tested. Appendix M describes the process.

2.5.4 Developer/Analyst Perspective

Frequently, the user of an ABS may also be considered the model developer, since many ABS systems are environments, software libraries, or other software constructs that enable a user to build an ABS representation of something with little or no programming. From that perspective, the ABSVal framework can and should be applied at least twice during the course of an analysis, once at the beginning, and once again near the end.

2.5.4.1 ABSVal at the Beginning

The ABSVal framework should lead the Developer/Analyst (D/A) through the analysis plan and model selection process and help identify the requirements for analysis once an ABS application is constructed. The D/A should identify a Conceptual Model that is

independent of available ABSs, identify the critical elements of the Conceptual Model and measures of goodness that may be required or desired, and compare those to the capabilities of various candidate ABS systems. In some cases, the ABS may have a workaround that surrogates one capability for another, enabling an effect to be modeled even though its cause is not the same as it is in reality (Example: In P-OCM, bomb hit point agents were considered to be 'leaders' that the mine agents were to move away from). Similarly, the framework can help identify data sources, means to convert those sources into values that are meaningful to the ABS, and ABS output that can be used to measure outcomes. Thus, by reviewing the ABSVal process early in an analytic effort the D/A can increase the likelihood that the risk of using the model is reduced. An independent assessment of the analysis plan and model at this stage will determine if the plan is likely to provide usable results, the selected model(s) have sufficient fidelity to provide useable results, and the team of analysts is experienced enough to accomplish the task at hand.

Note also, that there may not be a Mathematical Model, in the form of equations, but only a Conceptual Model of a behavioral theory and a logical mapping of the required capabilities of that theory to the available features of the selected ABS. This logical mapping itself may result in a 'simulation' of the theory, since the ABS may be representing the theory, and not actually directly implementing it. If the theory itself is inaccurate, only matching a fraction of the observed data, then the validation team must assess if the added inaccuracy of applying the ABS will reduce the representation of the theory, or merely look at the actual system in a slightly different, maybe even a better, more realistic way. Frequently, a collection of algorithms, which are basically what ABS systems are, can be used to capture behaviors that are difficult to represent in mathematics alone.

2.5.4.2 ABSVal at the End

ABSVal must also be applied near the end of the study, with enough time remaining to correct, or at least mitigate, deficiencies that may be uncovered. Reasons to repeat, or at least re-look at the validation assessment may include a change of personnel, a lack of good or complete data, or the existence of seemingly unexplained emerging behavior. Frequently, an analysis will not only attempt to answer the original question, but will also explore interesting phenomena uncovered during the analysis. These might include emerging behavior, unanticipated results, or previously unidentified cause and effect relationships. The ABSVal process should not only assess the risks of accepting the results of the analysis, given the limitations imposed by reality, but should also provide additional risk assessments of the ancillary findings of the analysis, if any. For example, the P-OCM study indicated that the result of the analysis of using bombs to clear shallow water obstacles is feasible (low risk), but that the ancillary finding of the number of bombs required is significantly higher risk, and should be studied in more detail, more fidelity, or with a different model.

The validation at the end can provide the D/A with additional analytic activities that can be performed to reduce the risk of believing the results of the analysis. The validation team or D/A may also identify areas for further study. These 'weaknesses' may be

acceptable in the current study to answer the question at hand, however, they might be interesting areas for further research that would also extend the use of the model. Care must be taken to avoid measuring results with precision when the input data itself is imprecise. What is important, however, is accuracy. Are the trends in the right direction and of about the right magnitude? Do one verse one interactions exhibit the expected behavior? It can be argued that if all of the one verse one interactions are valid and logical, then any emerging behavior that was developed in a complex many on many engagements must also be valid. In that case, the relative strength of the effects of each one versus one engagement must also be compared and evaluated for risk.

Finally, the D/A should not accept ‘Not valid’ as a conclusion from the validation team without conclusive proof that the analysis is flawed. A conclusion of ‘High risk’ is acceptable, but should include suggestions on how the analysis can be improved, either through a change in the modeling approach, the use of another or different model, or additional analysis cases to be tested to reduce risk and increase confidence.

3 CREATE MEASURES OF VALIDITY: ASSESSMENT OF RISK (TASK 1)

The identification of the risk of using a simulation for a specified analysis problem was an integral part of communicating to the decision maker the limits of valid use for that simulation. This included discussing the ability of the simulation to meet identified validation criteria, representing all necessary system elements at the needed levels of fidelity, and achieving accuracy and precision requirements. The Study Team approached identification of risk of use for a simulation in two ways. The first, theoretical, was embodied by a review of the literature, which revealed the Validation Process Maturity Model. The second was practical, through the assessment of risk for the framework applications and assessments of the process maturity through the use of a survey in the Phase II Workshop #4.

3.1 VALIDATION PROCESS MATURITY MODEL (VPMM)

Validity in the context of this maturity model is the assessment of whether the simulation is fit for a user's purpose.[1] The Validation Process Maturity Model (VPMM) moves through six levels starting with no validation attempt; moving through increasing levels of specificity in the validation criteria available, the information in the referent(s), and the objectivity and rigor of the validity analysis; and ending with a fully automated validation capability. Harmon and Youngblood base this capability maturity model for the validation of simulation models on assessing the validation process and the completeness of the evaluation as proxies for the quality of the information provided. The proposed maturity model has six levels. These levels range from Level 1, where Subject Matter Experts (SMEs) give an opinion on a simulation's validity usually through unstructured observation of the simulation results and face validation, to Level 5 where rigorous formal validity assessments are applied in an automated fashion. At Level 0 of the maturity model, no validity assessment is made and no information is provided to assess a simulation's validity for a purpose. From Level 2 through Level 4 of the maturity model, there is a corresponding increase in the specificity of the validation criteria available, the information in the referent(s), and the objectivity and rigor of the validity analysis.

The maturity of the validation process as proposed by Harmon and Youngblood depends not on the results of any given validation effort but rather on the quality of the validation information obtained by the validation process, with greater maturity gained through increasing detail, objectivity, and repeatability. Harmon and Youngblood propose three main characteristics of quality information from the validation process: completeness, correctness, and confidence.

Completeness: The validation processes determines the degree to which the simulation representation contains all of the elements required (e.g., parameters, entities, relationships, and factors) and highlights any missing user representation requirements in the simulation.

[1] *A Proposed Model for Simulation Validation Process Maturity*, S. Y. Harmon an d S. Youngblood, 2005

Correctness: The validation process determines the degree to which the simulation matches its "simuland," identifying any error in that match and the referent used in the comparison.

Confidence: The validation process provides an assessment to the user on the confidence (in the statistical sense) of the results from the validation assessment, specifically to error assessments, identifying areas where the simulation does not meet user confidence specifications and suggesting methods to improve confidence levels.

Validation assessment reports should also include sources of information used during the assessment process, including SME credentials, if used. These characteristics support risk assessment.

Harmon and Youngblood identify several elements that ideally should be included in any validation assessment. These can be broken into two main categories: 1) the validation criteria and 2) an assessment of the correspondence between the requirements found in the validation criteria and the simulation. The validation criteria used in the assessment includes descriptions of the entities, properties, and dependencies required to meet the user's purpose; acceptable error tolerances for a given set of input ranges, and required confidence values. In addition to these, validation reports should provide descriptions of the entities, properties represented in the simulation, errors in the representation output for specified input values, confidence levels for the output, and any differences between the validation criteria and the simulation capabilities. The six tiers of validation assessment (discussed in Section 3.1.2) describe the information evaluated in this list, while the levels in the validation process maturity model (discussed in Section 3.1.3) describe the process of that assessment. If the user does not specify the validation criteria for the validation agent, then the agent must determine what these criteria are, formally or informally depending on the validation process level applied, from the information that is provided in order to make the validity assessment. In the higher levels of the validation process, it is expected that independent validation agents, or validators, would arrive at the same conclusions.

3.1.1 Description

By creating the VPMM, Harmon and Youngblood sought to identify characteristic process elements of validation and provided a benchmark for improved application of validation efforts. They hoped this would lead to greater understanding of the models validated, through more thorough assessments and more quality information, and provided a basis for improvement in the field of simulation validation just as software and systems communities have improved their field of practice. Structuring the problem in this way allows discussion and evaluation of validation efforts in terms of what was assessed and how that assessment was made. It also hinted at areas for improved methods.

Their VPMM has six levels and five information artifacts: validation criteria, referent, Conceptual Model, development products, and simulation results. Validation criteria are the requirements for the simulation to support the user's intended use. The referent is

the description of the real world against which the Conceptual Model and simulation results are compared. Conceptual Models are descriptions of what will be instantiated in code (against which the code is verified). "Development products include … detailed design information, software development products, and implementation components" (Harmon and Youngblood 2005). The VPMM moves through the six levels starting with no validation attempt; moving through increasing levels of specificity in the validation criteria available, the information in the referent(s), and the objectivity and rigor of the validity analysis; and ending with a fully automated validation capability. These levels of validity are tied to six tiers of validity assessments, shown in Figure 3.

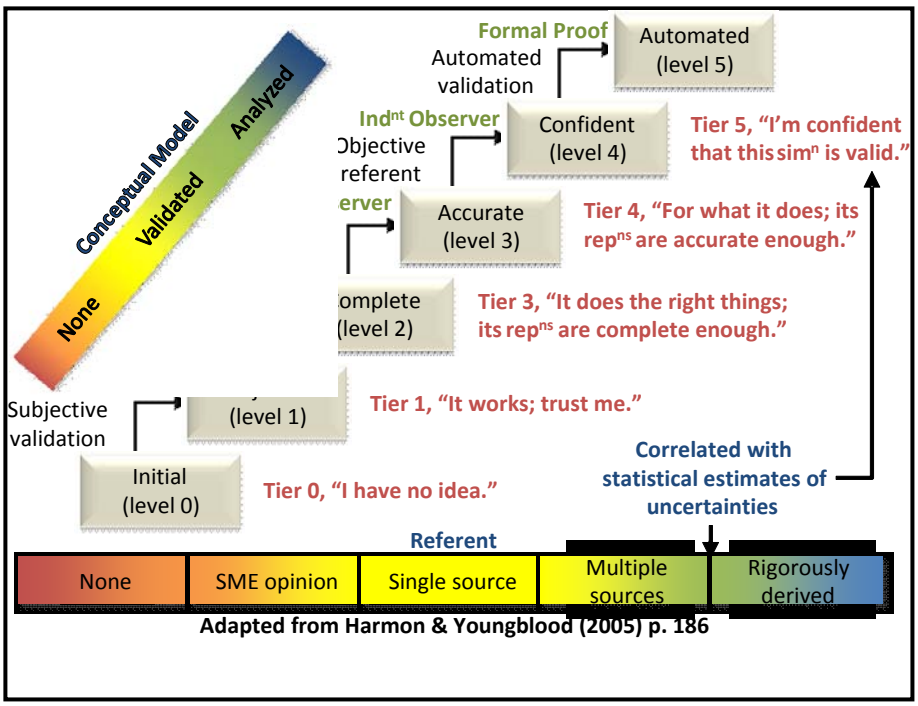


Figure 3 Validation Process Maturity Model – Levels of Validation Process

3.1.2 Tiers of Validity Assessment

Harmon and Youngblood propose six tiers of validity assessment based on the type of information upon which the assessment is made and the type of validity statement that can be made as a result of that assessment. The data available in the validation criteria, referent, and Conceptual Model and the analyses made in the results validation reflect the supporting information tied to these statements of validity.

At Tier 0, no validity statement is issued since there is no information available upon which to base an assessment. This is the base case. A Tier 1 assessment, usually based on subject matter expert (SME) opinion, is one of whether the simulation worked. A Tier 2 assessment is one of representation; that is, it is a static assessment of whether the simulation represents the necessary entities and attributes. At Tier 3, the completeness of the representation and the behavior of the simulation are assessed; that is, it looks to the dynamic representation and the relationships between elements in the simulation. Tier 4 is an assessment of whether the representations [and the

resulting behaviors from those representations] are sufficiently accurate. Lastly, Tier 5 is an assessment of [statistical] confidence in simulation results.

3.1.3 Validation Process Levels

The six levels of the VPMM proposed by Harmon and Youngblood are tied to improvements made in the available data and analyses applied to the specified information artifacts. As the VPMM progresses through the levels the specificity and support for the data improves as does the objectivity and rigor of the applied validation processes. The VPMM levels as given in (Harmon and Youngblood 2005) are described in the following paragraphs.

Level 0 (Initial) provides a baseline and is included for completeness. It is applicable to those simulations having no validity assessment.

Level 1 (Subjective) is an informal assessment of a simulation's validity using SME opinion for determining the validation criteria and referent and face validation of the simulation results. At this level of validation, no Conceptual Model is created. Generally, the validity assessment is provided as a statement with little or no supporting documentation (i.e., Tier 1 validity assessments).

Level 2 (Complete) has validation criteria that describe the required entities, properties and dependencies and a referent derived from SME opinion. SMEs validated the Conceptual Model and simulation results using the validation criteria and the development products are verified against the Conceptual Model. At this level, validity assessments are made at both Tier 2 and Tier 3 since both the states and the behaviors are evaluated. At Level 2 the validation process increases the objectivity of establishing the simulation's validity requirements.

Level 3 (Accurate) refines the validation criteria by including acceptable ranges of input, domains of output, and error tolerances; improved the referent by having a source other than SME opinion; and uses an objective party rather than a SME to assess the validity of the Conceptual Model and the simulation results. Referents used to support the accuracy assessments needs to be kept separate from those used to support the validation criteria selection and the simulation development. At this level, validity assessments are made at Tier 4. A Level 3 validation process improves the objectivity of the results assessment.

Level 4 (Confident) makes further improvement to the process by including required confidence levels in the validation criteria and uses multiple, independent, correlated sources for the referent to include estimates of uncertainties. The development products included sufficient verification to guide sampling space selection for results validation. An objective party completes the Conceptual Model validation and provides analysis to suggest sampling space selection for results validation. An objective party also completes the results validation using guidance derived from the Conceptual Model validation and the development verification. At this level, validity assessments are made at Tier 5. A Level 4 validation improves the referential information used in the

accuracy assessment, by incorporating multiple referent sources and confidence information.

Level 5 (Automated) moves to a mathematically formalized structure and automated assessments. Achieving this level is dependent on the development of underlying mathematics to enable its use and on the trust that the users have in the results of the process's application.

3.2 FRAMEWORK APPLICATION RISK ASSESSMENTS

A survey was given the participants to assess both the maturity of the process used to assess the validity of the models that illuminated the ability of the oral presentations to meet the needs of various audiences, and obtained additional audit information. The questions related to the maturity of the process were developed from a comprehensive survey developed by Scott Harmon in his VPMM, graciously provided to this study team as a courtesy. The audit questions were a subset of the questions asked by the audit team.

While the answers to the survey were sparse and inconsistent with many participants declining to provide a survey response in total or to many of the questions, some feedback was provided. The following gives the feedback provided in the surveys. Preceding the main feedback items are brief summaries of the comments. Following the feedback items is a recommendation (if identified by the ABSVal Team) for the ABSVal Framework based on the feedback provided.

(1) Although some participants were able to identify the validation and accuracy criteria used during the validity assessments, for others, the criteria applied was unclear.

Comments:

(a) Criteria identified were not (nor intended to be) comprehensive; variations in output were not examined [with respect to their effect on accuracy].
(b) Criteria were not linked to the constraints, limitations, and assumptions of the model and their effects on the intended use – analysis. They were limited, solely, to the model's capabilities or representation. Questions that should be addressed in this realm for analysis are
  (i) Is the model credible within its limits?
  (ii) What are the limits and do they match well?
(c) It seems unreasonable to expect that a positive validity assessment can be achieved without validation criteria present and explicitly identified. Without the potential of acceptability, then why proceed? Although accuracies were not statistical, rationale for assessments were clear.
(d) Although criteria were at times addressed, their application to the assumptions testing was not clearly applied.

Recommendation: Validation reports and presentations should make the criteria applied during the validation analysis clear, even if, or especially, if the validator develops these criteria or these criteria are subjective. The assessment methods for these criteria must also be clear.

(2) The participants overall found the referents unclear, unspecified, or poorly matched to the applications. The participants, in general, found that the P-OCM had a referent (data) and that the P-COIN model did not have a referent (no data). Although, some participants recognized that the P-COIN model referent was SME-based.

Comments:

(a) "Referent does not apply to assumptions testing"
(b) "No *successful* V&V methodology can be overly dependent on an accessible and 'apt' referent"
(c) "No referent available of the sort that could support results validation"
(d) "Moderately useful – but missing data confounds"
(e) "Mapping SME responses to aggregate behavior is tough. The referent being human behavior, it would be quite difficult to [identify] acceptable behaviors, etc. I wish I had some suggestions to offer, but I have none."

Recommendation: The role of referent in validation is critical. The referent is the best knowledge about the system being modeled and simulated; it is the thing against which simulation results are compared. The lack of a referent based in data does not remove the referent from the validation cycle. Without data, referents might become the validator's or developer's best knowledge (however good or poor that might be), but the referent does not disappear from the validation problem. The referent coupled with the needed validation and accuracy criteria for the intended use drive the validation assessments. As the ABSVal Framework is more fully developed, the role of the referent and ways to describe the referent, especially in the absence of observational or empirical data, needs to be more clearly specified. Although there were few comments on the Conceptual Model question in the survey, the responses indicate similar development is required in understanding Conceptual Models.

(3) Results validation requires a referent to against which assess the results, and any counter-intuitive results need to be investigated and understood.

Comments:

(a) While SME assessments of results are valuable, additional simulation case runs should also be run (to determine the effect on the results [to get at robustness and unexpected behaviors]).
(b) The validation process must also examine the methodology to achieve results (e.g., assessment measures of effectiveness, use of statistics, and design of experiments) for sound practices and appropriate

applicability for the intended use (i.e., analysis). The analysis question is also under scrutiny in the process.

(c) Analysis data generated during a study should be saved for posterity: "1st Rule: Don't discard, misplace, or ignore *any* data."

(d) Improve research on the input data [referent]

(e) More sensitivity analysis

Recommendations: Results validation has been a challenge throughout the course of this study in developing the ABSVal Framework. The problems the Study Team were approaching in building this process did not lend themselves to fully fleshed out referents with validated results data against which to compare simulation results. Discussions during the workshops tried to address this problem, with the usual recommendation of face validation. The comments provided give insight into the way ahead for results validation in this domain. The key to results validation lies in describing expected results under specified conditions (e.g., what the face validator would be looking for). The ability to collect this data limits the number of assessments that could be collected; however, designs of experiments can be used to target data collection. First order interactions and input boundaries can provide first targets. Core capabilities, sensitivity areas, and interesting cases can provide additional areas for SME data collection. Results need to be explainable and understood by the analyst.

3.3 RISK MEASURES

Conveying the risk of using a model and its simulation results was the primary purpose of validation. Early in the ABSVal Framework Study process, the Team developed some notional graphics (Figures 4 through 6) for displaying risk:

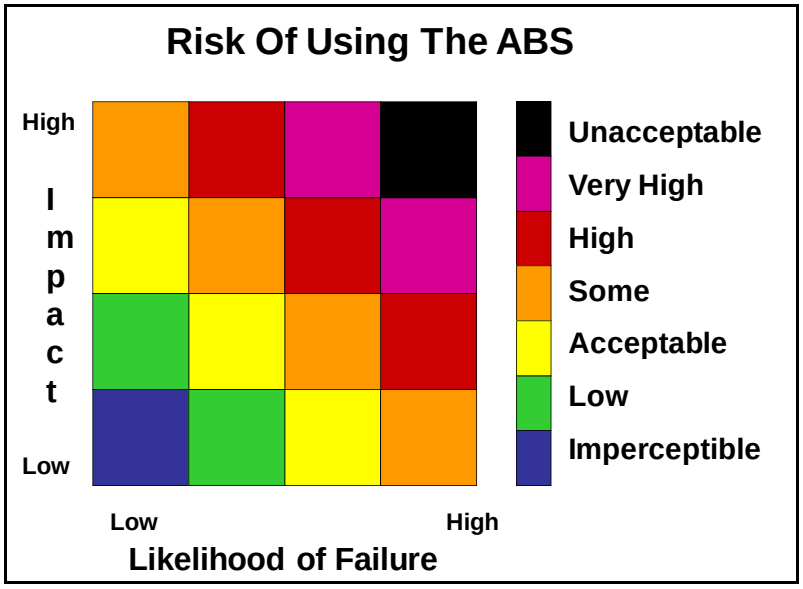


Figure 4 Notional Risk Assessment (Scaled from Low to High)

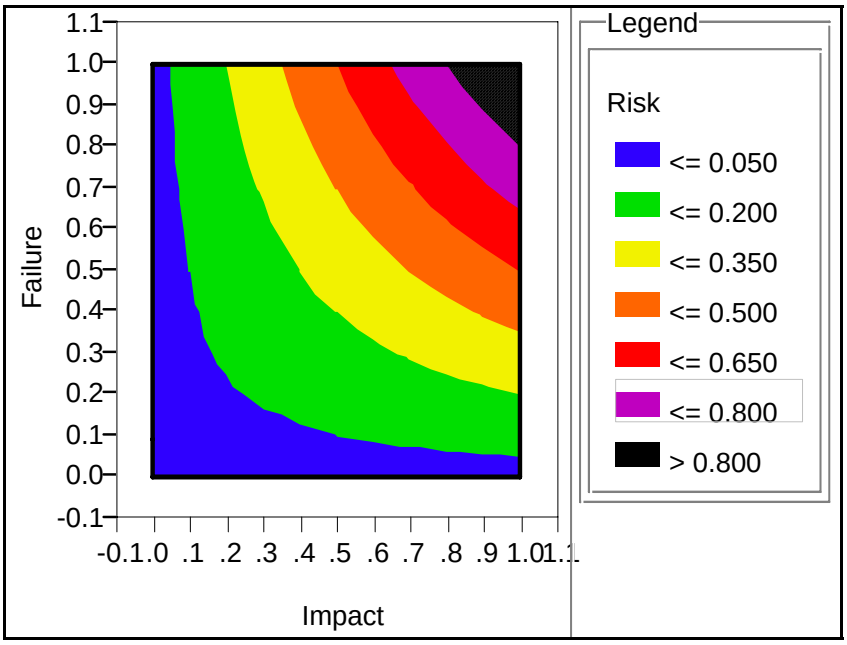


Figure 5 Notional Risk Assessment (Using Probability Assessments)

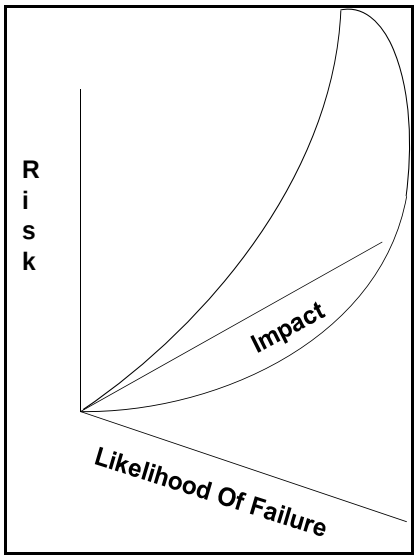


Figure 6 Notional Risk Assessment (3-Dimensional)

3.3.1 Improved Risk Measurement

Each of the validation efforts assessed risk at some point during the process (although these assessments did not make it into the reports against which the audits were based). Some of these assessment efforts resulted in the revised graphic, shown in Figure 7, which appeared in the P-OCM report and the P-COIN Conceptual Model report.

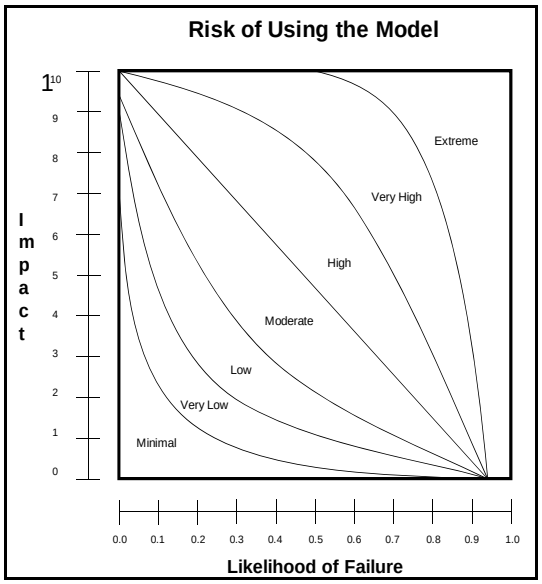


Figure 7 Risk Measurement

Several of the risk graphic's features are notional. For example, the x-axis intercept assumed the supported decision-maker would be unwilling to use a model if he or she were told its likelihood of failure were at least 93%, regardless of the potential impact level of the supported decision. While we doubt many decision-makers would set the threshold any higher, some might set it lower. We expect the risk assessment tool to be refined during post-contract work as part of academic endeavors.

3.3.2 Additional Work on Risk Assessment and Communication

The determination of risk and its communication to the users of models and the consumers of their simulation results was a critical area in the validation process that still requires additional work. The VPMM gets to one aspect of risk by assessing the process, including the types of data available to support the model and evaluation of its results. In addition, work during the ABSVal Framework Study identified risk as a function of error and consequence as shown in Section 3.3.1. Error was a function of the accuracy of the model and its simulation results (according to a specified measure). Consequence was a function of the intended use. Errors in the core elements of the model or resulting for the absence of needed model interactions that cannot be otherwise mitigated indicated a higher risk level. Additional research into the development of risk assessment methodologies in support of validation assessments was planned (see Section 5.4).

4 TEST CASE WORKSHOPS (TASK 3)

An integral part of the research effort supported by the Agent-Based Simulation Verification, Validation, and Accreditation Framework Study (Phases I and II) is the sharing of the work with the broader community. This is done through sponsored workshops and publications. Section 5 discusses the publication plan; this section discusses the workshops.

4.1 WORKSHOP #2

At the first workshop in Phase II, held 1-2 October 2007, Workshop #2, the study team presented the ABS validation framework to the larger DoD, industry, and academic community. The workshop was divided into three main sessions: a plenary session, working sessions, and an out-brief session. The plenary session included discussions of the workshop's purpose, an introduction to the Pilot ABS Validation Methodology, and briefings on the theory of validation, the framework for validation, application of the Pythagoras model to Counter-insurgency (COIN) Operations, and the Study's Phase II approach. The participants were encouraged to provide comments and suggestions on the framework, cautions and caveats on its application, and other potentially useful information. The results of this workshop are included in Appendix B.

4.2 WORKSHOP #3

Workshop #3 was held at the Northrop Grumman, Fair Lakes facility in Fairfax, VA on 25-26 March 2008. The purpose of the workshop was to present the initial results of the ABS validation framework's application to members of the DoD, academic, and industry community for critical evaluation of the validation efforts to date in light of the developed framework. The workshop was small so that participants focused attention on the mid-term results of the validation efforts. Comments during the workshop were significant to the direction of the validation framework. While the agenda was loose, it covered the following topics:

1) Overview of the framework
2) Description of the Pythagoras modeling environment
3) Description of the ABS used in the validation framework tests
4) Preliminary results of the validation experiments applied
5) Mid-term assessments of the framework
6) Critical evaluation of the validation experiments and the validation framework tests

The results of this workshop are included in Appendix C.

Discovered in the workshop, a frequently overlooked aspect of ABS validation is the *intended use* of the model. While the accuracy requirements for a simulation in terms of precision, level of detail, and fidelity of the real world representation embedded in the model usually consider the intended use of a simulation, the specifics of *using a model in a specific application* are often overlooked. That is, validation efforts will frequently

attack the Conceptual Model of a simulation without assessing whether the identified deficiencies are core elements to the application at hand, for instance, to answer the posed analysis question, or if these so-called problems with the model could be mitigated in some way, say, through sensitivity analysis. This was the key insight gained in Workshop #3: Validation in an analysis application is an analysis of the analysis. This insight required a mid-course correction of the P-COIN validation effort in the full application of the ABSVal Framework.

In addition, the Naval Post Graduate School (NPS) attended the ABSVal Workshops and prepared a report with recommendations following Workshop #3 (in Appendix D). Some of their observations and recommendations are as follows:

(1) A software platform like Pythagoras is not validated (rather its code is verified to work correctly) since there is no control over the platform's use.
(2) A validation toolkit should consist of tests to "prod" a model in an attempt to invalidate it. For example, probe the end-points of the input parameter space or develop robust statistical designs within the input parameter space to assess if the model behaves as expected.
(3) Investigate the extant simulation validation literature for tests and concepts that might be applicable to this specific validation context.
(4) Investigate how hypothesis testing concepts could be applied directly to the validation paradigm.

Lastly, in the section of the NPS report titled, "Credible Uses of Unvalidated Models," an argument is made that models that are not valid can be used in some contexts. The question becomes of appropriate use of the model "in situations in which the verisimilitude of the model is difficult to assess." This highlights the major recognition of this study: that all validation efforts need to take into stronger account the *intended use* of the model and not just concentrate on the capabilities of the model and its representational accuracy of the model with respect to the real world. For the purposes of this study, the intended use was analysis. Therefore, the validation assessments need to be done with respect to the models' support to the analysis efforts.

4.3 WORKSHOP #4

This last workshop was held at the Northrop Grumman, Fair Lakes facility in Fairfax, VA on 8-10 July 2008. During this workshop a final outbrief of the ABSVal Framework Study was made to the larger community. The validation test cases and framework were presented, and there was significant discussion on needed elements for the framework.

The context of the workshop was the presentation of the results from the study with the intent to get final commentary from the larger community when presented with the results of the efforts. The discussions underscored the insight obtained during the previous workshop. In particular, discussions centered on the need to identify the "core" elements of analysis and the methodology for using the model and its simulation results to support that analysis when engaging a validation effort. Failure to explicitly address the needs of the analysis will result in a failure of the validation effort to support the

intended use of analysis no matter how strong the capability assessment of the model. In particular, validation efforts need to explicitly explain the effect of its findings on the ability of an analyst to use the model and of the decision maker to trust the results. This is not a “trust/don’t trust” assessment. Rather, it is an explanation of the effect of the finding on the analysis. Deficiencies and assumptions along with their effects must also be discussed in validation reports that support analytical uses. Three questions were posed as useful in the context of assessing the validity of M&S in analysis uses:

(1) Does the model represent what it is advertised to represent?

(2) Are the limitations of the model explicit?

(3) Does the analysis (using the model as a tool) answer the analytical question(s)?

Collateral to number three above is an assessment of the analytical question and assessment methodology and whether either changed or ought to change as a result of the validation findings.

The full Workshop #4 summary report can be found in Appendix E.

5 PUBLISHABLE DOCUMENTATION AND REPORTS (TASK 4)

As a research project, sharing and communicating with the broader analysis, validation, and ABS communities was a large part of this study. The workshops held as part of this project as well as the publication of materials to the community website supported this goal. In addition, early in the project, an initial Publication/Media Plan was developed. The plan provided some suggested research areas arising from the Agent Based Simulation Verification, Validation, and Accreditation Framework Study with team member interest areas identified when possible, as well as some potential paper topics and publication venues.

Since the development of that publication plan, several team members have submitted papers or presentations to some of the conferences identified in the plan. In addition to these submittals, future publications will include the results of the ABSVal efforts towards validating the two ABS applications, as well as a critique of the framework and suggestions for the framework's improvement. Abstracts for the current submittals are provided below.

5.1 THE DIFFICULTIES WITH VALIDATING AGENT BASED SIMULATIONS OF SOCIAL SYSTEMS

Authors: L.J. Moya

Conference: Agent-Directed Simulation Symposium (ADS'08) Part of the 2008 Spring Simulation Multi-Conference (SpringSim'08), 14-17 April 2008, Ottawa, CANADA

Abstract: In previous work, we discussed the various definitions of validation and how these definitions apply to validation of agent based simulations (ABS) [1] and began to identify objectives of an ABS validation framework [2]. While results validation and face validation are often used methods for validating simulations, the difficulties with this approach for simulations having sensitivity to initial conditions, or chaotic/emergent effects, and the difficulties with validating human based representation models is well known. A methodology is needed that will provide effective validation for these complex behavior models. To this end, the Marine Corps Combat Development Center (MCCDC) Operations Analysis Division (OAD) commissioned an Agent Based Simulation Verification, Validation, & Accreditation Framework Study to develop general, institutionally acceptable processes and criteria for assessing the validity of agent-based simulations used as part of DoD analyses. This paper describes the some of the issues with validating ABS and describes the framework developed as part of this study for validating these types of models for analysis applications.

Status: Delivered

5.2 CLARIFYING VALIDATION FOR AGENT BASED SIMULATIONS

Authors: L.J. Moya & S. Youngblood

Conference: 2007 Fall Simulation Interoperability Conference, Orlando FL

Abstract: Validation has long been recognized as critical to the credible use of any model and simulation. The U.S. Department of Defense, Department of Energy, American Institute of Aeronautics and Astronautics, American Society of Mechanical Engineers, and the U.K. Ministry of Defense among others, all have a definition for validation. While these definitions differ slightly in language and application, each has three main components: the model, the thing being simulated, and a set of bounding principles. However, even with these common concepts within the definitions, these terms are not widely understood. In a recent workshop on the validation requirements for agent based simulations in military applications, it became clear that elaboration of validation terms was required before the definition could be applied to develop an agent based simulation validation framework. This paper discusses validation in terms of the prevailing definitions and suggests ways to interpret the validation definition for agent based simulations.

Status: Delivered

5.3 A VALIDATION FRAMEWORK FOR VALIDATING AN IRREGULAR WARFARE (IW) SIMULATION USING PYTHAGORAS

Authors: L.J. Moya

Conference: 76th Military Operations Research Society Symposium, 10-12 June 2008, Coast Guard Academy, New London, CT

Abstract: The Marine Corps Combat Development Command (MCCDC) Operations Analysis Division (OAD) is considering simulation models for future entry into the USMC Irregular Warfare (IW) Analytic Baseline. One simulation paradigm under consideration is agent based simulation.

Agent based simulations present a challenge for validation in support of analytic applications, especially in the realm of Irregular Warfare IW, due to scarcity of data and simulation complexity. To evaluate the feasibility of agent based simulation in analytic applications the MCCDC OAD funded Phase I of the U.S. Marine Corps (USMC) Agent-Based Simulation (ABS) Verification, Validation, and Accreditation (VV&A) Framework Study with Phase II funded by the Modeling and Simulation Coordination Office (M&SCO). The purpose of Phase I was to create a framework for performing VV&A on models. The study's primary effort was on the validation process with verification and accreditation addressed with respect to their interdependencies with the validation process. Phase II of the study elaborated on the framework and tested it against the Pythagoras implementation of a Columbian IW scenario for an analytical application as a proof of concept. The result from this study is a transparent, traceable, and reproducible methodology of validating these simulations based in the scientific method. This briefing gives an overview of the framework using results for the Columbian IW scenario to illustrate the methodology.

Status: Delivered

5.4 FRAMEWORK FOR UNDERSTANDING SIMULATION ANALYSIS REQUIREMENTS

Authors: L.J. Moya

Venue: Dissertation, Old Dominion University, College of Engineering; various conferences

Abstract: Validation is widely recognized as critical to the effective use of modeling and simulation. There are three key elements to the validation definition: real world, accuracy, and intended use. While in math, physics, and engineering applications where the intended use of the modeling or testing of system validity is well understood with respect to all three of these definitions, in other domains and other applications methods for validating a model and assessing the risk of its use are less understood. There is guidance in the Verification, Validation, and Accreditation Recommended Practices Guide and many techniques have been identified (e.g., boundary value analysis). However, missing from the literature is a definitive guide for determining which techniques to be used in specific circumstances. Further, there is little to help the user assess the risk of use based on the findings of a validation analysis. Recent work has highlighted the specific needs in validating a model's use in analysis application. This work gives an overview of necessary elements to include in the validity assessment of a simulation's use in an analysis context.

Status: In development

5.5 USING AN IRREGULAR WARFARE (IW) WARGAME TO FRAME A PYTHAGORAS SCENARIO AND ISSUES

Authors: R. Marling, R. Clinger, P. Rossmaier, B. Sheldon, and E. Bitinas

Conference: 76th Military Operations Research Society Symposium, 10-12 June 2008, Coast Guard Academy, New London, CT

Abstract: The Marine Corps Combat Development Command (MCCDC) Operations Analysis Division (OAD) IW study team selected the agent-based simulation Pythagoras as its primary modeling tool. To help frame the Pythagoras scenario, we conducted a Colombia wargame. The wargame highlighted areas of interest and critical decisions made that were later incorporated into a Pythagoras scenario for detailed analysis.

The emphasis of our briefing will be on capturing the essence of an IW wargame, translating the concepts from the wargame into Pythagoras inputs, and exploring the added value this technique offers.

Status: Delivered

5.6 DEMONSTRATION OF IRREGULAR WARFARE (IW) PYTHAGORAS MODELING SUITE

Authors: R. Marling, E. Bitinas, and B. Sheldon

Conference: 76th Military Operations Research Society Symposium, 10-12 June 2008, Coast Guard Academy, New London, CT

Abstract: This demonstration will include the latest version of Pythagoras and its supporting tool suite, i.e., Pythagoras 2.0, the Rapid Scenario Generation (RSG) tool, and the Design of Experiments (DOE) tool.

Pythagoras is an agent-based modeling environment, providing the user with a host of optional capabilities, rules, and behaviors to describe an agent. The new capabilities that it introduces include soft decision rules, dynamic sidedness, behavior-change triggers, non-lethal weapons, and variable attributes. Variable attributes, new to version 2.0.0, can be used to trigger new behaviors, and can be changed by weapons, communications, events or the terrain itself.

The RSG tool reduces the time required to develop an executable scenario file through the reuse of developed and approved simulation objects. The intent of this effort is to develop a generic front-end scenario development tool that might be used with any number of simulation models.

The DOE tool reduces the development and execution time for computational experiments that involve large numbers of factors by providing a generic front end interface to guide the analyst through the construction of an experimental design and facilitate that design execution in a high performance computing (HPC) environment.

Status: Rejected

5.7 SENSIBLE VALIDATION FOR IW SIMULATIONS

Authors: M. Bailey and V. Middleton

Conference: 76th Military Operations Research Society Symposium, 10-12 June 2008, Coast Guard Academy, New London, CT

Abstract: The Marine Corps, supported by other government, agencies, academia, and industry, has led a year-long effort to tailor the concept of simulation validation to analytical applications used to explore irregular warfare issues. The result of this effort has produced a set of guidelines that require actions from model developers and from analysts using models – specifically the mapping of essential elements of analysis to their associated representations in the simulation and its data. We provide a vernacular and a structured framework for exposing the adequacy of a simulation to support a typical irregular warfare analysis task, and give guidance concerning the assessment of positive and negative validation results.

Status: Delivered

5.8 MODEL VALIDATION AND SIR KARL POPPER: WHAT THE OLD AUSTRIAN CAN STILL TEACH US

Authors: R. Eberth

Venue: Phalanx

Abstract: Sometimes we have to step back in order to move forward.

Our industry has for several years been trying to come to grips with the problem of validating a class of models that has been widely held to be impossible to validate – agent-based simulations (ABSs). One of the attractive traits of ABSs is that they can display collective behaviors that are remarkably similar to what we observe in the real world. That same trait, however, constitutes "emergent behavior" – results that the model users could not predict a priori and that could not be expected to be repeated in subsequent iterations. Traditional validation techniques seemed unable to deal with such models.

Dr. Mike Bailey of MCCDC's Operations Analysis Division, with OSD funding, sponsored the 2007-08 "ABSVal" study to develop a new framework for validating ABSs. Debates on alternative techniques waxed long and eloquent among study team members and study workshop participants. The effort even included developing a new theory of simulation validation. But the study team eventually realized that we really needed to look back to a very old dialectic in the philosophy of science.

The step back is to 1919, and the setting is Vienna, Austria, shortly after the fall of the Austro-Hungarian Empire. Europe was swirling with revolution in several scientific spheres – Einstein's theory of relativity, Marx's theory of historical materialism, Freud's theory of psycho-analysis, and Alfred Adler's theory of individual psychology. Sir Karl Popper (1902 – 1994), then a young student, found himself in a philosophical quandary. He had flirted with Marxism, and had worked briefly as an assistant to Adler. He also, however, had joined a group of students deeply fascinated by Einstein's work, work that had just gained an enormous boost that same year by an experiment that readily could have refuted his theory of gravitation but instead confirmed it. While he found himself hugely impressed with the confirmation of Einstein's work, Popper also found himself increasingly unsatisfied with the other three theories. His issue was not with their truth per se, but – compared to the Einstein experiment – with how every observation seemed predestined to confirm their "truth" in their respective fields. It became apparent to Popper that a theory that could not be refuted was made not stronger but weaker thereby. He termed such theories "pseudo-scientific."

Ultimately, Popper developed a "line of demarcation" between science and pseudo-science. That line is *falsifiability.*

Popper's falsifiability criterion defines modern scientific method, and differentiates it from the "old" scientific method that relied almost exclusively on inductive reasoning.

The parallels to our validation problem were striking. Traditional validation techniques are largely exercises in inductive reasoning – pseudo-science. *Fortunately,* they wouldn't work for us and caused us to look around – and back. We saw we needed to apply modern scientific method. We didn't need to prove or even demonstrate that an ABS is valid for some particular application. We "only" needed to try, hard, *to falsify the hypothesis that it was valid for that application*. And if we couldn't falsify that hypothesis, then we would have strong evidence that it could be *accepted as valid for that application.*

The "ABSVal" study already has demonstrated that modern scientific method can work very effectively to assess the validity of an ABS for a particular application. The larger question for the community, though, is whether it should be applied across-the-board. Should we listen to the old Austrian? Should science trump pseudo-science in validity assessments? I hope the answer is obvious.

Status: In Progress

5.9 SOME COMMENTS ON MODELS

Authors: E. Visco

Venue: TBD

Abstract: Not available

Status: In Draft (paper available in Appendix G)

6 SUMMARY

The key insight gained by this project was that the validation of models in support of analysis resides within the analysis process itself. That is, validation cannot be decoupled from the analysis plan, process, and results. Validation in this intended use identifies the limitations and boundaries of the analysis itself, lending clarity to that process for the recipients of the simulation and analytical results.

6.1 ANALYSIS CONTEXT

Critical to the validation of a model and its simulation results in an analysis application lies in understanding the decision context for the analysis. This includes the core elements for supporting the analytical decisions as well as the decision's relationship to the influencing elements within the modeling environment. The validation process may uncover dynamic elements that need to be addressed more completely, mitigation techniques that ought to be taken or additional test cases that should be included within the analysis. A critical part of this is assumptions testing.

It is not enough to identify limitations in the model. These limitations must be explicitly linked to the analysis context and discuss the risk to using the model within that context. A conclusion of "Not Valid" is an unacceptable conclusion from the developer / analyst point of view without conclusive proof that the analysis is flawed. A conclusion of 'High risk' is acceptable, but should include suggestions on how the analysis can be improved, either through a change in the modeling approach, the use of another or different model, or additional analysis cases to be tested to reduce risk and increase confidence.

6.2 RESULTS VALIDATION

Results validation has been a challenge throughout the course of this study in developing the ABSVal Framework. The problems the Study Team were approaching in building this process did not lend themselves to fully fleshed out referents with validated results data against which to compare simulation results. Discussions during the workshops tried to address this problem, with the usual recommendation of face validation. Exercising the framework and comments from the workshop participant has provided insight into the way ahead for results validation in this domain.

6.2.1 Referents

The role of referent in validation is critical. The referent is the best knowledge about the system being modeled and simulated; it is the thing against which simulation results are compared. The lack of a referent based in data does not remove the referent from the validation cycle. Without data, referents might become the validator's or developer's best knowledge (however good or poor that might be), but the referent does not disappear from the validation problem. The referent coupled with the needed validation and accuracy criteria for the intended use drive the validation assessments. As the ABSVal Framework is more fully developed, the role of the referent and ways to describe the referent, especially in the absence of observational or empirical data,

needs to be more clearly specified. Conceptual Models also require better specification for a useful validation process.

6.2.2 Methods

While results validation is not generally accessible in the traditional manner as that of math, engineering, and physics based models, it is still possible. It, however, relies on eliciting expected outcomes for specified test scenarios and the creation of these test experiments. The key to results validation lies in describing expected results under specified conditions (e.g., what the face validator would be looking for). The ability to collect this data limits the number of assessments that could be collected; however, designs of experiments can be used to target data collection. First order interactions and input boundaries can provide first targets. Core capabilities, sensitivity areas, and interesting cases can provide additional areas for SME data collection. Results need to be explainable and understood by the analyst. That is, results validation requires a referent against which to assess the results and any counter-intuitive results need to be investigated and understood.

While SME assessments of results are valuable, additional simulation case runs should also be run to determine the effect on the results to get at robustness and unexpected behaviors. The validation process must also examine the methodology to achieve results (e.g., assessment measures of effectiveness, use of statistics, and design of experiments) for sound practices and appropriate applicability for the intended use (i.e., analysis). The analysis question is also under scrutiny in the process.

Validation reports and presentations should make the criteria applied during the validation analysis clear, even if, or especially, if the validator develops these criteria or these criteria are subjective. The assessment methods for these criteria must also be clear.

6.3 REPORTS

Validation reports should clearly describe the qualitative questions investigated and tests conducted. These descriptions should include the reasoning for the investigation and tests, as well as expected and desirable results. It should also state necessary conditions for “passing” the tests for the intended use (i.e., validation criteria) and conditions that would constitute a “failure.”

The process of making qualitative assessments would be more objective with less reliance on the subjective opinion of the validator if the process were more structured, expectations were identified, and if the reasoning for the validators conclusions were made explicit.

The significance of the results of all validity assessments (qualitative and quantitative) on the intended use and application of the model needs explicit explanation. That is the effect of failing the test on the intended use needs to be clear.

The report should identify mitigation strategies: in use (e.g., sensitivity analysis), in corrective development, or in model improvement.

6.4 FRAMEWORK

The framework should include the minimum sets of information and materials sufficient to reach a final validity determination using the ABSVal Framework.

A practitioner's guide to validation resulting from the ABSVal Framework should include rules, stepwise approach, branches, and criteria for determining necessary or sufficient information to support a given conclusion. Rules might include necessary information and model and simulation artifacts to embark on a validation investigation. The availability of material might influence the risk of use.

6.5 RISK

Conveying the risk of using a model and its simulation results is the primary purpose of validation. However, the determination of risk and its communication to the users of models and the consumers of their simulation results is a critical area in the validation process that still requires additional work. The VPMM gets to one aspect of risk by assessing the process, including the types of data available to support the model and evaluation of its results. In addition, work during the ABSVal Framework Study identified risk as a function of error and consequence. Error is a function of the accuracy of the model and its simulation results (according to a specified measure). Consequence is a function of the intended use. Errors in the core elements of the model or resulting from the absence of needed model interactions that cannot be otherwise mitigated may indicate a higher risk level. Additional research into the development of risk assessment methodologies in support of validation assessments is needed.

6.6 CONCLUSIONS

The Study resulted in significant modifications to the ABSVal framework developed in Phase I. In Phase II of the Study, through the application of the framework, it was discerned that an analysis-centric approach focusing on the intended use was essential to validation and was subsequently integrated into the framework. The result was a framework that provided a foundation for ABSVal that is a useful academic tool, which ideally could be vetted through further application to a broader set of ABSs. Notionally, this would allow some of the framework improvements implemented in the latter stages of the Study (especially to mitigate concerns from the audit team and the developer) to be applied and tested. A shortcoming of the Study in Phase II was the lack of suitable model application pairs on which to apply the framework, specifically applications displaying emergent behavior or whose intended use was primarily focused on an analysis decision and not on exploring the concept of ABS itself. A more robust and mature set of model applications would allow the full ABSVal framework to be tested in a more comprehensive manner. Also, it was apparent through both spirited discussions at the workshop forums and the audit recommendations received that the topic of

validation, like the ABS field of study itself, is ripe with emerging ideas and concepts that will provide a dynamic platform for future framework refinements.

APPENDIX A LIST OF ACRONYMS

AAV	Amphibious Assault Vehicle	MCCDC	Marine Corps Combat Development Center
ABS	Agent Based Simulation	MCWL	Marine Corps Warfighting Laboratory
ABSVal	Agent Based Simulation Validation	MEB	Marine Expeditionary Brigade
BZ	Beach Zone	MEU	Marine Expeditionary Unit
C2	Command and Control	OAD	Operations Analysis Division
CLAs	Constraints, Limitations, and Assumptions	ONR	Office of Naval Research
COA	Course of Action	P-COIN	Pythagoras Counter Insurgency
COIN	Counter Insurgency	POC	Point of Contact
DoD	Department of Defense	P-OCM	Pythagoras Obstacle Clearing Model
DoE	Department of Energy	RPG	Recommended Practices Guide
FARC	Revolutionary Armed Forces of Colombia	SEAS	Synthetic Environment for Analysis and Simulation
GOVT	Government (Colombia)	SME	Subject Matter Expert
HA/DR	Humanitarian Assistance/Disaster Relief	SZ	Surf Zone
IW	Irregular Warfare	USMC	United States Marine Corps
LPP	Littoral Penetration Point	V&V	Verification and Validation
M&SCO	Modeling and Simulation Coordination Office	VPMM	Validation Process Maturity Model
MAGTF	Marine Air-Ground Task Force	VV&A	Verification, Validation, and Accreditation
MANA	Map Aware Non-Uniform Automata		

APPENDIX B WORKSHOP #2 SUMMARY

This workshop report describes presentations and feedback on the Agent-Based Simulation (ABS) Verification, Validation, and Accreditation (VV&A) framework as it was during the workshop. Since the workshop, many of the ideas have been refined.

B.1 WORKSHOP #2: OVERVIEW

The Marine Corps Combat Development Command (MCCDC) Operations Analysis Division (OAD) hosted a two-day workshop on a pilot validation methodology for Agent Based Simulation (ABS) on 1-2 October 2007. Northrop Grumman sent workshop invitations to members of Government, industry, and academia. The goal was to present the process, principles, and the desired end state for developing a validation methodology for ABS. The workshop had a total of 51 participants from Government, industry and academia (excluding the Study Team and Study Sponsor) with a broad-based set of experience in both the academic and applied modeling areas. Approximately 15 of these participants also attended Workshop #1 (during Phase I of the Study), held in May 2007.

The workshop's purpose was to present our ABS validation framework to the attending practitioners and academics, attempt to apply the framework to other ABS simulations, and learn from that process. The workshop was divided into three main sessions: a plenary session, working sessions, and an out-brief session. The plenary session included discussions of the workshop's purpose, an introduction to the Pilot ABS Validation Methodology, and briefings on the theory of validation, the framework for validation, application of the Pythagoras model to Counter-insurgency Operations (COIN), and the Study's Phase II approach.

In addition to formal briefings in the auditorium, the workshop included working sessions on three topics. Discussion topics and session leaders were:

Topic 1: Constructive Critique of the Framework. [Ms. Lisa Jean Moya and Mr. Edd Bitinas (both on the Study Team supporting OAD) each led one working group]

Topic 2: Invalidation Tools and Techniques [Ms. Lisa Jean Moya and Dr. Eric Weisel led one working group; Mr. Edd Bitinas led the second working group (all session leaders are on the Study Team supporting OAD]

Topic 3: ABS Application Pairs [Dr. Michael Bailey, USMCCDC/OAD, led the session]

The workshop agenda was designed to encourage maximum collaboration and discussion, allowing workshop participants to provide input on each topic and gain knowledge from other participants. The session topics were addressed sequentially; topic 1 was addressed in the afternoon of the first day and the remaining topics were covered on the second day.

For each working session, participants were split into two groups to address a single topic with different session leaders. Working groups switched session leaders after

each session. After 75 minutes of discussion, all participants reconvened for a group discussion on the topic. During this discussion, the session leaders summarized the key points made during the working sessions.

B.2 PLENARY DISCUSSIONS

This section contains a summary of the workshop discussions on the plenary session topics.

B.2.1 Pilot Validation Methodology for Agent-Based Simulations Workshop: Where Are We? (Presented by Dr. Michael Bailey, Deputy Director, MCCDC OAD)

Dr. Bailey presented a brief introduction to the workshop and provided a recap of OAD's ABS study activities for participants who may not have attended the first workshop on this topic. Dr. Bailey defined Agent-Based Simulation (ABS) for the purposes of this workshop as a simulation that produces surprises and emergent behavior and one that focuses on Irregular Warfare applications. Dr. Bailey explained that he wants to use ABS to provide analysis and to provide information to decision makers. Dr. Bailey characterized validation as answering the question of whether a simulation is useful for answering the analytical question. He expressed that his goal was to develop a methodology and the basis of declaring a simulation inappropriate and defining the framework for analysis. He also stated that users/consumers must match the analysis with the simulation, which requires carefully defining the analytical question. Lastly, Dr. Bailey wants to identify ABS simulations that have successfully supported analysis.

B.2.2 Introduction to the Pilot ABS Validation Methodology (Presented by Mr. Edd Bitinas, Study Team Support to OAD)

Mr. Bitinas began his presentation by addressing the goals of the Framework and its desired result – that the Framework should be applicable to all models and simulations, and that it was specifically developed for ABS and irregular warfare (IW) applications. He also discussed what is missing from the current verification and validation process with respect to irregular warfare.

In his discussion, Mr. Bitinas stated that there are several types of model/simulation validation. Expected value, physics-based simulations are verifiable through experimentation, and random effects introduce predictable error. In stochastic, probability-based models, the distribution of model outcomes matches the distribution of observed outcomes. The model-generated and observed distributions are identical if they cannot statistically be proven otherwise. In ABS, the agents attempt to replicate, at least in part, the human decision making process. However, agents are limited in that humans may have more information than the agents may have, humans may include emotions and experience, humans may think/plan ahead, humans may anticipate the actions of others, and two humans, given the same information, may make different decisions. Thus, traditional validation may not be meaningful for ABS.

Another complication is that traditional models can be validated for a class of problems (e.g., campaign models), but ABSs are not necessarily models of anything in particular. Rather, the user assigns agent behaviors and capabilities in ABS, via input, for a specific application.

Mr. Bitinas stated that an ABS may be valid for a specific application over a limited range of inputs if the decisions it makes could be made in real life and/or if the emerging complex behavior can be traced to a realistic root cause or causes. However, an ABS is not valid if one can prove that it is invalid. Trying to invalidate an ABS for an application (and failing to do so) may result in lower risk in using the ABS for the application.

The Framework's goals are to determine the required accuracy and, more specifically, the sufficient accuracy for the intended ABS application. The Framework must also apply techniques to attempt the invalidating of models because validation itself may not be possible. The Framework must also establish the boundaries of validity because an ABS may be applicable to only a portion of the intended use. The Framework also seeks to ensure that validation is not a resource-intensive process. The validation process must be transparent, traceable, reproducible, and communicable to ABS model users and/or consumers.

B.2.3 Theory of Validation (Presented by Dr. Eric Weisel, Study Team Support to OAD)

Dr. Weisel began his presentation by addressing some basic questions in simulation science: What is simulation and what are its basic structures and the properties of those structures? How do simulations relate to ABS validation? What's the composability – if you build a simulation from parts and pieces, how does that work? He identified the objective of simulation science as developing useful theorems about simulations using the foundational sciences of mathematics, computability theory, logic, model theory, and systems theory.

Dr. Weisel then provided a survey of the theoretical framework for modeling in which a model is a computational function and a simulation is a method for implementing the model over time. The simulation is represented by a deterministic labeled transition system.

There are three key elements embedded within the U.S. DoD validation definition: accurate, real world, and intended use. We want to have confidence (or lack thereof) that our model represents the "real world." Since we cannot prove validity, we must rely on the scientific method to build confidence/assess risk. The validity of composition models must show that the simulation's relation exists for the composition of models in an ABS; that is, how the whole model works when you put the pieces together.

Figure 8 represents the Validation and Verification Continuum. In this continuum, the natural system is a real thing. To develop a model/simulation, one must abstract from

the natural world to get the ideal simulation. To accomplish this, we capture what we care about in the natural system/real world to develop the Conceptual Model.

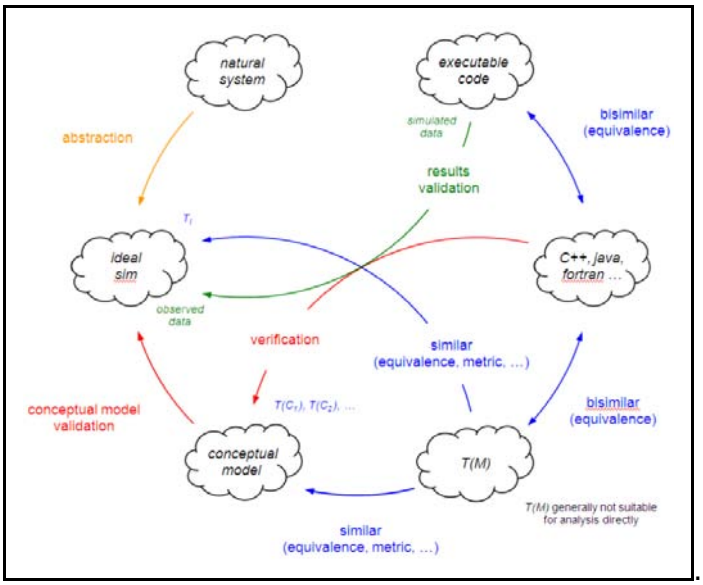


Figure 8 The Validation and Verification Continuum

Dr. Weisel indicated that one way to validate results is if you have observed data. A second way to verify results is to validate the linkages between the coded model and the Conceptual Model and between the coded model and the natural system. Since you cannot prove validity to the idea simulation, you can apply the scientific method to build confidence and assess risk of using the model. To apply scientific method, the validation authority must take one of two paths: assess the risk of a Type II error in application or prove the null hypothesis (i.e., these results are assumed valid until proven otherwise if one cannot invalidate them).

B.2.4 Framework for Validation (Presented by Ms. Lisa Jean Moya, Study Team Support to OAD)

Ms. Moya began her presentation by reviewing various definitions, including DoD's definition, for validation. Ms. Moya reminded the audience that DoD requires verification and validation (V&V) of military simulations by order of DoD Directive (DoDD) 5000.59 and DoD Instruction (DoDI) 5000.61. DMSO/MSCO provides guidance in their Verification, Validation, and Accreditation (VV&A) Recommended Practices Guide (RPG). While the last of these documents provides high-level descriptions of validation techniques and processes for various stages of simulation development (e.g., legacy and new development), the first two documents address mainly the responsibilities associated with V&V. None really address the "how" of V&V or the criteria for determining V&V sufficiency. The current version of the framework still lacks methods for validating ABS, with its specific validation challenges. One of the break-out sessions had as its purpose to eliciting potential techniques from the participants.

Three main areas are important for consideration when validating ABS: the Conceptual Model, the thing being simulated, and a set of bounding principles. Ms. Moya stressed that architecture of the ABS, such as the decision rule-set, elements must be assessed during the validation process. Two pieces of particular importance are the Conceptual Model and results validation as shown in Figure 9.

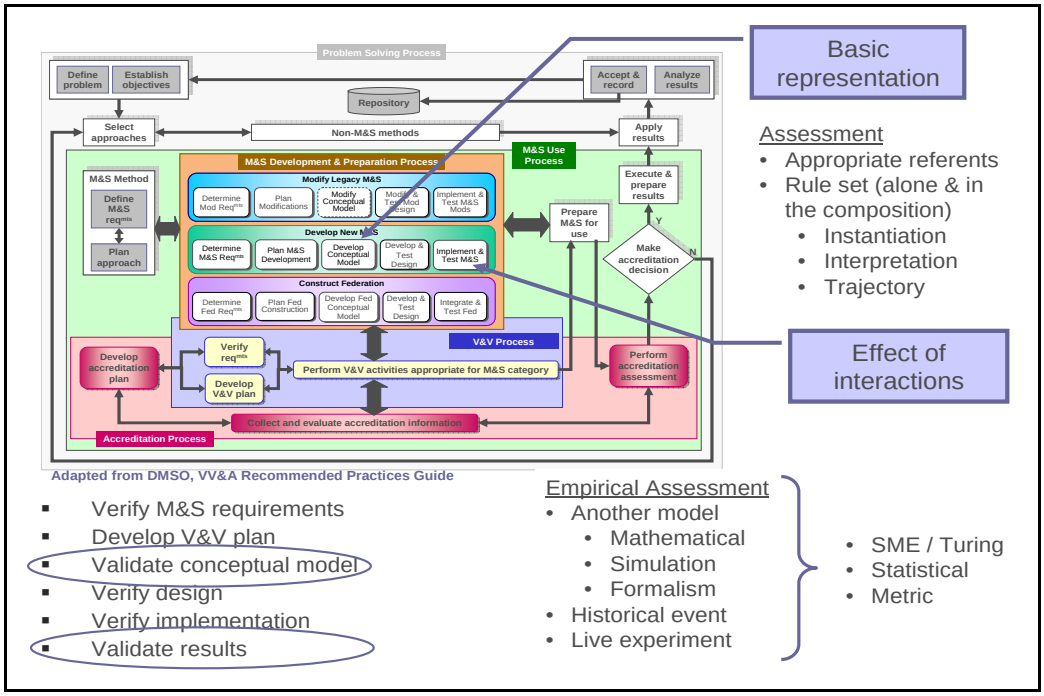


Figure 9 General Validation Process (Adapted from (Defense Modeling and Simulation Office 2004a))

Ms. Moya indicated that it is important for ABS to validate the Conceptual Model, the simulation results, and the interactions between agents and the agents with their environment. She stressed that due to one-to-many mapping inherent in ABS, multiple rule sets can be used to model the same system. Thus, one cannot decouple the results validation from validation of the Conceptual Model.

In physics-based modeling, the Conceptual Model validation is inherent in the acceptance of the mathematical equations by the scientific community. The desired level of accuracy in the solution determines the selection of solver algorithm and difference equations. Results validation of this type of model uses empirical data, experimental testing, predictive capabilities, and acceptable error tolerance. With agent-based models, there is little empirical data, so validating authorities must evaluate the Conceptual Model design, the knowledge base, the behavior engine and knowledge base implementation, and the integration with the simulation environment.

Ms. Moya postulated a spiral methodology for assessing the validity of an ABS. In this methodology, the validating agent assesses the risk of using the ABS for the specific/ intended use, communicates that risk to the consumer of the results of the ABSVal process, and applies the scientific method using validation experiments.

Given resource constraints in any model development activity, validating agents should conduct a cost-benefit trade-off to determine techniques that will invalidate the ABS quickly, or will significantly reduce risk in using the ABS for the specific/intended use. If the null hypothesis is rejected at any point, the ABSVal process is done. If the null hypothesis is not rejected, some decreased degree of risk can be conveyed to the consumer. If the null hypothesis is not rejected but the ABSVal performer does not have a high degree of confidence in the validity of a given piece of the model, the ABSVal performer can attempt to use another technique to invalidate that particular piece and/or can convey a higher level of perceived risk to the consumer. By communicating the level of perceived risk in an ABS after failing to invalidate it, the consumer is afforded a means to assess the ABS's applicability to hard-to-quantify, non-traditional areas or activities, such as Irregular Warfare (IW).

B.2.5 Phase II – The Way Ahead (Presented by Mr. Edd Bitinas, Study Team Support to OAD)

Mr. Bitinas addressed the planned Study Phase II, in which the Phase I-developed pilot framework for VV&A of Agent Based Simulations (known as ABSVal) will be applied to model applications being considered for future entry into the USMC Irregular Warfare Analytic Baseline. Phase II goals include: testing the viability and utility of the pilot ABSVal framework in a realistic institutional setting, evaluating ABSVal in a seminar setting combining communities of ABS users and developers, developing methodologies for applying ABSVal to future ABS development efforts, and producing informational products useful for the M&S community.

When testing the viability and utility of the pilot ABSVal framework, the Study Team will determine whether the Framework is useful and complete, and whether improvements can or should be made to the Framework. After selecting candidate ABS models, the Study Team will determine specific applications by examining whether the model results will be meaningful/useful, whether the model has identifiable limitations, and whether there are work-arounds to those limitations. The Study Team will demonstrate techniques that can invalidate the candidate ABS models, and will document everything to produce information products useful for the M&S community. During this process, the Study team will expand and/or modify ABSVal framework, as required.

Mr. Bitinas asked for audience assistance to identify candidate ABS–application pairs. Candidate models can have been used in completed studies, may be used in on-going studies, or planned for future studies. He also asked workshop participants to identify invalidation techniques the Study Team can use.

B.2.6 Preview of Pythagoras COIN (Presented by Mr. Edd Bitinas, Study Team Support to OAD)

Mr. Bitinas began his presentation by describing the Counter-Insurgency Operations (COIN) scenario for Irregular Warfare to which Pythagoras will be applied. Pythagoras will be used to model population dynamics and the influence of various actors on population segments – those actors include the Marine Air-Ground Task Force

(MAGTF) Commander's Courses of Action (COAs) and insurgency actions. Mr. Bitinas explained that population segments were broken up into five orientation sectors: insurgent, pro-insurgent, indifferent, pro-COIN, and COIN. The population distribution within each segment may change over time. He also indicated the scenario is based in Colombia.

Mr. Bitinas provided details of the scenario, known as Operation Pacific Breeze. The scenario involves a Humanitarian Assistance/Disaster Relief (HA/DR) operation after a volcanic eruption, earthquake, and tsunami on Colombia's coast. A Marine Corps deploys a Marine Expeditionary Unit (MEU) and a Marine Expeditionary Brigade (MEB) with the mission of carrying out HA/DR in response to the tsunami. A Colombian insurgent group, the Revolutionary Armed Forces of Colombia (known by the acronym FARC), is predicted to take advantage of the unstable situation, and the Marines will make every effort to prevent FARC activities in the area. The Marine commander must decide between two courses of action: minimize the footprint ashore by having Marines conduct HA/DR operations during the day and return to ship at night, or deploy ashore to provide 24/7 HA/DR operations. The measure of effectiveness for the simulation will be increase or decrease in insurgent activity and support.

B.2.7 Sample Methodology Approach (Applying the Framework to Pythagoras COIN) (Presented by Mr. Edd Bitinas, Study Team Support to OAD)

To apply the ABSVal Framework to Pythagoras, Mr. Bitinas examined whether the model is built correctly. In terms of model results, the validation questions are whether the citizens in the model change affiliations, whether this change is at the expected rates, whether the courses of action (CoAs) change the affiliation rates, and whether this change is in the expected amounts. Another question to consider is whether the "narrative paradigm" is applicable. For this study, validating the theoretical model may not be possible; in this case, the Study Team will assume the theoretical model is valid. To validate the model for this COIN application, reviewers must also consider whether the data are reasonable by examining the data's source, exploring whether real-world examples are available, and deciding if those real-world examples are relevant to the specific application.

Validating authorities must also consider whether all the known interactions are present in the model. If interactions are missing, the question is whether they can be added to the model easily. Validating authorities must also consider whether there are any interactions in the model that are inappropriate to the specific analytical question; if so, those interactions should be removed from the model, if possible. Finally, validating authorities must examine and identify the model's limitations and any bounds on the inputs.

Several possible invalidation techniques are available, including assumption testing, black box testing, Turing test, results validation, and comparison to other models, such as the prototype Excel model (much less functionality) and JAVA model (less functionality). Mr. Bitinas suggested some possible guidance for users and/or decision makers when validating the model. The first suggestion is to use Design of

Experiments, performing data farming on all variables. The second possible option is to identify chaos points, if any exist. If small inputs make great changes in the model's outcome, the model may not be valid near these points. The third possible option is to break the problem into pieces, using different assumptions for different population segments.

B.2.8 Constraints for V&V of Agent Based Simulation: First Results (Presented by Dr. Andreas Tolk, Old Dominion University)

Dr. Tolk began his presentation by addressing two caveats when considering the broader perspectives on agents and their application domain. The first caveat is that the real power of these models (and money) lies in the command and control (C2) market, particularly providing C2 M&S services, providing support to operations, conducting analysis for the warfighter in his headquarters, and performing decision support functions. The second caveat is that agents are more than human behavior. From the system perspective, everything that acts can be modeled as an agent. Moreover, agents are not simple, but can be as complicated as traditional simulation systems. In sum, agents are more than just human behavior; they represent aggregate systems. Given these caveats, the V&V framework needs to address more than just human behavior.

Figure 10 depicts Dr. Tolk's concept of agent verification and validation. Dr. Tolk described the chain of quality today as transitioning from data to information to knowledge to awareness, and suggested that if developers can bring systems into net centric environment, they can make the leap directly from information to knowledge. To do this, developers need more behavior based on C2 to increase the decision process for behaviors.

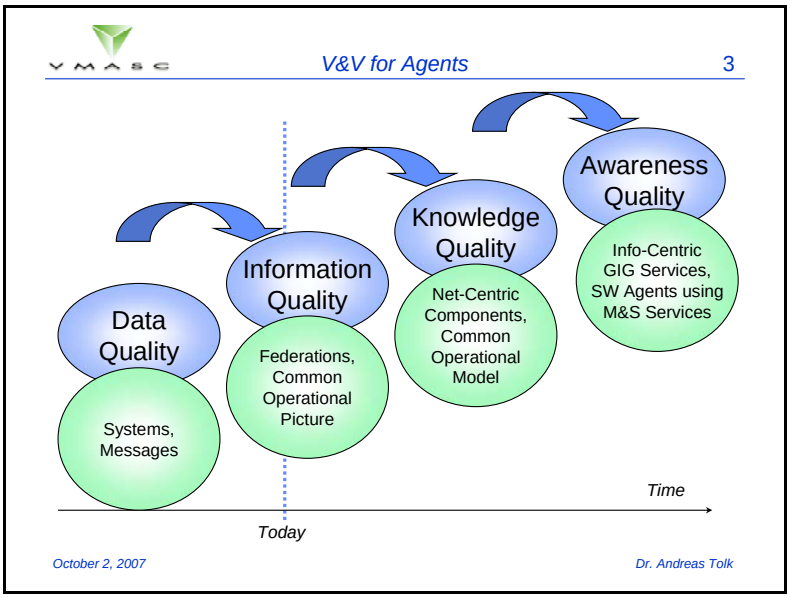


Figure 10 Verification and Validation of Agents

Dr. Tolk states that an agent and its environment are systems, involving a data model (input, output, and controls), a process model (functions), and a behavior model (mode and state changes). Agents are systems-of-systems, which require V&V authorities to implement a cascading V&V process to conduct:

- V&V for the agent components,
- V&V for the agent
- V&V for the agent in the environment
- V&V for the agent within the population
- V&V for the agent within the population in the environment
- V&V for the agent population in the environment

A multi-layered cascading framework is necessary because classical V&V methods fall short regarding these requirements.

B.2.9 ABMs and the Validation Hurdle - An Illustration (Presented by Mr. Paul Wehner, MITRE)

Mr. Wehner described the project he and his colleague, Dr. Alfred Brandstein, are working on for the Department of Homeland Security (DHS). The project involves evaluating immigration control measures along various portions of the U.S. border.

Mr. Wehner described models as a representation of reality, and stated that validation is a judgment of the model in terms of accuracy for its intended use. He acknowledged that resolving the accuracy problem is easy to say but difficult to do, particularly for ABS because it is an application area that does not lend itself to easy description or direct evaluation. Since ABMs are often used to explore the causal relationships of complex interactions or to help analysts in their efforts to develop bounded solutions to these challenging application areas, evaluating the 'appropriate' degree of accuracy (in the context of intended use) is problematic. Direct evaluation or observation of the targeted-application area may yield discrete results; however, the ability to consider those results as representative is at best limited due to the degrees of freedom involved.

Mr. Wehner suggests that can clear the traditional validation hurdle; however, one must evaluate whether the ABMs capable of doing so are sophisticated enough to explore the type of questions key to our interests. If an ABM's primary measure of success is an evaluation of whether it can and has been used to demonstrate (both within the community and externally) the need for and the consequence of changes, then "accuracy" is likely the wrong measuring stick. Rather than force-fit that construct, Mr. Wehner and his colleagues propose applying the following two criteria to determine if an ABM should be deemed valid:

- Is every model outcome possible?
- Is every possible outcome realizable by the model?

The Defense Modeling and Simulation Office recommends that validation activities are ongoing activities performed as part of a model/simulation's overall development process. While Mr. Wehner and his colleagues also contend that the degree of user (i.e., customer) involvement and the transparency of the representation must increase as a function of the perceived complexity of the target-application area; otherwise, the

validation hurdle becomes increasingly more challenging, if not insurmountable for ABMs.

Expected model outcomes include the ability to support discussion, debate, and decision making. What makes the agent-based modeling approach unique is its transparency - its ability to capture interactions provided by subject matter experts and the ability to visualize immediately what is happening. Typical uses include demonstrating the consequences of various actions and exploring means of encouraging or discouraging various outcomes. The hallmark of these model types is that they can produce non-intuitive results in a manner that convinces users that their intuition needs to be honed.

Mr. Wehner said that one should be hesitant to draw firm conclusions about the probability of occurrence for one or more outcomes. This capability, much like the capability to predict with any certainty how an individual or groups of individuals will react to an event, is largely beyond our reach. Moreover, especially in "complex" situations, there is no rigorous method available to determine sampling distributions to determine these probabilities.

B.2.10 Tools & Techniques (Presented by Ms. Lisa Jean Moya, Study Team Support to OAD)

Ms. Moya described a variety of techniques that could be used in agent-based models validation experiments. Some of these techniques are depicted in Figure 11, although this list is not exhaustive. After briefly discussing the list, workshop participants agreed that Risk Assessment Techniques might be a more appropriate title than Invalidation Techniques.

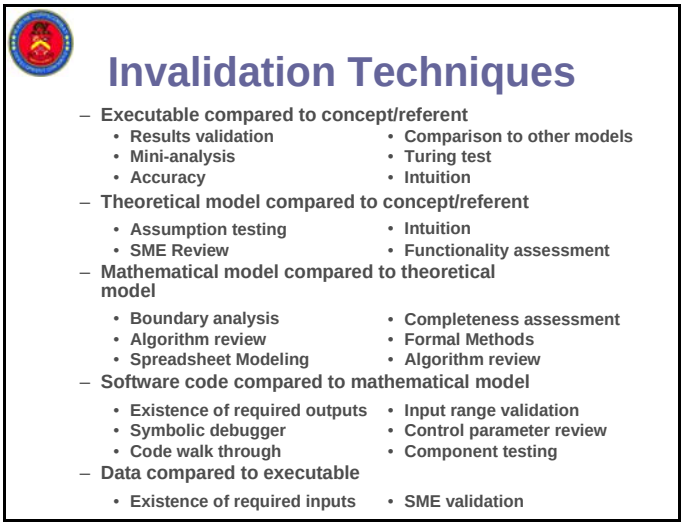


Figure 11 Invalidation/Risk Assessment Techniques

B.3 TOPIC DISCUSSIONS

This section summarizes the workshop discussions on the three topics of interest.

B.3.1 Topic 1: Constructive Critique of the Framework

B.3.1.1 Group 1

Participants of Ms. Moya's workshop suggested the Framework must describe the possible decision contexts to which the model can be applied in terms of the necessary elements, what other models the results might feed, required accuracy, inputs and outputs, and what part of the ideal simulation is being represented. One participant recommended that the model context should be part of model's meta-data. Participants also suggested that users must understand how the model should not be used (i.e., are certain contexts inappropriate for model application?) and what its limitations are. There appeared to be general consensus that excluded uses should be validated, as well.

Participants spent some time discussing basic scenario descriptors and drivers of the rate of change for the population segments represented in the COIN scenario. Participants suggested that data farming and sensitivity analysis could be used to validate assumptions Pythagoras developers used to derive rates of affiliation and how those rates change.

Some discussion ensued about the COIN model's population categories and whether these categories are mutually exclusive; that is, whether actors can be part of more than one group. Participants suggested that developers must justify their parameters and theory about population and individual behavior in terms of values, relationships, and effects.

During a discussion of cause and effect, participants suggested the Framework must explicitly describe the underlying assumptions, map those to the decision context, and understand the effect on the simulation's results.

One audience member suggested that the validating authority must settle on the context for the question being asked and the range of acceptable answers and/or results. He suggested users must define where they are, where they want to be, and the limits of acceptability for model results. Other audience members did not agree with this suggestion because ABMs provide non-intuitive results and the limits of acceptability might not include those results.

Participants also discussed how the model would represent the environment. One USMC participant was especially interested in the model's level of environmental detail and whether the model included "second order" effects (e.g., logistic routes and infrastructure problems). This participant opined that if second order effects are not explicitly modeled, emergent effects cannot be observed. He further stated that if you omit the environment, users cannot ascertain the relationships between action and the environment (particularly whether the environment is causing effects). In this case, he stated, a systems dynamics solution may be a more appropriate tool than an ABS.

Several participants mentioned that the Framework must match the decision context to the model, identify expected and unexpected cause-effect relationships, and test/validate those cause/effect relationships. Some discussion ensued about the level at which those relationships should be validated. Participants suggested evaluating

cause/effect relationships at the agent level under simple circumstances before evaluating them against more complex circumstances; that is, stimulate the agent to validate the agent.

Workshop participants also discussed how to describe the agent – what does it mean and how you interpret it over time (e.g., attributes). They addressed the necessary level of abstraction and how to interpret those abstractions (weighting, averaging). Ms. Moya suggested that having valid agents is necessary but not sufficient for validating the model. Participants indicated that model developers must provide decision makers a description of the agent (level of abstraction) and explain how to interpret results and agent behavior. They also agreed that developers must show decision makers traceability of why agents (e.g., the population in the COIN scenario) are supporting the insurgency. They must provide the context for why decisions are being made, conduct sensitivity analysis to describe the results, and build a story of the results, including how to interpret results and how rules are triggered in the model (agent interaction description).

Ms. Moya mentioned that the Conceptual Model can be constructed in multiple ways. She asked participants whether results validation is meaningful if empirical data does not exist, and suggested that users would need a source of confidence given the lack of empirical data. One audience member suggested conducting some wargaming to support validation. If game results are similar to modeling results, and if gamers believe model results are realistic, the model is valid. Another participant suggested using a consensus technique between simulations to see what is important. He indicated that two modeling approaches will both model pieces incorrectly; where they overlap, however, could be important.

Ms. Moya asked participants to consider the case in which model results are compared against a historical event. If the model produces one result out of 100 runs that matches the historical event, can we conclude the model is valid? Audience consensus was that the validating authority should examine the other outcomes to see which came close to replicating the historical event and how close it came. If, in that one case, the validating authority can trace why that one occurrence happened, this could lead to validity. The audience suggested examining the near misses to see how close they came to replicating the historical event, but agreed that standards for “close” would have to be established beforehand.

B.3.1.2 Group 2

Participants in Mr. Bitinas’s workshop discussed the fact that design patterns may affect results. For example, if a model always instantiates agents in the same order (e.g. alphabetic or numeric), it may give different results than if the agents are instantiated randomly. For this reason, the validation Framework should consider model design.

Participants also discussed sensitivity analysis’s role in understanding the effect of model parameters. This can be done using Design of Experiments (DoE) to determine the effect of changing input variables and the interactions between variables. Using DoE will also highlight the result space, giving insight into outliers and areas where the parameters are not valid.

There also was some discussion about aggregation within models. When objects are aggregated, sometimes low-level object characteristics that the analyst overlooked are lost in the aggregation. If the analyst is not interested in the low-level details (and therefore, not modeling at that level), it may not matter that these details are missing. However, overlooking these characteristics might matter if the result is that the model fails to capture the cause and effect relationships of these characteristics.

The participants also discussed the traceability of agent behavior to inputs and interactions. A participant suggested that model variable meaning should be "anchored" and then users/developers/validators can trace how the results developed from those inputs.

Component-level validation was also discussed because it is possible that individual components are valid but the way they interact is not valid. In this case, integration testing is necessary.

Participants recognized some difficulty determining whether a model is valid for the timeframe it is using. For example, if the (validated) model data is represented in seconds of time but the analyst is modeling decades of time, then the model may not be valid. The validation needs to be in the same time and space as the model.

When emergent behavior results from a model, analysts must determine if the behavior is a true surprise, an artifact of how the model is constructed, or a software bug. Participants agreed that emergent or unexpected behavior is not invalid just because it is counter-intuitive to the subject matter experts.

B.3.2 Topic 2: Tools and Techniques

B.3.2.1 Group 1

Participants in the session led by Mr. Bitinas discussed why the term "invalidation technique" was used instead of "validation technique." Participants agreed this was a semantic issue because one can apply techniques to validate a model and then use the same techniques to invalidate it. Moreover, one cannot validate using the scientific method, but one can invalidate it.

Model bias was addressed by the audience with general agreement that every model has some bias. The challenge is to understand the model's bias and mitigate its effect on model results. The Framework must address model bias and ways to overcome it, and users must know how to interpret the results properly given the model's bias.

Participants expressed concern that the modeling and simulation community has never properly addressed bias. One suggestion was to use a chi-squared statistical test between the mathematical model and the theoretical model. Another suggestion was to use more than one set of historic data to see if/how the data sets drive the results.

An effort must be made to ensure that validation methods are applied to the model's different levels of aggregation, not just at one level.

One participant highlighted the fact that model developers only code in the model the relationships they understand. By definition, all other relationships are omitted. There

is often an effort to employ “reduction” techniques but, in this instance, “expansion” techniques are more appropriate.

Participants also discussed the importance of sampling, especially when reviewing results and determining what outliers are important and what they mean. Participants agreed in the necessity to determine the set of interactions that caused the outlier, identify a probability of outlier occurrence, and document it. Participants agreed the analyst has a responsibility to communicate to the user the probability of the outlier and the reason behind it.

Participants were asked what technique(s) they would employ given a constraint of resources/time. One participant said he would select results validation against some referent, although he acknowledged that it is possible the results were hard-wired into the model. Another participant said his approach would depend on the tool/model. Some tools belong to the model developer and others belong to the accreditor/evaluator after the fact. If the accreditor/evaluator does not own the model or have intimate access to it, validation becomes harder and more expensive. Another participant said he would focus on evaluating whether the model was executable compared to the concept/referent and whether the theoretical model matched the concept/referent (see techniques for these evaluations in Figure 11 above). In the verification process, there must be traceability and documentation at the coding level in order to conduct validation.

One participant said that intuition should not be included in the Framework because ABSs generate “emergent behavior,” which is often counter-intuitive. Some audience members disagreed, saying intuition is a type of pattern recognition so one can validate using intuition. Intuition should be a flag for further testing. The group consensus was to include intuition in the Framework.

B.3.2.2 Group 2

Audience participants in the session led by Ms. Moya discussed how SME qualifications play into model validity. Audience opinion was that as long as developers use an excellent SME, validators must accept expert opinion. The audience discussed SMEs and how developers decide whom to use as experts. Because ABS often involves modeling without observable data, developers must rely on expert opinion. Audience members discussed ways to scientifically capture SME participation by “accrediting” SMEs and their intuition. The audience agreed on the value of carefully identifying SMEs, documenting their expertise, and providing that to users/consumers and validating authorities. Participants also suggested tying expert opinion to the model’s intended use; that is, ask the experts the same question being asked of the simulation. They also recommended using observed data and expert opinion to evaluate whether model data accurately represented the real system. Audience members indicated that, when using expert opinion, it helps to write their opinion down before the experts see the model results. In the participants’ opinion, user confidence in the model will increase as developers use higher level data and more/better SMEs.

B.3.3 Topic 3: ABS-Application Pairs

Two workshop participants agreed to review their models as prospective ABS application pairs. The first was Dr. Debbie Duong from Office of the Secretary o Defense (OSD), Program Analysis and Evaluation (PA&E). The second participant to review a model was Mr. Dave Holdsworth of Alion, who supports USSOCOM (SORR-J8-S).

B.3.3.1 Nexus

Dr. Duong presented her model called Nexus, which she stated is using a scenario that is similar to Pythagoras COIN. Nexus has the following characteristics:

One agent per social group acting with a group mind

All agents have access to past events and the feelings they engendered, and know how much each other agent supports its group

Computed group support operates as a loop

Dr. Duong used a Necker cube structure in the model. She believes this structure is appropriate because it is based on the narrative paradigm in which one identifies important neural networks and the model recognizes cognitive dissonance and associated rationalization. She suggested that Nexus could be used to address the type of analytic questions for which Pythagoras is being applied, such as the perception of the population's support for the government or an insurgent group.

When asked how developers/users would know whether the model had enough layers of representation, Dr. Duong indicated there is no way to know but that the model could be adjusted to make recent events more important in determining group affiliation than past events. When asked how this model could be validated given its inherently qualitative nature, Dr. Duong responded that neural network models are robust and deal with issues of trust that users intuitively understand. She indicated neural network modeling had explanatory power because people understand these ideas.

B.3.3.2 MCO-1 CAT

Next, Mr. Holdsworth presented the Major Combat Operation-1 Catastrophic Study (MCO-1 CAT) model for discussion. Figure 12 replicates the diagram he drew during his discussion. Mr. Holdsworth indicated the model's goal was not to find emerging behavior but to see how checkpoint operations behave in the MCO-1 model compared to checkpoint operations in the real world, such as in Iraq.

Mr. Holdsworth indicated that the model cannot replicate everything in event/behavior chains, but can replicate stylized situations in the behavior chains (analytic baseline).

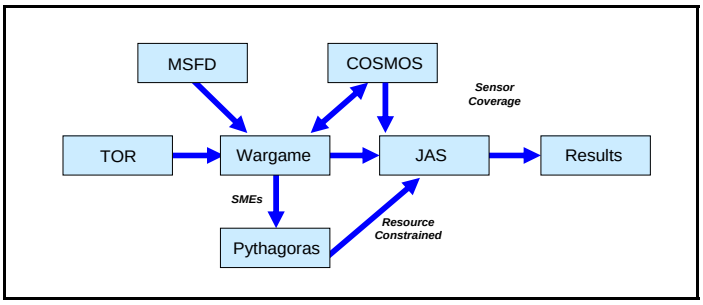


Figure 12 MCO-1 CAT

Mr. Holdsworth indicated that the use of experimental design should not be limited to changes in parameter values. Although time consuming to implement, changing context, such as target types, type of checkpoints, level of HUMINT cueing, type of terrain (restricted or open), with/without false positives, may be of equal or greater value than just changing parameter settings up or down. One issue regarding model validation is changing the right set of inputs sufficiently to convince oneself that the model is valid or good enough. A robust model's conclusions may be insensitive to changes in sensor values, response times, or other parameter settings.

Models used for training rather than testing may have different validation needs. This topic has been discussed by the Study Team, and should fall under the initial step in the framework, which includes determining "how good is good enough (accuracy)."

The use of data farming techniques, making many runs that span the trade space using design of experiments techniques to reduce the total number if needed, may be the shift in the analysis paradigm that has been anticipated. The days of a few subject matter experts selecting the six or ten cases that would be run in a highly detailed model over the next year may be replaced by three or four formulations of the same problem using different models, run thousands of times across the landscape of possibilities in a few weeks, and then analyzed both for similarity and for uniqueness (outliers). It is the outliers that may be of interest, since their emerging behaviors may be the key to further analysis. Outliers should not be simply discarded.

B.3.4 Collaborative Internet Environment

Livelink has been less than successful. Audience suggestions for replacing Livelink included MSIACS Community of Interest, and Groove. Subsequent discussions by the Study Team have determined that pbwiki hold the most promise. As a result, the Study Team constructed http://orsagouge.pbwiki.com/ABSVal.

B.3.5 Other Topics

Additional topics that generated brief participant discussion were the need for a conference specifically dedicated to Irregular Warfare (there is a MORS-sponsored workshop to be held in December 2007 at the Naval Post-graduate School), and the introduction of self-educating or evolutionary algorithms into our framework.

B.4 KEY POINTS MADE DURING WORKSHOP #2

This section contains the Study Team's summary of key points made during the workshop.

B.4.1 Validation of ABS Has Inherent Challenges

Based on discussions during this workshop and our own in-depth analysis, assessing the validity of ABS has inherent challenges not possessed of other simulation paradigms. In the case of many ABS, and certainly in the COIN application, ABSs model human behavior and interactions among people with different motives, perspectives, and behavioral norms. There is little doubt, if any, that modeling these interactions is especially challenging. Model validation is particularly problematic because little or no observable, empirical data exists to support modeling these interactions; when data does exist, it is often perishable and not necessarily representative of the entire system, negating it broad applicability for traditional results validation.

B.4.2 Validation Framework Based on Scientific Method

The Framework is based on the scientific method; that is, it is based on gathering evidence subject to specific principles of reasoning, collecting data through observation and experimentation, formulating and testing hypotheses, and correcting and integrating previous knowledge. In this case, the validation process first proposes a hypothesis that a model is valid enough (sufficiently accurate) for its intended or specific use. Next, the process develops appropriate experiments to attempt to invalidate that hypothesis. Upon application of those experiments, the process either conclusively demonstrates that the model is invalid for that application, or the process fails to invalidate the model. Upon failure to invalidate, the relative strength of the experiments applied to the model is used to assess the risks of using the model for its intended purpose, which are weighed against the importance or impact of the intended application.

As is required by the scientific method, these steps will be repeatable in order to predict future results dependably. The process will be objective to reduce biased results interpretation. The Study Team will also document, archive, and share all data and methodology so it is available for scrutiny by members of the M&S community and others. This full disclosure will allow model developers, academicians, researchers, and model users to verify results by attempting to reproduce them. Documentation will be especially important to ABSVal Framework activities because the dearth of empirical data necessitates heavy reliance on subject matter expert opinion.

The Study Team's goal is to improve the ABS modeling process by applying a rigorous validation framework. As evaluators/validators identify areas of concern within a specific ABS-application pair, model developers and users/consumers will be driven to try to address and resolve these concerns. This may involve a new analytic paradigm, intentionally exploring the model's trade space in greater depth than has been done with legacy simulations. In the end, this process should result in better ABS model

development, more appropriate applications of ABSs to specific uses, and decreased risk of using ABS to answer the difficult questions facing decision makers today.

B.4.3 Failure To Reject and the Implication for Validity

While a model can be demonstrated to be invalid, it cannot be shown to be valid, since the complexity of these simulations makes examining every trajectory at best intractable. There is general agreement that failure to invalidate does not necessarily mean the model is valid. Failure to invalidate may mean, however, that there is less risk of using the ABS tool for a specific application. As was highlighted many times during this workshop, the risk of using any ABS tool must be clearly communicated to the user/consumer. Any reduced risk gained from failure to invalidate must also be communicated to the user/consumer.

B.4.4 Tools and Techniques

While there was discussion about tools and techniques for assessing the validity of ABS for use both in breakout sessions and in the general session, there was little specific discussion about how to assess the validity of ABS in specific. There was general agreement that validation efforts need to evaluate and trace cause and effect relationships, apply sensitivity analysis, and evaluate the level of aggregation used in the model to ensure that the correct relationships are captured. However, there were few suggestions of specific techniques to use beyond SME based validation for assessing the validity of agent based simulations for use. The specific characteristics of ABS, such as their rule bases, knowledge bases, qualitative parameters, and referent, were not addressed. Although, one participant commented on the influence of the agents' decision-making architecture on the resulting behavior of the simulation, and Dr. Tolk's presentation identified the importance of a tiered assessment of these simulations, evaluating the agent, the agent and its environment, etc., techniques for assessing these aspects of an ABS were not identified in the workshop.

B.5 STEPS FORWARD

The Study Team has not reached a decision on which ABS-application pairs to examine using the ABSVal Framework. The Framework will be definitely applied to the Pythagoras-COIN application pair. Other candidate ABS-application pairs have been examined and solicitations from M&S community are still being made.

In the next phase of the study, the Study Team will synthesize the ideas generated from this workshop along with other materials we have gathered. The Study team will finalize selection of ABS-application pairs for validation. We will develop an implementation plan for applying the ABSVal Framework, and will apply the Framework to the selected ABS-application pairs. The Study Team will report to MCCDC/OAD on the validation Framework process. The Study Team will also write and publish academic papers on the validation Framework process and its application; we hope to present these papers to audiences within the M&S community to obtain their insights and improve our process.

APPENDIX C WORKSHOP #3 SUMMARY

C.1 INTRODUCTION

The Marine Corps Combat Development Command (MCCDC) Operations Analysis Division (OAD) hosted a two-day workshop to review the results of the initial application of the Agent-Based Simulation (ABS) Verification, Validation, and Accreditation (VV&A) Framework to the Pythagoras-COIN model on 25-26 March 2008. The group was kept small intentionally in order to allow for critical review, comment, and discussion on the framework and the preliminary results of its application. This workshop report describes presentations and feedback given during the workshop. Since the workshop, many of the ideas have been refined.

C.2 WORKSHOP #3 OVERVIEW

The agenda for the workshop included an update to the framework as developed through its application to the Pythagoras-COIN model, a description of the Pythagoras-COIN model for the workshop participants, an overview of the validation results, and a description of what is next for the study. The agenda for the workshop was followed loosely in line with the purpose of the workshop. During the workshop, there were nine main topics of discussion:

1. Validation Framework Update
2. Description of Pythagoras
3. The Irregular Warfare Project
4. Description of the Pythagoras-Coin Implementation
5. Overview of Assumptions Testing for Pythagoras-Coin
6. Overview of Preliminary Validation Results of Pythagoras-Coin
7. V&V Auditing: Objectives and Methods
8. Some Thoughts on ABS V&V
9. Surf Zone/Beach Zone (SZ/BZ) Obstacle Reduction Simulation

This report contains summaries of the discussions during the workshop. Briefings presented during the workshop can be found at http://orsagouge.pbwiki.com/ABSVal.

Following the workshop an IPR was held with Dr. Mike Bailey. As a result of the discussions during the workshop and the discussion during the IPR, several adjustments to the overall validation approach and framework were seen as necessary. The final framework will reflect these adjustments.

C.3 VALIDATION FRAMEWORK UPDATE (DR. ERIC WEISEL)

Dr. Weisel briefed the workshop participants on the current validation framework with an emphasis on the fundamental concepts and components that define it. The presentation generated significant discussion regarding the place and purpose of the theoretical elements to the ABS VV&A Framework. The discussion also generated some useful clarifying and guiding points for the VV&A framework.

C.3.1 The Validation "Cloud" Diagram

The validation "cloud" diagram formed the foundation for the discussion, shown in Figure 13. In the diagram, simulation elements are on the right hand side; the real-world elements being simulated are on the left hand side. The outside nodes of the cloud diagram represent elements of practical validation (tangible), while the inside nodes generally represent elements of theoretical validation (theoretical / mathematical).

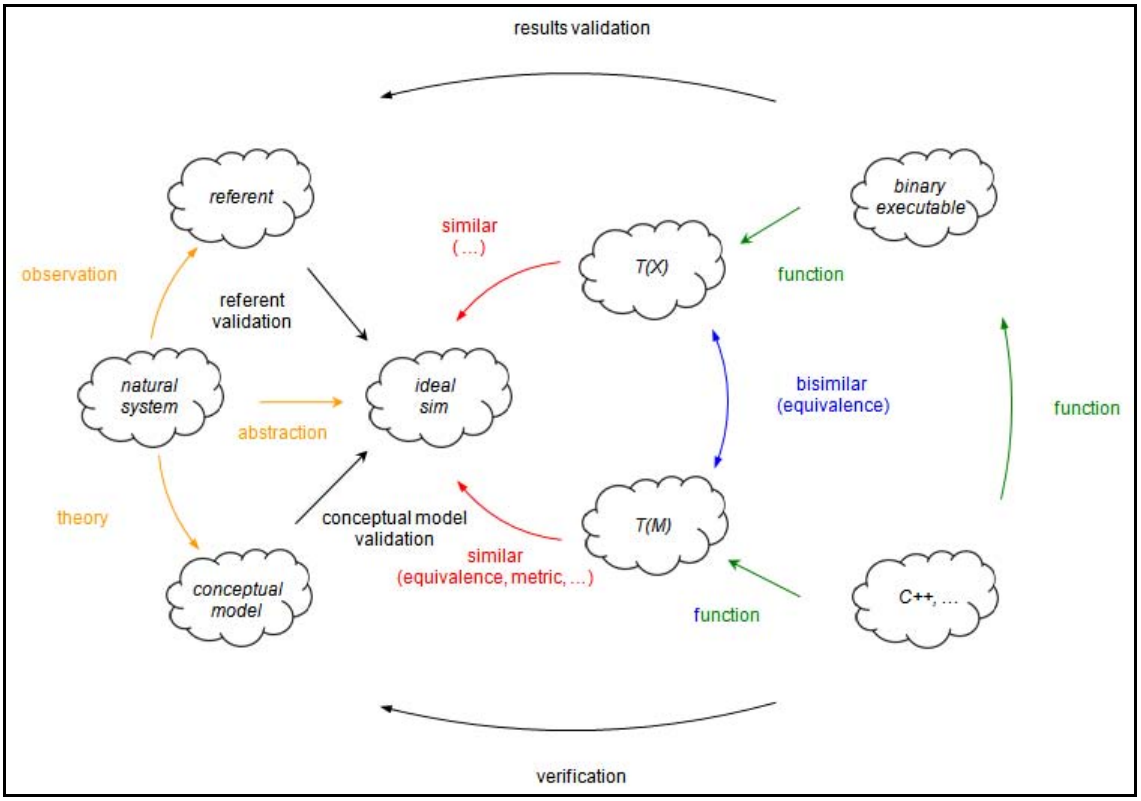


Figure 13 Validation "Cloud" Diagram

Validation may be defined as a comparison of transition systems. A digital computer running a programmed simulation creates one transition system, whose trajectory can be compared to the transition system of what is being simulated.

The Natural System (aka the Target System or the System of Interest) is derived as a subsystem of the universe in its entirety from the perspective of a perfect observer at a snapshot in time. If the universe from the perspective of a perfect observer was depicted as an infinite vector, one could derive the natural system/system of interest for a simulation by taking a notional "slice" of that vector. Since, pragmatically, not all elements in this Natural System can be modeled; an abstraction process is used to focus on elements of interest in the Natural System for the purposes of simulation. In mathematical terms, this abstraction amounts to the bounding of the state vector and the input vector. Abstraction is applied to create non-deterministic transition systems, creating the Ideal Simulation as depicted in the cloud diagram. For the Ideal Simulation,

the values of vectors match exactly with the corresponding values in the Natural System with no loss of accuracy (it extrapolates perfectly), but the elements of the Ideal Simulation are limited to what has been derived by abstraction. Theoretically, if one could not answer a posed analytical question using the Ideal Simulation, then one could not answer that question with any simulation representation.

The Code element in the cloud diagram represents a transition system in the form of a binary executable program. What the validation effort attempts to demonstrate is that the transition system that the code generates simulates the Ideal Simulation (every move the ideal simulation makes, the coded simulation can match). However, there are problems of tractability in generating a mathematical proof that shows that one transition system simulates the other, and there is no algorithm to evaluate these transition systems trajectory by trajectory.

Thus, for the inner loop (theoretical elements) of the cloud diagram validity cannot be proven (in an exclusively mathematical sense), and so other techniques of validation must be employed. Results validation is often used to provide evidence of the validity of a simulation; the binary executable is run to acquire an array of trajectories which can be compared to a referent, real-world empirical data from observations. Often for social systems, the referent is not solid; however results validation can still be employed (although much more subjectively) to assesses if simulation output is sensible and useful. The referent data, in turn, requires validation as well, to determine accuracy and suitability to the application.

In the framework, the intended use of the simulation defined by the analytical application drives the abstraction process to determine what details and elements in are critical for the simulation. Intended use also drives validation criteria e.g.; “Do the states in the simulation need to be within some tolerance compared to the Ideal Simulation?” or can more subjective methods be applied to determine the validity/usefulness of the simulation?

The Conceptual Model arises out of the Natural System by theory—it is the “best that is known” about the Natural System from a theoretical sense. An aspect of validation may be to provide evidence that the elements in the Conceptual Model are true with respect to the Ideal Simulation. This often may occur with some degree of subjectivity, especially when simulating human systems. Because of the inability to comprehensively validate a human system, the scientific method is employed in the framework as a method in which validation techniques are applied to the simulation in an attempt to disprove or “poke holes” in the integrity of the simulation with respect to the application. With rigorous validation techniques, the inability to invalidate the simulation subsequently provides evidence of its validity and usefulness toward an analysis application.

C.3.2 Commentary

Commentary to the briefing centered on two concepts: one was usefulness of the theoretical elements described in the diagram and the other was the nature of validation itself.

C.3.2.1 Diagram Elements

Several comments were provided that questioned the necessity of the elements of the cloud diagram. Dr. Bailey emphasized the need for a "distillation" of the framework that focuses on the critical elements. Other comments suggested a simplification of the diagram that was more oriented to a decision-maker's perspective in which several of the more theoretical elements could be collapsed. Dr. Weisel responded that the intention of the VV&A process as depicted by the elements of the cloud diagram was to provide a framework that could be used as a sort of notional reference to the validators of such simulations to organize the components or artifacts of simulation so that a comparison of such components could then take place, and that by providing this foundational framework the intention was to make validation of these complex systems more achievable.

He mentioned that the disparate nature of the existence and completeness of such simulation artifacts creates challenges in using the framework as it exists. There were several comments that illuminated that in any given application of the framework, elements in the diagram could be applied in multiple ways. For instance, open to interpretation is the referent for Pythagoras-COIN as is the Conceptual Model. The highlighted need to specify the elements in the cloud diagram during a validation effort is an insight in the framework application.

C.3.2.2 Nature of Validation

Several perspectives on the nature on validation were proffered:

- Validation must communicate the risk of using the model for a specific analytical application to the decision maker
- Validation should prove that the simulation's usefulness is better than a 50/50 coin flip, to what degree is it more useful, and why.
- Validation must account for the fact that many of these ABS are used for exploration and insight; i.e. trying to gain insight into the way natural systems behave, instead of attempting to mimic a natural system that we don't fully understand.
- Validation needs to employ tests of the simulation that are rigorous enough to provide some level of confidence that results from such a simulation are useful for analysis.

C.4 DESCRIPTION OF PYTHAGORAS (MR. EDD BITINAS)

As background to the validation application of the Pythagoras-COIN model, background information on Pythagoras was provided to the workshop participants. Mr. Bitinas

explained that Pythagoras is a modeling “toolkit” that can be used to build an array of different kinds of computer models. The Pythagoras environment accommodates the ability to model leadership, human factors, and physics, three tightly linked factors that were not often considered (as a whole) in traditional combat modeling.

Pythagoras is an Agent-Based distillation that uses decision-making entities, employs probabilities (where appropriate) to define agent actions, and often has unpredictable outcomes that can offer insight into analytical problems. The structure of the tool employs a user-friendly interface and allows the functionality to evolve as new operational requirements are identified. Pythagoras employs software objects that:

- are capable of choice
- have autonomous behaviors and are distinct from the background or environment and each other
- perform actions like:
  - Sense
  - Order
  - Communicate
  - Move
  - Influence

Pythagoras puts the complexity of creating behaviors in the hands of the analyst, not in the software. It offers the analyst the ability to:

- Mirror Actual Groups, Including Their Characteristics, Beliefs, and Standards of Behavior
- Examine Outcomes Based on Behavior
- Data Farm Over a Trade Space Using High Performance Computer Clusters
  - Execute a Range of Runs
  - Similar to Sensitivity Analysis
  - Design of Experiments Can Reduce Runs and Preserve Interactions
- Assess Impact of:
  - Uncertainty in Inputs
  - Some Idea of the Minimum-To-Maximum Range
  - Unclear Interactions Among Inputs

Some of Pythagoras’ key features are:

- Soft Decision Rules (allows for variability)
- Agent’s Decisions (e.g., movement)
- Generic Attributes (characteristics to measure controls on behavior)
- Dynamic Affiliation (Same Unit/Same Side/Enemy/ Neutral)
- Rule-Based Influencers (e.g., lethal, non-lethal)
- Triggers that Change Behavior Rules When Triggered
- MOEs
- Sensors (Three Bands)
- Comms (Three Channels)

- Terrain Features (Mobility – Protection – Height)
- Concealment

Pythagoras offers the ability to characterize agents in a variety of ways (e.g., Affiliation, Vulnerabilities, Leadership, Obedience, Detectability, Target Value, etc.)

Agents in Pythagoras operate per time step on an Agent Time Cycle as depicted in Figure 14 below:

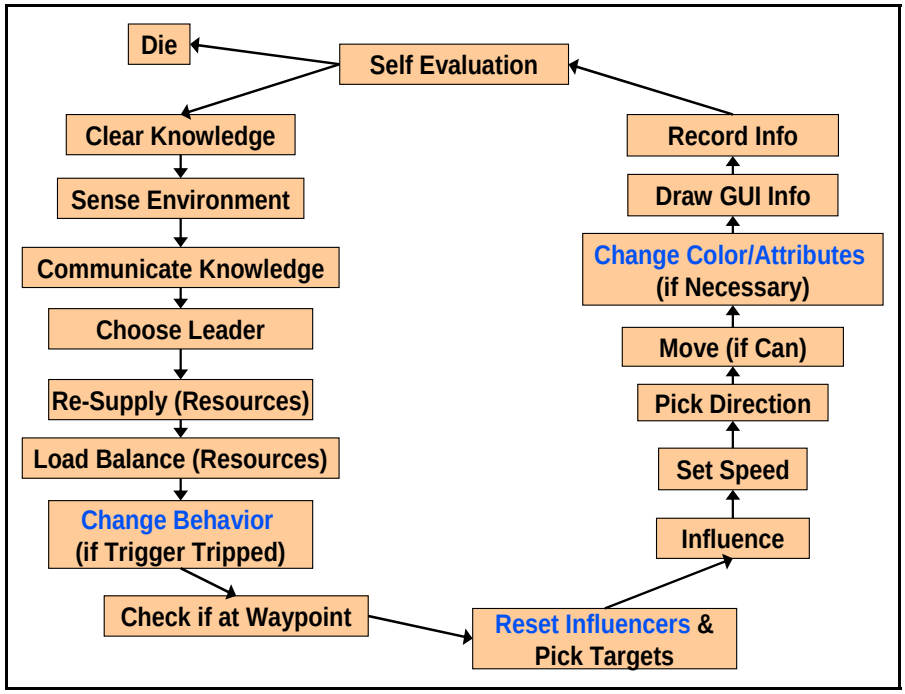


Figure 14 Pythagoras Agent Update Cycle

A multitude of time-series MOEs (e.g. Agent Attributes# Known Units (Friend, Enemy, Neutral), Kills Scored, Shots Fired, etc.) and end-of-run MOEs (e.g., Initial, Minimum, Maximum and Final Number of Agents Near the Final Objective, etc.) may be output from Pythagoras simulation runs.

In the Pythagoras COIN simulation, the following were set by design in Pythagoras:

- Agents do not “die” in the Pythagoras COIN simulation
- Agents clear knowledge at every time-step
- Each agent is aware of all other agents
- No resources are assigned
- There is only one event in the simulation: MAGTF arrives and leaves.
- Communications devices are used as “influencers”
- MOE is the number of agents that change affiliation

Several functional updates have been made in Pythagoras 2.0.0, including the ability for agents to shoot at multiple targets per time step, the introduction of ten generic

attributes and the ability to change them via weapons, communications, events or terrain, the ability of agents to carry up to ten each of sensors, weapons, and communication devices, and many others. Modernization updates in Pythagoras 2.0.0 include an upgrade to Java 1.5 and the enhancement to a 4000x4000 pixel playbox.

Pythagoras has been used by an array of individuals and organizations for a host of different applications, including MAGTF Optical Requirements (Night Vision Lab/Ft. Belvoir), Homeland Defense (Northrop Grumman), Thermobaric Weapons Assessment (MCCDC), and others.

C.5 THE IRREGULAR WARFARE PROJECT (DR. BOB SHELDON)

Dr. Sheldon briefed the audience on the Irregular Warfare (IW) project, the background for the analysis objectives explored using the Pythagoras-COIN simulation, and the methods employed to derive input data for the Pythagoras COIN simulation. The goal of the Irregular Warfare Project is to develop a prototype methodology for analyzing a USMC IW problem in-house. Several challenges exist to modeling Irregular Warfare. The inputs and MOE's for these models are different from traditional combat models. Examples of traditional combat model inputs include Weapon Probability of Kill, Armor Thickness, and Vehicle speed (more objectively measurable); examples of IW inputs include social factors such as Influence and Susceptibility (less objectively measured). Likewise MOE's for IW models are different from traditional combat models. Examples of traditional combat model MOEs include Lethality and Survivability; examples of IW MOEs include Population Response and Behavior. With this "soft" or less objective aspect of modeling human behavior, some expectation management for these models should be considered, since they will inherently involve a higher level of uncertainty (lower statistical correlation) than modeling traditional force-on-force combat.

C.5.1 Conceptual Model

For the Pythagoras-COIN Conceptual Model, the civilian population in Colombia was conceptually decomposed into population segments, and within each segment, five Insurgency Behavior orientations were assigned, to define subgroups within the segments' relative affiliation to the Revolutionary Armed Forces of Colombia (insurgent) or to the Government of Colombia. Influencing factors and events determine if the population shifts generally towards or away from insurgency.

The Pythagoras COIN scenario consisted of a MAGTF Mission to provide Refugee Camp Security and Humanitarian Assistance / Disaster Relief in the region, with two possible courses of action: Sea-Based Operations or Shore-Based Operations. The analysis sought to determine the plausible range of civilian population behaviors for these courses of action. The background for this scenario focused on two provinces on the Pacific coast of Colombia. The primary city in this area is Buenaventura, a seaport that is a predominant thru-way for drug traffic. Although it has a small upper class and a growing middle class, Buenaventura is populated mainly by urban poor and displaced persons who have been driven from their villages by the insurgency and crime that is gripping Columbia. The two key players in the insurgency are 1) the insurgents, the

Revolutionary Armed Forces of Colombia (FARC), and 2) the counterinsurgents, the Government of Colombia (GoC). Other critical players are the militias, the drug traffickers, the Colombian Army, and the police. All have a presence in Buenaventura. It is a fomenting hot bed with a crime rate many times higher than New York City.

In the problem scenario, a tsunami has struck the area indicated in red on the map, destroying much of Buenaventura, and a Marine Air-Ground Task Force (MAGTF) has been sent to the area as part of a Joint, Combined, Interagency Task Force at the request of the GoC with the mission as shown here.

C.5.2 Input Data

Input data for the model was acquired through SME Interviews, a process that was met with several challenges:

1. Analyst & cultural SME communication challenge
2. Analysts need numbers, e.g., probabilities, percentages ,cultural SMEs are non-quantitative thinkers
3. Note that the cultural data as acquired was narrowly focused on a specific region, and therefore, the data is not accurate for the rest of Colombia.
4. To define population segments, data was elicited for each population segment and sought to determine:
5. Prevalence of current behavior patterns
6. Perceived needs are affected based on three factors (using Narrative Paradigm)
7. Natural tendency of the population segment (the population segment's narrative with respect to the insurgency)
8. Effect of current events on population segment (impact) – how the population segment reacts to a given COA
9. Effect of other population segments on a population segment (influence) – How the population segment reacts to the narratives offered by other population segment

C.5.3 Population Segments

The population segments were defined as follows:

1. Illicit Organizations
2. Catholic Church
3. Police
4. Military
5. Displaced Persons
6. Urban Poor
7. Urban Middle Class
8. Old Money

Cultural behavioral data was sought on the population orientation (initial and natural tendency), the impact of MAGTF COAs, and the influence of population segment interactions.

The following question was posed for the Initial orientation data:

> "How do the actions of this population segment support the insurgency (FARC) or the Government of Colombia (GoC)?"

The following question was posed for the Natural Tendency orientation data:

> "Given no external influences, over time, how would the actions of this population segment change to support the FARC or the GoC?"

Quantitative data was derived from the SME interview responses through a Markov Chain matrix. This matrix is input to the Pythagoras agent-based simulation environment. Pythagoras does not carry out Markov matrix computations but uses the values to represent the 'dynamic sidedness' of the agents as it interacts with events and other agents.

This Data Elicitation required a process that translated SME words to a quantitative measure. Osgood's Semantic Differential method was used to provide three major factors or dimensions of judgment:

> EVALUATIVE (good - bad)
> POTENCY (strong - weak)
> ACTIVITY (active - passive)

The data to define the impact of the proposed COAs was elicited through SME interviews that posed the question:

> What words would this population segment use to describe MAGTF 'sea-based/shore-based' operations?"

'Positive words' averaged to measure leaning more towards GoC; 'Negative words' averaged to measure leaning more towards FARC. This data was input into another Markov matrix that displays new behavior patterns reflecting the impact of the MAGTF. This matrix was input to the Pythagoras agent-based simulation environment.

To obtain data to define the influence of other population segments was elicited through SME interviews that posed the question:

> "What words would this population segment use to describe another population segment?"

The Influence of Population Interactions data was input into another Markov matrix that displays new behavior patterns reflecting the influence of other populations. This matrix was input to the Pythagoras agent-based simulation environment.

C.6 DESCRIPTION OF THE PYTHAGORAS-COIN IMPLEMENTATION (MS. BRITTLEA SHELDON)

Ms. Sheldon briefed the audience on the implementation of the COIN scenario in the Pythagoras modeling environment.

C.6.1 COIN Scenario

Fundamentally, the implementation of the COIN scenario in Pythagoras is focused on a theoretical perception of COIN and insurgency in which defined population segments are attributed with an array of orientations towards insurgency (e.g., Insurgent, Pro-Insurgent, Indifferent, and Pro-COIN); as the simulation runs, these orientations may drift, which in effect cause a particular population segment to become more or less insurgent. Figure 15 depicts this concept:

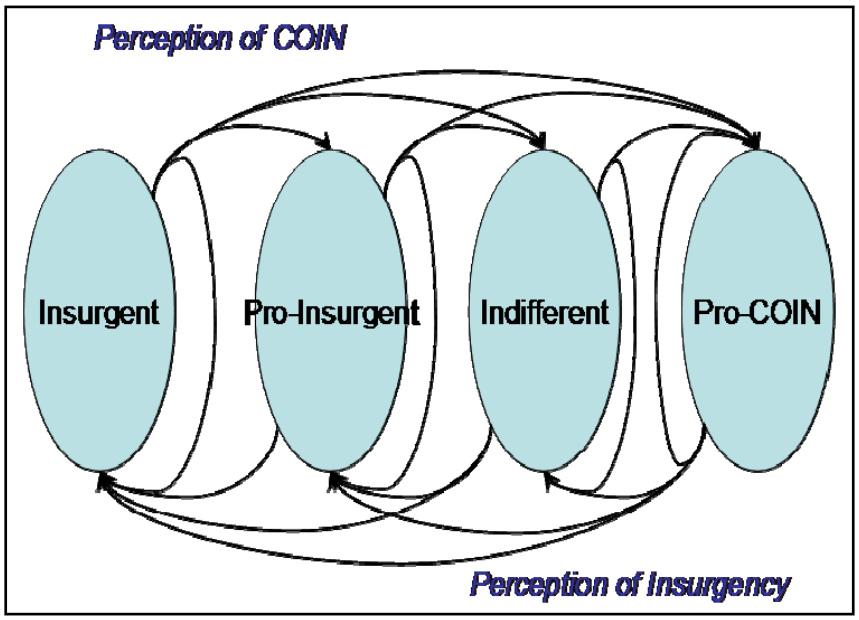


Figure 15 Theoretical Perception of COIN

C.6.2 Conceptual Model

The Conceptual Model defined for this simulation requires each agent within each population segment's orientation classes to shift per time step so that the exit arrows out of each population segment "bubble" shown in Figure 15 sum to 100%. Conceptually, agents remaining within an orientation bubble return to the bubble through a feedback arrow as shown in Figure 15. The initial distribution of orientations within each defined population segment is defined by researched demographic data, and the initial value of each arrow is defined by the conceptual "insurgency susceptibility" of the subpopulation within the bubble. This insurgency susceptibility can be further defined with three variables: the Interaction Estimation Transition Effect on the targeted population (the Direct Effect of the singular MAGTF arrival event), the Salience Transition Effect on population segments receiving information about events (the Indirect Effect of the interaction between population segments), and the

Background Susceptibility Transition (the Ongoing Effect of population tendencies). The brief mentioned a precedence configuration for these effects; however, it was clarified that this was a legacy configuration from a previous version of Pythagoras, which is no longer relevant in Pythagoras 2.0.0, in which all of these effects can occur simultaneously on an agent within the model (the averaged effect is used to set the agents' orientation status per time-step). For the COIN scenario, the only event that affects the simulation is the arrival of MAGTF. Additionally, there are some soft-rules random elements that influence these population drifts (not every agent interacts with every other agent per time-step).

The initial population in each of the eight defined population segments in the COIN scenario was scaled to be 100 agents, so that there was equal opportunity for all population segments to affect one another. Each agent within a particular population segment is 1% of that segment's population. Each agent has a set of 5 Attributes that define the insurgency orientation of that 1% of the population (1 being insurgent, 5 being COIN). In each time step, the sum of attributes normalizes to equal 1000. This normalization reduces the potential for simulation inaccuracies caused by round-off error. Attribute Changers represent the population tendencies, the influence between population segments, and the influence of the MAGTF actions. Communication devices represent interactions and possess the Attribute Changers which will do the influencing. Each agent can carry up to 10 Communication Devices; each communication device has 3 channels. Each of these channels contains an attribute changer that represents interaction and the ability to change another agent's orientation attribute through interaction, which allows the interaction between agents to be very specific.

The Background Susceptibility Transition Effect (Vulnerability) of an agent to an orientation change, described as a Markov effect, has been implemented in Pythagoras COIN as an incremental Attribute Changer which increments the Attributes for each agent (1% of the population) in each time step per the Markov chain matrix values, normalized in Pythagoras.

The Salience Transition Effect is implemented as a relative Attribute Changer. An average Orientation for the two interacting population segments is calculated (using 1 = FARC through 5 = COIN); a Delta value is calculated based on the difference of the average of the two population segments; the Delta value determines the direction of the influence.

The influence effect of the MAGTF arrival was modeled as an attribute changer for the simulation, which acted as multipliers upon the original agent attribute values.

C.6.3 Commentary

Several comments, questions, and digression discussions relating to the Pythagoras COIN implementation provided in Ms. Sheldon's brief generated some useful clarifying and guiding points for the validation study:

It was questioned why agent Movement was not modeled into this simulation as an element that could affect interactions and population orientations, and it was considered that the abstraction of Movement out of the simulation may be a potentially critical flaw in the simulation. Ms. Sheldon responded that the model is primarily based on the influence aspect of these population segments and that agent Movement may not be critical as suggested.

The following question was posed to Mr. Bitinas: "What did running the (Pythagoras COIN) model tell us that putting together the model did not?" Mr. Bitinas responded that running the model did not result in significant new information and results supplementary to the input data to the model; however, if the model were to be validated, it could provide a baseline platform for the scenario to which other more complex elements could be added to further pursue analysis objectives.

There was general consensus that the assumptions incorporated into a simulation (such as the assumption to abstract Movement out of the Pythagoras COIN simulation) must be examined with respect to the intended analytical use of the simulation to determine if such assumptions pose a risk to the model's usefulness, and to do so, there must be a context that is instantiated to provide bounds for the analysis, the simulation, and the validation effort. It was stated that it may be a useful exercise to examine assumptions within a simulation by altering them to see if simulation output is affected in order to determine the impact of such assumptions. It was reiterated by Dr. Weisel that the referent as defined in the current framework can often be extremely muddy (e.g. the Pythagoras COIN referent) and may not be useful in providing context for the validation activity. Pertaining to this intended use-bound examination of assumptions; Dr. Bailey referenced a format called Constraints, Limitations, and Assumptions that provides context and bounds for these types of analysis activities and to some extent defines the critical elements of a simulation for the analysis.

C.7 OVERVIEW OF ASSUMPTIONS TESTING FOR PYTHAGORAS-COIN (MR. ROBERT EBERTH)

Mr. Eberth briefed the audience on an overview of assumption testing that was performed on the Pythagoras COIN simulation. It was noted that the validation work that Mr. Eberth performed was independent of the work that the WernerAnderson validation team was performing, and that the focus of Mr. Eberth's work was the Conceptual Model.

C.7.1 Methodology

The methodology for validating the Conceptual Model for the Pythagoras-COIN simulation was heavily reliant on assumption testing; assumptions made for the development of the simulation would be examined with respect to the analysis objectives, with findings communicated to the decision maker. In an ideal sense, if one could identify every assumption that was made in building the model, then one would understand every way in which that model departs from reality. Here, the assumption

testing techniques applied to Pythagoras-COIN served to exercise the validation framework developed during this study.

The validation activities applied to the Pythagoras COIN simulation are based on the scientific method. The null hypothesis (research hypothesis) for validation is that the model/simulation is valid for the intended use, and the validation activities (including assumption testing) attempt to falsify the null hypothesis. An inability to falsify the simulation results in more confidence in the usefulness of the model towards the intended use; the degree of confidence then depends on the rigor and power of the tests applied.

The plan outlined to perform the validation of the Conceptual Model for Pythagoras COIN was as follows:

1. Identify the analytic questions at hand, their metrics, and degree that results are expected to shape decisions
2. Detailed review of all related documentation
3. Interview Application Sponsor
4. With the Application Sponsor, identify the referent; i.e., the proxy for the real world for accuracy comparisons
5. With the Application Sponsor, determine the accreditation criteria: How "accurate" must the model be? How can/will accuracy be determined? (quantitatively/qualitatively)
6. Criteria must establish lower bounds of acceptability for the model

It is desired to assess the validity of the referent; by confirming that no preferable referent could be made available, assumption testing the referent (for other than empirical datasets), determining the operational implications of the assumptions, determining the bounds of validity imposed on the application's problem space and on the model's validity assessment by the referent's assumptions, and determining whether the operational implications and bounds of validity are acceptable to the Application Sponsor.

However, in the case of the Pythagoras COIN simulation, it was determined that there was no solid referent that could be validated, as is often the case with ABS. For the Pythagoras COIN model, it was often difficult to determine where to segregate different aspects of the framework, and if specific critical assumption (e.g., the Markov Chain assumption for the Background Susceptibility Transition Effect (Vulnerability)) were part of the referent or the Conceptual Model. For the Pythagoras COIN simulation it was determined for the purposes of validation that the Theoretic Model and the referent were one in the same, but it was stated that a more detailed clarification of these elements is needed in the framework.

It was stated that the Conceptual Model generally includes the Theoretic Model, the Mathematical Model, and the Algorithmic Model. Each in turn receives same assessment techniques:

1. Logical verification – determining sub-model is an adequate and correct implementation of its predecessor.
2. Assumption testing
3. Identify/derive the assumptions that are inherent to/embedded in the sub-model
4. Determine the operational implications of the identified assumptions in the context of the particular application, Determine the bounds of validity of the model that are the result of the identified assumptions
5. Determine whether the operational implications and bounds of validity are acceptable to the Application Sponsor for the intended application
6. For some models, it may prove necessary to reverse-engineer one or more sub-models from later models. It may even be necessary to reverse-engineer the conceptual model, or portions of it, from source code (Not the case with Pythagoras COIN).
7. Independent SME reviews

The operational implications of any assumptions used in the simulation must be examined to determine if the assumptions made have biased the results in a way that compromises its usefulness toward the analytical objectives.

Reverse engineering of the Algorithmic model may be used as a technique to derive assumptions that have not been clarified or stated in the Conceptual Model, but this may prove challenging.

In terms of this validation effort as an exercise of the proposed VV&A framework, it was stated that the framework "is working," but improvements are needed on how to define the referent and assess the validity of the referent when empirical data are not available for use as the referent. It was also stated that linear, checklist-oriented templates would be useful for the framework.

For the Pythagoras COIN simulation, the Theoretic Model was the focus. The Mathematical Model, for this simulation, does not exist, as it went "straight-to-code" during development, and the Algorithmic Model had not yet been validated at the time of the workshop. It was acknowledged that the Pythagoras COIN Algorithmic Model may depart in some ways from the Conceptual Model.

The effort to validate the Conceptual Model began with a review of the Pythagoras User's Manual and related detailed discussions with Mr. Edd Bitinas, the simulation developer. However, Pythagoras itself was not assessed. Interviews were conducted with LT Robin Marling, USN, the COIN study's Project Officer, and several study-related documents were made available and reviewed. An interview with Dr. Akst, the Application Sponsor was conducted, and resulted in the following considerations for validation:

- Purpose was to "make headway in developing a COIN model."
- Did not specify an ABS, let alone Pythagoras
- Approved recommendation of using "sea versus land basing" as study's analytic question, but did not specify it at the outset of the analysis

- Approved stated Marine missions, and O.K. with implied mission
- Insisted study must use real-world dataset.

One of the findings of these interviews was that there were multiple, conflicting objectives for the Pythagoras COIN simulation:

- OAD was to "make headway" in developing a COIN model
- NGMS was tasked to determine whether and how Pythagoras could be used to support IW analyses
- Study at hand had the analytic objective of determining whether it was best to leave the MAGTF ashore or afloat in a Columbian Humanitarian Assistance/Disaster Relief/Security scenario

There are several USMC missions in the Colombian scenario. (Refugee camp security, Humanitarian Assistance, Disaster Relief), but the collective study team (all stakeholders) found no way to directly evaluate the effectiveness of mission performance. Thus, it was decided to use allegiance changes of population segments among several distinct affiliation possibilities – thus producing an "implied mission" of keeping the insurgents from gaining strength (stated as "Do not allow illicit organizations to take advantage of situation"). This may also imply that ABSs in general and Pythagoras in particular cannot support traditional MOEs of mission performance; however, the MOE that emerged from this simulation (changes in population orientations) may be a novel and worthwhile analysis consideration.

Several assumptions in the Pythagoras COIN Conceptual Model had a large impact:

1. Modeling the transitions among affiliations as a Markov process (a "memoryless" process). This is a significant assumption and may be very limiting to the potential validity of the application.
2. Constant transition probabilities across all time steps (except those during the Marines time in-country).
3. Constant transition probabilities across all time steps while the Marines were in-country (although different probabilities from the baseline). This assumption must be relaxed to increase the potential of validity.

Initial indications are that the above assumptions absolutely pre-determined the results and in a predictable way (i.e., the model became deterministic if allowed to run to steady-state). Unfortunately, that may mean that OAD cannot make a solid determination on the usefulness of Pythagoras in the IW or COIN context from this particular application

It also may mean that the answer to the one analytic question (afloat or ashore) depends entirely on the methodology used to develop the transition probabilities – the "influence estimation" and "salience" parameters, and those are suspect because of potential bias in data collection/analysis methodology (semantic differential), because the SME's used for data collection were not from the region of interest, and that the distinction between "data" and "context" could result in bias in the input data.

Initial indications with respect to the study (again, only from assessment of the theoretical model, so subject to change) were as follows:

- Probably cannot yet give a defensible answer to the afloat/ashore analytic question
- Implementation assumptions in the Pythagoras COIN Conceptual Model too limiting
- Semantic differential data collection/analysis methodology far too suspect

However, Mr. Eberth postulated that the study may represent a huge leap forward in IW analysis in the following ways:

- Could/should cause a re-evaluation of COGs and MOEs for IW environments
- Could/should lead to a series of studies on semantic differential and alternative methodologies for capturing the propensity of persons to change affiliations, particularly in response to actions/events rather than just presence

C.7.2 Commentary

Several comments, questions, and digression discussions relating to the assumption testing and validation of the Conceptual Model for Pythagoras COIN topics provided in Mr. Eberth's brief generated some useful clarifying and guiding points for the validation study:

It was asked, "How did the (COIN analysis stakeholders) make the decision to use Pythagoras (for this simulation)". Mr. Bitinas answered that an exploration was conducted of several simulations. Pythagoras was chosen, and the inherent malleability of Pythagoras as a modeling tool was a factor.

Dr. Bailey asserted that the validation of the Conceptual Model presented in Mr. Eberth's brief proved that the framework techniques provide the ability to criticize a model; however, acknowledging that all models are flawed, the results of this validation did not 1) Illuminate what the usefulness of the model was in the analytical context. 2) Provide an assessment that determines if the critical elements to the analysis are present in the model.

Mr. Eberth asserted that Validation is a model improvement process, and as an improvement process with respect to Pythagoras/COIN, there are things in the model that must be changed for it to be useful

There was general consensus on several topics posed during the brief as follows:

1. An improved format for validation results may be needed to illuminate the usefulness of the model with respect to the analytical objectives.
2. The framework needs to provide process steps early in the validation activity to determine the critical analysis aspects or concepts so that the model/simulation can ultimately be examined with respect to these analysis aspects in subsequent

validation processes; e.g. a bottom up, emergent approach is preferable to a top-down approach to population dynamics---does the Model implement this?

3. In an irregular warfare context, the role of mission MOEs in the VV&A process is missing and needs to be considered.

C.8 OVERVIEW OF PRELIMINARY VALIDATION RESULTS OF PYTHAGORAS-COIN (MS. LISA JEAN MOYA)

Ms. Moya reviewed the validation results to date. Part of the discussion highlighted that many of the concerns expressed during previous discussions regarding the Conceptual Model were not implemented by the Pythagoras-COIN implementation. This discussion illuminated the need to define the framework elements for a specific validation effort. This briefing generated a lot of open discussion.

C.8.1 The Nature of Validation/ABS

There were several opinions offered on the nature of validation brought up by the workshop team for consideration:

It was questioned: What should be communicated to the decision maker when a validation effort is unable to invalidate any aspect of the model? What does the format for that report look like? Is there still some risk that needs to be conveyed? On the other hand, if there are flaws in the model, given that ABS validation in general may not be a binary valid/invalid solution, what can be said about its usefulness? Does the framework accommodate this non-binary schema?

It was postulated that the manner in which the analytical question is posed may create a circumstance in which even the ideal simulation could not answer the question. In those cases (note that it was hypothesized that the MAGTF ashore/afloat question posed for Pythagoras/COIN may be an example of this), for the purposes of validation, it may be necessary to refine the analytical question or problem statement in a way that the model can feasibly answer (if designed and implemented properly).

It was postulated that the nature of ABS analysis, in general, may be more exploratory and experimental, utilized to identify patterns or insights from output data, and the validation framework should accommodate these objectives, perhaps by linking the validation process more intrinsically with the analytical process such that there is a more complementary relationship between these two processes.

C.8.2 Validating for Critical System Elements

Several times, it was postulated that the validation framework must contain within it a process for determining the critical elements of the system being modeled that must be implemented in order to support the analysis objective. In the case of the Pythagoras/COIN validation exercise, many felt that the element of “memory” was a critical element of the real world system that had not been considered in the Markov Chain implementation of the Salience Effect.

C.8.3 Identifying the Real World Proxy

There was much discussion about what can be used as the real world proxy that a developed simulation could be validated against, and that many of the techniques that the validation framework employs require something to serve in this regard as the counterpoint for validation comparisons. Operational narratives and use cases derived from SME or researched knowledge that bound the "best we know" about the real-world system being modeled and contain the presumed critical elements were some of the suggestions that were proffered to serve as this real-world proxy in the case of human factors-based ABS, where often there is no solid referent to use for validation. Having this real-world proxy would facilitate the validation activity so that the model's parameters can be stressed to determine if the results match what is surmised about the real world within acceptable thresholds. It was stated that it may also be worthwhile to decompose the real world proxy to some extent to arrive at some easily defendable piece of referent data to which both aspects of the Conceptual Model and simulation results can be compared against during the validation activity.

C.8.4 Simulation Results and Emergent Behavior

For simulation results in general, it was stated that it is desired to examine the simulation output to look for trends, relationships, aggregate behavior, and emergent behavior and examine their relevancy to the "best we know" about the real world.

In analyzing the Pythagoras-COIN simulation, no emergent behavior was found. However, the nature of emergent behavior in the context of the VV&A framework is something that several participants discussed.

Emergent or aggregate behavior may be a worthwhile analysis factor in ABS model/simulation studies. It was postulated that people may be inherently locally focused in their thinking and analysis. If a simulation captures a set of generally accepted ("best we know about the real world") social rules that are defined for local (micro-level) interactions, then this may allow for a valuable analysis of any aggregate or emergent behavior that results from simulating those local level rules and interactions in replication.

The question was posed: If emergent behavior does exist, then what does that mean for validation? How would you go about validating emergent behavior, as by nature it may be a divergence from what we generally know about the natural system? Results validation is most likely the exclusive technique that can be used, and most likely will be an analysis to determine if any emergent behaviors are either 1) worthwhile as real potential outcomes of the system (may be subjective to some extent) or 2) anomalous behavior created by some flaw in the model or 3) aggregate behavior forced by pre-defined model assumptions.

C.8.5 Validation Thresholds and Replication Analysis

It was considered that that the bounds in which a model could be considered valid could theoretically be represented by some tolerance ε; and if the real world was represented by μ_0 and the simulation output was represented by μ_1, then if the variance between μ_0 and μ_1 was less than ε, then the model could be considered valid. It was further considered that, in this context, the way one might invalidate the null hypothesis is to have a sufficient number of simulation replications that result in a threshold Type II error probability that the variance between μ_0 and μ_1 exceeds ε.

Subsequently, it was voiced that the challenge in this context was to identify μ_0, the real-world proxy. It was hypothesized that for Pythagoras COIN, μ_0 may be the "sand chart" described in Dr. Sheldon's brief

C.9 V&V AUDITING: OBJECTIVES AND METHODS (MR. AMOS KENYON)

Mr. Kenyon briefed the audience on the methods and objectives of auditing the V&V activity from and accreditation authority perspective. The following is a summary of some of the key points provided in Mr. Kenyon's brief:

The auditor acts in a role as an accreditation authority who must determine whether the model should be relied upon to support the analysis of a specified class of problems. The auditor should be an experienced consumer of modeling in support of past studies and can engage in discussions of theory, but remain in a practical orientation.

A general approach for auditing is as follows:

1. Focus on the V&V report as a reflection of the process.
2. Formulate, more or less independently of the V&V process, a series of questions we would want answered about the model in support of an accreditation decision.
3. For each of these questions, ask how clearly and convincingly the V&V report answers it, and how readily the answer is extracted from the report.
4. Use the V&V report to answer the second set of questions.
5. Comment on both the report and the process based on this exercise.

The accreditation authority (SPA) for the Pythagoras/ COIN VV&A effort developed a set of review questions to perform the audit. At the time of the workshop, the audit process was underway but not at completion. The review questions were presented to the audience along with some preliminary observations from the audit activity.

C.9.1 Audit Questions

The following are a list of those questions and the response from the accreditation authority based on preliminary observations as they were presented in Mr. Kenyon's brief, and the feedback provided by the workshop audience where given:

Audit Question #1: Does the validation report clearly identify the class of applications (study questions) for which the model is being validated, and the model's role in addressing those questions?

Yes. The report states that the application is a course-of-action analysis, and that we don't necessarily need accurate point estimates of outcomes, but a reliable ranking of those point estimates under the alternatives. We assume the validation uses Colombia as a test case and that the purpose of validation is to determine whether the model could be used in any country, given reliable data.

Here, feedback was provided that the assumption above is wrong as it relates to supporting an accreditation decision – from the accreditor's perspective, the application is specific to Colombia and includes the data. The question of general validity "given good data" along with what it takes to obtain and validate data could still be relevant in the model improvement context.

Audit Question #2: Does the validation report clearly describe the tests that were performed on the model, the possible outcomes for each test, and the criteria for passing?

Some of them, the significance of the test to the accreditation decision is not always clear.

Audit Question #3: For each test performed, is the result clearly presented in a way that relates directly to the specified pass/fail criteria?

Generally not (in the draft), but the structure of the report seems to reflect the intent to do this.

Audit Question #4: Does the validation report make a convincing argument that the tests conducted collectively provide a sufficient basis for the recommended accreditation decision?

It would be helpful if, for each validation step, all the validation questions were presented in one place, each with a reference to the section in which the associated tests and their results are presented. Then, the auditor could see the chain of logic and perhaps recognize any gaps. When the questions are scattered throughout the report, this is less readily done.

Audit Question #5: Does the validation report make clear what elements (parameters) were treated as part of the model and what elements were treated as data subject to variation?

Not in its present form. The auditor must assume from the target application that all numbers relating to drift, salience, and the effects of interventions are data.

Here, feedback was provided that in the context of accreditation, all data are part of the thing being accredited.

Audit Question #6: For those parameters identified as data that might vary within the scope of the application for which accreditation is recommended (if it is recommended), has the V&V team identified a general approach to validating the data for applications other than the test case?

No. This means that should the model be accredited for a given class of applications, users would have no guidance for judging whether they have, or what it takes to develop, suitable data for another instance of that class of problem. The open questions about the scales for representing salience and influence make it difficult to solve this.

Here, feedback was provided that in the context of accreditation, all data are part of the thing being accredited.

Audit Question #7: Does the validation report provide a recommended decision for the accreditation authority?

There is a placeholder for this. This may be useful up-front information as well in the report.

Audit Question #8: To how broad an audience does the report make its findings readily accessible?

The reader should meet some standard of familiarity with M&S practice in order to use the information conveyed, but this may have been gained from practical experience. The report reflects an assumption that mathematical notation is the most efficient way to communicate with the reader.

Corollary question: Does ABSVAL specify its target consumer?

Audit Question #9: Does the report appear to reflect an efficient plan of attack?

Working around the lack of documentation on salience and influence suggests otherwise.

The relevance of some tests was not made clear.

This question addresses the practical aspects of the ABSVAL process. An inefficient plan of attack could still provide good support for an accreditation decision. This question is included because we assume it is desired that ABSVAL enable an efficient attack of the V&V problem for any given application; i.e., start with the low hanging fruit and the make-or-break questions. However, these are fundamentally unfair questions to ask about a test case, where the V&V team might want to exercise as much of the process as possible.

C.9.2 Commentary

Mr. Kenyon's briefing generated useful discussion in the application of the VV&A framework:

1. In the discussion of the validation approach, the report does not distinguish between validation and failure to invalidate.
2. From the perspective of the accreditation authority, it may be preferable to structure the presentation more like a proposal or an attorney's presentation, and less like an academic paper.
3. The report appears targeted at someone with a checklist mentality and limited time to extract key information.
4. For every question raised, there should be an answer and a reference to where the answer can be found. For every test, the associated validation question should be identified.
5. No discussions of what might also be interesting for someone else to investigate later.
6. The report should use more examples, less use of mathematical notation to make key points.
7. With COA analysis as the intended application, the availability of (or ease of developing) valid supporting data for a given problem instance is a practical question of equal interest to the accreditation authority as the model's validity.
8. Failure of any ABM to "fit" the validation framework should be considered a limitation of the framework, not the model.
9. Model features should not generate exceptions in the process.

A key SPA takeaway from this presentation was that because V&V also supports model improvement programs, the audit approach should also adopt the point of view of a developer and assess the value and actionability of reported V&V findings from that perspective. Most of the same auditing questions would apply, but we might also ask whether any recommended changes to the model are clearly identified, if the model is not assessed to be irredeemably useless.

C.10 SOME THOUGHTS ON ABS V&V (MR. VIC MIDDLETON)

Mr. Middleton briefed the audience on some thoughts on the ABS V&V process. He contended that the changing spectrum of operations is driving a demand for ABS models that may be used to derive useful information to address non-traditional analytical questions. There is a changing view of conflict in which human social systems and interaction effects are significant factors to analyze and consider, as well as the physical, cognitive, and informational aspects of network-centric warfare, for which these ABS models are often better suited.

Modeling human interactions may require a different perspective on the expected outcomes and the analysis approach for modeling-based analysis. Operational narratives may be a useful tool to bound and define attributes for modeled agents. Ultimately we would like a continuum of Human Factors, at one end, sensory and

psycho-physiological factors, at the other end, will, morale, culture etc. In the middle of the continuum, there is analysis for Situation Awareness and Decision-Making, rooted in concepts with real “face validity,” that are based on some accessible and accepted theory, from which analysis measures and metrics can be derived and can be expressed concretely in simulation.

Human Systems Representation can create many challenges for validation, for example:

1. Resolution & Fidelity Issues
2. Individual Human interaction with Terrain & the environment
3. Representation of complex terrain (MOUT)
4. Methodologies unique to soldier operations, e.g.:
5. Target acquisition in urban areas and inside structures (complex backgrounds, varying light levels, etc.)
6. Target engagement process at short ranges
7. Representation of “bad” information: incomplete, inaccurate, inconsistent
8. Representation of the integration of hardware/equipment with human systems

Intelligent Agents may have several characteristics that are important to consider from a validation perspective:

1. Perception: can sense their environment (key point, perceptions can be subjective, incomplete, or just wrong!)
2. Action: can effect change on their environment
3. Knowledge: can relate perceptions to world object "states" and make inferences to supplement perceptual data
4. Autonomy: can act based on current perceived world state instead of following only pre-programmed actions

C.10.1 Emergent Analysis

Mr. Middleton presented his approach for Emergent Analysis, which is defined by the following analytical phases:

The Problem Statement: The problem statement must be well-defined and well-scoped to establish the desired content and form of the answer. It should state the critical elements to the analysis question at hand, preferably in quantifiable terms, employing MOPs and MOEs.

The Operational Narrative: The operational narrative frames the necessary and salient elements of the analysis. A set of standard scenarios is a key element to the ABS Val process. These may take the form of use cases of operational scenarios developed from an operational perspective (versus a modeling perspective). Face validity should be used to determine credibility and completeness of the narratives with respect to responding to the problem statement.

The Conceptual Model: Like the operational narrative, the conceptual model must be intelligible to both the analyst and to the decision maker. The Conceptual Model is derived from the operational narrative to provide a conceptual foundation on which the computer simulation can be developed. The Conceptual Model facilitates validation by providing a comparator for determining the completeness, fidelity, and resolution of the developed simulation.

Representational Model/Simulation: The appropriate choice and appropriate application of tools for the development of the simulation are both required for sound results. The simulation can be validated by determining how well it corresponds to the requirements of the problem statement, operational narratives, and the Conceptual Model.

Experimental Instantiation: Executing the simulation and analyzing the results can perform simultaneous roles in analysis and validation. Results can be vetted against any available referent data (ideally empirical data) or SME knowledge. The simulation may produce results that may be invalidated or explored as useful emergent behavior. Parameters can be adjusted to determine if the simulation is sensitive to such adjustments.

C.10.2 Validation Steps

Mr. Middleton defined the following steps for the validation process, shown in Figure 16:

Step	Validation Criteria
Definition of a valid problem statement	a well-formulated problem statement expressed in the "right" MOPs &MOEs
Construction of a set of valid operational narratives/ scenarios/ use cases.	1) face validity and credibility derived from the developer(s)/ authoritative agencies 2) internal consistency 3) intensity/ depth of individual cases, breadth of the entire set
Selection/ adaptation/ development of a conceptual model	Degree of completeness, fidelity, resolution
Selection and validation of a reliable simulation	Representation of scenarios/ use cases based on the fit between the simulation and conceptual model - fidelity, resolution, completeness
Validity of the experimental design and its implementation	Determined by feedback, assessment & interpretation of simulation results.

Figure 16 Middleton Validation Process

C.11 SURF ZONE/BEACH ZONE (SZ/BZ) OBSTACLE REDUCTION SIMULATION (MR. R.W. PATERSON)

Mr. Paterson briefed the audience on the Surf Zone/Beach Zone (SZ/BZ) Obstacle Reduction simulation. This simulation, built in Pythagoras, is the second candidate model on which the Study team will exercise the VV&A framework.

The operational threat that was central to the analysis question addressed by the model is depicted by the graphic in Figure 17. The standard mine threat was provided to us by Marine Corps Intelligence Activity. It includes various mines, obstacles and concertina wire.

The background for this particular analysis was that at the time, there was no program of record for clearing shallow water obstacles, especially in the surf zone. Precision guided munitions were thought to hold some promise in clearing these obstacles, but required US Air Force support. The question posed was: Could agent based simulations help with this problem?

As input data, the analysis had Newton's Laws of Motion and results from actual, live fire tests conducted at Eglin Air Force Base. For these tests, a pond was built and various controlled experiments were performed. The tests included varying both the water depth and the depth of the bomb in the water when it detonated. After the test, the pond was drained to see what happened. All measurements for displacement distances were recorded. All three MK series bomb sizes were used for the tests. Mines near the blast's center were detonated. These results were used to create the probability that the bomb 'killed' the mine through sympathetic detonation. Through regression analysis, a predicted lethal radius was determined, which showed that there is a trade off between mines being killed versus mines being moved or tossed aside as a function of water depth.

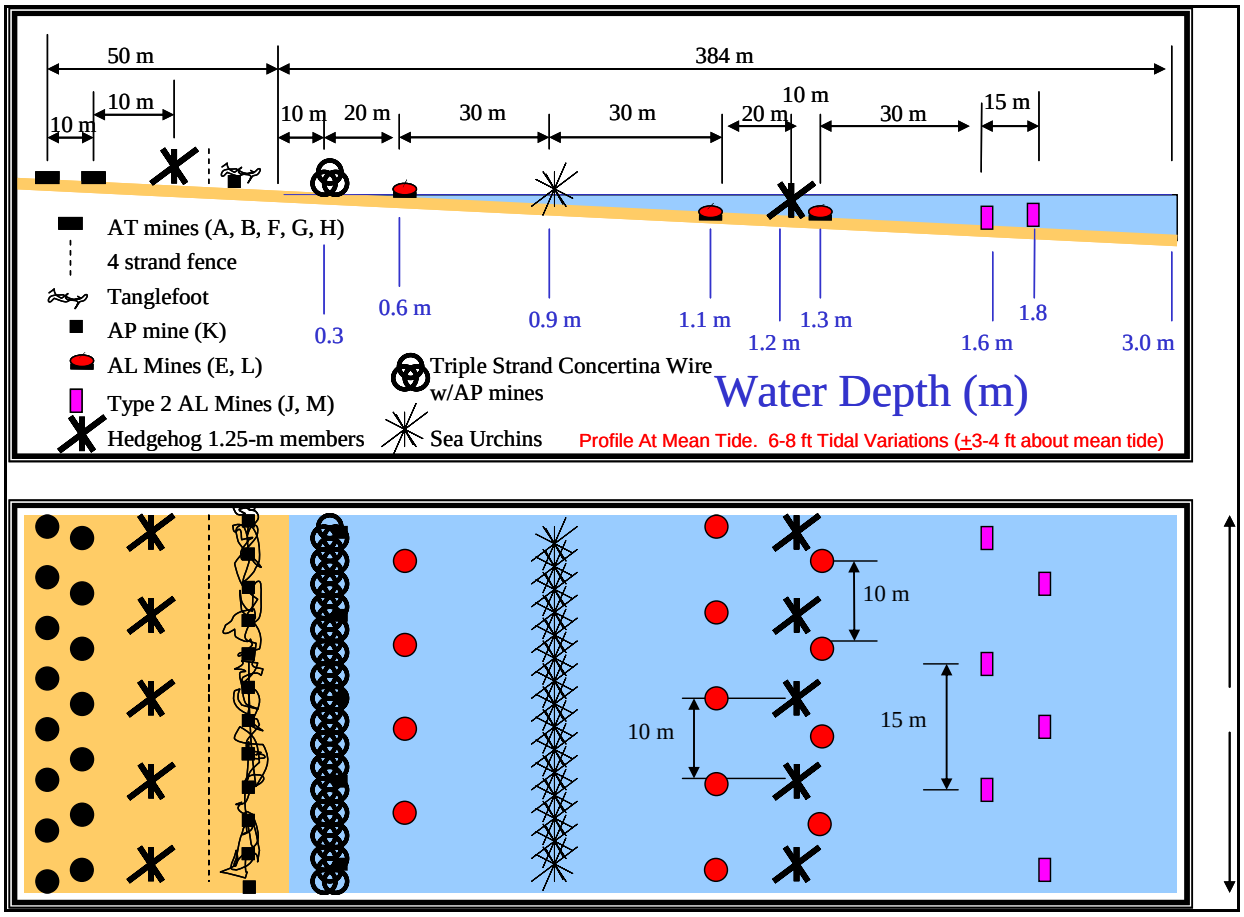


Figure 17 Obstacle Clearing Model

Given this test data, an agent based simulation was set up to attempt to simulate the physics of the problem. A number of operational questions remained:

- How well placed (accurate) do the bombs need to be?
- How well to the landing craft need to know the location of the lanes being cleared?
- Do the obstacle belts need to be located in advance?
- How many weapons are required per lane?

Using the Pythagoras simulation of the Surf Zone/Beach Zone (SZ/BZ) Obstacle Reduction scenario, alternative tactics were explored, such as channeling, belt strikes, and point strikes. Accuracy requirements were investigated, using both precision JDAM and regular gravity bombs. It was determined that the output should include the number of lost amphibians, rather than a measure of the size of the breach. This would create an operational effectiveness measure.

Through the simulation, the number of lost amphibians was analyzed. By varying mine location and aim point location proximity, it was determined that there was some sensitivity to these parameters. It was also determined Bomb type three provided the best results (fewest lost amphibians).

The conclusions drawn from the simulation-based analysis was as follows:

- Precise weapons cleared a narrow lane.
- Imprecise weapons cleared a much wider lane.
- There was a tendency for the mines to be shoved along the lane in the precision case, while in the imprecise case, if a mine was shoved along the lane by one bomb; it was likely to be moved laterally by another.
- More bombs were better than fewer bombs

The unexpected finding that resulted from the simulation analysis of the Surf Zone/Beach Zone (SZ/BZ) Obstacle Reduction scenario was that precision was not required. This experiment was repeated using live ordnance and a B-52. The Pythagoras results were confirmed in the live fire test.

C.12 SUMMARY OF WORKSHOP (IPR DISCUSSIONS)

The purpose of the meeting was to summarize the lessons learned and the concepts to be considered from the VV&A collaborative workshop held on 25-26 March 2008. Specifically, items that should be addressed as deficiencies in the current VV&A framework and potential areas for improvement were discussed.

The primary theme of the discussion centered on the need (expressed by Dr. Bailey) for the VV&A framework to disengage somewhat from the focus on validating the nuances of the Conceptual Model and to add content to the framework that provides an assessment of the usefulness of the model toward answering or examining an analytical question. The bullets below provide a brief summary of items discussed and key points made during the discussion.

- For any model, substantial criticism can be brought to bear against the aesthetics of the Conceptual Model and a model's implementations of abstractions and assumptions, especially for human-system based models. The framework should not provide a foundation for the validation activity to get bogged down in these criticisms, but rather, it should provide an avenue for examining if the assumptions made in designing the model compromise its usefulness toward answering the analytic question, and if so, how much. The usefulness of a model may only be to explore and eliminate some operational options, depending on the analysis objectives/questions that the model's application addresses.
- In the framework, there needs to be an equal emphasis on the analysis question and the usefulness of the model being validated towards analysis objectives. The model may have significant assumptions and abstractions that may render it invalid in a broader or alternative context, but the framework process needs to be able to determine if those things actually affect the model's usefulness toward the specific analytical question being posed. The decision context allows a validator to accept a level of abstraction in the model that might be completely unacceptable given some other context.

- Greater emphasis on the application of the model to analysis question/objectives drives a need for analysis question/objectives to be firmly scoped, with well understood requirements and demands for the application at hand.
- Validation cannot be decoupled from its intended use, especially when that use is analysis. Therefore, validation is part of the analytical process.

Below (Figure 18) is a notional illustration that Dr. Bailey provided for bridging the gap between the analysis-driven approach and the model-driven approach. Dr. Bailey referenced this notional illustration when asked what was lacking in the current framework.

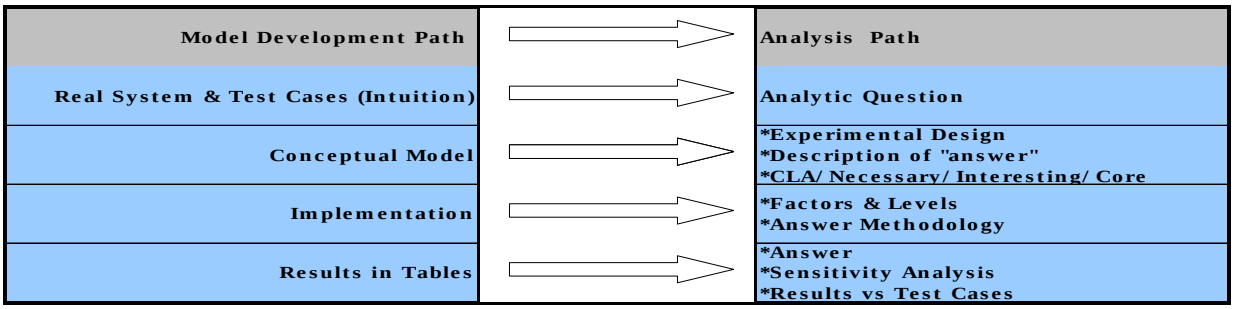


Figure 18 Dual Paths of Model Development and Application Analysis

The project framework, thus far, has emphasized the model and needs additional work to address the application of the model within the analysis process.

APPENDIX D NPS OBSERVATIONS FROM WORKSHOP #3

The following pages contain the document submitted by the Naval Post Graduate School participants providing their observations on Workshop #3.

Observations and Suggestions from the 25-26 March 2008 ABS VV&A Framework Study Workshop

David Kelton, Gary Horne, Ed Lesnowicz, and Tom Lucas
(wdkelton@nps.edu, gehorne@nps.edu, ejlesnow@nps.edu, twlucas@nps.edu)
SEED Center for Data Farming, Operations Research Department, Naval Postgraduate School
(http://harvest.nps.edu)
7 April 2008

We appreciate the opportunity to participate in this important and interesting effort. This document summarizes our observations from the 25-26 March 2008 workshop held at Northrop Grumman's Fair Lakes facility in Fairfax, Virginia, and offers several suggestions ranging from quite general to quite specific. Throughout, we try to stay focused on the VV&A effort, with special attention to validation.

There appeared to be general agreement at the workshop that a simulation model can never be 100% validated, and we agree with this. We also feel that a software platform like Pythagoras is not even subject to validation, since how various and sundry people choose to use (or misuse) it cannot be controlled. Rather, a practical validation toolkit should consist of a variety of spears with which to "prod" a specific simulation model in an attempt to *invalidate* it; failure to invalidate in this way builds confidence in the model's validity. Some such spears are rather obvious, like probing the corners of the feasible input-parameter space to "stress" the system and (hopefully) observe that it responds in a predictable and intuitive way. Other, more sophisticated methods might also be considered, such as attaining good, yet efficient, coverage of the input-parameter space through the use of Nearly Orthogonal Latin Hypercube (NOLH) designs in order to check validity throughout the parameter space, not for just a few anecdotal examples that were likely chosen haphazardly; see, e.g., Kleijnen, Sanchez, Lucas, and Cioppa (2005), or the "Software downloads" section of http://harvest.nps.edu/. Higher-resolution validation depends on having sharper and longer spears, and in this report we suggest how such might be constructed. In addition, it is important to note that there are plenty of good uses of simulations that have not been (or cannot be) validated.

Section 1 tries to put the current validation effort in the context of extant simulation research literature on the topic. Section 2 offers what might be a somewhat different orientation with respect to validation metrics. Section 3 comments on a couple of specific features in the current design of Pythagoras with reference not only to validity, but also to execution accuracy and efficiency. Section 4 contains a specific idea, apparently novel in the validation context, for sample-size determination sufficient to probe deeply in the invalidation effort (which, if the model survives, should substantially enhance confidence in both its validity and credibility). Section 5 briefly discusses credible uses of simulations that have not been validated. Section 6 concludes and recapitulates.

1. ABS Validation in the Larger Simulation-Research Context

Starting with the first computer-simulation efforts in the 1950s and 1960s, simulation analysts, practitioners, and researchers have always been concerned with the issue of model validation, i.e., whether the conceptual model (apart from its executable computer-code implementation) is a valid representation of the real-world system being simulated, for the purposes of the study. Thus, a substantial body of research and a literature has been developed on simulation validation.

While we appreciate that the internal logic, operation, and perhaps intent of ABS models are different from more general simulation models (e.g., discrete-event simulations for queueing-based systems), the fundamental questions of validation is really the same—do the results from the simulation agree with the results that would be obtained from the real-world system? For this reason, it seems that there might be more attention paid in this effort to prior published research, and that the focus should be reoriented to building on that prior work and adopting it as needed for ABS projects, rather than reworking it altogether.

Many of the basic definitions still used in validation, including in the workshop, go back as far as Fishman and Kiviat (1968). Probably the most recognized and cited author on validation of simulation models is Robert G. Sargent; a recent work of his on the topic is Sargent (2007), which, in turn, contains many (56, to be exact) references back to the prior validation literature. Standard simulation textbooks, such as Banks, Carson, Nelson and Nicol (2005) and Law and Kelton (2000) contain chapters on validation and verification. Moreover, research on validation is ongoing, including ABS, such as that found in Champagne and Hill (2007). These works, while general in the sense that they apply to all different kinds of simulation (not just ABS), are

nevertheless quite specific about recommendations, procedures, statistical methods, and best-practices case studies. We recommend that the project make ample use of this previous work; certainly not all of what is contained in this literature will apply, but a good deal of it probably will, including general frameworks, terminology, concepts, and procedures.

2. Validation of Multiple Simulation Output Processes

Few, if any, simulations result in just a single output performance measure. Indeed, simulation has been called "antistatistics" in the following sense. In basic statistics, we take a possibly large data set of numbers, unintelligible in their raw form, and stew them down to a handful of numbers or graphics that can be easily digested and interpreted. In simulation, we take a few input parameters and distributions, and produce from them a blizzard of output data from which we need to make some kind of sense.

What usually complicates this problem even further is that most simulations produce multiple output sequences over time (assuming a dynamic, not static, simulation). This output can be represented as a vector-valued stochastic process $\mathbf{Y}(t) = [Y_1(t), Y_2(t), \ldots Y_k(t)]$ where each $Y_i(t)$ is a scalar-valued output stochastic process. In turn, then, we often wish to estimate some parameter of each output process, such as $\lim_{t\to\infty} \mathrm{E}[Y_i(t)]$, the steady-state (i.e., long-run) expectation of the ith output process, or $\lim_{t\to\infty} \mathrm{P}[Y_i(t) > c]$, the steady-state probability that the ith output process will exceed some constant c of interest (short-run, or terminating, versions of such outputs are also certainly possible and of interest).

Research in simulation output analysis has focused on estimating such output parameters, usually via a confidence interval of some sort. In this context, multivariate statistical methods must be used. Turning to the validation side of this, we need to recognize that a model might produce valid results for some output processes, but not for others. If all output processes are of interest (as we presume they are, else they would not be simulated in the first place), then this is not good enough. A model that is valid for some, but not all, output processes is simply not valid. Complex military simulations involving combat, logistics, or communications, will likely have multiple outputs, and these all need to be validated as part of the overall model-validation effort.

How might multivariate validation be done? Again, looking back to extant simulation-research literature, there are many suggestions for statistical validation methods, assuming that

referent data are available. For instance, confidence intervals and hypothesis tests are often used in this context. For multivariate output, we must use the vector-valued versions of these. Such versions might be specialized multivariate statistical methods (e.g., Hotelling's T^2 test), or could be conservative adaptations of univariate methods, using the Bonferroni inequality to adjust the component-wise confidence levels. Unfortunately, this usually implies far greater data sets, but one cannot expect to achieve an ambitious goal like multivariate, vector-valued validation without some effort.

3. Some Specific Suggestions on Future Directions for Pythagoras With Respect to Model Validation and Efficiency

Since the primary software platform of interest to the group is Pythagoras, we offer two specific suggestions on how future versions might be improved with respect to validation as well as statistical efficiency (i.e., precision).

3.1 Time-Stepped Logic

The time-stepped nature of this platform, and of most other ABS combat-modeling platforms as well, presents problems; indeed, it was noted at the workshop that changing the size of the time step can definitely change the results, which, of course, is undesirable. The appeal of the time-stepped logic is strong and intuitive, perhaps because it is easy to understand in principle, and also is an outgrowth of physics-based models from which many combat models have grown.

However, there are two well-known criticisms of time-stepped models. Since all "action" takes place at the instant of a time step, the occurrence of events through continuous time must be rounded up to the next time step. This creates a modeling inaccuracy (and thus makes the model less valid), especially if the time step is relatively large.

This inaccuracy could be ameliorated by making the time step smaller, which leads to the second well-known criticism: Time-stepped models tend to be slow. And, of course, reducing the time step in an effort to make the model's event occurrences more accurate only exacerbates this run-time problem.

An attractive alternative for at least some aspects of Pythagoras would be to move to event-driven logic. This solves both of the two problems mentioned above inherent in time-stepped logic. Things will happen in the simulation exactly when they are supposed to (or, at least up to

the floating-point resolution of the computer, which is typically quite fine-grained), so that no model distortion is introduced by rounding off event occurrences to the next time step. In addition, since event-driven logic does not bother to represent the system at all between events (when nothing is happening anyway), they tend to be considerably faster. Though many people apparently believe that military-relevant activities like movement and detection can be simulated only in time-stepped logic, Buss and Sanchez (2005) demonstrated that this is simply not the case in general.

So, while it might not be possible to move Pythagoras completely over to event-driven logic, it seems that at least some aspects could move in that direction. Such aspects include most logistics activities, as well as at least some movement-and-detection activities, as noted above.

3.2 An Improved Random-Number Generator, and Its Implications

Many (not all) models for which Pythagoras is used are stochastic, i.e., involve at least some inputs that are draws from input probability distributions. Such models rely on an underlying random-number generator as the "engine room" from which draws from other input probability distributions are formed.

Stochastic simulations produce output that is also stochastic, which makes it uncertain and subject to variance. The less output variance present, the more precise the results. A "brute-force" method of variance reduction is always available—just run the simulation more, meaning either longer or more replications (or both). However, many combat simulations take a long time to run, so such brute-force variance reduction may not be practical.

There are, however, a variety of so-called *variance-reduction techniques* (VRTs) that have been developed, which reduce output variance with little or no additional computational effort. VRTs typically achieve this via judicious and careful *re*use of the basic underlying random numbers. Perhaps the best-known (and most widely used) example is *common random numbers* (CRN), applicable when two or more alternative simulated scenarios are to be compared (a typical simulation situation—change a weapons-system configuration and see what difference it makes). Key to making CRN and most other VRTs work properly is the ability to *synchronize* random-number use across the different scenarios, so that the same random number is used *for the same purpose* in different scenarios. Synchronization relies on the ability to precisely control the allocation of random numbers.

A proven route to random-number synchronization is to segment the underlying random-number generator into *streams* and *substreams*, which are nothing but subsegments of the underlying generator. Then one stream could be assigned to a specific purpose across all scenarios, another stream to a different purpose, and so on; the substreams are used to advance all streams in each replication so as to maintain synchronization past the first replication.

Currently, Pythagoras uses the built-in Java random-number generator, which does not support streams or substreams. A better choice would be one of the more modern and thoroughly tested generators, such as that developed by L'Ecuyer, Simard, Chen, and Kelton (2002), which supports streams and substreams, and has C++ code that is freely available. A side benefit is that the underlying random-number generator has a far longer period and provably better statistical properties.

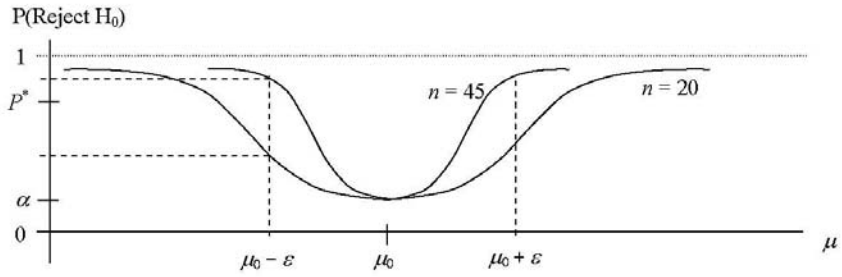
4. Simulation Sample-Size Determination for Achieving Desired Statistical Power to Reject H_0: Validity

As noted in the workshop, and in the validation literature referred to above, a useful setup for validation testing is to state the null hypothesis as H_0: this model is valid. Since one can never "prove" a null hypothesis, one can never "prove" that a model is valid. The best we can do is subject the model to a battery of tests in an attempt to invalidate it; while failure to invalidate does not constitute validation, it does build confidence.

But how much confidence? One situation in which we can specifically address this was mentioned in the workshop—we have a "tolerance" $\varepsilon > 0$ on some output parameter, within which we are willing to declare the model as, say, *practically valid.* We also must have available a referent, and a particular parameter μ_0 that constitutes "exact" validity.

So we are willing to declare the model as practically valid if its output parameter μ is no more than ε away from μ_0. The idea is to come up with a sample size (number of replications) n of the simulation model that will produce an acceptable power (type II error probability) of at least, say P^*, that we will reject H_0: validity, if the model's actual μ is more than ε away from the desired μ_0 (P^* could be anything here, at the discretion of the decision maker). Put another way, we really do want to invalidate the model if its mean is too far away from the target; on the other hand, if the mean is close enough to the target we will accept the model as practically valid.

How would such a sample size be determined? As a start, if the parameters are means and standard normal-theory statistics can be robustly applied, the figure below illustrates the idea.



The horizontal axis is the true value of the mean of the model, and the vertical axis is the probability that we will reject H_0: model validity, as a function of the true mean of the model. Thus, this is a standard power curve of the validity hypothesis test. The size (type I error probability) of the test is α, so that we wrongly reject a true null hypothesis with probability α, which is the standard setup in hypothesis testing. Just as an example, in this figure there are two sample sizes, 20 and 45, and the two curves give the probability of finding the model invalid for the corresponding value of m on the horizontal axis (larger sample sizes produce a "sharper" power curve). The desired value of P^* is depicted on the vertical axis. We see that a sample size of 20 is inadequate to give us the power we want, since at $\mu = \mu_0 \pm \varepsilon$, the power of detecting model invalidity of this magnitude is not up to the desired P^*. On the other hand, a sample size of 45 is more than adequate; by standard normal-theory manipulations and a reasonably crafted spreadsheet, it should be easily possible to solve for the smallest sample size that will detect model invalidity of a magnitude about which we care ($\pm \varepsilon$) with at least the specified probability (P^*). The analyst could then be assured of an invalidity "spear" that is just good enough to detect model invalidity as desired.

This is surely not the only such framework for such an analysis. For example, a one-tailed version could be easily constructed for situations in which we care about a response being too big, but not too small (or the reverse). If we are concerned about the normality assumption, we could turn to nonparametric tests such as those described by Conover (1971). And, if we are

dealing with an autocorrelated time series, as is characteristic of the output from many simulations, we could turn to time-series estimation techniques.

5. Credible Uses of Unvalidated Models

George Box (1979) famously remarked that "all models are wrong, but some are useful." Indeed, as discussed above, most validation efforts are attempting to estimate how wrong they are in various dimensions. However, it is important to note (see Hodges and Dewar 1992) that models can be useful even when it is difficult or impossible to get sufficient data to accurately measure reality—such as, e.g., models of global nuclear war. Indeed, people have recently been interested in agent-based simulations to explore hard-to-quantify issues such as a population's allegiance to their government or their overall frustration level.

The key question is how do we effectively use models in situations in which the verisimilitude of the model is difficult to assess? Hodges and Dewar list seven credible uses of unvalidated models and the criteria for evaluating those uses. For example, an unrealistic model, perhaps deliberately so, can be an effective training aid if it helps induce a desired behavior. For analytical purposes, if a model is consistent with the salient information at hand, as determined by subject matter experts (SMEs), and is deemed to produce "plausible outcomes"—i.e., outcomes that experts cannot refute as possible—then it can support good decision making. For instance, if a model suggests that a particular course of action has a high probability of a catastrophic outcome that cannot be discounted by SMEs, then it likely makes sense to avoid that course of action. Unvalidated models can also be effectively used to assess the consequences of various assumptions (sometimes called logic tracing), generate hypotheses, communicate ideas, and help organize debates. It is our view that many of the effective uses of agent-based simulations will be along these lines. Indeed, it is sometimes said that there are no valid force-on-force models, but there are valid uses of them.

6. Conclusions

We've tried to make some specific, concrete, and practical suggestions about how ABS validation, specifically using the Pythagoras platform, might be attempted. We wish to emphasize, however, that no software platform, be it Pythagoras or anything else, can ever be absolutely validated. Rather, we prefer to think of assessing the degree of validity, and of

specific models constructed with Pythagoras, and have made some concrete suggestions about how this might be quantified and improved.

We feel that it could well be profitable to look more at the extant simulation-validation literature, since ABS models have much in common with more general simulation models from this standpoint. We also feel that it would be helpful to be more precise in setting up and analyzing the specific output processes and hypothesis-testing framework. This could lead to some specific kinds of analyses, like the sample-size determination we described. We also offered a couple of concrete suggestions on future development of Pythagoras itself. Finally, with proper usage, we think implementations in Pythagoras can have value even if the situation is such that the model has not been or cannot be validated.

References

Banks, J., J.S. Carson II, B.L. Nelson, and D. Nicol (2005), *Discrete-Event System Simulation*, 4th ed., Englewood Cliffs, NJ, Prentice-Hall.

Box, G. (1979), "Robustness in the Strategy of Scientific Model Building," *Robustness in Statistics*, R. Launer and G. Wilkinson (Eds.), Academic Press, New York, p. 202.

Buss, A.H. and P.J. Sanchez (2005), "Simple Movement and Detection in Discrete Event Simulation," *Proceedings of the 2005 Winter Simulation Conference*, pp. 992-1000, open access at http://www.informs-sim.org/wsc05papers/118.pdf.

Champagne, L.E. and R.R. Hill (2007), "Agent-Model Validation on Historical Data," *Proceedings of the 2007 Winter Simulation Conference*, pp. 1223-1231, open access at http://www.informs-sim.org/wsc07papers/144.pdf.

Conover, W.J. (1971), *Practical Nonparametric Statistics*, Wiley, New York.

Fishman, G.S. and P.J. Kiviat (1968), "The Statistics of Discrete-Event Simulation," *Simulation*, 10, pp. 185-195.

Hodges, J. and J. Dewar (1992), "Is It You or Your Model Talking? A Framework for Model Validation," R-4114-A/AF/OSD, RAND, Santa Monica, CA.

Kleijnen, J.P.C., S.M. Sanchez, T.W. Lucas, and T.M. Cioppa (2005), "A User's Guide to the Brave New World of Designing Simulation Experiments," *INFORMS Journal on Computing* 17, pp. 263-289 (with online companion available open-access at http://www.informs.org/site/IJOC/article.php?id=22).

Law, A.M. and W.D. Kelton (2000), *Simulation Modeling and Analysis*, 3rd ed., New York, McGraw-Hill.

L'Ecuyer, P., R. Simard, E.J. Chen, and W.D. Kelton (2002), "An Object-Oriented Random-Number Package with Many Long Streams and Substreams," *Operations Research*, 50, pp. 1073-1075.

Sargent, R.G. (2007), "Verification and Validation of Simulation Models," *Proceedings of the 2007 Winter Simulation Conference*, pp. 124-137, open access at http://www.informs-sim.org/wsc07papers/014.pdf.

This report can also be found at http://orsagouge.pbwiki.com/ABSVal.

APPENDIX E WORKSHOP #4 SUMMARY

This report can also be found as a stand-alone at http://orsagouge.pbwiki.com/ABSVal.

E.1 INTRODUCTION

The Marine Corps Combat Development Command (MCCDC) Operations Analysis Division (OAD) hosted a three-day workshop to review the results of the application of the Agent-Based Simulation (ABS) Verification, Validation, and Accreditation (VV&A) Framework to the set of models considered during the Agent Based Simulation Verification, Validation, and Accreditation (VV&A) Study Joint/DoD Phase II study. This workshop presented the final outbrief and results of Phase I and Phase II of this project including an overview of the test cases and the developed Agent-Based Simulation Validation Framework. Topics included presentations on the framework, the models used in the framework's testing, the validation experiments applied to those models in application of the framework, and a presentation of the results of those experiments. The purpose of the workshop, held on 8-10 July 2008, was to accept feedback and criticisms of the framework. This summary report describes presentations and feedback given during the workshop.

E.1.1 OVERVIEW

The agenda for the workshop included an update to the framework as developed through its application on the analyses, a description of the models under consideration for the workshop participants, an overview of the validation results, and some general commentary from Subject Matter Experts regarding the topic of ABS and the concepts of validation. The workshop included ten topical briefings:

(1) "Sensible Validation for IW scenarios" (Dr. Michael Bailey)
(2) "Some Comments on Models" (Mr. Eugene Visco)
(3) "How Simulation Theory Supports the Validation Framework" (Dr. Eric Weisel)
(4) "ABS Validation Framework Overview" (Ms. Lisa Moya)
(5) "Description of Pythagoras Obstacle Clearing Model (P-OCM) and Analysis Application" (Mr. Edd Bitinas)
(6) "Pythagoras Obstacle Clearance Model: Validity Assessment" (Mr. Robert Eberth)
(7) "Overview of Pythagoras COIN Scenario, Model and Analysis" (Mr. Edd Bitinas)
(8) "Pythagoras COIN Conceptual Model Validation" (Mr. Robert Eberth)
(9) "Pythagoras COIN Analysis Validation" (Ms. Lisa Jean Moya)
(10) "Results from Validation Audits" (Mr. Amos Kenyon)

This report summarizes the activities of the Agent Based Simulation (ABS) Verification, Validation, and Accreditation (VV&A) Framework Study Phase II Workshop #4 (the fourth workshop held during the entirety of the study, including both phases). Materials from this workshop can be found at http://orsagouge.pbwiki.com/ABSVal.

Following each discussion of a validation application (presentations (6), (8), and (9) above), a survey was given which petitioned workshop participants for comments and opinions on the validation analysis. Subsequently, several open discussion forums were initiated during the workshop to communicate these comments and opinions, brainstorm with respect to the validation process, and identify any gaps perceived in the validation framework process. While questions regarding the validation assessments were entertained during the briefings, participants' questions, answers, and comments regarding the briefings presented were deferred to these open discussion forums following the completion of surveys. This report also documents these discussions.

E.1.2 Sponsor Perspective: "Sensible Validation for IW Scenarios" (Bailey)

Dr. Bailey briefed the workshop participants on the study sponsor perspective on the ABS VV&A Framework Study and some current thoughts on IW analysis/validation of ABS. The discussion that follows is taken from minutes recorded during his presentation.

Dr. Bailey characterized both the analysis and validation of ABS as being at its "best" when conducted as an exploration activity. With respect to the simulation validation processes, challenges are ubiquitous because of the nature of analysis coupled with validation, where unknowns exist and ideal data and simulation dynamics are often hard to come by.

From ABS experiments such as ISAAC, it can be seen that seemingly organized emergent behavior can be achieved through the tweaking of simulation parameters. The challenge for validating ABS experiments such as these is to understand if these results are an item of analysis interest or a simulation anomaly. First principles, Conceptual Model validation, while a useful tool, typically falls short as a testing method for ABS with emergent behavior, adding emphasis and relevance to Results validation as a validation tool, which inherently is also problematic when dealing with unexpected results. The validation activity must always consider the analytic questions at hand to frame the application; the objective is to "match" the simulation to analytic goals.

Dr. Bailey postulated a method of dealing with the challenges of validating and examining these ABS models by defining analysis elements that are characterized by simulation dynamics coupled with required data. The Yost scale was presented as a tool for characterizing these analysis elements, where the level of abstraction in a simulation is coupled to an associated level of data integrity. Simulation analysis was presented conceptually as having 4 layers:

(1) Core (drive the results of your experiment, align with the key elements of analysis)
(2) Habitat (has impact on the circumstances relevant to exercising the core model dynamics, create situations, not elements of analysis)
(3) Orientation (details necessary to support the model, cases to be considered to achieve analytical goals)

(4) Constraints, Limitations, and Assumptions (necessary to give context, scope the analysis, and interpret the results).

With respect to validation, analysts should have the most confidence in the core elements, having high-quality data and well-studied simulation dynamics (high on the Yost scale, which is a qualitative scale of simulation dynamics and data) without displaying uncontrolled emergent behavior (emergent behavior should arise out of habitat elements). Negative information about a model could be characterized as the following:

(1) Elements not data-driven
(2) Elements not controllable
(3) Element displays undesired emergent behavior
(4) Element displays unexplainable 1st-order influence
(5) Element is not in the anticipated level of influence with unanticipated dynamics

The analyst's understanding of the model, the data, and the relationship of that model's core elements to the analytical question are key to ensuring confidence in the product of the analysis. Negative information scopes the analytical power of results, and it is the "art" of the analyst to expand this scope responsibly.

E.2 BACKGROUND BRIEFINGS

The following are summaries taken from minutes of briefings that were offered during the workshop as background for the developed ABS VV&A Framework.

E.2.1 Comments on Modeling and Validation (Visco)

Mr. Visco briefed the audience on some of his thoughts on the validation of ABS. One of Mr. Visco's initial points focused on trying to change the attitude that has fostered notions such as the commonplace quote "all models are wrong …", contending, instead, that models are not intended to replicate reality as that statement may imply, but are designed to be useful tools for the analysis of complex phenomena.

Through an examination of the history of analytical models, Mr. Visco drew a relationship between the advancement of technology and the potential loss of model integrity that ensued: as models began to be more complex, including many more variables, entities and assumptions, less attention was paid to the expression of assumptions and the impact of assumptions on model behavior. Mr. Visco cited that validation was not a focus historically, and that some of the challenges present with modeling IW, ABS, and validation have persisted and remain somewhat unresolved—"…still struggling with the concept of the truth and applicability of models, rife with assumptions, both stated and unstated." Mr. Visco emphasized that flawed assumptions about phenomena that exist in models often perpetuate and become generally accepted, detracting from the value of the analysis based on these assumptions and delivering inaccurate results. Mr. Visco proposed that analysts must "Take back the responsibility for our work—take it away from a pro-forma, mechanistic, devoid-of-substance process and put it where it belongs: in the hands of the analyst,"

recognizing that models are tools that must be vetted and carefully examined in each application of the tool by the analyst to drive confidence in results. In doing so, these advancements in technology, which may have "softened" rigorous and carefully considered analysis, may be harnessed in the right way, and that a "model improvement" paradigm may be achieved, in which the quality and usefulness of these tools can progress through this iterative/collective cycle of improvement.

E.2.2 "How Simulation Theory Supports the Validation Framework (Weisel)

Dr. Weisel briefed the workshop participants on the fundamental theory that supports the validation framework. The DOD definition for validation has three key elements: accurate, real world, and intended use. The goal of developing a theoretical foundation for ABSVal was to tie the simulation to the real world in some mathematical way. The presentation generated significant discussion regarding the place and purpose of the theoretical elements to the ABS VV&A Framework.

E.2.2.1 The Validation "Cloud" Diagram

The validation "cloud" diagram formed the foundation for the discussion, shown in the diagram below in Figure 19. Simulation elements are on the right hand side; the real-world elements being simulated are on the left hand side. The outside nodes of the cloud diagram represent elements of practical validation (tangible), while the inside nodes generally represent elements of theoretical validation (theoretical / mathematical).

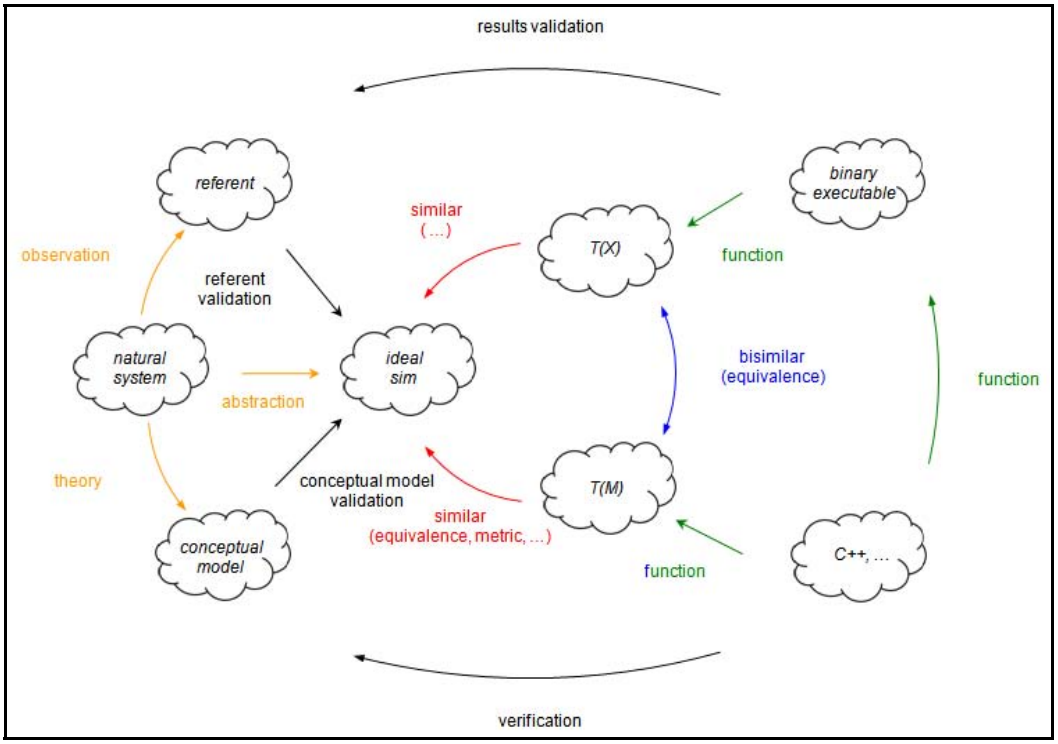


Figure 19 Validation "Cloud" Diagram

Validation may be defined as a comparison of transition systems. A digital computer running a programmed simulation creates one transition system, whose trajectory can be compared to the transition system of what is being simulated.

E.2.2.2 Natural System

The Natural System (aka the Target System or the System of Interest) is derived as a subsystem of the universe in its entirety from the perspective of a perfect observer at a snapshot in time. If the universe from the perspective of a perfect observer was depicted as an infinite vector, one could derive the natural system/system of interest for a simulation by taking a notional "slice" of that vector. Since, pragmatically, not all elements in this Natural System can be modeled; an abstraction process is used to focus on elements of interest in the Natural System for the purposes of simulation. In mathematical terms, this abstraction amounts to the bounding of the state vector and the input vector. Abstraction is applied to create non-deterministic transition systems, creating the Ideal Simulation as depicted in the cloud diagram. For the Ideal Simulation, the values of vectors match exactly with the corresponding values in the Natural System with no loss of accuracy (it extrapolates perfectly), but the elements of the Ideal Simulation are limited to what has been derived by abstraction. Theoretically, if one could not answer a posed analytical question using the Ideal Simulation, then one could not answer that question with any simulation representation.

E.2.2.3 Code

The Code element in the cloud diagram represents a transition system in the form of a binary executable program. What the validation effort attempts to demonstrate is that the transition system that the code generates simulates the Ideal Simulation (every move the ideal simulation makes, the coded simulation can match). However, there are problems of tractability in generating a mathematical proof that shows that one transition system simulates the other, and there is no algorithm to evaluate these transition systems trajectory by trajectory.

Thus, for the inner loop (theoretical elements) of the cloud diagram validity cannot be proven (in an exclusively mathematical sense), and so other techniques of validation must be employed. Results validation is often used to provide evidence of the validity of a simulation; the binary executable is run to acquire an array of trajectories, which can be compared to a referent, real-world empirical data from observations. Often for social systems, the referent is not solid; however results validation can still be employed (although much more subjectively) to assesses if simulation output is sensible and useful. The referent data, in turn, requires validation as well, to determine accuracy and suitability to the application.

In the framework, the intended use of the simulation defined by the analytical application drives the abstraction process to determine what details and elements in are critical for the simulation. Intended use also drives validation criteria (e.g., "Do the states in the simulation need to be within some tolerance compared to the Ideal Simulation?" or "Can more subjective methods be applied to determine the validity or usefulness of the simulation?).

E.2.2.4 Conceptual Model

The Conceptual Model arises out of the Natural System by theory—it is the "best that is known" about the Natural System from a theoretical sense. An aspect of validation may be to provide evidence that the elements in the Conceptual Model are true with respect to the Ideal Simulation. This often may occur with some degree of subjectivity, especially when simulating human systems. Because of the inability to comprehensively validate a human system, the scientific method is employed in the framework as a method in which validation techniques are applied to the simulation in an attempt to disprove or "poke holes" in the integrity of the simulation with respect to the application. With rigorous validation techniques, the inability to invalidate the simulation subsequently provides evidence of its validity and usefulness toward an analysis application.

E.3 ABS VALIDATION FRAMEWORK OVERVIEW (MOYA)

Ms. Lisa Moya briefed the audience on the VV&A Framework developed for the Study. Ms. Moya presented a list of framework requirements as follows:

(1) Understanding the meaning of Valid Enough for an Intended Use Application of Analysis
(2) Techniques for uncovering validation shortcomings in the presence of a weak referent
(3) Expressing validation boundaries
(4) Methods to communicate risk
(5) Being conservative with VV&A resources

In summary, Ms. Moya stated that through the validation process, we want to be able to provide a report (to decision makers, analysts, developers, etc.) that elucidates the limitations of the model/analysis. Since one cannot typically determine that a simulation is valid as a matter of fact, this report must demonstrate whether the model is "good enough"/not "good enough" with respect to the intended use of the model in the analysis context. The framework, therefore, must contain the tools and techniques for discovering deficiencies in the model in that regard.

The framework developed for this study has been an evolving process. In the study, the study team initially researched the DMSO handbook as a potential foundation for the framework and discovered that the techniques and processes therein fell short in validating ABS models in several ways:

(1) DMSO guidance did not focus on intended use
(2) DMSO guidance did not give information on how or when to apply specific validation techniques
(3) DMSO validation techniques were not applicable to ABS
(4) Conflicting guidance in DMSO: the handbook states that face validation is the worst technique for validation, but subsequently states that for ABS, Face Validation is the only technique known

The study then went on to develop an initial VV&A framework, which focused on the validation Conceptual Model against the referent. This approach still seemed to have shortcomings with respect to the analysis context. A conclusion drawn from the prior workshop during this Study is that validation is an analysis process, and it is intertwined with the analysis being conducted for which ABS is employed as a tool. This is challenging from a framework development perspective because often these analysis processes and techniques are difficult to describe in "handbook" form —there is an "art" involved driven by analyst intuition and subject matter expertise. Some aspects of this analysis are understanding parameters and how they change, understanding the surprise elements in the model, and removing those surprise elements (i.e., uncontrolled emergent behavior).

Also evolving is the notion of what needs to be communicated to the decision maker, the context of ABS results and why the results are credible/not credible. The concept of Constraints, Limitations, and Assumptions (CLAs) has become more of a focus as a method to communicate this information, recognizing that the validation of ABS models is typically not a pass/fail endeavor, but a communication of these CLAs, the dynamic factors influencing the model, and the impact of CLAs on the model's feasibility as a tool for its intended use. Validation is an "analysis" in and of itself, where validators are putting on an "analyst's hat" and attempting to be as comprehensive and well documented as possible with the examinations and results.

The evolution of the framework with this emphasis on analysis introduces new potential validation questions such as follows:

(1) Under what conditions am I evaluating alternatives?
(2) What are the core elements of the simulation analysis?
(3) What is causing variability in test case results?
(4) How are dynamic influencers impacting the results?

An Assessment of Risk may be communicated using validation results that express the potential consequences of using a particular model for the analysis.

All this information needs to be contextualized for the decision maker, focusing on the intended use (note that the analytical questions/intended use may need to be refined iteratively for the purposes of validation), and the framework must contain a description of what a really good analysis report would provide, recognizing that there are levels of rigor for validation that may or may not be achieved depending on the data available, the complexity of the model, and the validation resources available.

E.4 PYTHAGORAS OBSTACLE CLEARING MODEL (P-OCM)

This section summarizes the briefings given with respect to the validation of the Pythagoras Obstacle Clearing Model (P-OCM).

E.4.1 Description of the Pythagoras Obstacle Clearing Model and Analysis Application (Bitinas)

Mr. Bitinas briefed the audience on the Surf Zone/Beach Zone (SZ/BZ) Obstacle Reduction simulation. This simulation, built in Pythagoras, is the second candidate model on which the Study team exercised the VV&A framework.

The operational threat that was central to the analysis question addressed by the model is depicted by the graphic in Figure 20. The standard mine threat was provided by the Marine Corps Intelligence Activity. It includes various mines, obstacles, and concertina wire.

The background for this particular analysis was that at the time, there was no program of record for clearing shallow water obstacles, especially in the surf zone. Precision guided munitions were thought to hold some promise in clearing these obstacles, but required US Air Force support. The question posed was: "Could agent based simulations help with this problem?"

As input data, the analysis had Newton's Laws of Motion and results from actual, live fire tests conducted at Eglin Air Force Base. For these tests, a pond was built and various controlled experiments were performed. The tests included varying both the water depth and the depth of the bomb in the water when it detonated. After the test, the pond was drained to see what happened. All measurements for displacement distances were recorded. All three MK series bomb sizes were used for the tests. Mines near the blast's center were detonated. These results were used to create the probability that the bomb 'killed' the mine through sympathetic detonation. Through regression analysis, a predicted lethal radius was determined, which showed that there is a tradeoff between mines being killed versus mines being moved or tossed aside as a function of water depth.

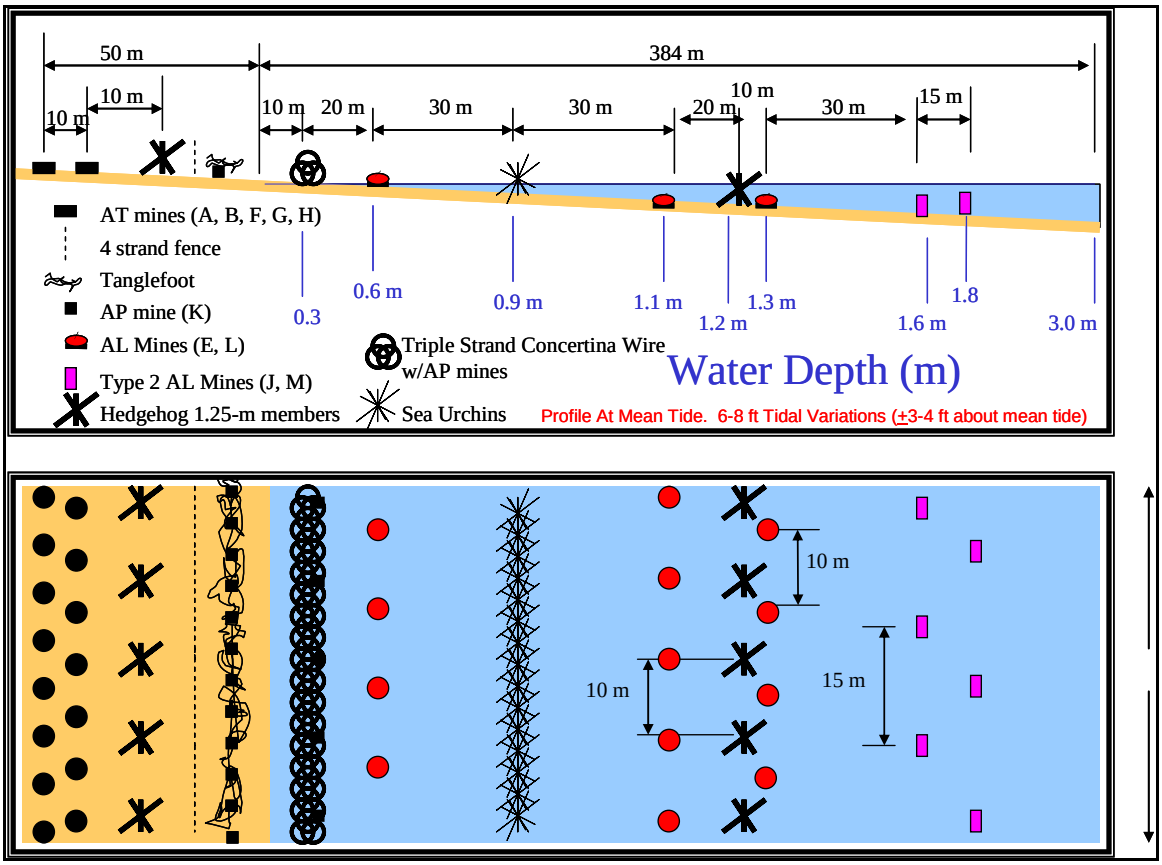


Figure 20 Pythagoras Obstacle Clearing Model

Given this test data, an agent-based simulation was set up to attempt to simulate the physics of the problem. A number of operational questions remained:

(1) How well placed (accurate) do the bombs need to be?
(2) How well to the landing craft need to know the location of the lanes being cleared?
(3) Do the obstacle belts need to be located in advance?
(4) How many weapons are required per lane?

Using the Pythagoras simulation of the Surf Zone/Beach Zone (SZ/BZ) Obstacle Reduction scenario, alternative tactics were explored, such as channeling, belt strikes, and point strikes. Accuracy requirements were investigated, using both precision JDAM and regular gravity bombs. It was determined that the output should include the number of lost amphibians, rather than a measure of the size of the breach. This would create an operational effectiveness measure.

Through the simulation, the number of lost amphibians was analyzed. By varying mine location and aim point location proximity, it was determined that there was some sensitivity to these parameters. It was also determined Bomb type three provided the best results (fewest lost amphibians).

The conclusions drawn from the simulation-based analysis was as follows:

(1) Precise weapons cleared a narrow lane.
(2) Imprecise weapons cleared a much wider lane.
(3) There was a tendency for the mines to be shoved along the lane in the precision case, while in the imprecise case, if a mine was shoved along the lane by one bomb; it was likely to be moved laterally by another.

The unexpected finding that resulted from the simulation analysis of the Surf Zone/Beach Zone (SZ/BZ) Obstacle Reduction scenario was that precision was not required. This experiment was repeated using live ordnance and a B-52. The Pythagoras results were confirmed in the live fire test.

E.4.2 P-OCM Validity Assessment (Eberth)

Mr. Eberth briefed the audience on the validation of the Pythagoras obstacle clearing model and analysis application. The following is a summary of some of the key points provided in Mr. Eberth's brief:

The Study objective of performing the validation exercise was to test the VV&A Framework, but to do this; both the analysis and the application had to be examined. Assumption testing was the focus of the validation efforts, relying on this technique of identifying and vetting assumptions to discriminate the model's usefulness with respect to the intended use. With respect to ABSVal, an assumption may be defined as the "mechanism of abstraction from reality." (If you could list ALL assumptions, you would know all the ways the model deviates from reality). Subsequently, it must be determined if those are acceptable per the analysis context. One possible method of determining acceptability of the assumptions is by allowing the decision maker to determine acceptability of abstractions.

The scientific method of attempting to falsify the null hypothesis that the simulation is valid was employed. Using this method, there is no absolute validity, but by failing to falsify the null hypothesis, evidence is provided of the usefulness of the model. Analytical questions and objectives frame the intended use of the model, which must be considered throughout the validation process. For P-OCM, there were no specific accreditation criteria.

Some challenges with assumption testing were recognized during the validation exercise: assumptions are often not written down but are imbedded in the algorithms of the model, which requires sometimes tedious reverse engineering of assumptions from algorithms.

It was recognized that there were two slightly different sets of analytical questions, (which created some ambiguity during validation), as listed below:

Initial Study Questions:

(1) Are there current weapons that can be used differently to defeat obstacles?
(2) Are there promising new technologies that Project Albert can model?
(3) Is agent based simulation modeling a realistic tool for this problem?

(4) Is there anything we need the models to do differently for this problem

Analytic Study Questions:

(1) What accuracy is best?
(2) Do simultaneous or sequential detonations play a role in developing a better "lane?"
(3) Is there significant difference between using precision bombs versus conventional bombs?
(4) How many bombs should be dropped at each aim point?

One of the primary conclusions from the validation exercise was that the model results were based on an MOE (# of AAVs killed as they attempted to transit the lane after the clearing operation) that was possibly not ideal for the analytical problem at hand. The MOE chosen added a lot of variables/complexity to the problem. A better choice may have been obstacle density/displacement before and after the clearing activity. The MOE choice drove the design of the model, which ultimately omitted potentially valuable and necessary data items for analysis. For example, data that described the displacement of mines and obstacles were not available, resulting in an inability to examine the effects of the obstacle clearing operation scenario in any more depth.

The model underwent numerous runs, using the volume of input data created during the live "pond tests." It was observed in examining input data that significant variation existed in the empirical real world pond tests. For example, in examining the "plume" effect of the bomb, some obstacles were pulled toward center of the plume (vacuum effect), while others were pushed away from the plume. In the design of the model, it was recognized that the target boxes on belts were of variable sizes, and that Bombs could push obstacles/mines out of belts.

Assumption testing of the P-OCM revealed several assumptions as shown below:

Overall:

- For both this assessment and the one for P-COIN, Pythagoras was assumed valid
  (Effect is indeterminate, but in this particular case there is reason to question the validity of Pythagoras. In general, assuming the validity of the tool can be a mistake)

Scenario:

- Assault planners know the locations of the mine and obstacle belts
- Planners place the aim points only within the boxes oriented on top of the belts

(The primary operational impact of the two assumptions, taken together, is that it leaves major portions of the assault lane unaddressed. Incorrect intelligence, inaccurate placement of mines and obstacles by the enemy, or their physical displacement could create a high-risk transit for the AAVs.)

Structural:

- P-OCM is two-dimensional. While water depth "changes," it's only with respect to determining differing displacements for differing notional depths (the vertical dimension is never represented, let alone used)
- Time steps are established only to separate detonations sufficiently to allow all displaced mines and obstacles to come to rest before the next detonation
- No modeling of the detonation plume
- Each run had three double-strand mine belts, one Hedgehog belt, and one TSC belt
- No slope or slope effects (but handled the equivalent through assigning different water depths of mines/obstacles)
- No treatment of bottom composition
- No cratering

Variables:

- Number of bombs per target box (6, 9, 12, or 15)
- Size of the target box
- Aim point Accuracy. Aim points are distributed on a uniform random basis throughout the target boxes. This parameter enables the hit points to vary from their aim points.
- Minefield Placement. The mine and obstacle belts are assumed to have a uniform random distribution of their respective mines/obstacles. This parameter allows variations of those uniform placements.
- Minefield Intelligence. This parameter permits addressing imperfect intelligence by allowing the target boxes to vary up to 12.5m off the centerline of the targeted belt.
- AAV Maneuver Accuracy. This parameter permits examining the effects of AAV navigational errors.
- Bombing patterns. These had to be set up by the analyst. Patterns could be a straight line of equidistant bombs, set up by defining a very narrow target box, to wide boxes with uniform random bomb distributions, to zigzag patterns set up by forcing a bomber to drop bombs in an alternating pattern to two parallel narrow target boxes.

Causal

- When a bomb hit occurs (a hit point is determined), any mines within the lethal area radius of the hit point are killed
- All other mines and obstacles move directly away from the hit point, with the distance moved determined by an internal look-up table developed from the Air Force data (the entering argument is the initial distance from the hit point)
- All mines and obstacles have time to come to rest from one detonation before the next one hits (the operational impact of this one is not at all clear, but we suspect it causes a somewhat narrower channel than would be the case if the detonations were closer together in time)

- AAVs must maneuver around any damaged or sunken AAVs and any Hedgehogs or TSC they encounter, then set a course straight to the LPP once clear (this assumption could cause AAVs to stray out of their planned transit lane; an alternative assumption could have the AAVs setting course for the next waypoint in order to regain the center of the assault lane)

Mathematic (the most important mathematic assumptions deal with assigning damage and/or displacement):

- Lethal Area Radius used to determine which mines are killed. Lethal Area Radius damage models generally tend to assign kills too generously. The conservative approach would be to use as the correct kill radius that one which is known to produce 100% kills. The conservative approach tends to underestimate kills, however, with the result (in this case) that resource estimates would become much higher.
- Use of empirical data in internal look-up tables to determine displacement. This should be a very low-risk approach, provided that the table is drawn from directly analogous experiments. If not, it is sure to yield invalid results.
- The hit points for precision weapons (JDAM series) were assumed to be directly on the aim points – no dispersion.
- The hit points for imprecise weapons were presumed to have a circular normal distribution about the aim point.
- Desired depth of burst was achieved with certainty in all cases

The validation exercise concluded that four items in the model design (including assumptions) stand out within the Conceptual Model as potentially creating problems:

- Choice of MOE. We see the selected MOE as unnecessarily introducing too many new and difficult variables into the basic problem. The most difficult are AAV navigational errors and AAV maneuver doctrine, either of which could have a huge impact on reported results and could suppress otherwise useful information related to the assigned analytic questions.
- Assumption of knowing the location of mines and obstacle belts. We see the assumption as highly suspect in the real world in the first place. It also injects a great deal of risk into the analysis in the second place. Finally, it enables the use of a counter obstacle tactic that is extremely risky.
- Use of a counter obstacle tactic that injects a high level of risk. Specifically, attempting to place the “target boxes” directly on the mine/obstacle belts, leaving gaps between the belts. We believe this tactic alone could be responsible for one set of counter-intuitive results (discussed later).
- Use of a Lethal Area Radius approach to determining damage to mines by bombs. Even when done carefully, such approaches will consistently under or over estimate damage.

Results Validation of the P-OCM Model led to the following conclusions:

- In many instances as bombing precision increases, AAV survivability decreases
- Fewer bombs can be more effective than more bombs.

Both of these model results were questionable, given that those phenomena are physically and logically counterintuitive. Further analysis of these simulation phenomena was not possible, because of the data omissions previously mentioned (no data describing displacement of obstacles/mines). The following are the stated overall conclusions for the P-OCM validation application:

- Although assumption testing failed to falsify the Conceptual Model, the results of the model appear almost certainly invalid
- Unfortunately, the data that could have helped determine just what happened and why was not retained (the actual damage and displacement data)
- Cannot recommend accreditation

Also there was a finding regarding the ABSVal framework itself:

> Separate assessments of the Theoretical/Mathematical/Algorithmic parts of the Conceptual Model appear to be needed only when a simulation is being developed, and after development, the model "it is what it is," and a single assessment that looks at the theoretic underpinnings, the mathematic sub-model, and the coded algorithms holistically should be fully as effective and far more efficient.

E.4.3 Open Discussion

Mr. Eberth's briefing generated useful discussion in the application of the VV&A framework. Several notional validation questions were postulated as follows:

(1) Does the model represent what it is advertised to represent?
(2) Are the limitations of the model explicit?
(3) Does the analysis (using the model as a tool) answer the analytical question(s)?

Dr. Bailey, the Study sponsor, expressed that there was a lot of information provided in the P-OCM Validation briefing to assist answering question #1 above, but the results provided were deficient with respect to answering Questions #2 and #3. Mr. Bitinas, who was also the simulation developer for P-OCM, stated that the approach was to fix the P-OCM model to the empirical data, and that the model was not suited for assessing scenarios where those assumptions stated in the validation briefing are critical. Several other questions were broached for consideration:

(1) Does the framework ask the questions that will make the limitations explicit? (Bailey: "In this case, possibly not").
(2) How do we communicate those model limitations to the end users of the model?
(3) How do the limitations of the model impact the analysis?
(4) Did the analytical question change as you were proceeding thru the analysis? (i.e., the Operational question of obstacle clearing vs. the academic question of "Is this worth pursuing as a simulation-based analysis?") Can we categorize the analysis questions as Predictive vs. Operational questions to facilitate validity in some way?

(5) At what point is the model so limited that it is too risky to use?
(6) Do we understand the analysis context?

A subsequent dialogue attempted to frame the simulation analysis in the layered configuration postulated in Dr. Bailey's briefing, resulting in this notional decomposition for the P-OCM model:

- Core: movement of mines and obstacles, location of bombs and forces
- Habitat: $location_0$ of mines, navigation errors
- CLA: assumptions listed in the validation briefing, e.g., water body floor, mine lane strategy, no overwatch, beach perpendicular, navigation not affected by current

Several suggestions were made regarding the framework.

E.4.3.1 Non-Boolean, CLA-Based Analysis and Validation

Several comments by the workshop participants concurred that validation of ABS should not be perceived as an attempt to draw a good/bad or valid/invalid Boolean type conclusion, but instead to communicate to the end user of the model an assessment of its capabilities and limitations. The assessment of limitations must inform the model user/decision maker of the impact, effect, and criticality of that limitation. The results of validation also may define subsequent, second tier validation questions that could further provide evidence of the model's usefulness that could not be answered due to technical or even resource constraints. These subsequent questions could be especially useful in attempting to determine if emergent behavior in the model is of analytical importance or a model anomaly.

E.4.3.2 Communication with the Decision Maker

There should be consideration given to the language and semantics used to describe this validation and analysis results, recognizing that the decision maker may not be either an ABS or validation SME. Definitions such as "assumption" need to be clarified and the verbiage for the postulated "Core," "Habitat," etc. elements may be reviewed toward that end.

E.4.3.3 Assessment of Risk

It was noted that an assessment of risk was listed as a framework aspect but not represented as part of the briefing in detail. It may useful to attempt to assess the risk of type II error, if possible. It was noted that, for the validation of P-OCM, if the analytical question is the displacement of bombs, then one may be able to define the risk of type II error, because there is ample data.

E.4.3.4 Empirical Data vs. Model Results/Model Complexity Issues

There was general concurrence that whenever possible, the model should be tested against empirical data. If P-OCM was designed only to model the pond experiment, validation could have included a testing of model data vs. pond experiment data for

(1) Goodness of fit
(2) Statistical test
(3) Comparisons/hypothesis tests

Subsequently, learning cases and test cases could be run on the simpler model.

Reducing the complexity in the model may facilitate a more thorough validation in this manner.

E.4.3.5 Analysis Plan, In-Phase Validation, Validation Resources

An analysis plan that explicitly considers validation would be a useful artifact; in fact, an ideal case may be a situation where validation occurs in phase with development, so that assumptions can be vetted and tested in the course of development and so that the necessary artifacts are generated to describe the CLAs present in the model. This plan also may describe what level of validation is actually achievable given the resources and information available. Also, it was broached whether the validators themselves need to be accredited in some manner as having suitable SME to perform the task at hand. Standardization of the reporting format and consistency of the information delivered were noted as important.

E.5 PYTHAGORAS COUNTER INSURGENCY (P-COIN)

This section summarizes the briefings given with respect to the validation of the Pythagoras Counter Insurgency (P-COIN).

E.5.1 Overview of P-Coin Scenario, Model, and Analysis (Bitinas)

The goal of the Irregular Warfare Project is to develop a prototype methodology for analyzing a USMC IW problem in-house. Several challenges exist to modeling Irregular Warfare. The inputs and MOE's for these models are different from traditional combat models. Examples of traditional combat model inputs include Weapon Probability of Kill, Armor Thickness, and Vehicle speed (more objectively measurable); examples of IW inputs include social factors such as Influence and Susceptibility (less objectively measured). Likewise MOE's for IW models are different from traditional combat models. Examples of traditional combat model MOEs include Lethality and Survivability; examples of IW MOEs include Population Response and Behavior. With this "soft" or less objective aspect of modeling human behavior, some expectation management for these models should be considered, since they will inherently involve a higher level of uncertainty (lower statistical correlation) than modeling traditional force-on-force combat.

E.5.1.1 System Conceptual Model

For the Pythagoras-COIN Conceptual Model, the civilian population in Colombia was decomposed conceptually into population segments, and within each segment, five Insurgency Behavior orientations were assigned, to define subgroups within the segments' relative affiliation to the Revolutionary Armed Forces of Colombia (insurgent)

or to the Government of Colombia. Influencing factors and events determine if the population shifts generally towards or away from insurgency.

The Pythagoras COIN scenario consisted of a MAGTF Mission to provide Refugee Camp Security and Humanitarian Assistance / Disaster Relief in the region, with two possible courses of action: Sea-Based Operations or Shore-Based Operations. The analysis sought to determine the plausible range of civilian population behaviors for these courses of action. The background for this scenario focused on two provinces on the Pacific coast of Colombia. The primary city in this area is Buenaventura, a seaport that is a predominant thru-way for drug traffic. Although it has a small upper class and a growing middle class, mainly urban poor and displaced persons who have been driven from their villages by the insurgency and crime that is gripping Colombia populate Buenaventura. The two key players in the insurgency are 1) the insurgents, the Revolutionary Armed Forces of Colombia (FARC), and 2) the counterinsurgents, the Government of Colombia (GoC). Other critical players are the militias, the drug traffickers, the Colombian Army, and the police. All have a presence in Buenaventura. It is a fomenting hot bed with a crime rate many times higher than New York City.

In the problem scenario, a tsunami has struck the area indicated in red on the map, destroying much of Buenaventura, and a Marine Air-Ground Task Force (MAGTF) has been sent to the area as part of a Joint, Combined, Interagency Task Force at the request of the GoC with the mission as shown here.

E.5.1.1.1 Input data

Input data for the model was acquired through SME Interviews, a process that was met with several challenges:

(1) Analyst & cultural SME communication challenge
(2) Analysts need numbers, e.g., probabilities, percentages, cultural SMEs are non-quantitative thinkers
(3) Note that the cultural data as acquired was narrowly focused on a specific region, and therefore, the data is not accurate for the rest of Colombia.
(4) To define population segments, data was elicited for each population segment and sought to determine:
(5) Prevalence of current behavior patterns
(6) Perceived needs are affected based on three factors (using Narrative Paradigm)
    - Natural tendency of the population segment (the population segment's narrative with respect to the insurgency)
    - Effect of current events on population segment (impact) – how the population segment reacts to a given COA
    - Effect of other population segments on a population segment (influence) – How the population segment reacts to the narratives offered by other population segment

E.5.1.1.2 Population Segments

The population segments were defined as follows:

(1) Illicit Organizations
(2) Catholic Church
(3) Police
(4) Military
(5) Displaced Persons
(6) Urban Poor
(7) Urban Middle Class
(8) Old Money

E.5.1.1.3 Orientation

Cultural behavioral data was sought on the population orientation (initial and natural tendency), the impact of MAGTF COAs, and the influence of population segment interactions.

The following question was posed for the Initial orientation data:

> "How do the actions of this population segment support the insurgency (FARC) or the Government of Colombia (GoC)?"

The following question was posed for the Natural Tendency orientation data:

> "Given no external influences, over time, how would the actions of this population segment change to support the FARC or the GoC?"

Quantitative data was derived from the SME interview responses through a Markov Chain matrix. This matrix is input to the Pythagoras agent-based simulation environment. Pythagoras does not carry out Markov matrix computations but uses the values to represent the 'dynamic sidedness and attributes' of the agents as it interacts with events and other agents.

E.5.1.1.4 Influences

This Data Elicitation required a process that translated SME words to a quantitative measure. Charles Osgood's Semantic Differential method was used to provide three major factors or dimensions of judgment:

- EVALUATIVE (good - bad)
- POTENCY (strong - weak)
- ACTIVITY (active - passive)

These three factors were combined into a single 'Salience' factor used as the influence factor between population segments.

E.5.1.1.4.1 MAGTF Influence

The data to define the impact of the Impact of the proposed COAs was elicited through SME interviews that posed the question:

> What words would this population segment use to describe MAGTF 'sea-based/shore-based' operations?"

'Positive words' averaged to measure leaning more towards GoC; 'Negative words' averaged to measure leaning more towards FARC. This data was input into another Markov matrix that displays new behavior patterns reflecting the impact of the MAGTF. This matrix was input to the Pythagoras agent-based simulation environment.

E.5.1.1.4.2 Salience – Group Influences

To obtain data to define the influence of other population segments was elicited through SME interviews that posed the question:

> "What words would this population segment use to describe another population segment?"

This data was input into another Markov matrix that displays new behavior patterns reflecting the influence of other populations. This matrix was input to the Pythagoras agent-based simulation environment.

E.5.1.2 COIN Scenario

Fundamentally, the implementation of the COIN scenario in Pythagoras is focused on a theoretical perception of COIN and insurgency in which defined population segments are attributed with an array of orientations towards insurgency (e.g., Insurgent, Pro-Insurgent, Indifferent, and Pro-COIN); as the simulation runs, these orientations may drift, which in effect cause a particular population segment to become more or less insurgent. Figure 21 depicts this concept.

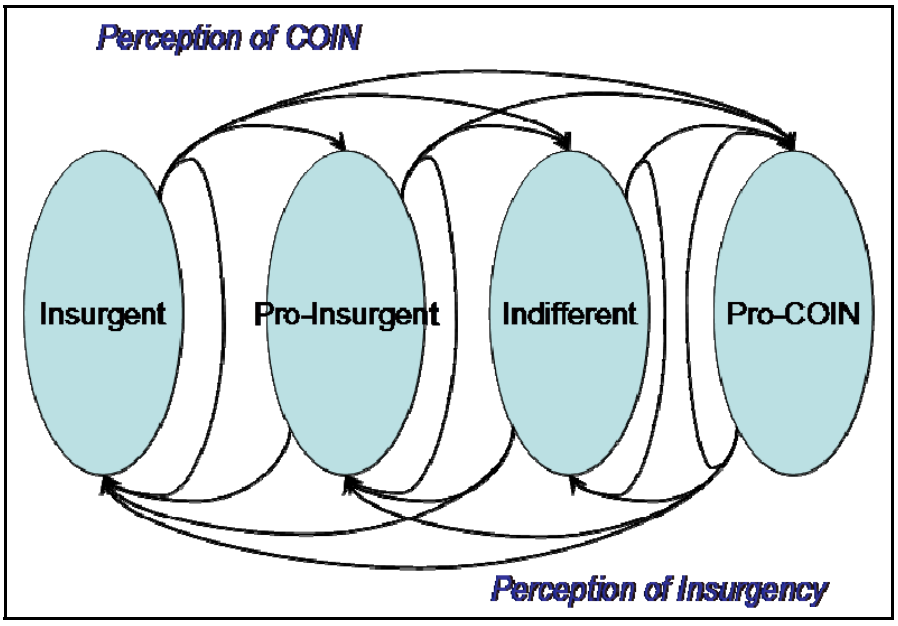


Figure 21 Theoretical Perception of COIN

E.5.1.3 Simulation Conceptual Model

The Conceptual Model defined for this simulation requires each agent within each population segment's orientation classes to shift per time step so that the exit arrows out of each population segment "bubble" shown in Figure 21 sum to 100%. Conceptually, agents remaining within an orientation bubble return to the bubble through a feedback arrow as shown in Figure 21. The initial distribution of orientations within each defined population segment is defined by researched demographic data, and the initial value of each arrow is defined by the conceptual "insurgency susceptibility" of the subpopulation within the bubble. This insurgency susceptibility can be further defined with three variables: the Interaction Estimation Transition Effect on the targeted population (the Direct Effect of the singular MAGTF arrival event), the Salience Transition Effect on population segments receiving information about events (the Indirect Effect of the interaction between population segments), and the Background Susceptibility Transition (the Ongoing Effect of population tendencies). The brief mentioned a precedence configuration for these effects; however, it was clarified that this was a legacy configuration from a previous version of Pythagoras, which is no longer relevant in Pythagoras 2.0.0, in which all of these effects can occur simultaneously on an agent within the model (the averaged effect is used to set the agents' orientation status per time-step). For the COIN scenario, the only event that affects the simulation is the arrival of MAGTF. Additionally, soft-rules random elements influence these population drifts (not every agent interacts with every other agent per time-step).

E.5.1.3.1 Population

The initial population in each of the eight defined population segments in the COIN scenario was scaled to be 100 agents, so that there was equal opportunity for all population segments to affect one another. Each agent within a particular population segment is 1% of that segment's population. Each agent has a set of 5 Attributes that define the insurgency orientation of that 1% of the population (1 being insurgent, 5 being COIN). In each time step, the sum of attributes normalizes to equal 1000. This normalization reduces the potential for simulation inaccuracies caused by round-off error. Attribute Changers represent the population tendencies, the influence between population segments, and the influence of the MAGTF actions. Communication devices represent interactions and possess the Attribute Changers that will do the influencing. Each agent can carry up to 10 Communication Devices; each communication device has 3 channels. Each of these channels contains an attribute changer that represents interaction and the ability to change another agent's orientation attribute through interaction, which allows the interaction between agents to be very specific.

E.5.1.3.2 Orientation Vulnerability

The Background Susceptibility Transition Effect (Vulnerability) of an agent to an orientation change, described as a Markov effect, has been implemented in Pythagoras COIN as an incremental Attribute Changer which increments the Attributes for each agent (1% of the population) in each time step per the Markov chain matrix values, normalized in Pythagoras.

E.5.1.3.3 Salience

The Salience Transition Effect is implemented as a relative Attribute Changer. An average Orientation for the two interacting population segments is calculated (using 1 = FARC through 5 = COIN); a Delta value is calculated based on the difference of the average of the two population segments; the Delta value determines the direction of the influence.

E.5.1.3.4 MAGTF Influence

The influence effect of the MAGTF arrival was modeled as an attribute changer for the simulation, which acted as multipliers upon the original agent attribute values.

E.5.2 P-COIN Conceptual Model Validation (Eberth)

Mr. Eberth briefed the audience on an overview of assumption testing that was performed on the Pythagoras COIN simulation. It was noted that the validation work that Mr. Eberth performed was independent of the work that the WernerAnderson validation team performed, and that the focus of Mr. Eberth's work was the Conceptual Model.

E.5.2.1 Methodology

The methodology for validating the Conceptual Model for the Pythagoras-COIN simulation was heavily reliant on assumption testing; assumptions made for the development of the simulation are examined with respect to the analysis objectives, with findings communicated to the decision maker. In an ideal sense, if one could identify every assumption that was made in building the model, then one would understand every way in which that model departs from reality. Here, the assumption testing techniques applied to P-COIN served to exercise the validation framework developed during this study.

The validation activities applied to the P-COIN simulation are based on the scientific method. The null hypothesis (research hypothesis) for validation is that the model/simulation is valid for the intended use, and the validation activities (including assumption testing) attempt to falsify the null hypothesis. An inability to falsify the simulation results in more confidence in the usefulness of the model towards the intended use; the degree of confidence then depends on the rigor and power of the tests applied.

The plan outlined to perform the validation of the Conceptual Model for Pythagoras COIN was as follows:

(1) Identify the analytic questions at hand, their metrics, and degree that results are expected to shape decisions
(2) Detailed review of all related documentation
(3) Interview Application Sponsor
(4) With the Application Sponsor, identify the referent; i.e., the proxy for the real world for accuracy comparisons

(5) With the Application Sponsor, determine the accreditation criteria: How "accurate" must the model be? How can/will accuracy be determined? (quantitatively/qualitatively)
(6) Criteria must establish lower bounds of acceptability for the model

Referent validity is a desirable end state. This is determined by the following:

(1) Confirming that no preferable referent could be made available
(2) Assumption testing the referent (for other than empirical datasets)
(3) Determining the operational implications of the assumptions
(4) Determining the bounds of validity imposed on the application's problem space and on the model's validity assessment by the referent's assumptions
(5) Determining whether the operational implications and bounds of validity are acceptable to the Application Sponsor

However, in the case of the P-COIN simulation, it was determined that there was no solid referent that could be validated, as is often the case with ABS. For the P-COIN model, it was often difficult to determine where to segregate different aspects of the framework, and if specific critical assumption (e.g., the Markov Chain assumption for the Background Susceptibility Transition Effect (Vulnerability)) were part of the referent or the Conceptual Model. For the Pythagoras COIN simulation, it was determined for the purposes of validation that the Theoretic Model and the referent were one in the same, but it was stated that a more detailed clarification of these elements is needed in the framework.

It was stated that the Conceptual Model generally includes the Theoretic Model, the Mathematical Model, and the Algorithmic Model. Each in turn receives same assessment techniques:

(1) Logical verification – determining sub-model is an adequate and correct implementation of its predecessor.
(2) Assumption testing
(3) Identify/derive the assumptions that are inherent to/embedded in the sub-model
(4) Determine the operational implications of the identified assumptions in the context of the particular application, Determine the bounds of validity of the model that are the result of the identified assumptions
(5) Determine whether the operational implications and bounds of validity are acceptable to the Application Sponsor for the intended application
(6) For some models, it may prove necessary to reverse-engineer one or more sub-models from later models. It may even be necessary to reverse-engineer the Conceptual Model, or portions of it, from source code (Not the case with Pythagoras COIN).
(7) Independent SME reviews

The operational implications of any assumptions used in the simulation must be examined to determine if the assumptions made have biased the results in a way that compromises its usefulness toward the analytical objectives.

Reverse engineering of the Algorithmic model may be used as a technique to derive assumptions that have not been clarified or stated in the Conceptual Model, but this may prove challenging. .

In terms of this validation effort as an exercise of the proposed VV&A framework, it was stated that the framework "is working," but improvements are needed on how to define the referent and assess the validity of the referent when empirical data are not available for use as the referent. It was also stated that linear, checklist-oriented templates would be useful for the framework.

For the P-COIN simulation, the Theoretic Model was the focus. The Mathematical Model, for this simulation, does not exist, as it went "straight-to-code" during development, and the Algorithmic Model had not yet been validated at the time of the workshop. It was acknowledged that the P-COIN Algorithmic Model may depart in some ways from the Conceptual Model.

The effort to validate the Conceptual Model began with a review of the Pythagoras User's Manual and related detailed discussions with Mr. Edd Bitinas, the simulation developer. However, Pythagoras itself was not assessed. Interviews were conducted with LT Robin Marling, USN, the COIN study's Project Officer, and several study-related documents were made available and reviewed. An interview with Dr. Akst, the Application Sponsor was conducted, and resulted in the following considerations for validation:

- Purpose was to "make headway in developing a COIN model."
- Did not specify an ABS, let alone Pythagoras
- Approved recommendation of using "sea versus land basing" as study's analytic question, but did not specify it at the outset of the analysis
- Approved stated Marine missions, and O.K. with implied mission
- Insisted study must use real-world dataset.

One of the findings of these interviews was that there were multiple, conflicting objectives for the Pythagoras COIN simulation:

(1) OAD was to "make headway" in developing a COIN model
(2) NGMS was tasked to determine whether and how Pythagoras could be used to support IW analyses
(3) Study at hand had the analytic objective of determining whether it was best to leave the MAGTF ashore or afloat in a Colombian Humanitarian Assistance/Disaster Relief/Security scenario

There are several approved USMC missions in Colombian scenario. (Refugee camp security, Humanitarian Assistance, Disaster Relief), but the collective study team (all stakeholders) found no way to directly evaluate the effectiveness of mission performance. Thus, it was decided to use allegiance changes of population segments among several distinct affiliation possibilities – thus producing an "implied mission" of keeping the insurgents from gaining strength (stated as "Do not allow illicit organizations to take advantage of situation"). While the MOE that emerged from this simulation

(changes in population orientations) may be a novel and worthwhile analysis consideration, it was noted as a validation concern that it does not align with the stated and approved set of missions.

Several assumptions in the P-COIN Conceptual Model had a large impact and reflected negatively on the usefulness of the model in the analytical context:

(1) Modeling the transitions among affiliations as a Markov process (a “memory less” process). This is a significant assumption and may be very limiting to the potential validity of the application.
(2) Constant transition probabilities across all time steps (except those during the Marines time in-country).
(3) Constant transition probabilities across all time steps while the Marines were in-country (although different probabilities from the baseline). This assumption must be relaxed to increase the potential of validity.
(4) Buenaventura is the only area considered, may not be representative of the operational scenario
(5) Insertion of MAGTF was the one and only stressor in the simulation and does not consider how MAGTF actually operates/interacts with the population
(6) Focus on Center of Gravity as the civilian population is questionable for the analysis
(7) Semantic differential data seemed inherently flawed because its origin was not connected to the area of interest

Initial indications are that the above assumptions absolutely pre-determined the results and in a predictable way (i.e., the model became deterministic if allowed to run to steady-state). Unfortunately, that may mean that OAD cannot make a solid determination on the usefulness of Pythagoras in the IW or COIN context from this particular application

It also may mean that the answer to the one analytic question (afloat or ashore) depends entirely on the methodology used to develop the transition probabilities – the “influence estimation” and “salience” parameters, and those are suspect because of potential bias in data collection/analysis methodology (semantic differential), because the SME’s used for data collection were not from the region of interest, and that the distinction between “data” and “context” could result in bias in the input data.

Initial indications with respect to the study (again, only from assessment of the theoretical model, so subject to change) were as follows:

(1) Probably cannot yet give a defensible answer to the afloat/ashore analytic question
(2) Implementation assumptions in the Pythagoras COIN Conceptual Model too limiting
(3) Semantic differential data collection/analysis methodology far too suspect

However, Mr. Eberth postulated that the study may represent a huge leap forward in IW analysis in the following ways:

(1) Could/should cause a re-evaluation of COGs and MOEs for IW environments
(2) Could/should lead to a series of studies on semantic differential and alternative methodologies for capturing the propensity of persons to change affiliations, particularly in response to actions/events rather than just presence

E.5.2.2 P-COIN Validation (Moya)

Ms. Moya briefed the workshop audience on the application of the validation framework with respect to the Pythagoras COIN scenario. The analysis context was presented with the information concerning this scenario. This analysis was focused on the four main aspects of the model, which consisted of the core, different cases, the dynamic influences, and background of the data.

E.5.2.2.1 Background

The P-COIN scenario was created and analyzed on the basis of eight different population segments that could have up to five different political orientations. The population segment that an individual belonged to remained constant, but the political orientation could change based upon the salience and natural drift factors.

E.5.2.2.2 Description of Analysis

First, Pythagoras had to be able to model population dynamics. Then, in order to evaluate the model in the best possible way, the question being asked needed to be changed so that Pythagoras could formulate a credible answer. A new question was created so that only a better answer was chosen rather than attempting to determine the best course of action. Once the question was established, the measure of the effectiveness and the method of measuring were analyzed. The two measurements utilized in the analysis were that there is no increase in insurgency activity and that the MAGTF support improves the backing of the pro-government political orientation. This was shown in a box-and-whisker plot that depicted the results with no MAGTF, MAGTF ashore, and MAGTF afloat. See Figure 22 and Figure 23.

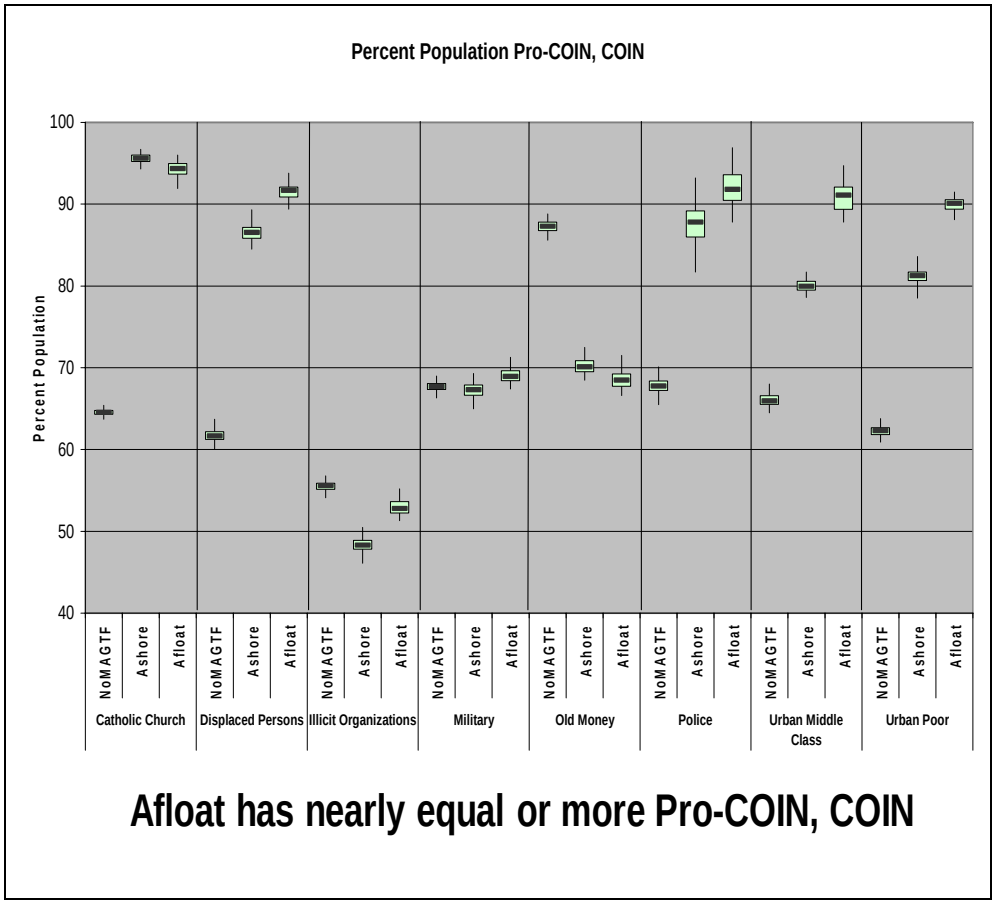


Figure 22 Measure of Pro-COIN and COIN

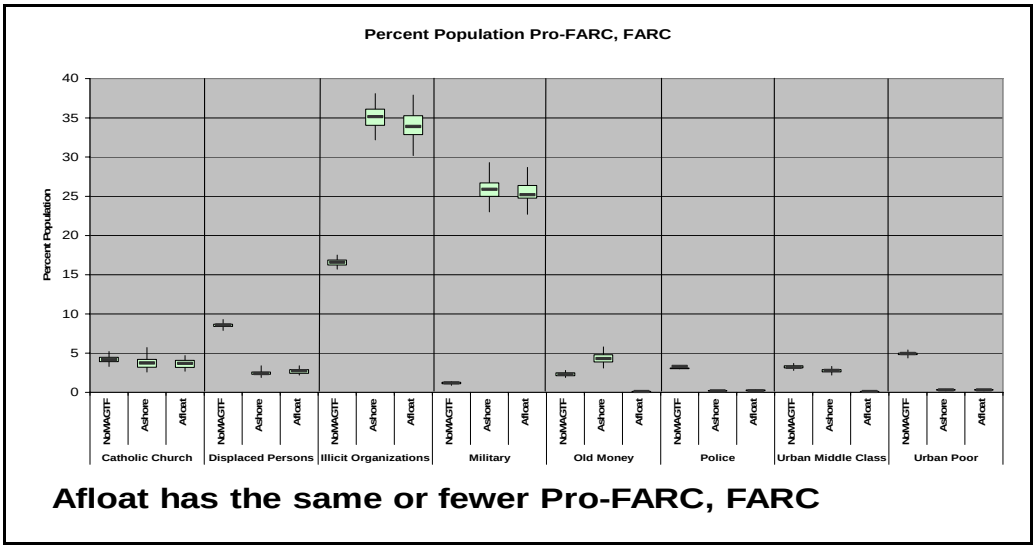


Figure 23 Measure of Pro- FARC and FARC

E.5.2.2.3 Areas of Interest

The four areas of interest that were discussed were the core, different cases, the dynamic influences, and the background of the data. The core consisted of the MAGTF influence on the insurgency orientation. The different cases were then MAGTF versus no MAGTF and ashore MAGTF versus an afloat MAGTF. The dynamic influences were those that affected the orientation of the population segments regarding their political orientation. This consisted of the natural drift of a population along with the salience of the individuals. This, however, results in only a first order assessment and does not take into account the second and third order interactions that affect the changing distribution of the population.

E.5.2.2.4 Results of Analysis

The analysis resulted in determining the problems that are presented with the Pythagoras COIN model along with recommendations for improving the reliability of the model's results. There were multiple problems regarding the use of the data. First, the data is considered to be perishable in that if a major event were to occur, the population dynamics would no longer be consistent with the given data. Along with the fact that the data is considered perishable, it is also inexact. This is due to the large tolerances were used in the trials and the data was collected over multiple runs since the results were constantly changing. Secondly, the salience and natural drift of an individual were only utilized with respect to the individual's initial state, and thus did not change with changing orientation throughout the experiment. Lastly, the data does not attain the second and third order interaction effects on the population segments.

E.5.2.2.5 Recommendations

Recommendations were made regarding the model. It was suggested that influencers be applied more forcefully to the scenario so that more accurate information could be determined regarding the change in population diversity. The model was determined to be invalid in establishing the changing population dynamic long after the MAGTF arrival. Along with this, the model had no statistical comparisons and no real description of how to determine the optimum course of action.

E.5.3 Open Discussion

The briefs given by Mr. Bitinas and Mr. Eberth led to both a discussion on the necessary components of a validation and developing a better understanding of the process used to obtain a convincing validation. Using Eberth's briefing as a reference, the question was posed as to what needs to be provided to determine if the end user can make use of the model with confidence in the result. Along with this, new terminology is needed to translate the terms for habitat, core, and orientation for the final decision maker so that the model can more easily be utilized.

When analyzing a model, many components are taken into consideration. The impacts of these components affect the quality of the results, and as such need to be reviewed carefully. Looking at the CLA, it needs to be determined what the influence is on the model. Questions such as "what is the quality of the data?", "what is the criticality with

respect to intended use?" and "is there a negative impact?" need to be asked. These questions are crucial to determining the impact of the CLA and its affect on the use of the model. Along with the CLA, the orientation needs to be examined to decide if a stronger foundation can be formulated as well as checking the tolerances of the attributes so that they still form a valid answer.

When giving a recommendation for use of a model, there needs to be certain aspects that are present as to convince the decision maker that it can be utilized for the necessary situation. This can be done by explaining why certain negative impacts do not stop the model from producing a valid result as well as showing why the other positive impacts do make it a valid result. To accomplish this, a complete understanding of all constraints, limitations, and assumptions is needed. However, questions including "what are the deficiencies in the model CLA at present?", "what is important to determining the structure of this model?", and "are the same assumptions carried out throughout the model?" must be asked. After these questions have been answered it may also be necessary to go back to the person requesting the work to clarify that all assumptions are accurate and the question being asked is the correct one. If the correct question is not being evaluated for the model it would lead to inconclusive results for use of the model.

When validating a model it is best to conduct a validation of the Conceptual Model (algorithmic) as well as a results validation when possible. In all cases, a validation of the Conceptual Model is possible, but a results validation may not be. A results validation is accomplished by comparing model results to a real world problem. In most cases this can not be done exactly, but similar scenarios will suffice to validate the model's results. If this is possible, it provides a much stronger and more convincing validation.

Ultimately the critical information involved in the scenario that is constructed in the model needs to be determined along with the operational implication of the assumptions that are acceptable to the decision-maker for the validation of the model.

E.6 VALIDATION AUDITING: OBJECTIVES AND METHODS (KENYON)

Mr. Kenyon briefed the audience on the methods and objectives of auditing the V&V activity from and accreditation authority perspective. The following is a summary of some of the key points provided in Mr. Kenyon's brief:

The auditor acts in a role as an accreditation authority who must determine whether the model should be relied upon to support the analysis of a specified class of problems. The auditor should be an experienced consumer of modeling in support of past studies and can engage in discussions of theory, but remain in a practical orientation.

A general approach for auditing is as follows:

(1) Focus on the V&V report as a reflection of the process.

(2) Formulate, more or less independently of the V&V process, a series of questions we would want answered about the model in support of an accreditation decision.
(3) For each of these questions, ask how clearly and convincingly the V&V report answers it, and how readily the answer is extracted from the report.
(4) Use the V&V report to answer the second set of questions.
(5) Comment on both the report and the process based on this exercise.

The accreditation authority (SPA) for the VV&A effort developed a set of review questions to perform the audit. The review questions were presented to the audience. Note that audit questions were added and modified resulting from feedback given during the prior Workshop session regarding model improvement recommendations.

E.6.1 Artifacts Evaluated

The following artifacts were evaluated during the audit activity:

P-COIN: Two reports covering separate phases of V&V, as an artifact of the validation team structure.

(1) “Assumption Testing Report” (Verification of Conceptual Model)
(2) “Validation Report” (Validation of Instantiated Model)

These reports treated as a single V&V decision package.

P-OCM: Validity Assessment Report

E.6.2 P-COIN Audit Questions

The following are a list of those questions and the response from the accreditation authority as they were presented in Mr. Kenyon’s brief.

Audit Question #1: Do the reports clearly identify the application (set of study questions) for which the model is being validated, and the model’s role in addressing those questions?

> Yes. Report states that the application is course-of-action analysis (afloat vs. ashore) and that we don’t necessarily need accurate point estimates of outcomes, but a reliable ranking of those point estimates under the alternatives.

Audit Question #2: Does the validation report clearly describe the tests that were performed on the model, the possible outcomes for each test, and the criteria for passing?

> It was reasonably easy to understand from the reports what tests (or questions) P-COIN was subjected to, once we found them. They were more clearly enumerated in the Assumption Testing report. In the Validation Report, we were able to infer the tests (which related to methodology) from observations presented in a section titled “Data Considerations.”

Audit Question #3: For each test performed, is the result clearly presented in a way that relates directly to the specified pass/fail criteria?

This question would be more applicable to results validation tests, of which none were available for our review. The tests performed were generally qualitative in nature.

When problems were found, the reports sometimes stopped short of assessing their implications for acceptance.

Audit Question #4: Do the reports provide a recommended decision for the accreditation authority?

Yes, two different ones! Assumption Testing: "…we have to advise against reporting out any results from the model as actionable in any sense" (sect. 6.5.4) Unambiguous, not repeated in recommendations summary.

Validation: "Subject to the caveats, limitations, and cautions listed above, P-COIN can answer (the analysis) question…" (sect. 8.3)

Also: "…there is little risk in using the results…since the analysis does not advocate a change in current…procedure…" (sect. 8.3)

This "after the fact" risk assessment leaves open the question of whether the model is more useful than a magic 8-ball, and the remainder of the quoted sentence seems to acknowledge this.

Can the model increase our confidence that the default COA is right?

We used Boolean logic along with our own assessment of the information provided to infer that yes and no means no, don't use it in its current state to support the afloat vs. ashore decision.

Audit Question #5: Do the reports make a convincing argument that the tests conducted collectively provide a sufficient basis for the recommended accreditation decision?

Yes. The tests and results presented were sufficient to cause us to doubt the model's ability to reliably support the afloat-vs.-ashore decision at this moment, so the practical mandate (regarding the Accreditation decision) is fulfilled. Supporting a positive accreditation recommendation, had that been "the right answer," would probably have required much more work on the validation side. Where problems were found in assumption testing, these were assessed not to invalidate the Conceptual Model. Although we believe in the value of assumption testing, we have to wonder how squarely ABSVal is actually founded on challenging a null hypothesis of validity, given the appearance that effort was expended on zero-power tests. Question: Is it at least possible in theory that an assumption could be shown to be bad enough to invalidate the Conceptual Model?

Audit Question #6: Are recommendations provided that are actionable by a model improvement program?

Yes, there is one in the Validation Report but the report asserted elsewhere* that options for pursuing it within the Pythagoras framework may be limited. Recommendations for the ABSVal framework were also provided (Assumption Testing Report).

E.6.3 P-OCM Audit Questions

Audit Question #1: Do the reports clearly identify the application (set of study questions) for which the model is being validated, and the model's role in addressing those questions?

Yes. Impact of:

- Accuracy (intelligence and targeting)
- Timing/sequencing of detonations
- Bomb type
- Number of bombs used

On AAV survivability against MIW threat in the surf zone

Audit Question #2: Does the validation report clearly describe the tests that were performed on the model, the possible outcomes for each test, and the criteria for passing?

Yes. Again, the tests generally did not have quantitative thresholds or other features that called for any elaborate description of the test and its purpose prior to giving the result.

Audit Question #3: For each test performed, is the result clearly presented in a way that relates directly to the specified pass/fail criteria?

Yes, except that the acceptance criteria were implicit. Counterintuitive results validation findings without the ability to drill down on them.

Audit Question #4: Do the reports provide a recommended decision for the accreditation authority?

Yes. "P-OCM has to be considered invalid for analytic applications due to the near-impossibility of some of its results." (Section 6.6.1)

Given that it was not possible to drill down on the causes of those results. Doing so might have produced recommendations for model improvement, recommendation to increase numbers of replications, an attribution of the counterintuitive result to something other than a problem with the model, and/or a recommendation for changing the set-up to avoid the observed problem

Audit Question #5: Do the reports make a convincing argument that the tests conducted collectively provide a sufficient basis for the recommended accreditation decision?

> Yes. The tests and results presented were sufficient to cause us to doubt the model's ability to reliably answer the study questions.
>
> Again, supporting a positive accreditation recommendation, had that been "the right answer," would probably have required much more work on the validation side. Where problems were found in assumption testing, these were again assessed not to invalidate the Conceptual Model.

Audit Question #6: Are recommendations provided that are actionable by a model improvement program?

> No, but the model was unavailable for diagnosis of counter-intuitive results, the activity that would produce most such recommendations as its by-product.

E.6.4 ABS VV&A Framework Recommendations

Recommendations for the ABSVAL framework were provided. The following stoplight chart (Figure 24) summarized the findings of the audit activity:

Validation Report	P-COIN	P-OCM
Application clearly identified		
Tests clearly explained	*Some a little under-the-radar*	
Test results clearly presented	*Validity implications of adverse results*	*Validity implications of adverse results*
Recommendation on Accreditation	*Some post-processing required*	
Support for Accreditation Recommendation	*Validity implications of adverse results*	
Model Improvement Recommendations		*N/A*

Figure 24 Audit Summary

A slide was presented that illuminated a distinction between the validation approaches for the P-COIN application vs. the P-OCM application. While both approaches employed assumption testing as the method for examining the Conceptual Model, the P-COIN validation approach performed a subsequent validation of the instantiated model, while in the P-OCM application attempted a referent validation against the empirical data collected, showing that often details of the validation approach will be

determined by the model design and data availability. The following insights and recommendations were generated out of the VV&A audit activity:

(1) The Assumption Testing Report included material describing the foundations of ABSVAL and some components of the process. However, it was difficult to infer from the reports much about the overall ABSVAL "recipe" (or whether there is one).
(2) Probably in part because material supporting an adverse Accreditation recommendation was found before much ground needed to be covered.
(3) Demonstrations suggested that subjective elements (e.g., assumption testing) are difficult to eliminate from V&V altogether without material degradation to the process outcome.
(4) In Assumption Testing, there is no way to guarantee a complete enumeration of the key assumptions, but decomposition of the Conceptual Model and assumption types is a good way to ensure a reasonably diligent search. (Fundamentally, it's brainstorming)
(5) Reports reflect significant effort to develop and demonstrate broadly applicable V&V techniques and principles, but test cases did not appear to showcase or exercise any features of ABSVAL specifically responsive to the particular challenges of ABS, especially emergent behavior.
(6) Limited selection of real world ABS applications available on which to test the process.
(7) Casting the process to emphasize model improvement and building credibility (vice trying to invalidate) might help draw more test cases out of the woodwork
(8) Process as defined is inherently adversarial, and sponsor reluctance to expose completed work to scrutiny can limit opportunities to apply it.
(9) How would results validation be done for an ABS that had actual emergent behavior? Test cases did not force this question
(10) We saw a collection of validation techniques described and applied, but only vague indications of any rules governing their application, or guidelines for identifying what steps are necessary or sufficient in a given situation.
(11) It would be helpful to clarify the minimum body of information/ material to which the V&V team must have access in order to reach a favorable accreditation recommendation.
(12) When accreditation support (vice model improvement) is the goal, why start if no chance for up-check?
(13) Need to adjudicate "potential problems," caveats, etc. as, e.g., "show stopper" "limits scope of valid uses" or "not really a problem after all."
(14) When the instantiated model does not conform to the Conceptual Model, test the assumptions underlying the "offending" portion of the IM for impact.
(15) Third party, after-the-fact validation is fundamentally a game of catch-up. We believe that much of the Validation effort (the part not specific to a given use of the model) could be accomplished in conjunction with model development and would be largely reusable in support of multiple Accreditation decisions.
(16) Suggest a future test of ABSVAL be structured with a reusable component and an application-specific component, applied in sequence.

(17) Reusable component could be developed as a concurrent task within model development, if an appropriate test case were available.

In summary, the following points were presented:

a. The accreditation recommendation was tentatively adverse for both applications, and neither test case exercised all parts of the ABSVAL process.
b. Sufficient information to support a favorable accreditation recommendation for either application may not have been available to the V&V team
c. ABSVAL response to challenges introduced by ABS, emergent behavior in particular, not really demonstrated by the test cases

E.7 FINAL DISCUSSIONS

The workshop briefs led to a final discussion on what needs to be examined when validating a model and what the corresponding report entails. There is a basic framework that is used when completing the validation process which includes an in-depth look at the model being used. The framework can be considered a series of questions to ask, different things to look at, or even a set of descriptions of the critical analysis requirements.

E.7.1 Model

When validating a model it was determined that there are various items that need to be scrutinized. One question that needs to be answered in this process is about the emergent behavior a model exhibits. There has to be a legitimate reason as to why the specific emergent behavior is appearing and this can typically be found by doing a backwards analysis on the data. It must be determined that the behavior makes sense with the data and as a result even though it may not be an expected behavior it is still a reasonable one. However, in order to be able to make use of the data when validating a model, the quality of the data needs to be taken into account. This can be done by utilizing the Yost scale, where most of the high quality data goes into the core and the lower quality data is in the separate cases. This is since the core is the most important piece of the model and as such requires the most accuracy. As a result, certain precautions must be taken into account when choosing the model for a question. If the problem is found in the core, it means that an incorrect model was chosen for the given problem. If the problem is a result of the habitat, the conclusions that were drawn were not strong enough to help formulate a valid solution. If the problem is an effect of the cases, the wrong model is being used and there was poor experimental design factored in to the issue.

When preparing to provide a result on the validation, three general steps must occur. The first being that all appropriate documentation must be obtained. This consists of everything from the system design document to the source code and the test results. This should then be used to determine if the model is correct. The next step would be to define the question; figure out what is important and how the question could be

answered. The last step is to establish what the required inputs and outputs are for the given question.

E.7.2 Report

The report that is generated from a model validation is a key component to being able to utilize the model. It must be in straightforward enough terminology such that an O-3 or middle manager can understand the document and decide how to communicate to higher ups. Additionally all information that may pose question needs to be in some sort of appendix to the report. This includes having access to the code and data from the model. The report must answer questions such as “what kind of quality is presented with the results?” and “how precise is the information?” Along with the important questions that must be answered, all assumptions need to be explained and justified. This comprises of structural, casual, mathematical and scenario assumptions. All of these will factor into the validity of the model, and as such need to be reported.

APPENDIX F VV&A PHASE II PUBLICATION/MEDIA PLAN

F.1 PURPOSE

This document provides suggested research areas as requested by Dr. Mike Bailey, MCCDC/OAD, and identifies potential venues for presenting this material.

F.2 RESEARCH TOPICS

The Agent Based Simulation Verification, Validation, & Accreditation Framework Study has produced a wide body of research areas, some of which are of interest to individual team members, others of which are collaborative team efforts. This document describes some of these areas in two sub-sections. The first sub-section discusses the area Ms. Moya would like to utilize for her dissertation research as requested by Dr. Mike Bailey. The second sub-section provides other potential research areas that may be of interest to team members. In the many team working-sessions, individual team members have expressed interest for some of these areas. This interest has been identified where known.

F.2.1 Ms. Moya's Dissertation Research Area

Ms. Lisa Jean Moya sees that establishing the importance of validating the rule set of agent based simulations and the linkage shown between the ideal simulation and the Conceptual Model as her core research area for her dissertation research proposal. The research problem to be posed in Ms. Moya's proposal exists in three parts:

F.2.1.1 Necessary Condition for Simulation Validity

The ideal simulation as described in our framework is not available to simulationists for validation. In practice, simulationists have only a proxy for this (referent). Consider the rule set for an agent based simulation in a more general way as a set of well-formed formulas. Then using the concept of model from model theory, a necessary condition for $T(C) \sqsubseteq T_I$ (i.e., the transition system of the ABS model simulates the ideal simulation) is that the proxy for the ideal simulation is a model (in the context of model theory) for the well-formed formulas. Ms. Moya sees establishing this as critical for establishing the importance of validating ABS rule sets.

F.2.1.2 Developing Validation Experiments for Conceptual Model Validation

The team has established the scientific method as its basic approach to validation. That is, we agree that we cannot prove that a model is valid; rather, all we can do is through the failure to demonstrate a model is invalid is to build evidence for validity. This is in large part to the intractability conjecture posited by Dr. Weisel. Mr. Bitinas and Mr. Eberth have collected several validation techniques from their extensive experience and brought them to the team. None of these is specifically oriented to validation experiments of the well-formed formulas with respect to the proxy for the ideal simulation (although some may be appropriate; e.g., assumption testing). Ms. Moya discusses in (Moya et al. 2007) one technique that could be appropriate in general, if

developed further. Ms. Moya sees as core to her research problem the development of a set of techniques or validation experiments that could be used to support validation of the well-formed formulas in the Conceptual Model.

F.2.1.3 Demonstration of Validation Experiments for Conceptual Model Validation

It is insufficient to posit hypothetical experiments for Conceptual Model validation or to list the requirements for these experiments. It is also necessary to demonstrate the viability of these experiments to demonstrate the claim through a proof of principle. The exercising of the Validation Framework against the Conceptual Model validation of the COIN model is a critical element of Ms. Moya's research topic.

F.2.2 Other Topics

There are many other publishable research areas in this work, some of which are discussed below. This list is not comprehensive.

F.2.2.1 The Foundations of Simulation Science

Dr. Eric Weisel has been working on developing a theory of simulation science since 2001. His Ph.D. dissertation contains work on this topic. The "cloud diagram" constructed as part of this Framework Study illustrates many of the foundational concepts in this theory as well as how all the theoretical elements are connected. As part of this foundational theory, several conjectures have been posited. Dr. Weisel would like to continue to expand and publish in this research area by, for instance, developing and publishing proofs of these conjectures.

F.2.2.2 Scientific Method

The team came to a consensus that the scientific method is the only viable method for validation with the agreement that we are really seeking to invalidate a simulation with each experiment. This is in part due to the intractability conjecture. Mr. Eberth posed this thought directly in November 2002 presentation at 2002 Annual USMC M&S Conference held by MCMSMO.

F.2.2.3 Risk Assessment

The team agreed that risk assessment is a critical element of validation. This is a function of the error of using an invalid model and the impact of its use. The team has suggested methods for calculating (e.g., utility theory) and display (e.g., a surface and stoplight assessments) of risk. At least one workshop participant commented that the estimation of engineering risk that might be appropriate for consideration.

F.2.2.4 Simulation Matching

There are two aspects to establishing that a simulation suitable for one use is suitable for another use. The first is theoretical: if $T(M_1) \sqsubseteq T(M_2)$ and $T(M_2) \sqsubseteq T(M_3)$ does this mean that $T(M_1) \sqsubseteq T(M_3)$. The second is practical: how are the well-formed formulas established and described to allow a match assessment.

F.2.2.5 Results Validation

Results validation is a necessary condition for ABS validity. If simulation results and expected results from the proxy (e.g., empirical data) do not match, then the simulation cannot be valid. (We conjecture though that it can be shown that matching even a large set of results does not guarantee this is true for all possible results in the proxy; further, we conjecture that if only a single trajectory in the proxy needs to be matched an exhaustive search may be required to find a simulation trajectory to match it.) Investigating approaches for results validation in the context of "soft referents" is a large part of Phase II of the Validation Framework Study.

F.2.2.6 Application of the framework

Several papers are possible resulting from the application of the framework to various ABS. The results of the application may form the basis for the discussion of any of the above topics or may result in its own paper published by the team.

F.3 POSSIBLE VENUES FOR PRESENTATION AND PUBLICATION

Below are some possible conference venues.

- BRIMS 2008
  - 14-17 April 2008, papers due approximately February 2008, (Providence, RI)
- SpringSim 2008
  - 14-17 April 2008, papers due November/December 2008 (Ottawa, CANADA)
  - Agent Directed Simulation, paper submitted October 31, 2008
- Spring SIW 2008
  - 14-17 April 2008, abstracts 3 December 2007 (Providence, RI)
- 76th MORSS
  - 10-12 June 2008 (Coast Guard Academy, New London, CT)
- I/ITSEC 2008
  - end of November 2008, abstracts due approximately February 2008 (Orlando, FL)
- Winter Simulation Conference 2008
  - beginning of December 2008, papers due approximately April 2008

F.4 CONCLUSIONS

This document produces a list of suggested research areas arising from the Agent Based Simulation Verification, Validation, and Accreditation Framework Study with team member interest areas identified when possible. This list is not a comprehensive list. Specific paper topics, authors, and venues will be identified when known as the project progresses.

APPENDIX G SOME COMMENTS ON MODELS (VISCO)

A Polemic

(Well, perhaps not so fierce as all that)

E. P. Visco

Visco Consulting

June 2008

Introduction

Did Shakespeare say, “First, kill all the lawyers”? If so, he was wrong. First, we must kill all the bumper sticker philosophies. We should start with “All models are wrong” and its adjunct “Some models are useful.” Like most universals, the phrase is wrong and misleading; note that I said “most universals.” That’s the chicken in me, since I can’t possibly review all bumper sticker-like phrases, I cop out by saying most universals. It may be that almost all bumper sticker-like phrases are wrong and misleading. But, let’s stay with most, or be even more chicken-hearted and say many such phrases are wrong. Some models are right; maybe many models are right. I doubt that anyone, including the originator of the phrase about all models, can truthfully say that all models have been reviewed and assessed for their degree of rightfulness. The devil is still in the details. The ground issues are the definitions of “right” and “wrong” and, even, “usefulness.” One argument supporting the bumper sticker statement is that models are not the reality, that is, they are representations of reality (or, at least, they are attempts at representations of reality). By definition, then, models are incomplete, since they are not the reality itself. If they are incomplete, therefore they are wrong. A photograph of a house is a model. While it is not the reality of the house, it serves the purpose for which it was designed. The photo can be used to market the house, showing many of its desirable features, or it can be used to help someone identify the house for a visit. The issue is that the model serves a designed purpose quite well. Hence, it is not “wrong.”

And thus we are back to validity or its inverse, the Popper-ism of “falsification” or attempting to prove that something alleged is wrong. Failure to falsify a proposition (or model), after due diligence in the attempt, is tantamount to accepting the proposition (or model), with reservations. Taleb (have you read *The Black Swan. The Impact of the Highly Improbable* (recommended by Peter Perla)?) suggests that Popper’s more important contribution is the emphasis “…on skepticism as a *modus operandi*, refusing and resisting definitive truths.” [Taleb, p. 56] Right On, Popper!

My conclusion: some models are right (depending on one’s definition of right, related to the applications of the models); some models are wrong (similar definitional concerns); some models are useful (depending on the intended applications); and some models are not useful (ever).

Which Came First?

A feature article in *Phalanx* (Vol. 40, No. 4, December 2007) by George Akst focuses on data, in the on-going "chicken and egg" debate of models and data. Before going on, I must state one of my biases. I have said that the United States Marine Corps is presently the most erudite of the Services. I say that after serving the US Army as a civilian operations analyst and researcher for over 50 years and having served in the US Navy during World War II. George Akst demonstrates the truth of that statement of the quality of thinking in the Corps. George is direct, clear-thinking, and a nice guy as well. Another *caveat*: there are fine thinkers and excellent minds in all the Services. It is just that the Marine Corps, being the smallest of the Services, seems to have a disproportionate percentage of smart folks.

In the case of his recent note in *Phalanx*, I think he is not wrong but also not right. Many years ago, the distinguished British comic, actor, and mimic Peter Ustinov recorded a routine about the Grand Prix of Gibraltar, if one can imagine a Formula One racing event up and down the Rock, avoiding the apes in the process. Ustinov played all the voices on the disc which consisted of a reporter from a racing journal interviewing drivers from the many countries in the race: the Frenchman (drawing on a *Galois*), the German, the Japanese, the American…When interviewing the American, the reporter asked: "What do you see as the most important part of the race car you're driving?" The American drawled his answer: "Waal, I think the engine is pretty important—and I think the steering wheel is important—and I think the four wheels are important—and I think the axles are important—and I think…" Well, you get the point. The issue is not whether data are more important than the model or the model is more important than the data or the data have to come first or the model has to come first. One is not much good without the other. On occasion, the model has come first (think of observations about the movements of the planets, perhaps—although one might argue that the observations were data-oriented). On other occasions, the data have come first (think of the work of the first US Army Air Force opsannies with the Eighth Bomber Command in England in the fall of 1942, with the mission of helping to double the number of bombs on target then being achieved). The issue here is the application or the problem under study. If one is attempting to get some handle on the future, not prediction but rather comparative analysis as George Akst has discussed in his unpublished note on musings about validation, then the model and the data must proceed side-by-side. When one is looking at a specific system behavior with an eye towards determining weaknesses or opportunities for improvement (the bombs on target problem of the Eighth), then the data come first. A model results from the data reduction process, designed to represent the system behavior, in simplified form, to allow for system tweaking to lead to improved (more efficient or less costly) performance. In his musings about validation, Dr. Akst also made the point that observations of behaviors and the resulting statistical descriptions are useful for cases falling within the boundaries of the observations—and perhaps very minor extrapolations.

What Is It About Models?

To elaborate, there are two types of models. One is developed from observations and data collection around some phenomenon of interest. To be more precise, a structure representing, at least crudely, the phenomenon in order to guide the data collection. Perhaps that initial view of the phenomenon might be what some refer to as a Conceptual Model, although that phrase seems, on occasion, to be used for highly detailed structures with important assumptions. My view of a Conceptual Model is a first cut, perhaps "back of the envelope" version of the phenomenon. Its use, as noted, is to guide data collection and processing. Statistical analyses of the data then leads to a more detailed and applicable model. That result is useful for helping to understand the behavior of the phenomenon, in a historical sense. By no means should such a model be used for prediction of behavior of the phenomenon beyond the range of the observations (the data). (Oh, maybe a little sneaky extrapolation can be allowed, with great care and only with clarity of exposition, when providing results and recommendations to those who have to make decisions and carry the risks attending the decisions.)

The second type of model is at the crux of much military operations research these days. It is an almost purely idealized (alleged) representation of a complex phenomenon (close ground combat, for example). Much of the design work on models of this class is carried out by analysts using historical examples and the advice of subject matter experts, who are often not well vetted. Models of this type often rely heavily on poorly stated or unsubstantiated assumptions. This type of model is occasionally, perhaps too often, used for prediction of future behavior of the complex phenomenon with limited caveats and without calling clear attention to the impact of the assumptions on the behavior of the model.[2] In this case, data collection generally becomes intense as the model is constructed and the variables seen as important, by the model designers, are identified. To a degree, the logic is circuitous: the assumptions critical to the model are made by the designers, who then identify the important variables (which often develop from the assumptions), which then determine data collection and application.

A Bit of History—More Than Enough to Bore You

A historical diversion to provide one argument as to how we got ourselves into this fix. There was a rush, in the US, immediately following World War II, to adopt the newly named field of operations research, which made so many important contributions to the war effort on the Allied side, particularly. Among the many groups formed to provide analytic services to the US defense community were Project RAND (originally supporting the newly created Air Force; later morphed into the RAND Corporation supporting different elements of the national defense structure); The Johns Hopkins University Operations Research Office (supporting the Army); the Operations Evaluation

[2] I am indebted to comments from A.E.R. Woodcock during Cornwallis XIII, for identifying the distinctions between the two types of models.

Group (derived from wartime groups supporting the Navy; later converted into the Center for Naval Analyses); the Institute for Defense Analyses (a somewhat later organization, derived from the Weapons System Evaluation Group and designed to support the Secretary of Defense and the emerging Joint Staff); and many smaller groups organized within the Services, along with a number of commercial firms acting on defense contracts. Many groups began serious research on the complex phenomena known as military operations to single out one; the ORO began work on developing a sound understanding of tactical ground combat operations, very complex phenomena. At one point in the process, a seminal paper was written by a young analyst, the late Richard E. Zimmerman. The paper, which took the 1956 Lanchester Prize (from the newly formed Operations Research Society of America) presented for the best English language OR paper of the year, is titled "A Monte Carlo Model for Military Analysis."[3] The paper argues that the requirements for use of a digital computer derive from the dimensionality (i.e., the complexity) of the model. Dimensionality is defined as the number of variables of interest and the time needed for solution. At the time of the development of the initial model (later named Carmonette—from the first syllables of the words Monte and Carlo, reversed, and given the suffix "ette" meaning small; the model represented small-scale tactical combat), 30 minutes of combat took about 20 minutes of computer time (on an ERA [Engineering Research Associates] IBM 1101 cathode tube computer. To make 100 runs of one combination of variables required 33 hours of computer time. The model was designed, not to provide predictions of the outcome of engagements between a US Army tank company, supported by infantry and mortars, and a Soviet tank company, supported by anti-tank guns and dismounted infantry. It was designed to allow for detailed research on the interaction of weapons at the tactical level of combat, to provide a basis for research on combat. The US side was represented by 20 elements, the Soviet force by 24 elements. Not overtly acknowledged at the time is the overwhelming importance of human behavior during such complex phenomena. Setting that aside for the moment, the history goes on.

The first generation of computers, vacuum tube machines, lasted until about 1959. Programming costs were high, instruction execution times were long, and mean time between failures was short. About 1959 the machines went solid state with transistors; costs were reduced, instruction execution times speeded up, and mean time between failures extended. Operations analysts who were devoted to digital models as tools of their trade became ecstatic. Models began to be more complex, including many more variables, entities and assumptions than before. Now even less attention was paid to the clear enunciation of assumptions and the identification of the potential impact of the assumptions on model behavior. The third generation arrived about 1964 with the integrated printed circuit boards (the era of the IBM 360) with the usual results of considerably reduced cost of execution, shorter instruction times for execution, and improved failure rates. The impact on military modeling was another leap forward, with only limited accompanying research on the finer interactions among things and people on the battlefield. Microchips, in 1975, not only continued the improvement in

[3] Joseph F. McCloskey & John M. Coppinger, eds., *Operations Research for Management. Vol. II. Case Histories, Methods, Information Handling*, The John Hopkins Press, 1956.

operational terms but also led to the revolution in the major reduction of the size of computers and hence the widespread use of personal computers with capabilities vastly exceeding those of the first generation of vacuum tube digital computers. With the ability to write detailed models while sitting at their desks, or in the airport waiting lounge, or even while airborne, modelers walked away from the task of doing the research to allow them to understand the phenomena they were attempting to represent in computer code. The community was enchanted—no—beguiled and seduced by the computer. Do I overstate the situation? Perhaps a bit, but only to emphasize important weaknesses that affect our ability to carry out our mission: to provide the best possible analyses of problems affecting the lives, well-being and performance of the young warriors who go in harm's way to defend our country.

A Little More History: Validation

When the first efforts at getting a handle on models and model development began, appropriately with the US Army, under the direction of Walt W. Hollis, then the Deputy Assistant Secretary (Operations Research), ultimately to become the last person to hold that position. [The position was the only position in the military Services at that level in the bureaucracy designated as Operations Research. Its passing, by action of the Secretary of the Army, says something about the view of operations research at the top level of the Department of the Army.] Returning to the management of modeling in the Army and the Services, the Army established the Army Model Improvement Program, with some staffing provided by the Training and Doctrine Command. Administration of the program was at Fort Leavenworth, Kansas, with Mr. Hollis providing the policy level leadership. Much of the Army's analytical community participated in meetings to discuss model management. At the outset (early 1980s) the word validation was rarely mentioned. Emphasis was on structuring a hierarchy of models in a coordinated way. The notion was that the entry point for the hierarchy would be the output from systems level models which would be the responsibility of the Army Systems Analysis Activity. Systems data would be fed into low-level, highly detailed models representing small unit combat (with responsibility assigned to the analytic teams of TRAC (for TRADOC Analysis Center), White Sands, and New Mexico). The output of those models would feed into the next level of models (brigade and division-level forces), the responsibility of the TRAC teams at Fort Leavenworth. Support unit and services would be the responsibility of TRAC at Fort Lee, Virginia. The then Concepts Analysis Agency (now the Center for Army Analysis) was responsible for theater and strategic level models, using the output of the lower-level, more detailed models. The full implementation of the hierarchy was never reached, however noble the idea was.

Shortly after the initiation of the Army Model Improvement Program (no other Service began such a focus on the management of models at that time) the Military Operations Research Society increased its interest in modeling of combat, recognizing that the subject was dwarfing the notion of analysis as established during and immediately following World War II. Other groups entered the fray also, notably The Military Conflict Institute. Paul Davis of RAND with colleagues began to raise questions about the basis

for model development.[4] A spirited challenge to the Davis paper was mounted by Phil Louer, then Deputy Director, US Army Concepts Analysis Agency.[5] In the same issue of *Phalanx*, Davis replied with a short rebuttal. Recall that the initial applications of the newly named practice of operations research consisted of paper and pencil (computational) efforts and Conceptual Models, laboratory experiments, field experiments and demonstrations, and operational data collection and compilation. It was never really as much a notion of the "mixed team" concept as it was the notion of a mixture of methods. A review of *Phalanx*, the bulletin of military operations research and once a primary source of information about the progress of military operations research in the US, reveals no significant reference to model validation until 1988. A paper by MAJ Flanagan of TRAC, titled "V&V: A TRAC Approach", defines validation as "a determination of whether the model/simulation reflects results expected in the real world." The paper continues with reference to a process ("relatively straight forward procedure designed to lead us to a sense of well-being with respect to the validation question."). In each step of the procedure, the word reasonable is repeated in reference to the results of the step. So we have "reasonable test," "combined effects reasonable," "results generally reasonable, intuitive and comparatively pleasing."

In 1986 MORS initiated what became a series of mini-symposia and workshops named More Operational Realism in the Modeling of Combat, acronym MORIMOC. MORIMOC I was structured to deal with three aspects of modeling: operations, mathematics, and physics-engineering. Very briefly, the findings of the workshop expressed particular weaknesses and difficulties with the first two aspects and less concern with the physics-engineering aspect. Among the key findings were:

Over-estimates of the lethality of almost everything;

Underestimates of the creativity of opponents in limiting damage; and

No accounting for the degree to which smart or dumb use of forces can dominate the outcome.

[Any improvement today?]

MORIMOC II, in January 1989, was a mini-symposium with considerably emphasis on some of the weaknesses found in the first workshop. Sessions were conducted on human factors in decision issues, human performance models, and applications, predicting human performance and availability in combat environments, combat as a data source, and representation of human performance in combat models and simulations.

Building off of MORIMOC II was MORIMOC III, titled Human Behavior and Performance as Essential Ingredients in Realistic Models of Combat, conducted in March 1990.

---

[4] Paul K. Davis & Donald Blumenthal, *The Base of Sand Problem: A White Paper on the State of Military Combat Modeling*, RAND N-3148-OSD/DARPA.
[5] *Phalanx*, Vol. 24, No. 4 (December 1991).

Shortly thereafter, MORS initiated a series of special meetings named SIMVAL, for Simulation Validation. A first mini-symposium was conducted in October 1990, as a forum for discussing on-going efforts in model validation. The working definition for the mini-symposium guidance was: Substantiation that a computer model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model. Note: domain of applicability; satisfactory range of accuracy, and intended application. The keynote address for the mini-symposium was delivered by MG George B. Harrison, USAF, and then the Air Force sponsor of MORS. One short quote from his message:

> "…As you know, MORS itself is not a policy-making…organization…But, recognizing the organizations represented here…I do not believe there is anyone who can come up with a better approach if we set our mind on an issue? Do you? So when I say we have an opportunity to *significantly affect* DoD policy, I may be guilty of understating our influence a bit…" [emphasis in the original]

Keep that quotation in mind!

Subsequently, MORS conducted a SIMVAL II, a workshop designed to formalize what has now become known as Verification, Validation, and Accreditation. The prophetic keynote words at the first mini-symposium on the subject came true. The Department of Defense, in its own development of a process and institution to coordinate and produce military modeling policy throughout the defense community essentially adopted the products of the MORS special meetings. At no time, did the phrase "unintended consequences" ever enter the dialogues. With the adoption of the MORS findings by the Department and the resultant institutionalization we were left with procedures and policies with heavy emphasis on form and major gaps in content. When a senior Defense official was queried as to why there was so little specification and standards introduced into the policy, the response was: "We will wait for the community to establish the quantitative standards" or words to that effect.

And, so here we are, still struggling with the concept of the truth and applicability of models, rife with assumptions, both stated and unstated.

Are We Nearing the End of This Monologue?

George Akst is certainly correct in arguing that we have moved into a different arena with respect to the types of data now required—but, we have also moved into a different arena with respect to the types of models required. George says: "…we must not forget that complicated analytical tools can be no better than the data that drive them." We must also not forget that quality data, particularly now the more rigorous data called for, is not particularly useful when incorporated into models that are either irrelevant with respect to the analytic questions asked or so hampered by poorly defined and clarified assumptions as to be ineffective. And thus we back into the underlying matter: it's analysis that binds the two, data, and model. Understand the questions of concern (and we acknowledge that difficulty, requiring multiple conversations with the leadership, prodding the leadership and pushing to get clarity on the "real" problem), determine the

most expeditious and relevant way to respond to those questions (models, in a generic sense), and, simultaneously with the model development or selection, identify and obtain the necessary, quality, rigorous data.

There is another cycle that is also relevant—perhaps even more so than the foregoing discussion. When faced with strange, new, interesting, or different phenomena, the first step is data gathering—or at least, that should be the first step. Often, unfortunately, the first step is stating hypotheses (actually, assumptions) about the phenomena, relying on the fond fallbacks, the Subject Matter Experts (often, the analyst is his or her own SME!). Thus, when concerned even with basic matters such as the interactions of small arms at the fire team or squad level for infantry operations, as an example, we blithely went ahead with assumed relationships which became set in concrete and were perpetuated over the years until they became as gospel never to be challenged or, worse, examined carefully by analysts relying on real data. So here we are, a full 50 plus years after the first practical discussions of digital computer representations of combat (now seen almost exclusively as models of interest to military analysis) without much of a clue as to what combat is really like, from an analytic standpoint.

We are on the brink of a new period (actually, we are already in the midst of the new period), that of agent based simulation modeling. This relative new arena takes advantage of the very computer characteristics there were so seductive and, I believe, so damaging during the past few decades of military analysis. It is only through the magnificent speeds, simplicity of programming, and low cost of computation that we can now begin to represent the most important elements of our analyses—human behavior. Our earlier failures, while attributed primarily to our failure to do the fundamental research into complex systems behaviors and relationships, also stemmed from our inability to account for the human on the battlefield or other military environments.

However, we cannot neglect the research and analysis necessary to provide definition to the human behaviors we simulate. If we follow our historical practice of making assumptions, as opposed to carrying out the research, we will continue to produce models that are ineffective, irrelevant, or even potentially damaging to the very institutions we serve. There is a wealth of applicable, quality data existent from a variety of sources not usually tapped by military operations analysts or modelers. There is little excuse for not digging in, surfacing and applying the needed human behavioral data, including interactions among humans in complex situations.

I nearly close this diatribe with an anecdote. Some months ago, at a symposium I heard a paper on the application of an agent based simulation model to a building clearing operation by Marines. A critical assumption had to be made because of a weakness in the model (the work was exploratory, the very kind of fundamental exercise needed to gain understanding of a complex phenomenon). The assumption quite possibly significantly affected the observations arising from the simulation. The two young analysts were quite forthright in their exposition and their highlighting of the assumption and its likely affect. A refreshing incident and perhaps a view of the future of military analysis.

Penultimate Comment

I relied heavily on the collection of *Phalanx* bulletins, compiled ably by Lee Dick and available on a compact disk from MORS. In strolling through the scanned archives covering the publication 1966 to 2000, I came across many interesting and exciting pieces (including some ramblings of my own over the decades). There is great value in going back over many of the issues, containing sound and detailed papers of wide variety—many of them terribly relevant to today's military analytic world.

I was particularly delighted to read a short note by Gene Woolsey, who was invited as a banquet speaker (when we had real banquets) at a MORS meeting. He introduced a series of questions that he planned on putting to every presenter he was to listen to. The questions represent Woolsey at his best but moreover they are fundamental to our practice. Herewith is the catechism due to Woolsey:

> Did you know what they were doing before you modeled it?
>
> If yes, how did you know? (The only acceptable answer is "Because I did it the old way first.")
>
> Is your model in use?
>
> If yes, how do you know?
>
> Does it work?
>
> If yes, is there a measurable, verifiable reduction in cost over what was done before or a measurable, verifiable increase in readiness?
>
> If yes, show it to me now.

Ultimate Comment

I came across a most pertinent statement that is very much worth remembering as we move along with attempts to improve the world. It comes from essayist Jonathan Yardley ("Victimization Strikes Out," *The Washington Post*, August 2, 1999, p. C2) and reads:

> "Consider, if you will, Yardley's Law of Unforeseen Consequences. Put as simply as possible: Bad always follow good. Put more elaborately: Any action, no matter how noble the intentions behind it, sooner or later has unanticipated ramifications that are mischievous at best, disastrous at worst, and if that action is legislative or judicial, the potential for unforeseen and undesirable consequences increases exponentially."

And, now the final point, for true. The saving grace is that for all of our efforts, weaknesses, and continued efforts to reach "truth," we have to face the humbling

observation that few if any significant military decisions are made solely on the basis of a model's output.

Annex: Taleb on Models[6]

"As I have said earlier, the world, epistemologically, is literally a different place to a bottom-up empiricist. We don't have the luxury of sitting down to read the equation that governs the universe; we just observe data and make an assumption about what the real process might be, and 'calibrate' by adjusting our equation in accordance with additional information. As events present themselves to us, we compare what we see to what we expected to see. It is usually a humbling process, particularly for someone aware of the narrative fallacy, to discover that history runs forward, not backward. As much as one thing that businessmen have big egos, these people are often humbled by reminders of the differences between decision and results, between precise models and reality.

What I am talking about is opacity, incompleteness of information, the invisibility of the generator of the world. History does not reveal its mind to us—we need to guess what's inside of it.

The above idea links all the parts of this book. While many study psychology, mathematics, or evolutionary theory and look for ways to take it to the bank by applying their ideas to business, I suggest the exact opposite; study the intense, uncharted, humbling uncertainty in the markets as a means to get insights about the nature of randomness that is applicable to psychology, probability, mathematics, decision theory, and even statistical physics. You will see the sneaky manifestations of the narrative fallacy, the ludic fallacy, and the great errors of Platonicity, of going from representation to reality."

The prepublication version of this paper can also be found at http://orsagouge.pbwiki.com/ABSVal.

[6] Nassim Nicholas Taleb, *The Black Swan. The Impact of the Highly Improbable*, Random House, 2007, p. 268.

APPENDIX H P-OCM VALIDATION REPORT

H.1 INTRODUCTION

In Phase II of the ABSVal Study, the Team was to assess at least two candidate simulations useful in developing an Irregular Warfare Analytic. The objective was to test the viability and utility of the ABSVal framework in a realistic institutional setting. The first simulation selected was the Pythagoras Counter-Insurgency (COIN) model developed by NGMS for use by the Operations Analysis Division (OAD) of the Marine Corps Combat Development Command (MCCDC). The second simulation selected was another Pythagoras application, an Obstacle Clearance Model developed by NGMS for the Marine Corps Warfighting Laboratory (MCWL) and the Office of Naval Research (ONR). Sanderling Research Corporation's (SRC's) role in both efforts was to apply a single technique, assumption testing, within that framework to assess the validity of the Conceptual Model of each Pythagoras instantiation.

This paper reports the conduct and results of SRC's assessment of the second simulation, the Pythagoras Obstacle Clearance Model (P-OCM). The report of the conduct and results of SRC's assessment of the first simulation is addressed in a separate paper. In order for the two papers to be able to stand alone, the remainder of this Introduction section and down through Section 6.2, "Plan," are essentially identical in each paper.

The potential usefulness of assumption testing as a validity assessment technique may be seen by considering the nature of models in general. George Box famously stated, "All models are wrong, some are useful." Box was absolutely correct in the literal sense – no model *is* reality. Rather, every model is an abstraction of reality to some extent. That aspect of a model is widely if perhaps not universally recognized. What is more rarely recognized and far more rarely appreciated is that the mechanism of abstraction is the *assumption.* Thus if we could identify every assumption used to create a given model, we would know how it deviates from reality or, in Box's terms, we would know just how "wrong" it is.

As a practical matter, however, we can not *explicitly* identify every assumption in even simple models. The good news is there is no need to identify all of them. We need only to identify, and "test," the assumptions that have significance to the *intended purpose* of the model and especially to the *analytic questions at hand.* That is, of course, more easily said than done. Part of the art, vice science, of assumption testing is to be able to recognize in at least broad terms which assumptions are likely to be significant, given only a description of the model, the context of the study, and the specific analytic questions at hand. Thus analysts generally have to cast a wider net than would be necessary if they had full knowledge going in as to which assumptions are significant. Assumptions having little or no apparent significance are set aside. Ones having apparent significance are tested as described later in this report.

H.2 ROLE OF SCIENTIFIC METHOD. [7]

The ABSVal framework approach is based in the scientific method, with the thrust being to find evidence that would reject (falsify) the null hypothesis that the model or simulation is valid for the intended purpose. Scientific method applies within assumption testing, but in an indirect fashion. A validity assessment attempts to determine whether a model or simulation is "sufficiently accurate," *vis-à-vis the real world*, for a particular application. But because assumptions represent purposeful *departures* from the real world, and sometimes quite significant departures, a direct application of scientific method – one that directly compared the assumptions to the real world -- could readily hold the model to be invalid without even considering the intended application. (Which is why Box's famous "All models are wrong" quotation often is cited, incorrectly, as evidence that validation is a waste of time and money.) The indirect scientific method approach used in assumption testing notes the departures from reality, but then determines their *operational implications in the context of the application at hand*. The final step of the indirect approach is for the Application Sponsor to decide whether those operational implications are acceptable for his or her application.

H.3 CONCEPTUAL MODEL

In the ABSVal framework, the term "Conceptual Model" encompasses three distinct sub-models: [8]

- ***Theoretic model*** is the initial expression, usually in textual and/or graphical form, of the context of the model and of the cause-and-effect relationships believed to be operative in the situation of interest and that are intended to be incorporated within the end model. In an ABS, it contains all of the agent behaviors and relationships.

- ***Mathematic model*** captures the specific logical structures and expressions (equations, conditional statements, logic tables, etc). Note that the relationship between the theoretic model and mathematic model is one-to-many; i.e., there are numerous ways a theoretic model could be represented mathematically.

- ***Algorithmic model*** is the coded form of the mathematic model. Again, note that the relationship between the mathematic model and algorithmic model is one-to-many.

[7] For a detailed discussion of scientific method, the falsifiability criterion, and research, null, and alternative hypotheses, see Section 4.3 of the ABSVal Phase I Final Report, Ref 1.

[8] The ABSVal Phase I Final Report, Ref 1, only recognizes the first two sub-models, the theoretic and mathematic models, as composing the Conceptual Model. However, it also neglects assumption testing the third sub-model, the algorithmic model, which actually is the most important of the three. For that reason, the algorithmic model is addressed in this report as part of the Conceptual Model.

H.4 TYPES OF ASSUMPTIONS

As stated in the Final Report of Phase I of this study [Youngs and Bitinas, 2007], there are four sets of assumptions of interest: causal, structural, mathematic, and scenario:

- ***Causal assumptions*** deal with cause-and-effect relationships among agents/objects/entities and with their environment(s) and other stimuli.
- ***Structural assumptions*** deal primarily with the processing order of stimuli, decisions, and actions within a model, but also may deal with starting, ending, and boundary conditions within a model.
- ***Mathematic assumptions*** deal with the myriad assumptions made to enable constructing a determinable mathematic abstract of real-world scenarios, processes, behaviors, and events; mathematic assumptions include the choice of algorithms and other logic structures, and thus assumption testing includes an assessment of those algorithms/structures.
- ***Scenario assumptions*** deal with bounding the real-world environment (which may be behavioral as well as geophysical) to be addressed within the model, with the geophysical features and environmental conditions contained therein, and with the entities and their characteristics to be "in play" in a particular scenario.

H.5 ASSUMPTION TESTING PROCESS

Also as stated in the Phase I Final Report, assumption testing is a three-step process:

- ***Step 1: Identify the assumptions.*** Assumptions, particularly causal and mathematic assumptions, are rarely if ever well-documented and may even have to be reverse-engineered from the source code. In some cases, even some of the algorithms may not be documented. This is *by far* the most difficult aspect of assumption testing.
- ***Step 2: Determine the operational implications of the assumptions.*** Accomplished as a cooperative effort between M&S and operational subject matter experts (SMEs).
- ***Step 3: Determine the acceptability of the identified operational implications to the decision-maker.*** During M&S system development, the decision-maker is the M&S sponsor. For a particular application of the M&S system, the decision-maker is the application sponsor (designated by DoD policy as the accreditation authority for that particular application).

H.6 ASSUMPTION TESTING APPLIED TO P-OCM

H.6.1 Precepts [9]

H.6.2 Modern Scientific Method

Throughout, the planned process is based on modern scientific method and most specifically on the *falsifiability criterion* contained therein.

As a convention for this assessment, the *null hypothesis* is the *research hypothesis* that the model being assessed *is valid ("sufficiently accurate") for its specific intended application*. The *alternative hypothesis* then is defined as the negation of the null; i.e., the model *is not sufficiently accurate for that particular application.* We then attempt to *falsify the null hypothesis.*

Every step of the planned process (except for writing the end-game report) is intended either to set the stage for falsification of the null or to execute falsification tests of the null.

A failure to falsify the null does not mean the model is *proven* valid, but it should greatly increase confidence in the model's validity for that particular application. The degree of confidence depends on the rigor and power of the tests applied.

H.6.3 Plan

H.6.3.1 Identify Analytic Questions

Identify the *analytic questions* the model is/was intended to address in the specific application at hand, the metrics applicable to those questions, and the degree to which model results are/were expected to shape the decisions to be made.

- Detailed review of all available documentation of the application
- Interview the Application Sponsor

H.6.3.2 Identify the Referent

In collaboration with the Application Sponsor, identify the *referent*; i.e., the proxy for the real world for the purpose of accuracy comparisons. See Sect 4.0 of the Phase I Final Report for the various forms such a proxy may take.

H.6.3.3 Identify the Accreditation Criteria

In collaboration with the Application Sponsor, identify the *accreditation criteria.*

- Establish just how accurate the end results of the model have to be when used in the particular application at hand

[9] As stated earlier, for a detailed discussion of scientific method, the falsifiability criterion, and research, null, and alternative hypotheses, see Section 4.3 of the ABSVal Phase I Final Report, Ref 1.

- Must include how “accuracy” will be determined, and may have both qualitative and quantitative aspects
- Criteria must be “pass/fail;” i.e., they must establish the *lower bounds of accuracy* that must be met for the model to be acceptable for the application at hand

H.6.3.4 Assess the Validity of the Referent

- Confirm that no alternative referent is available or could reasonably be constructed that would be preferable to the one identified (e.g., is an empirically-derived database available? Could one be made available?)
- Assumption testing:
  - Identify/derive the *assumptions* that are inherent to/embedded in the referent
  - Perform logical verification -- determining whether the referent as written *adequately and correctly implements* underlying theory and assumptions
  - Determine the *operational implications* of the identified assumptions in the context of the particular application and with respect to the remaining steps of the ABS Val framework
  - Determine *bounds of validity* imposed on the application’s problem space and on the *model’s* validity assessment by the referent’s assumptions
  - Determine whether the operational implications and bounds of validity are acceptable to the Application Sponsor
- Independent SME review(s) (ideally, these will be contrarian reviews from SMEs that would focus on any potentially falsifying aspects the referent)

H.6.3.5 Determine remaining workplan

Determine the most efficient sequencing of the remaining steps of the ABS Val framework, based on:

- Information developed to this point
- Estimates of the difficulty and costs of the individual remaining steps
- The relative power of each of those steps to falsify the null

H.6.3.6 Assess the Validity of the Conceptual Model

- Potentially as many as three separate assessments; in sequence:
  - Theoretic sub-model
  - Mathematic sub-model (if it exists)
  - Algorithmic sub-model
- Each in turn will have the same assessment techniques applied to it:
  - Logical verification -- determining whether the sub-model as written is an *adequate and correct implementation* of its predecessor (free of logical, mathematic, or algorithmic error)
  - Assumption testing:
    - Identify/derive the *assumptions* that are inherent to/embedded in the sub-model
    - Determine the *operational implications* of the identified assumptions in the context of the particular application

        - Determine the *bounds of validity* of the model that are the result of the identified assumptions
        - Determine whether the operational implications and bounds of validity are acceptable to the Application Sponsor for the intended application
    - For some models, it may prove necessary to reverse-engineer one or more sub-models from later models. It may even be necessary to reverse-engineer the Conceptual Model, or portions of it, from source code.
    - The referent serves as the predecessor for the theoretic sub-model

H.6.3.7 Assess the validity of the instantiated model

- Logical verification -- determining whether the instantiated model as coded is an *adequate and correct implementation* of its predecessor (the algorithmic sub-model)
- Data validation
    - Source
    - Data element definitions
    - Data values

H.6.3.8 Assess the validity of model results

- Comparison of model results to the referent
- Must address each accreditation criterion

H.6.3.9 Develop final validation assessment report

- Specific addressal of each accreditation criterion
- Incorporates, at least by reference, the dataset used to generate “results”
- Make accreditation recommendation
- All in the context of the specific application at hand

This last set of steps assumes the validity assessment process goes all the way through assessing the validity of model results. If at any earlier point the null hypothesis is falsified, the process may be cut short and the report written to reflect the findings at that point.

For the Pythagoras Obstacle Clearance Model, SRC applied only Steps H6.3.1 through H6.3.4 and H.6.3.6 above. For all the other models, the process *may* be tailored to each individual model/application pair. In general, however, we expect each step to apply.

H.7 RESEARCH

H.7.1 Project Documentation.

Mr. Ryan Paterson of The Praevius Group, who had been the Application Sponsor and Principal Analyst for the study while stationed at the Marine Corps Warfghting Laboratory as a Marine Corps captain, provided the following documentation:

- *Precision Guided Munitions versus Shallow Water Obstacles: The Project Albert Contribution,* undated briefing
- R. Paterson, M. McDonald, J. Eusse, T. Erlenbruch, E. Bitinas, *Shallow Water Obstacle Clearing*, briefing to 5th Project Albert International Workshop (PAIW5), Uberlingen, Germany, July 2002
- R. Paterson, E. Bitinas, *Modeling Obstacle Reduction with the Pythagoras Agent-Based Distillation,* Maneuver Warfare 2003
- R. Paterson, *Thoughts on 6th Project Albert International Workshop (PAIW6),* Unpublished Notes to the Surface Zone/Beach Zone Obstacle Group, March 2003
- R. Paterson, M. McDonald, J. Eusse, T. Erlenbruch, E. Bitinas, *Shallow Water Obstacle Clearing,* brief to 5th Project Albert International Workshop, July 2002
- R. Paterson, *Surf Zone/Beach Zone (SZ/BZ) Obstacle Reduction,* Marine Corps Warfighting Laboratory briefing, undated
- R. Paterson, *Precision Guided Munitions versus Shallow Water Obstacles: The Project Albert Contribution,* undated briefing

In addition, Mr. Paterson provided the input and results data files from the study's Pythagoras runs.

Interviews
The interviewees were:

H.7.1.1 Application Sponsor

Mr. Ryan Paterson of The Praevius Group, who as indicated above was at the time of the study a Marine Corps captain assigned to the Marine Corps Warfighting Laboratory, was the Application Sponsor. The problem he was working was as old as Amphibious Warfare itself: clearing a path through mine and obstacle fields to enable amphibious penetrations to and through a hostile beachhead.

Quoting from Reference 1 [Paterson and Bitinas, 2003], "The Navy and the Marine Corps currently have no program of record to reduce obstacles in the Surf Zone and Beach Zone (SZ/BZ). The Office of Naval Research (ONR) Assault Breaching System (ABS) program is taking a two-prong approach to develop capabilities to reduce these obstacles. The program has been divided into near term and far term efforts with the near term effort hoping to produce results by fiscal year 2006. The premise behind the near term effort is the use of precision guided munitions, Joint Direct Attack Munitions (JDAM), to reduce obstacles in the SZ/BZ. At the 5th Project Albert International

Workshop (PAIW5) (1-7 July 2002) a multinational team began using the Pythagoras distillation to explore factors such as weapon size, weapon number, aimpoint pattern, impact timing and others."

More specifically, Mr. Paterson was working with Code 32 of the Office of Naval Research and the U.S. Air Force to determine whether air-dropped bombs could be effective counter obstacle devices in the Surf Zone (SZ) and Beach Zone (BZ). The Air Force had conducted a series of live static tests of Mk 80 series bombs against different obstacles in a large pond constructed at Eglin Air Force Base. Tests were conducted both with the pond dry, to simulate a beach zone, and wet, to simulate a surf zone. Two water depths were used: 3 ft and 6 ft, and the bombs either were suspended just above the bottom of the pond or were buried to one-half their length in the bottom. One bomb had been suspended at a 45-degree angle from the vertical. Detonation tests can only establish whether the necessary damage and displacement results can be achieved against some target set. It's always still necessary to determine whether the weapons of interest could be employed tactically to achieve the desired end result. Thus Mr. Paterson needed a tool that would help him model the results of the live tests in order to answer the tactical questions, such as the ones cited above, about the employment of bombs for obstacle clearance operations. Going into the study, Mr. Paterson established four initial questions:

- Are there current weapons that can be used differently to defeat obstacles?
- Are there promising new technologies that Project Albert [10] can model?
- Is agent based simulation modeling a realistic tool for this problem?
- Is there anything we need the models to do differently for this problem?

When it became obvious that Pythagoras would be able to model the problem, he added several specific tactical questions:

- What accuracy is best?
- Do simultaneous or sequential detonations play a role in developing a better "lane?"
- Is there significant difference between using precision bombs versus conventional bombs?
- How many bombs should be dropped at each aim point?

Mr. Paterson also decided on a significant change from most studies that address clearing mines or obstacles from amphibious assault lanes. Most studies use as their Measure of Effectiveness the percentage of the lane that is cleared or the removal of a stated percentage of mines and obstacles (often stated as the confidence level of having cleared a threshold percentage of mines and obstacles). Mr. Paterson instead decided to use the number of Amphibious Assault Vehicles (AAVs) surviving the attempted transit of the assault lane after bombs had been dropped to clear it (although all the reports actually were stated in terms of the numbers of AAVs killed). Finally, he focused the study entirely on the Surf Zone.

[10] Project Albert was a Senate-sponsored High-Performance Computing R&D program that produced, inter alia, the Pythagoras agent-based simulation toolkit.

H.7.1.2 Northrop Grumman Project Leader

Mr. Edmund "Edd" Bitinas, the developer of Pythagoras, led the technical effort to develop a Pythagoras instantiation of the Obstacle Clearance problem. The initial thrust of the effort was to use an ABS to simulate the observed behaviors of various types of obstacles when subjected to bomb detonations in shallow water (at "surf zone" depths). The observed behaviors came from the Air Force tests mentioned earlier.

The remaining paragraphs of this section are drawn from the interviews with Mr. Bitinas, supplemented by various P-OCM reports and graphics from those reports.

The Air Force testing at Eglin had begun in 1999, well before the P-OCM study was initiated. Figure 25 is the "High-Threat" Obstacle Clearance scenario developed by the Marine Corps Intelligence Activity at Quantico. The threat used by the Air Force in its tests was considerably less robust than the "High Threat" scenario, and the placement of mines and obstacles was chosen to yield the best data rather than match the MCIA scenario placement. Subsequently, the threat array used in P-OCM was still simpler, using only one type of mine and two types of obstacles – Hedgehogs and Triple-Strand Concertina (TSC) wire.

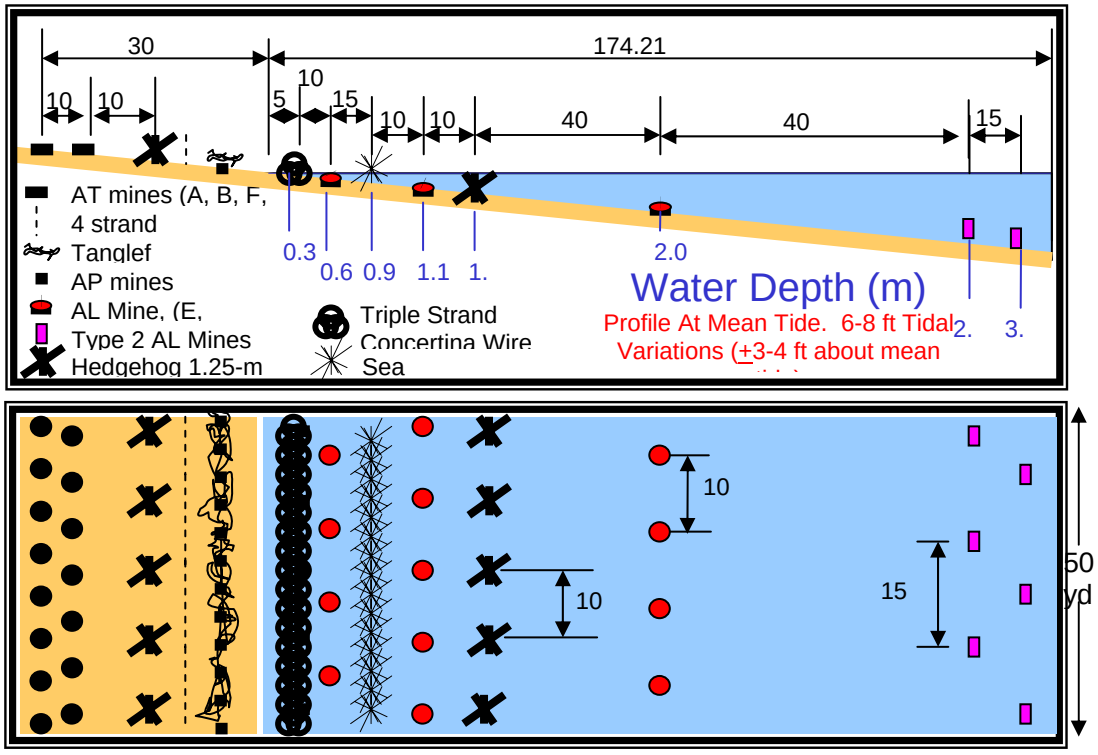


Figure 25 MCIA High Threat Laydown, Beach Gradient of 1:99

Figure 26 summarizes the live tests conducted by the Air Force. Figures 27 and 28 provide representative individual test results. Two things are worth noting in those last two figures: the fairly wide variation of displacements even in individual tests, and the fact that in Figure 28, two mines were "pulled" toward the detonating bomb rather than being either destroyed or displaced outward from it. Note that those two mines were

just outside the radius of the water plume created by the bomb detonation, possibly indicating a “suction” effect from the violent upwelling of the water mass.

LAND

- Mk 82 – Single, 5 Sequential, 5 Simultaneous, 2 Simultaneous, 5 Single against damaged obstacles
- MK 84 – 5 Single
- MK 83 – 3 Single

WATER

- Mk 82 - 2 Simultaneous, 6 ft of water
- Mk 82 - 3 Simultaneous, 6 ft of water
- Mk 82 - Single (vertical), 6 ft of water
- Mk 82 - Single (45°angle), 6 ft of water
- Mk 82 - Single (1/2 buried), 3 ft of water
- Mk 84 - Single (vertical), 6 ft of water

Note: There is also a database of Mk83 tests against obstacles on land and in water.

	Explosive Weight (lbs)	Case Weight (lbs)	Warhead Size(diam x length, in)	Overall Weight, incl fins	JDAM Variant Certified
Mk 82	192	301	10.8 x 60.6	LD 525/HD 567	
Mk 83	385	520	14.0 x 72.5	LD 985/HD 1020	F/A-18
Mk 84	945	935	18.0 x 97.3	LD 2000/HD 2030	F/A-18, B-1B/B-2A B-52F

LD = Low Drag, HD = High Drag *Ref: Joint Munitions Effectiveness Manual.*

Figure 26 Mk 80 Series Tests Conducted

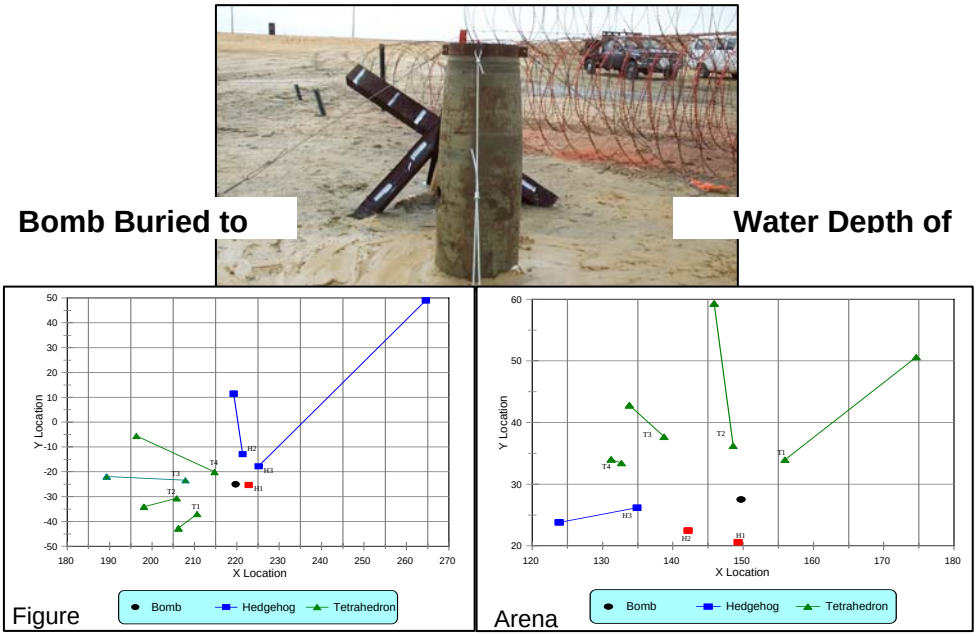
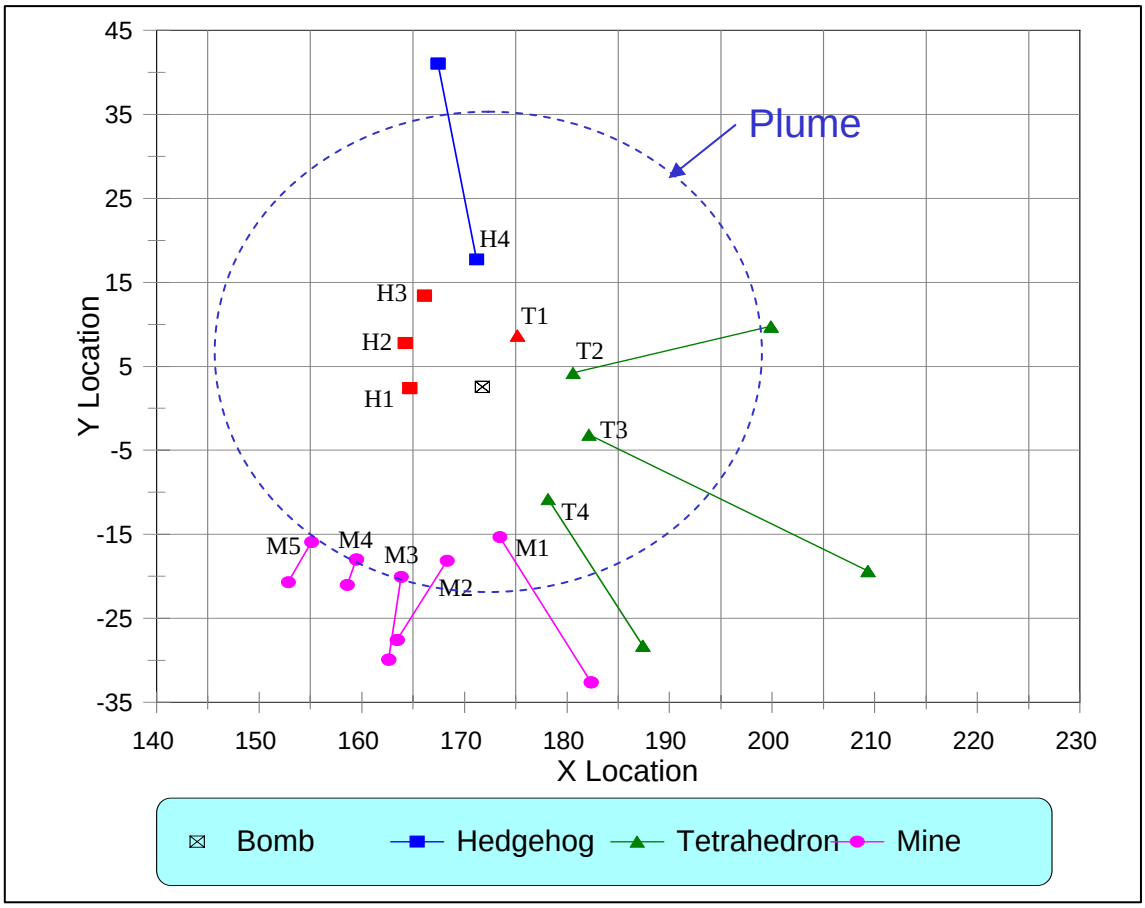


Figure 27 Surf Zone Tests Mk 82 Pond Test – Half Buried (March 2002)



- Note - M5 and M4 moved towards the bomb

Figure 28 Surf Zone Tests Mk 84 Pond Test – April 2002

Figure 29 summarizes the damage caused by the Mk80 bombs against the various types of obstacles used in the Air Force test. It established the basis for damage assumptions made in P-OCM (e.g., that bombs had no effect on TSC, could only displace hedgehogs, and could only destroy mines within a limited range). A second display of actual damage data from multiple sources, however, shows that the actual kill experience of high explosives against mines is quite complex (Figure 30).

	LAND	**WATER**
Simultaneous Detonation	Little or no effect	Little or no effect
Cumulative Damage	Can be a factor	No data
Craters	Minimal	Minimal
Tetrahedrons	Sensitive to weld strength Easy / hard to kill	Sensitive to weld strength Hard to kill; Displacement
Hedgehogs	Sensitive to weld strength Hard to kill; Easy to displace	Sensitive to weld strength Hard to kill; Some displacement
TSC Wire / Fence / Sea Urchin Barriers	Fragments cut up barrier Sea urchin remnants easily displaced Considerable debris	N/A
TSC Wire (Completely Submerged)	Fragments cut up barrier Sea urchin remnants easily displaced Considerable debris	Ineffective (no fragments)
Log Posts	Easy to kill	Easy to kill
Concrete Cubes	Considerable amount of debris	Interesting results
Mines	TBD	Some capability

Figure 29 Mk 80 Series Test Results (Preliminary) Obstacles *

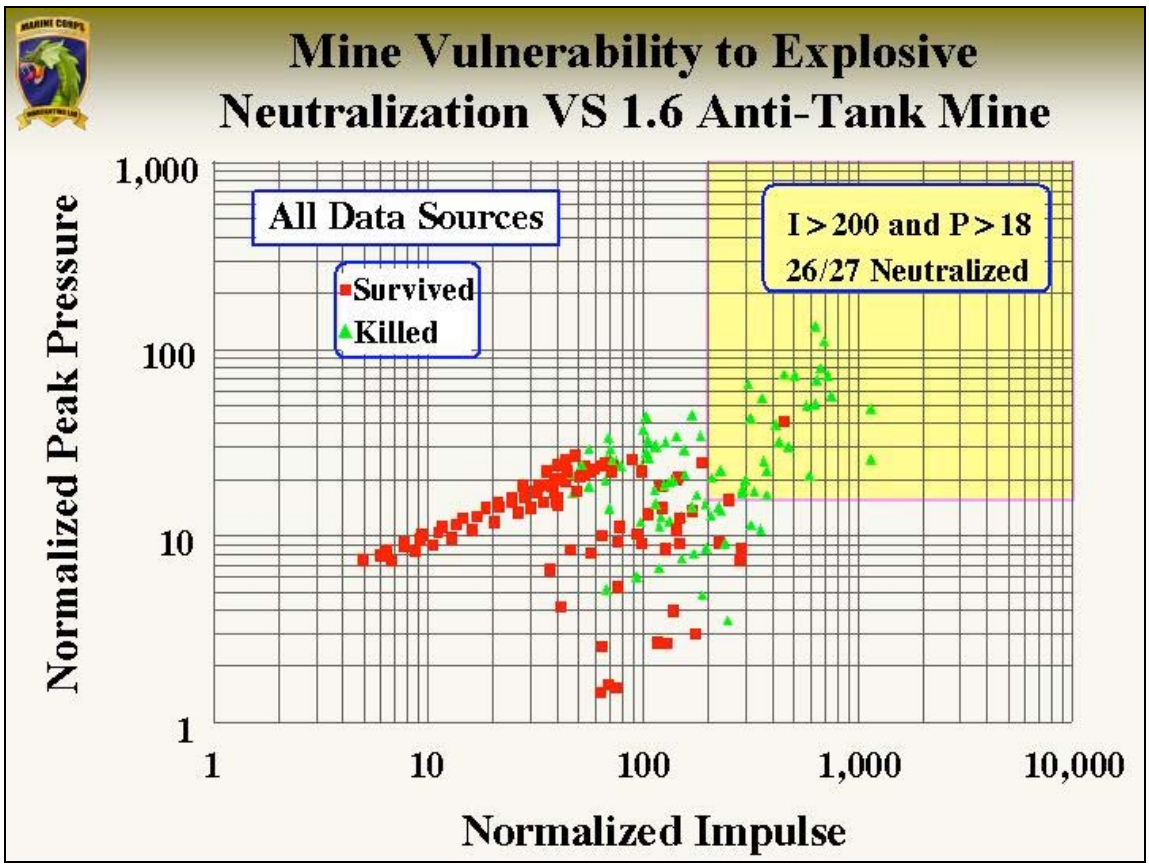


Figure 30 Pressure times Impulse mine neutralization graph

A separate analysis (i.e., part of the study, but separate from P-OCM) used the Air Force data to develop a table of the predicted "Lethal Area Radii" of various bomb type/mine type pairs. The results are displayed in Figure 31. In P-OCM, the bombs themselves were not modeled explicitly; rather, the type of bomb and type of mine determined the radius of the "lethal area" around a hitpoint.

Mr. Bitinas decided early on that "everything" would be agents – obstacles, bombers, aimpoints, hitpoints, and AAVs. The bombs themselves were not modeled explicitly; rather, the type of bomb, type of mine, and depth of water were used to enter the table of Figure 31 to determine the radius of the "lethal area" around the associated hitpoint. Mines that were outside the lethal area were assumed not to be destroyed and were subject to displacement. Mines that were inside the lethal area were assumed destroyed.

Mine Type	Bomb A		Bomb B		Bomb C
	6' Dept	**3'** Dept	**6'** Dept	**3'** Dept	**6'** Dept
A	20*	14	24	16	31
B	10	9	14	10	19
F	17*	12	21	14	28
G	17	12	21	14	28
H	17	12	21	14	28
* Killed out to 21 feet in summer 2000 tests ** Testing continued after this table was constructed					

Figure 31 MK 80 Series Bombs: Predicted Lethal Radius (Feet) for Mines**

The basic concept of P-OCM was quite simple: Mines could be destroyed or displaced; Hedgehogs could only be displaced; TSC would not be affected by bombs; all displacements would be directly away from the point of detonation (the hitpoint), and magnitudes of displacement would be determined from an internal look-up table constructed from the actual Air Force displacement data (with necessary interpolations).

After a bombing run had destroyed or displaced some number of the obstacles, the AAVs would attempt to use the just-created assault lane to reach their Littoral Penetration Point (LPP). The AAVs would be subject to navigational errors and might have to maneuver off their intended course to avoid damaged or sunken AAVs or TSC. If they did have to maneuver, they would again set course directly for the LPP as soon as they were again clear. (Note that a succession of maneuvers could cause some AAVs to be outside their planned assault lane. The problem could be exacerbated by navigational errors).

A series of P-OCM runs was made early on in an attempt to replicate the Air Force experimental trails and their results. Mr. Bitinas indicated the comparisons of P-OCM and Air Force data were "good," but acknowledged that while the comparisons were numerical, they were not statistical. The data supporting those comparisons apparently have not been retained.

H.8 P-OCM PROJECT OBJECTIVES AND ANALYTIC QUESTIONS

As indicated above, P-OCM was part of a much larger effort to find a near-term solution to breaching mines and other obstacles in the Surf Zone and Beach Zone in conjunction with amphibious landings. The specific initial questions were:

- Are there current weapons that can be used differently to defeat obstacles?
- Are there promising new technologies that Project Albert can model?
- Is agent based simulation modeling a realistic tool for this problem?
- Is there anything we need the models to do differently for this problem?

And again, when it became obvious that Pythagoras would be able to model the problem, several specific tactical questions were added:

- What accuracy is best?
- Do simultaneous or sequential detonations play a role in developing a better "lane?"
- Is there significant difference between using precision bombs versus conventional bombs?
- How many bombs should be dropped at each aim point?

H.9 REFERENT

The P-OCM study was unusual, especially for an ABS study, in that an *empirical* referent was readily available – the extensive Air Force experimentation dataset from their pond tests of Mk-80 series bombs against various mines and obstacles, which spanned several years.

H.10 ACCREDITATION CRITERIA

The topic apparently had not been addressed during the study, although a numerical comparison of P-OCM data to the referent was done.

H.11 VALIDATION OF THE REFERENT

The referent varied from the conditions of the "real world" addressed by the study in three major respects:

- The bottom of the Eglin AFB pond was packed earth rather than the sandy composition of many/most SZ/BZ environments
- The bottom of the Eglin pond was flat rather than having the notional slope of the threat scenario
- The only simultaneous or sequential-detonation experimental events were accomplished with the smallest bomb, the Mk-82 (see Figure 26)

We believe the composition of the bottom of the SZ could have a significant effect on the ability of bombs to displace any of the obstacles. While that should have no effect on the validity of the *model* within P-OCM, it could negatively affect the accuracy and usefulness of analytic answers derived from P-OCM. The flat bottom could have a similar problem, but the slope of the MCIA threat scenario in Figure 25 is only 1:99, which should have negligible effect. Experiments could establish the gradient at which "slope" becomes a significant factor. The third variation could be the most significant. If the weapon selected for the breaching study were anything other than the Mk-82, the results could not be considered valid since the only tactics addressed in the study used sequential and simultaneous detonations and all modeled displacements were taken from the experimental results (the referent).

H.12 ASSESSMENT OF THE VALIDITY OF THE CONCEPTUAL MODEL

We were able to use the project documentation and the interviews to identify the major points and the assumptions of the theoretic model. As for P-COIN, however, there was no mathematical model – the programmer went straight to code. Moreover, the coded model also was not available. So the assessment of the Conceptual Model was limited to the theoretic model.

H.12.1 Significant Assumptions

As in the earlier assessment of the validity of the P-COIN model, one significant assumption affected the conduct of the validation assessment rather than the P-OCM study – that Pythagoras itself is error-free and thus did not need to be subjected to scrutiny. Because the overarching purpose of this assessment was to test the viability and utility of ABSVal, assuming Pythagoras error-free was a matter of indifference. Had the purpose been to rigorously assess the P-OCM model for real-world use, however, the assumption would have been inappropriate and potentially counterproductive. In fact, as discussed later in this report, Pythagoras itself has to be suspect in this particular case.

H.12.1.1 Measure of Effectiveness (MOE)

As indicated earlier, most studies of obstacle clearance use as their MOE the percentage of the lane that is cleared or the removal of a stated percentage of mines and obstacles (often stated as the confidence level of having cleared a threshold percentage of mines and obstacles). P-OCM, however, used the number of Amphibious Assault Vehicles (AAVs) surviving the attempted transit of the assault lane after bombs had been dropped to clear it (although all the reports actually were stated in terms of the numbers of AAVs killed).

H.12.1.2 Scenario assumptions

P-OCM uses a very simple and purely geophysical scenario. A hostile featureless beach is defended by multiple belts of mines (one type only), Hedgehogs, and TSC, all parallel to the beachfront and each other. AAVs attempt to transit a pre-planned assault lane by use of waypoints to reach an LPP. Before they attempt the transit, an aircraft drops bombs on pre-planned aimpoints in rectangular boxes targeted on top of the mine and obstacle belts (whose locations are known) and along the centerline of the assault lane. The scenario is depicted in Figure 32.

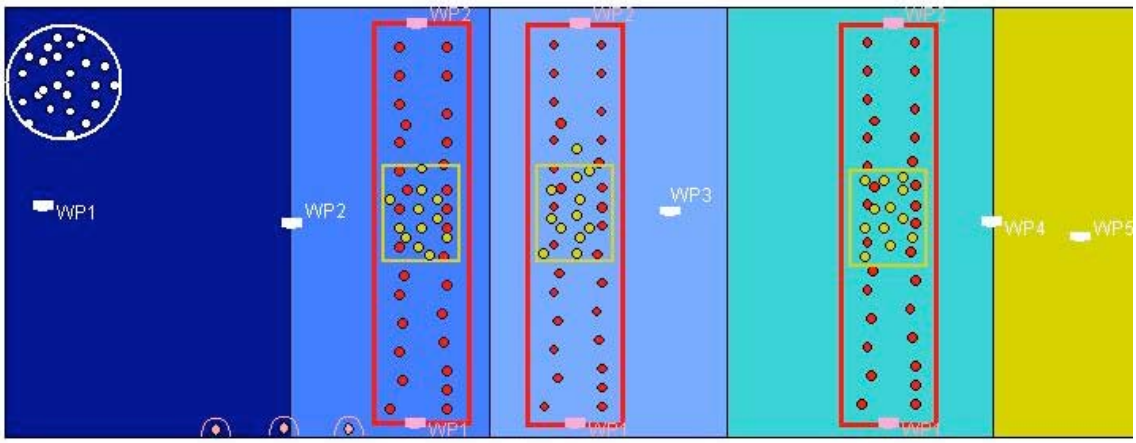


Figure 32 JDAM ABS Scenario (target boxes in yellow)

The critical scenario assumptions are:

- Assault planners know the locations of the mine and obstacle belts
- Planners place the aimpoints only within the boxes oriented on top of the belts

The primary operational impact of the two assumptions, taken together, is that it leaves major portions of the assault lane unaddressed. Incorrect intelligence, inaccurate placement of mines and obstacles by the enemy, or their physical displacement could create a high-risk transit for the AAVs.

H.12.1.3 Structural assumptions

First, note that the two critical assumptions above also could have been identified as structural assumptions. Other than those, the major structural assumptions are:

- P-OCM is two-dimensional. While water depth "changes," it's only with respect to determining differing displacements for differing notional depths (the vertical dimension is never represented, let alone used)
- Time steps are established only to separate detonations sufficiently to allow all displaced mines and obstacles to come to rest before the next detonation
- No modeling of the detonation plume
- Each run had three double strands of mines, one strand of Hedgehogs, and a strand of TSC
- No slope or slope effects (but handled the equivalent through assigning different water depths of mines/obstacles)
- No treatment of bottom composition
- No cratering

Variables:

- Number of bombs per target box (6, 9, 12, or 15)
- Size of the target box (12-25m)

- Aimpoint Accuracy (0-10m). Aimpoints are distributed on a uniform random basis throughout the target boxes. This parameter enables the hitpoints to vary from their aimpoints.
- Minefield Placement (5-25m). The mine and obstacle belts are assumed to have a uniform random distribution of their respective mines/obstacles. This parameter allows variations of those uniform placements.
- Minefield Intelligence (0-12.5m). This parameter permits addressing imperfect intelligence by allowing the target boxes to vary up to 12.5m off the centerline of the targeted belt.
- AAV Maneuver Accuracy (5-25m). This parameter permits examining the effects of AAV navigational errors.
- Bombing patterns. These had to be set up by the analyst. Patterns could be a straight line of equi-distant bombs, set up by defining a very narrow target box, to wide boxes with uniform random bomb distributions, to zigzag patterns set up by forcing a bomber to drop bombs in an alternating pattern to two parallel narrow target boxes.

H.12.1.4 Causal assumptions

The principal causal assumptions are:

- When a bomb hit occurs (a hitpoint is determined), any mines within the lethal area radius of the hitpoint are killed
- All other mines and obstacles move directly away from the hitpoint, with the distance moved determined by an internal look-up table developed from the Air Force data (the entering argument is the initial distance from the hitpoint)
- All mines and obstacles have time to come to rest from one detonation before the next one hits (the operational impact of this one is not at all clear, but we suspect it causes a somewhat narrower channel than would be the case if the detonations were closer together in time)
- AAVs must maneuver around any damaged or sunken AAVs and any Hedgehogs or TSC they encounter, then set a course straight to the LPP once clear (this assumption could cause AAVs to stray out of their planned transit lane; an alternative assumption could have the AAVs setting course for the next waypoint in order to regain the center of the assault lane)

H.12.1.5 Mathematic assumptions

The most important mathematic assumptions deal with assigning damage and/or displacement:

- Lethal Area Radius used to determine which mines are killed. Lethal Area Radius damage models generally tend to assign kills too generously. The conservative approach would be to use as the correct kill radius that one which is known to produce 100% kills. The conservative approach tends to underestimate kills, however, with the result (in this case) that resource estimates would become much higher.

- Use of empirical data in internal look-up tables to determine displacement. This should be a very low-risk approach, provided that the table is drawn from directly analogous experiments. If not, it's sure to yield invalid results.
- The hitpoints for precision weapons (JDAM series) were assumed to be directly on the aimpoints – no dispersion.
- The hitpoints for imprecise weapons were presumed to have a circular normal distribution about the aimpoint.
- Desired depth of burst was achieved with certainty in all cases.

H.12.2 Conceptual Model Findings

Only four things stand out within the Conceptual Model as potentially creating problems, with two of those as validity issues:

- Choice of MOE. We see the selected MOE as unnecessarily introducing too many new and difficult variables into the basic problem. The most difficult are AAV navigational errors and AAV maneuver doctrine. Both of which could have a huge impact on reported results and could suppress otherwise useful information related to the assigned analytic questions.
- Assumption of knowing the location of mines and obstacle belts. We see the assumption as highly suspect in the real world in the first place. It also injects a great deal of risk into the analysis in the second place. Finally, it enables the use of a counter obstacle tactic that is extremely risky.
- Use of a counter obstacle tactic that injects a high level of risk. Specifically, attempting to place the "target boxes" directly on the mine/obstacle belts, leaving gaps between the belts. We believe this tactic alone could be responsible for one set of counter-intuitive results (discussed later).
- Use of a Lethal Area Radius approach to determining damage to mines by bombs. Even when done carefully, such approaches will consistently under or over-estimate damage.

None of the above, however, serves to invalidate P-OCM.

H.12.3 Assess the validity of the instantiated model

The instantiated model was not available for examination or test.

H.12.4 Assess the validity of model results

Figures 33 and 34 provide evidence that something is amiss. Figure 33 indicates that in many instances as bombing precision increases, AAV survivability decreases.

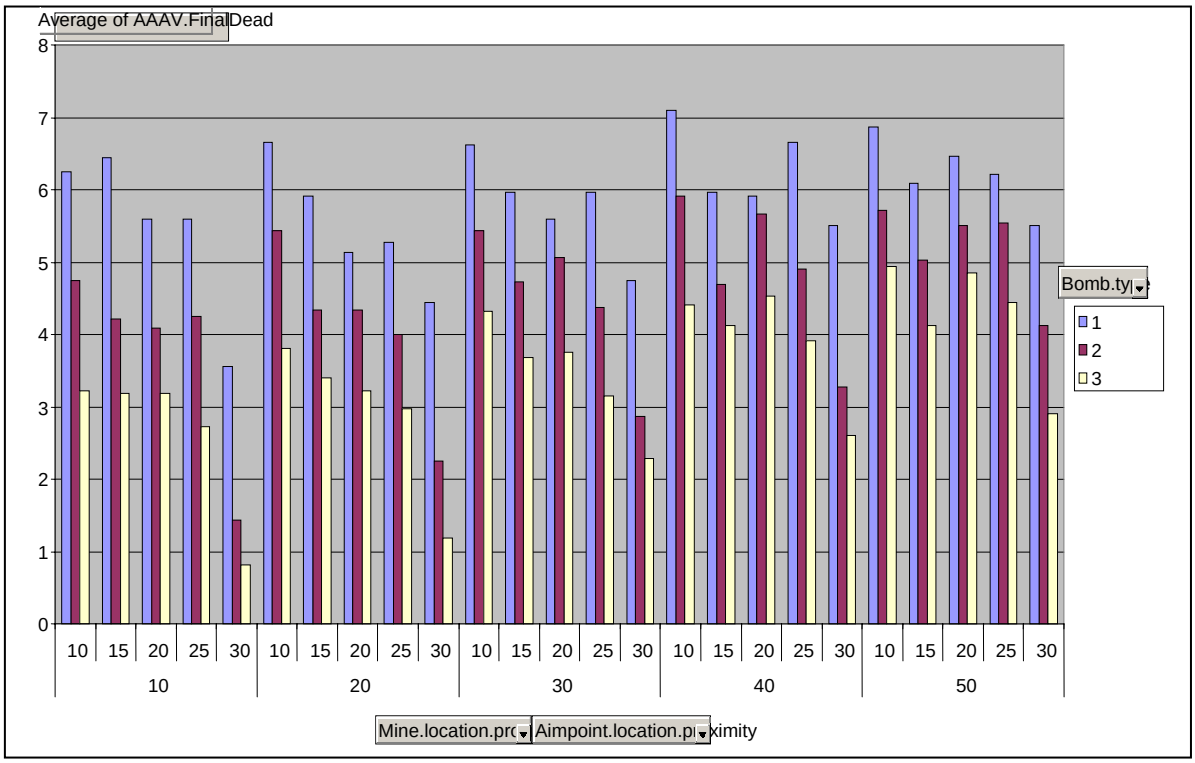


Figure 33 AAVs Killed vs. Mine Location Precision and Aimpoint Location Precision

One possible explanation goes back to the issue about how the target boxes were placed directly on the mine and obstacle belts (Figure 32). Mines could be getting displaced out of the targeted belts and into the "open" areas between belts – where no bombs are targeted. In addition, a very tight dispersion of bombs along a straight line could be displacing mines to the side, where the next bombs would have little or no effect. Conversely, a less-accurate spread could be destroying some of those displaced mines and displacing others still farther from the center of the assault lane. All that, however, is sheer conjecture. Neither the displacement data nor the AAV kill location data were captured and retained.

Figure 34 presents an even more unlikely proposition, that fewer bombs can be more effective than more bombs.

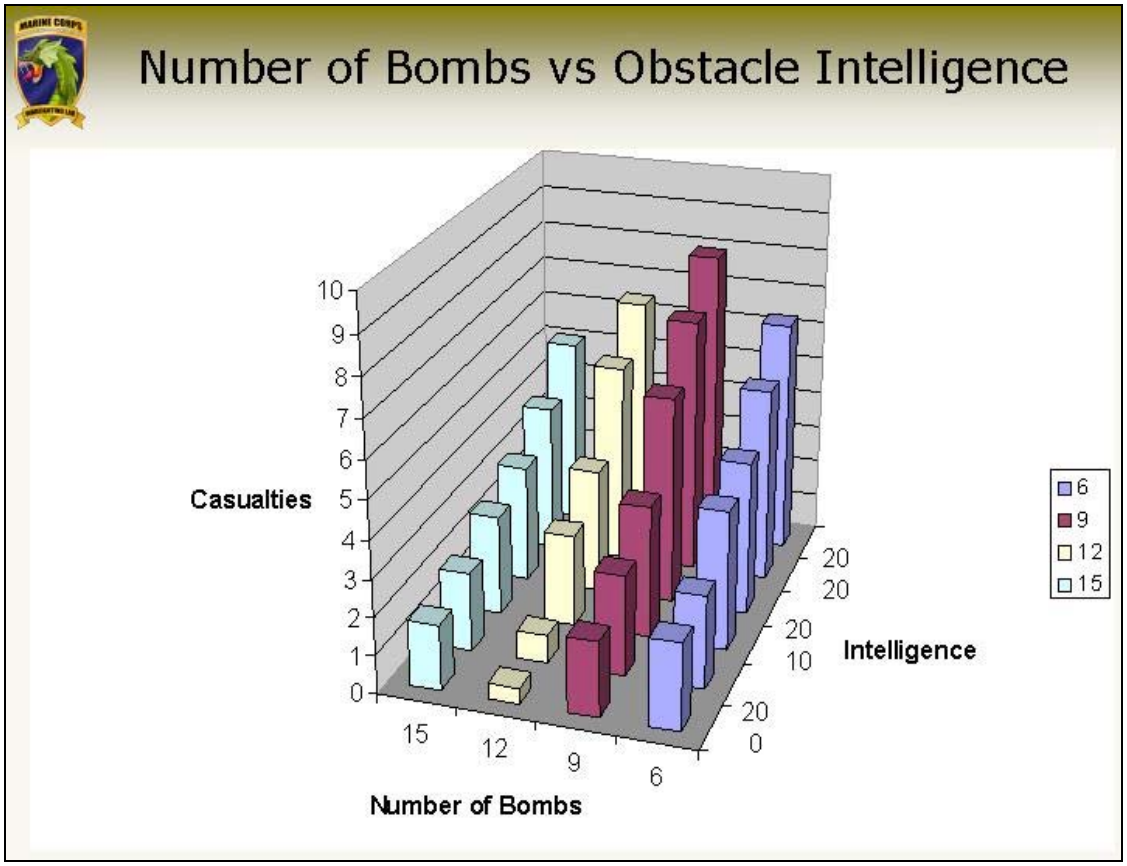


Figure 34 AAVs Killed vs. # Bombs & Intel Accuracy

In Figure 34, as intelligence accuracy decreases, AAV losses go up, as would be expected. However, it also indicates that 6 bombs are almost always better than 9 bombs, and can even be better than 12 bombs. It also says that when intel is very good, 12 bombs are better than 15. Those instances seem to go far beyond being counterintuitive and into the realm of physical impossibility. Something is wrong somewhere, and not just in the George Box sense of "wrong."

Unfortunately, nothing in the assumption testing of the Conceptual Model gives even a hint of anything that could cause the results seen in Figure 34. The problem could be with the dataset, or with the P-OCM instantiation, or with Pythagoras itself. But there is a problem, and it appears to be serious.

We believe finding and correcting the problem will require a "do over," with an eye to capturing and analyzing detailed data of all destruction and displacement of mines and obstacles, and of AAV damage and destruction. The best starting point from a cost perspective, though, would be with the data.

Until the causes of the Figure 34 results are found and corrected, and the causes of the Figure 33 results are at least satisfactorily explained (if not also corrected), P-OCM has to be considered invalid for analysis.

An application of the risk assessment tool developed during the ABSVal study reinforces that finding. The tool is shown in Figure 35 below. [11]

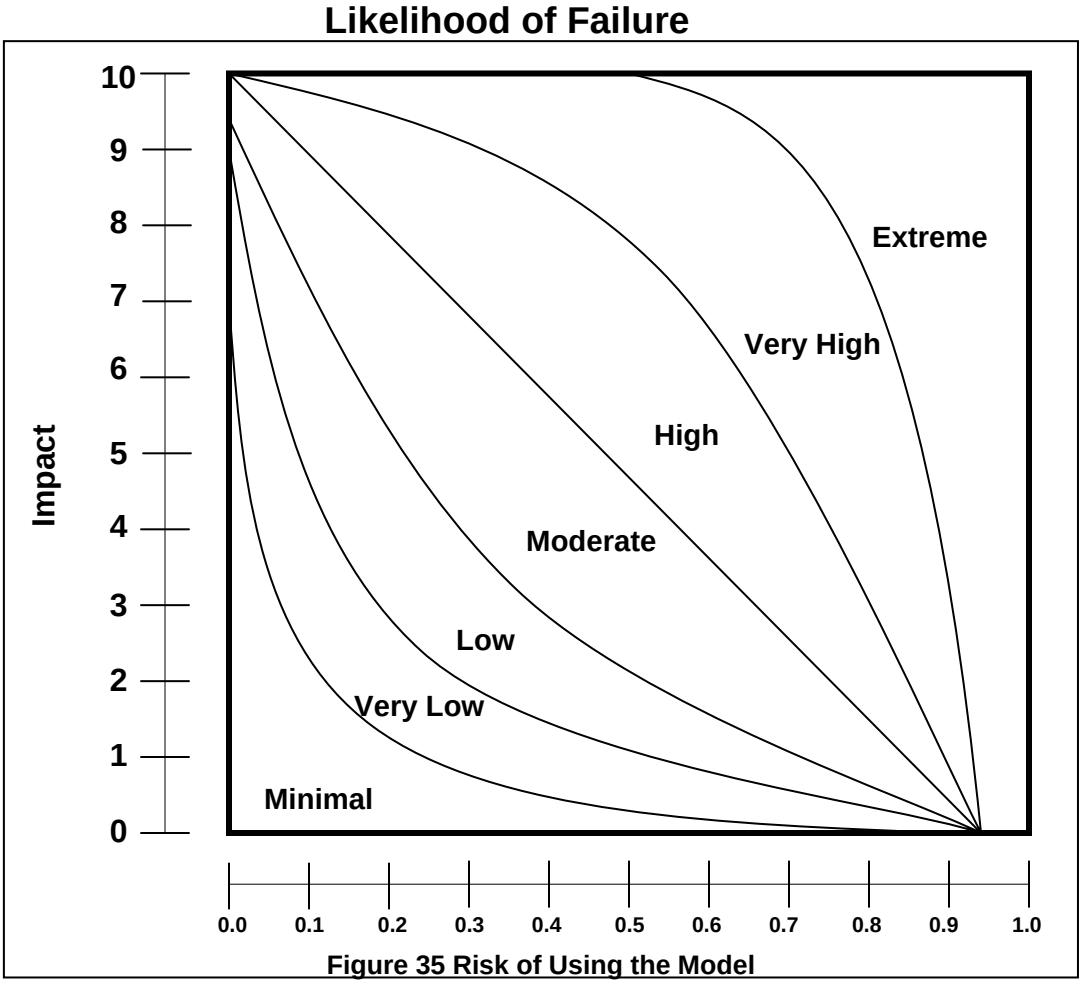


Figure 35 Risk of Using the Model

In light of the earlier discussion of Figures 33 and 34, we assessed the likelihood of failure of the model to be high, at least 0.6. The impact of the supported decision also is quite high because invalid model results could lead to the unnecessary loss of AAVs and Marines in actual assaults. We rated that impact in the 8 – 10 range. The

[11] Figure 35 remains a work-in-progress. Several of its features are notional. For example, the x-axis intercept assumes the supported decision-maker would be unwilling to use a model if he or she were told its likelihood of failure were at least 93%, *regardless* of the potential impact level of the supported decision. While we doubt many decision-makers would set the threshold any higher, some might well set it lower. We expect the risk assessment tool to be refined during post-contract work as part of academic endeavors.

combined effect was a risk assessment rating of Very High to Extreme for the P-OCM model in its current form and in this particular application.

The "Conclusions" section below indicates a significant change that could be made to the P-OCM model and the MOE in any future study of the same problem. We *believe* those changes would yield a far more favorable risk rating.

H.13 CONCLUSIONS

H.13.1 P-OCM

As indicated above, P-OCM has to be considered invalid for analytic applications due to the near-impossibility of some of its results. We also strongly recommend any future work on the problem addressed by P-OCM that the "AAVs surviving" MOE be discarded in favor of MOEs that directly measure the effectiveness of clearance TTPs on mine and obstacle density.

We note, moreover, the strong possibility that the obstacle clearance model within P-OCM may in fact be fully valid for this application (valid in the sense that it would not be falsified if allowed to operate separately). We believe the problem that showed up in results validation could have been *entirely due* to the AAV lane transit model that had been imposed on top of the obstacle clearance model. Unfortunately, the data that could have allowed examination of that possibility (the positions of destroyed and displaced mines and of displaced obstacles after each bomb detonation, and the positions of destroyed AAVs) were not collected. If we are correct, however, stripping out the AAV lane transit model and changing the MOE as indicated above could yield a very useful obstacle clearance model.

H.13.2 ABSVal Framework

As stated in the Introduction, the objective of this exercise was to test the viability and utility of the ABSVal framework in a realistic institutional setting. In that regard, we found several areas of the ABSVal framework that appear in need of modification as noted below (note that this section is very similar to portions of the companion P-COIN report):

H.13.2.1 Separate assessments for Theoretic, Mathematic, and Algorithmic Sub-Models (Sect 2.5 of the Phase I Final Report)

The three separate assessments appear to be needed only when a simulation is being developed. After development, "it is what it is," and a single assessment that looks at the theoretic underpinnings, the mathematic sub-model, and the coded algorithms holistically should be fully as effective and far more efficient. In the P-OCM test case, we looked at only the theoretic sub-model because there is no mathematic sub-model – the developers went straight to code – and the algorithmic model was not available.

H.13.2.2 Definition of Accreditation

We currently define accreditation three ways in five different places in the Phase I Final Report (Sects ES.1 (2), 1.0 (2), and 1.1.1.3). Strongly recommend promoting only the last of those, but with the addition of the Analysis Plan, i.e.:

"The official certification that a model, simulation, or federation of models and simulations and its associated data, user team, and analysis plan are acceptable for the specific application at hand."

H.13.2.3 Definition of Credibility

One problem endemic to the VV&A industry is the loose use of the term "credibility," often supplanting the term "validity." In Sect 1.5, the Phase I Final Report offers a new definition of "credible," one which, if more fully developed, could go a long way to resolving the ambiguity associated with the term "credibility" in the industry.

H.13.3 Recommendations

- Modify the ABSVal framework to permit a holistic validity assessment of the Conceptual Model (vice as many as three sequential assessments) once the simulation has been developed.
- Change the definition of accreditation throughout to match the one given above. That one adds critical dimensions to the existing DoD definition and clarifies the specificity of an accreditation decision.
- Further develop and promulgate as part of the study's Phase II Final Report a definition of "credibility" along the lines of Sect 1.5.

H.13.4 H.13.4 References

1. R. Paterson, E. Bitinas, *Modeling Obstacle Reduction with the Pythagoras Agent-Based Distillation,* Maneuver Warfare 2003
2. R. Paterson, M. McDonald, J. Eusse, T. Erlenbruch, E. Bitinas, *Shallow Water Obstacle Clearing*, briefing to 5th Project Albert International Workshop (PAIW5), Uberlingen, Germany, July 2002
3. R. Paterson, *Thoughts on 6th Project Albert International Workshop (PAIW6),* Unpublished Notes to the Surface Zone/Beach Zone Obstacle Group, March 2003
4. R. Paterson, *Surf Zone/Beach Zone (SZ/BZ) Obstacle Reduction,* Marine Corps Warfighting Laboratory briefing, undated
5. R. Paterson, *Precision Guided Munitions versus Shallow Water Obstacles: The Project Albert Contribution,* undated briefing
6. B. Almquist, *Surf Zone/Beach Zone CMCO Future Systems,* ONR brief to Mine Warfare Association, 2005

The P-OCM Validation Report can also be found at http://orsagouge.pbwiki.com/ABSVal.

APPENDIX I P-COIN CONCEPTUAL MODEL VALIDATION REPORT

I.1 INTRODUCTION

In Phase II of the ABSVal Study, the Team was to assess at least two candidate simulations useful in developing an Irregular Warfare Analytic. The objective was to test the viability and utility of the ABSVal framework in a realistic institutional setting. The first simulation selected was the Pythagoras Counter-Insurgency (COIN) model developed by NGMS for use by the Operations Analysis Division (OAD) of the Marine Corps Combat Development Command (MCCDC). Sanderling Research Corporation's role in the effort was to apply a single technique, assumption testing, within that framework to assess the validity of the Conceptual Model of OAD's specific Pythagoras COIN instantiation.

The potential usefulness of assumption testing as a validity assessment technique may be seen by considering the nature of models in general. George Box famously stated, "All models are wrong, some are useful." Box was absolutely correct in the literal sense – no model *is* reality. Rather, every model is an abstraction of reality to some extent. That aspect of a model is widely if perhaps not universally recognized. What is more rarely recognized and far more rarely appreciated is that the mechanism of abstraction is the *assumption.* Thus if we could identify every assumption used to create a given model, we would know how it deviates from reality or, in Box's terms, we would know just how "wrong" it is.

As a practical matter, however, we can not *explicitly* identify every assumption in even simple models. The good news is there is no need to identify all of them. We need only to identify, and "test," the assumptions that have significance to the *intended purpose* of the model and especially to the *analytic questions at hand.* That is, of course, more easily said than done. Part of the art, vice science, of assumption testing is to be able to recognize in at least broad terms which assumptions are likely to be significant, given only a description of the model, the context of the study, and the specific analytic questions at hand. Thus analysts generally have to cast a wider net than would be necessary if they had full knowledge going in as to which assumptions are significant. Assumptions having little or no apparent significance are set aside. Ones having apparent significance are tested as described later in this report.

I.2 ROLE OF SCIENTIFIC METHOD

The ABSVal framework approach is based in the scientific method, with the thrust being to find evidence that would reject (falsify) the null hypothesis that the model or simulation is valid for the intended purpose. Scientific method applies within assumption testing, but in an indirect fashion. A validity assessment attempts to determine whether a model or simulation is "sufficiently accurate," *vis-à-vis the real world*, for a particular application. But because assumptions represent purposeful *departures* from the real world, and sometimes quite significant departures, a direct application of scientific method – one that directly compared the assumptions to the real world -- could readily hold the model to be invalid without even considering the intended

application. (Which is why Box's famous "All models are wrong" quotation often is cited, incorrectly, as evidence that validation is a waste of time and money.) The indirect scientific method approach used in assumption testing notes the departures from reality, but then determines their *operational implications in the context of the application at hand*. The final step of the indirect approach is for the Application Sponsor to decide whether those operational implications are acceptable for his or her application.

I.3 CONCEPTUAL MODEL

In the ABSVal framework, the term "Conceptual Model" encompasses three distinct sub-models: [12]

- ***Theoretic model*** is the initial expression, usually in textual and/or graphical form, of the context of the model and of the cause-and-effect relationships believed to be operative in the situation of interest and that are intended to be incorporated within the end model. In an ABS, it contains all of the agent behaviors and relationships.
- ***Mathematic model*** captures the specific logical structures and expressions (equations, conditional statements, logic tables, etc). Note that the relationship between the theoretic model and mathematic model is one-to-many; i.e., there are numerous ways a theoretic model could be represented mathematically.
- ***Algorithmic model*** is the coded form of the mathematic model. Again, note that the relationship between the mathematic model and algorithmic model is one-to-many.

I.4 TYPES OF ASSUMPTIONS

As stated in the Final Report of Phase I of this study [Youngs and Bitinas, 2007], there are four sets of assumptions of interest: causal, structural, mathematic, and scenario:

- ***Causal assumptions*** deal with cause-and-effect relationships among agents/objects/entities and with their environment(s) and other stimuli.
- ***Structural assumptions*** deal primarily with the processing order of stimuli, decisions, and actions within a model, but also may deal with starting, ending, and boundary conditions within a model.
- ***Mathematic assumptions*** deal with the myriad assumptions made to enable constructing a determinable mathematic abstract of real-world scenarios, processes, behaviors, and events; mathematic assumptions include the choice of algorithms and other logic structures, and thus assumption testing includes an assessment of those algorithms/structures.
- ***Scenario assumptions*** deal with bounding the real-world environment (which may be behavioral as well as geophysical) to be addressed within the model, with the

[12] The ABSVal Phase I Final Report, Ref 1, only recognizes the first two sub-models, the theoretic and mathematic models, as composing the Conceptual Model. However, it also neglects assumption testing the third sub-model, the algorithmic model, which actually is the most important of the three. For that reason, the algorithmic model is addressed in this report as part of the Conceptual Model.

geophysical features and environmental conditions contained therein, and with the entities and their characteristics to be "in play" in a particular scenario.

I.5 ASSUMPTION TESTING PROCESS

Also as stated in the Phase I Final Report, assumption testing is a three-step process:

- ***Step 1: Identify the assumptions.*** Assumptions, particularly causal and mathematic assumptions, are rarely if ever well-documented and may even have to be reverse-engineered from the source code. In some cases, even some of the algorithms may not be documented. This is *by far* the most difficult aspect of assumption testing.
- ***Step 2: Determine the operational implications of the assumptions.*** Accomplished as a cooperative effort between M&S and operational subject matter experts (SMEs).
- ***Step 3: Determine the acceptability of the identified operational implications to the decision-maker.*** During M&S system development, the decision-maker is the M&S sponsor. For a particular application of the M&S system, the decision-maker is the application sponsor (designated by DoD policy as the accreditation authority for that particular application).

I.6 ASSUMPTION TESTING APPLIED TO PYTHAGORAS COIN

I.6.1 Precepts [13]

I.6.1.1 Modern Scientific Method

Throughout, the planned process is based on modern scientific method and most specifically on the *falsifiability criterion* contained therein.

As a convention for this assessment, the *null hypothesis* is the *research hypothesis* that the model being assessed *is valid ("sufficiently accurate") for its specific intended application*. The *alternative hypothesis* then is defined as the negation of the null; i.e., the model *is not sufficiently accurate for that particular application.* We then attempt to *falsify the null hypothesis.*

Every step of the planned process (except for writing the end-game report) is intended either to set the stage for falsification of the null or to execute falsification tests of the null.

A failure to falsify the null does not mean the model is *proven* valid, but it should greatly increase confidence in the model's validity for that particular application. The degree of confidence depends on the rigor and power of the tests applied.

[13] As stated earlier, for a detailed discussion of scientific method, the falsifiability criterion, and research, null, and alternative hypotheses, see Section 4.3 of the ABSVal Phase I Final Report, Ref 1.

I.6.2 Plan

I.6.2.1 Identify Analytic Questions

Identify the *analytic questions* the model is/was intended to address in the specific application at hand, the metrics applicable to those questions, and the degree to which model results are/were expected to shape the decisions to be made.

- Detailed review of all available documentation of the application
- Interview the Application Sponsor

I.6.2.2 Identify the Referent

In collaboration with the Application Sponsor, identify the *referent*; i.e., the proxy for the real world for the purpose of accuracy comparisons. See Sect 4.0 of the Phase I Final Report for the various forms such a proxy may take.

I.6.2.3 Identify the Accreditation Criteria

In collaboration with the Application Sponsor, identify the *accreditation criteria.*

- Establish just how accurate the end results of the model have to be when used in the particular application at hand
- Must include how "accuracy" will be determined, and may have both qualitative and quantitative aspects
- Criteria must be "pass/fail;" i.e., they must establish the *lower bounds of accuracy* that must be met for the model to be acceptable for the application at hand

I.6.2.4 Assess the Validity of the Referent

- Confirm that no alternative referent is available or could reasonably be constructed that would be preferable to the one identified (e.g., is an empirically-derived database available? Could one be made available?)
- Assumption testing:
    - Identify/derive the *assumptions* that are inherent to/embedded in the referent
    - Perform logical verification -- determining whether the referent as written *adequately and correctly implements* underlying theory and assumptions
    - Determine the *operational implications* of the identified assumptions in the context of the particular application and with respect to the remaining steps of the ABS Val framework
    - Determine *bounds of validity* imposed on the application's problem space and on the *model's* validity assessment by the referent's assumptions
    - Determine whether the operational implications and bounds of validity are acceptable to the Application Sponsor
- Independent SME review(s) (ideally, these will be contrarian reviews from SMEs that would focus on any potentially falsifying aspects the referent)

I.6.2.5 Determine remaining workplan

Determine the most efficient sequencing of the remaining steps of the ABS Val framework, based on:

- Information developed to this point
- Estimates of the difficulty and costs of the individual remaining steps
- The relative power of each of those steps to falsify the null

I.6.2.6 Assess the Validity of the Conceptual Model

- Potentially as many as three separate assessments; in sequence:
  - Theoretic sub-model
  - Mathematic sub-model (if it exists)
  - Algorithmic sub-model
- Each in turn will have the same assessment techniques applied to it:
  - Logical verification -- determining whether the sub-model as written is an *adequate and correct implementation* of its predecessor (free of logical, mathematic, or algorithmic error)
  - Assumption testing:
    - Identify/derive the *assumptions* that are inherent to/embedded in the sub-model
    - Determine the *operational implications* of the identified assumptions in the context of the particular application
    - Determine the *bounds of validity* of the model that are the result of the identified assumptions
    - Determine whether the operational implications and bounds of validity are acceptable to the Application Sponsor for the intended application
  - For some models, it may prove necessary to reverse-engineer one or more sub-models from later models. It may even be necessary to reverse-engineer the Conceptual Model, or portions of it, from source code.
  - The referent serves as the predecessor for the theoretic sub-model

I.6.2.7 Assess the validity of the instantiated model

- Logical verification -- determining whether the instantiated model as coded is an *adequate and correct implementation* of its predecessor (the algorithmic sub-model)
- Data validation
  - Source
  - Data element definitions
  - Data values

I.6.2.8 Assess the validity of model results

- Comparison of model results to the referent
- Must address each accreditation criterion

I.6.2.9 Develop final validation assessment report

- Specific addressal of each accreditation criterion
- Incorporates, at least by reference, the dataset used to generate "results"
- Make accreditation recommendation
- All in the context of the specific application at hand
- This last set of steps assumes the validity assessment process goes all the way through assessing the validity of model results. If at any earlier point the null hypothesis is falsified, the process may be cut short and the report written to reflect the findings at that point.

For the Pythagoras Obstacle Clearance Model, SRC applied only Steps I6.3.1 through I6.3.4, I.6.3.6, and a brief variant of I.6.3.8 above. For all the other models, the process *may* be tailored to each individual model/application pair. In general, however, we expect each step to apply.

I.7 RESEARCH

I.7.1 IW Project Documentation

LT Robin Marling, USN, the Study Project Officer, and Mr. Edmund Bitinas, NGMS Study Leader and Pythagoras Developer, supplied all the project documentation used in the assumption testing portion of the Pythagoras COIN validity assessment. The documentation included:

- "*Irregular Warfare Project,*" NGMS Status Brief to MCCDC OAD, Fairfax, VA, Jan 2008
- "*USMC Irregular Warfare Project – Working Paper,*" NGMS, Fairfax, VA, 16 Aug 2007
- "*Insurgency Assessment Guide*," MCCDC OAD, Quantico, VA, 05 Jan 2007
- "*A Semantic Differential Approach to Incorporating Qualitative Data into Models*," article submitted for publication, LT Robin Marling, USN, Quantico, VA (undated)
- "*The Agent Based Simulation Verification, Validation, and Accreditation (VV&A) Framework Study*," Attachment A to NGMS Proposal No. R1-1H879-106.510, 23 Jan 2007
- "*Pythagoras User's Group Update*," Maj Jon Ault, USA, TRAC Monterey; LT Robin Marling, USN, MCCDC OAD; Mr. Edmund Bitinas, NGMS, (undated)
- "*The Semantic Differential and Attitude Research*," David R. Heise in *Attitude Measurement*, Gene F. Summers, ed., Chicago: Rand McNally, 1970, pp. 235-253
- "*Pythagoras Manual, Version 1.10.2*," NGMS, Fairfax, VA (undated)

I.7.2 Interviews

The interviewees were:

I.7.2.1 Study Sponsor

Dr. George Akst, MCCDC Senior Analyst:

- Study objective was "to make headway in developing a Counter-Insurgency model."
- Did not dictate the type of model to be used in the study, but approved the OAD recommendation to use an ABS, Pythagoras.
- Directed that the study use a real-world dataset, although it could begin with a notional "Country X" dataset.
- Did not have a specific analytic question in mind at the outset, but approved OAD's recommendation to address an important real-world analytic question: in a Disaster Relief/Humanitarian Assistance mission, is it better to base the MAGTF ashore or afloat?
- Approved the broad scenario details and three specific MAGTF missions
    - Area of interest would be the coastal Buenaventura region of Columbia, which struggles with the FARC insurgency and a large and growing drug trade
    - The external stressor would be a natural disaster, a tsunami that devastates the coast and displaces many of its residents.
    - Specific missions:
        - Provide refugee camp security
        - Provide Humanitarian Assistance
        - Provide Disaster Relief
    - Using as the Measure of Effectiveness the shift of members of various population elements among five possible affiliation groups:[14]
        - FARC
        - Pro-FARC
        - Neutral
        - Pro-Government of Columbia (Pro-GoC)
        - GoC
- Separately approved the "implied mission" resulting from the Conceptual Model itself:
    - Prevent the FARC and pro-FARC elements from gaining strength as a result of the tsunami and its aftermath (including the introduction of the MAGTF) [15]
- Had not established a referent or accreditation criteria because meeting his objective required neither.

[14] Note the selected MOE has no apparent direct relationship to any of the three specified missions. Conversely, nothing currently in the study attempts to evaluate the MAGTF's effectiveness in carrying out its specified missions.

[15] This implied mission alternatively was defined by study participants as building the strength of the Pro-GoC group. Note that the two alternative expressions are not equivalent because they would treat the Neutral group differently.

I.7.2.2 Study Project Officer

LT Robin Marling, USN, MCCDC OAD: [16]

- Pythagoras had been selected for the study by her predecessor, Mr. Steve Stephens, who subsequently deployed to Iraq.
- Directed the development of the "cultureware" for the study, the compendium of cultural and historical information for the area of interest that is required by Pythagoras for this particular type of study.
- With the help of four foreign area (Columbia) SMEs, divided the population of Buenaventura into eight segments:
    - Catholic Church officials
    - Displaced persons
    - Illicit organizations
    - Military
    - "Old Money" (landed gentry)
    - Police
    - Urban Middle Class
    - Urban poor
- Developed for each population segment a square (5X5) Markovian "susceptibility matrix" in which cell X_{ij} represents the probability that persons in affiliation group *i* will change to affiliation group *j* in the next time step. The susceptibility matrix probabilities were fixed under the assumption that they represented "natural inclinations" that would remain unchanged absent external influences, and would return to their natural state if an external influence were removed.
- Identified the population of Buenaventura as the Center of Gravity of the scenario and recommended as the Measure of Effectiveness the shifts in affiliations of members of the population.
- Identified Heise's "semantic differential" as a potentially useful concept to apply to the COIN problem and directed the research to implement it.

I.7.2.3 NGMS Study Leader

Mr. Edmund Bitinas, NGMS Chief Scientist, and Pythagoras Developer:

- Primary function and challenge was to instantiate an extremely complex theoretic model into Pythagoras
- When theoretic constructs could not be directly represented in Pythagoras (e.g., Markovian matrices), had to develop algorithmic alternatives

I.8 IW PROJECT OBJECTIVES

As indicated above, there are three different objectives being pursued simultaneously within the USMC Irregular Warfare (IW) Project.

[16] Information presented is derived from both the interview and subsequent telephone conversations and from LT Marling's paper, "*A Semantic Differential Approach to Incorporating Qualitative Data into Models*," Ref 2.

- "To make headway in developing a Counter-Insurgency model." (MCCDC Study Sponsor)
- "Can Pythagoras model the Buenaventura Disaster Relief/Humanitarian Assistance scenario?" (NGMS)
- "In the Buenaventura Disaster Relief/Humanitarian Assistance scenario, is it better to base the MAGTF ashore or afloat?" (MCCDC OAD)

I.9 SIGNIFICANT ASSUMPTIONS

One significant assumption affected the conduct of the validation assessment rather than the COIN study – that Pythagoras itself is error-free and thus did not need to be subjected to scrutiny. Because the purpose of this assessment was to test the viability and utility of ABSVal, assuming Pythagoras error-free was a matter of indifference. Had the purpose been to rigorously assess the Pythagoras COIN model for real-world use, however, the assumption would have been inappropriate and potentially counterproductive.

I.9.1 Scenario assumptions.

- The world beyond the immediate Buenaventura area is assumed away. While it could be argued that the geopolitical situation and developments in the whole of Columbia would in fact impact the attitudes and behaviors of the Buenaventura population, we find the assumption reasonable for the specific application at hand (i.e., the COIN study).
- The Buenaventura population could be divided into eight individually homogeneous segments as named earlier. We think it would be pointless for the application at hand to argue for more or fewer segments.
- As mentioned earlier, there are five possible affiliations for the population:
    - FARC
    - Pro-FARC
    - Neutral
    - Pro-Government of Columbia (Pro-GoC)
    - GoC
- The stressor occasioning the need for a MAGTF is a tsunami, also a very reasonable assumption.
- The only scenario "action" taken by any of the primary actors (The Gov't of Columbia (GoC), the insurgency, and the MAGTF) would be the insertion of the MAGTF, and the only variable in that regard would be whether the MAGTF performed its missions from the seabase or from a shore base. This is a very strong assumption, particularly with regard to the one analytic question of the study – is it better to base the MAGTF ashore or afloat? At the very least, contingency analyses should be conducted to determine whether alternative or additive action assumptions should be used (e.g., the insurgency takes actions intended to discredit the GoC by making the Marines appear as an enemy of the people).

I.9.2 Structural assumptions

- By far the most important assumption, and the one that shaped the entire study, is that in warfare the focus of modeling efforts should be on the Center of Gravity (COG).[17] That assumption is no longer revolutionary [e.g., Falzon and Priest 2004, Hetherington 2005] but it is still rare.
- A companion assumption, and one of nearly equal importance to the study, is that the civilian population of the Buenaventura region is the COG.[18] While the civilian population is not always the COG, not even in Irregular Warfare (consider an insurgency against an occupation force), we find the argument compelling in this case.
- The third of the most important assumptions is that the sole Measure of Effectiveness (MOE) is allegiance changes by members of the population segments among the five affiliation groups. The most immediate issue regarding this assumption is that it effectively dismisses the MAGTF's assigned missions (refugee camp security, humanitarian assistance, disaster relief) in favor of an implied mission (preventing the FARC and pro-FARC elements from gaining strength).
- Another extremely strong assumption is that the susceptibility of individuals in any of the population segments to move from one affiliation group to another can be represented by a Markovian matrix. The problem with that assumption is that a Markov process is memoryless. An individual's past affiliation history does not matter to his/her next affiliation decision, nor do any past events. The assumption is so "at odds" with human behavior that it has to be challenged.
- Compounding the above is a companion assumption that the transition probabilities in each Markovian matrix remain constant up to the point that the MAGTF is inserted, do a step-change at that point, and then remain constant at their new values until the Marines are redeployed. Once the Marines are gone, the transition probabilities return to their original values. Again, the assumption virtually demands a challenge. Opposing forces are both reactive and adaptive (and may try to be proactive). If something is working, they try to leverage it. If it's not working, they change. Reactions and adaptations could be expected to vary – possibly significantly – from time step to time step depending on perceived "success" up to that point, and the two COAs (afloat or ashore) could be expected to offer differing opportunities to both sides.

I.9.3 Causal assumptions

- The behavior of each population segment (its proclivity to support the GoC or the insurgency) is determined by three factors:
    - Natural tendency (established by the cultural SMEs)
    - Impact of current events (the segment's reaction to the introduction of the MAGTF, also established by SMEs)

[17] Marling, *op cit*
[18] *ibid*

  - Influence of other population segments (established by an application of Osgood's Semantic Differential [Heise, 1970]. It is this third assumed factor that may represent a real breakthrough in IW analysis, or may instead only inject a confusion factor. This is explained in greater detail immediately below.

- As stated by the Study Project Officer,[19] "Semantic Differential posits that human beings make judgments based on three factors: Evaluative, Potential, and Activity. Evaluative is a value judgment: Is object A good or bad? Potential is a potency judgment: is object A weak or strong? Activity is an action judgment: is object A active or inactive? … Semantic Differential theory also posits that words have meanings in addition to what they physically represent. It implies that word selection by social scientists is significant and conveys information. … Researchers have conducted several surveys of thousands of college students from multiple institutions about the Evaluation, Potential, and Activity values of thousands of nouns, actions, adjectives, and modifiers on a Likert scale of (-5,5)." Semantic Differential theory has proven its value in numerous marketing studies. Its use in the COIN study, however, appears to represent an entirely new application, and has some validity issues:
  - The COIN application was to attempt to understand in a quantitative way how each population felt about each other segment in order to determine how each was influenced by the others.
  - The cultural SMEs used single words to describe the feelings of each segment for the others, and the words were "scored" along E-P-A axes.
  - Both the SMEs and all the students were Americans in order to avoid or at least mitigate possible semantic mismatches. However, unavoidably, the *context* was certainly different. The students made their Likert scale responses in the context of American society and, within that, university society. The SMEs words were cast in the context of not only Columbian society, but a subset of that society that existed in the best of times in the middle of a violent insurgency and a burgeoning drug trade. At the minimum, some experiments need to be designed and conducted to show that the students' E-P-A responses are valid in the Buenaventura context. The likely case is that they are not.
  - Even absent the above issue, the SMEs themselves reportedly were concerned that the E-P-A methodology lost context. [20]
  - [Note: the following bullets also could be placed with Mathematic Assumptions, but are kept here to maintain the logical flow.] LT Marling indicated that OAD wanted to devise a single numerical value to represent all three factors. The method chosen is a variation of the "Polarization" parameter, *P*, described by Heise: [21]

$$P = \mathrm{SQRT}(e^{**2} + p^{**2} + a^{**2})$$

[19] *ibid*
[20] *ibid*
[21] David R. Heise, "*The Semantic Differential and Attitude Research*," in *Attitude Measurement*,

The variation is (using the same notation as above):

Influence = *e**(SQRT(*p**2* + *a**2*)),

but with a data transformation for *p* and *a* from the (-5,5) Likert scale to a (0,10) scale. While the result is given the name "Influence," and can be positive or negative depending on the sign of the Evaluation factor, it's by no means clear what it actually represents. For one thing, it equal-values the Potential and Activity factors, which is an assumption that both Osgood and Heise appear to have avoided. Perhaps more importantly, it suppresses the impact of negative Likert-scale responses, and accentuates the impact of positive responses. At the least, the transformation needs theoretical justification before it can lay claim to being a valid metric.

- The Influence parameter results then were accorded a "binning technique" that used the averages of negative values and of positive values to determine their influence on the Markovian transition probabilities. The appropriateness of using averages can't be determined from the theoretic sub-model. It will have to wait for an examination of the algorithmic model and the live data. In general, however, averaging destroys what can be critical information regarding modes and variances.

I.9.4 Mathematic assumptions

- Each population segment is represented by 100 agents, with each agent representing 1% of the population in that segment.
- The Pythagoras "playspace" for the COIN study is 1000X1000 pixels.
- Each population segment (its 100 agents) is uniformly distributed throughout the playspace.
- Each agent has 1000 "parts" spread among the five affiliations. At game start, all agents within a population segment have their 1000 parts spread the same among the five affiliations. They end up different only because of a "proximity" factor that determines the % of a population segment that will be affected by the presence of the MAGTF, and that varies across the population segments. The proximity factor remains constant across time steps as long as the MAGTF is present.
- We see no problems or issues with any of the above assumptions, but note that the last one appears to have been rendered unnecessary by other assumptions in this particular theoretic sub-model. Specifically, we believe sensitivity analyses would show that its effect is completely dominated by the assumptions of a Markovian transition matrix and constant transition probabilities, to the extent that it can make no difference to the "ashore/afloat" recommendation (I.e., it would affect the absolute end-state values of the affiliation groups, but not their relative values).

I.10 FINDINGS AND CONCLUSIONS

I.10.1 Multiple Study Objectives

The three study objectives are not inconsistent, but neither are they congruent. That sets up the need for separate findings with respect to the validity of the theoretic sub-model for each of the three objectives, repeated below:

- "To make headway in developing a Counter-Insurgency model." (MCCDC Study Sponsor)
- "Can Pythagoras model the Buenaventura Disaster Relief/Humanitarian Assistance scenario?" (NGMS)
- "In the Buenaventura Disaster Relief/Humanitarian Assistance scenario, is it better to base the MAGTF ashore or afloat?" (MCCDC OAD)

I.10.2 Validity of Pythagoras COIN Model for making headway in developing a Counter-Insurgency model

The most important assumptions relative to this objective are:

- Focus on the Center of Gravity (COG) of the scenario
- Identification of the population as that COG
- Adoption of the concept of using allegiance changes as the sole MOE
- Assuming that Osgood's Semantic Differential, a predominantly market-research technique, could be applied to a COIN scenario.

The only one of the above assumptions that has a validity problem with respect to this particular objective is that allegiance changes are the sole MOE. As noted earlier, that assumption ignores the actual assigned missions of the MAGTF in favor of an implied mission.

We note, however, that this is only a caveat with regard to validity for the stated objective. It means that even after it has matured (see Sect 6.5.4), the Pythagoras COIN model should not be used alone to address analytic questions dealing with anything other than the allegiance changes.

The bottom line for the objective of "making headway" is that, with the one caveat noted, the theoretic sub-model of Pythagoras COIN appears fully valid. We actually believe that, more than just making headway, it could represent a real breakthrough in conflict modeling in general.

I.10.3 Validity of Pythagoras COIN Model for modeling the Buenaventura Disaster Relief/Humanitarian Assistance scenario

The validity of the Pythagoras COIN Model for the scenario at hand appears to have been designed into the theoretic sub-model from the start. The clearest evidence of

that is the choice of affiliation changes as the sole MOE. As discussed earlier, that choice ignored the actual assigned missions of the MAGTF in favor of an implied mission. It also meant that the scenario and the analysis had become tightly circumscribed by the capabilities of the tool. While that is nearly always a bad mistake in addressing a real-world analytic question, we believe it was fully justified by the overriding objective of the Study Sponsor, as discussed above. It also, however, creates a validity issue for the third study objective as discussed below.

I.10.4 Validity of Pythagoras COIN Model for Answering the "Afloat vs. Ashore" Question

As it stands today, the theoretic sub-model of Pythagoras COIN has serious issues with respect to its validity for addressing the afloat/ashore question:

- Use of only a single MOE, with that one MOE addressing an implied MAGTF mission while ignoring the three assigned missions, all of which would undoubtedly be affected by the MAGTF basing decision.
- Use of the memoryless Markovian process model for a scenario in which human memory of past affiliations and events would be expected to play a significant role.
- Use of constant transition probabilities, thus effectively ignoring the inevitable react/adapt cycles of opposing forces. [22]
- Use of a suspect algorithm to force single values from the three-dimensional E-P-A metric of semantic differential.
- Use of semantic differential word-scoring data developed in one context with words selected for a radically different context, and without experimental justification.
- Loss of context in general with the use of the semantic differential.

As a result, we have to say the validity of Pythagoras COIN to answer "ashore versus afloat" is highly suspect at this point in time. None of the above issues, however, *proves* it to be invalid for that particular purpose. Thus, strictly speaking, our assessment of the theoretic sub-model has failed to falsify the null hypothesis that it is valid.

That above fact of "failure to falsify," however, is at best a very weak endorsement of the validity of the system to answer the ashore versus afloat question. We believe the entire question of validity in this case is best thought of as a maturation issue. More research is needed into each of the areas noted above. It may be that the research will provide solid justification for some aspects, or cause some or all to change. Additional

[22] While needing confirmation through experimentation, it appears that the combined effect of these last two assumptions is to turn the Pythagoras COIN simulation into a deterministic system, wherein the final results at equilibrium are absolutely and predictively determined by the cell values of the Markovian transition matrix. If confirmed, it would mean not only that there would be no emergent behaviors, but that it would not even be necessary to run the simulation to know the results.

research could be particularly valuable in the case of semantic differential, which we see as potentially the most important analytic tool emerging from this Pythagoras COIN study. For the present, though, we have to advise against reporting out any results from the model as actionable in any sense.

The most important determinant of the above recommendation is that the P-COIN model addresses *only* the mission implied by the selected MOE and effectively ignores the three missions specifically assigned to the MAGTF. *If* (and it's a *very big* "If") the MAGTF commander were to state that he is indifferent to shore versus sea-basing from the perspective of his primary missions, then the P-COIN model could reasonably be *considered* for use to inform the ashore/afloat decision. In that case, the following risk assessment tool, developed as part of the ABSVal framework, would apply: [23]

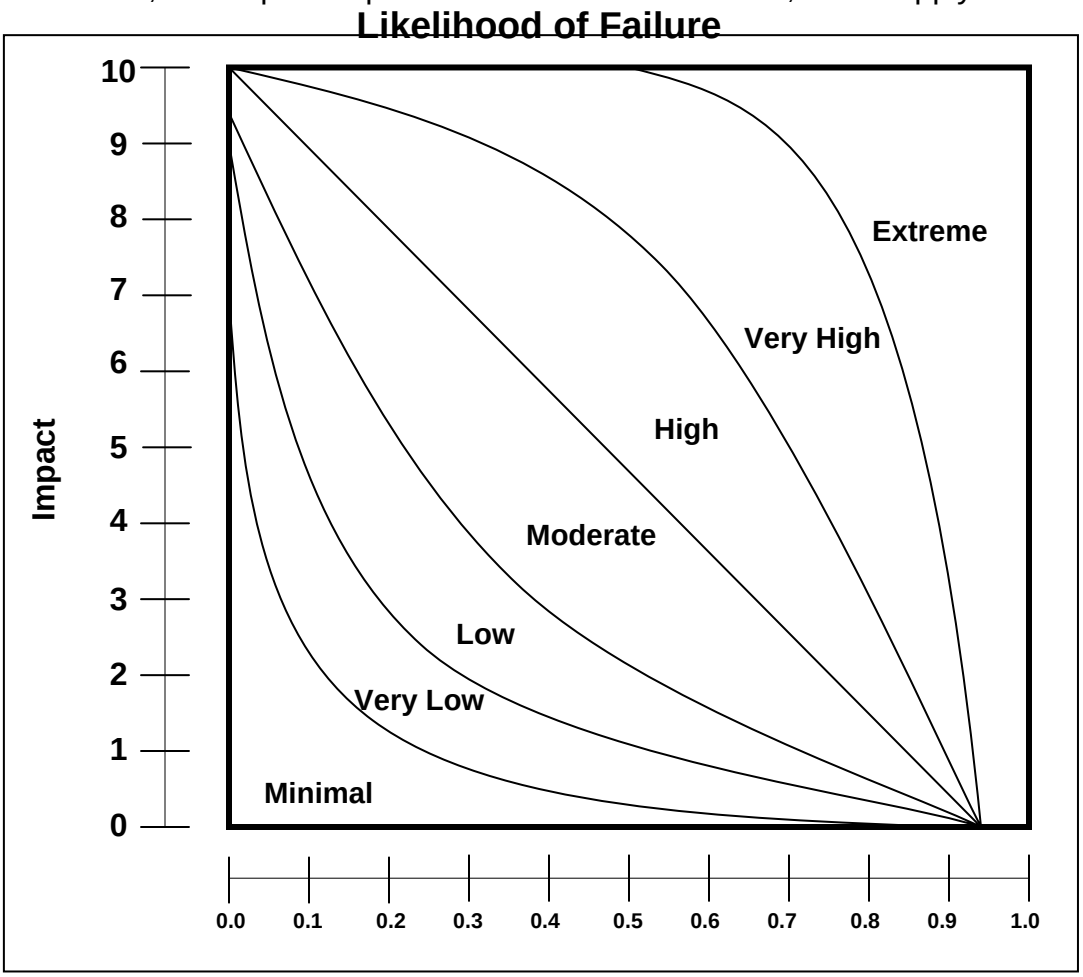


Figure 36 Risk of Using the Model

[23] Figure 36 remains a work-in-progress. Several of its features are notional. For example, the x-axis intercept assumes the supported decision-maker would be unwilling to use a model if he or she were told its likelihood of failure were at least 93%, *regardless* of the potential impact level of the supported decision. While we doubt many decision-makers would set the threshold any higher, some might well set it lower. We expect the risk assessment tool to be refined during post-contract work as part of academic endeavors.

In the P-COIN case, the issues listed at the start of this section lead us to assess the "Likelihood of Failure" as better than the toss of a fair coin, but perhaps not much better, say 0.35 – 0.45. However, the assumption that got us to this point is that the MAGTF commander is at the point of "afloat/ashore" indifference with respect to his primary mission areas and thus is willing to consider the implied mission area addressed by the P-COIN model as his principal decision factor. But buying into that implied mission means buying into its principal assumptions as well – i.e., that the civilian population of the Buenaventura region is the COG of the scenario, and affiliation change (Pro-FARC versus Pro-GOC) is the MOE. Because of that, "being wrong," i.e., recommending ashore when afloat is the right answer, or vice versa, portends failing in the implied mission. We assumed the MAGTF commander would feel that such a mission failure to be a serious impact, which we rated in the 7 – 10 range. The combined result then would place use of the P-COIN model in the High to Very High risk range.

Conversely, had we assumed that the MAGTF commander did not really care about the success or failure of an *implied* mission, then we would have rated its impact in the 0 – 2 range. That then would have yielded a risk assessment rating of Minimal to Low for the use of the P-COIN model.

The large difference between the above two results demonstrates the importance of the supported decision-maker and his or her perception of the importance of the supported decision with respect to assessing the risk of use of *any* model.

I.10.5 ABSVal Framework

As stated in the Introduction, the objective of this exercise was to test the viability and utility of the ABSVal framework in a realistic institutional setting. In that regard, we found several areas of the ABSVal framework that appear in need of modification as noted below:

I.10.5.1 Separate assessments for Theoretic, Mathematic, and Algorithmic Sub-Models (Sect 2.5 of the Phase I Final Report)

The three separate assessments appear to be needed only when a simulation is being developed. After development, "it is what it is," and a single assessment that looks at the theoretic underpinnings, the mathematic sub-model, and the coded algorithms holistically should be fully as effective and far more efficient. In the Pythagoras COIN test case, we looked first at the theoretic sub-model and, at this point in time, have not started on the algorithmic sub-model (there is no mathematic sub-model – the developers went straight to code). But from conversations with the second ABSVal team, which has looked at the coded model, it appears to have effectively changed the theoretic sub-model in implementation. If so, some unknown portion of this current assessment may itself be invalid. That problem would have been avoided by taking a holistic view of the full Conceptual Model (some programming work remains to be done, but the model appears to be sufficiently complete to enable such a view for the purposes of assumption testing).

I.10.5.2 Assumption of Theoretic Sub-Model validity

The most vexing problem throughout the ABSVal study has been how to define a referent for validity comparisons. The Phase I Final Report discusses that problem at length. Sect 2.5.1 implies that the theoretic model may simply have to be accepted as the primary referent (i.e., without having been subjected to a validity assessment itself). Our prior experience in conducting validity assessments indicates that the theoretic model can be hopelessly invalid for a specific application even when strictly physical phenomena are being modeled. The ABS modeling environment makes theoretic model validity issues more likely, not less, thus potentially placing a study and its sponsor at risk of producing invalid results if the theoretic model is not itself assessed. The results of this current validity assessment effort certainly support *never* simply assuming the validity of the theoretic model.

I.10.5.3 Definition of an ABS

At several points in the Phase I Final Report, we indicate that emergent behavior is a characteristic of an ABS (Sects 4.1, 4.2, C.2.1, C.2.5, G1.1, G.2.2.1, H). Pythagoras is undeniably an ABS simulation toolset. Yet in Pythagoras COIN, we appear to have created a deterministic ABS – and one exhibiting no emergent behavior. We need to resolve the conflict.

I.10.5.4 6.5.5.4 Definition of Accreditation

We currently define accreditation three ways in five different places in the Phase I Final Report (Sects ES.1 (2), 1.0 (2), and 1.1.1.3). We need to choose one and make the others match.

I.10.5.5 Definition of Credibility

One problem endemic to the VV&A industry is the loose use of the term “credibility,” often supplanting the term “validity.” In Sect 1.5, the Phase I Final Report offers a new definition of “credible,” one which, if more fully developed, could go a long way to resolving the ambiguity associated with the term “credibility” in the industry.

I.11 RECOMMENDATIONS

- Modify the ABSVal framework to permit a holistic validity assessment of the Conceptual Model (vice as many as three sequential assessments) once the simulation has been developed.
- Modify the ABSVal framework to eschew, or at least heavily caveat, ever simply assuming the theoretic model is itself valid.
- Change the definition of an ABS to delete emergent behavior as a required or an innate characteristic for a model to qualify as an ABS.
- Change the definition of accreditation throughout to match the one given in Sect 1.1.1.3. That one adds a critical dimension to the existing DoD definition and clarifies the specificity of an accreditation decision.
- Further develop and promulgate as part of the study’s Phase II Final Report a definition of “credibility” along the lines of Sect 1.5.

I.12 REFERENCES

1. Mitch Youngs and Edmund Bitinas, “The Agent-Based Simulation Verification, Validation, and Accreditation (VV&A) Framework Study,” Phase I Final Report, Contract #M00264-06-D-0001, Delivery Order #005, Northrop Grumman Mission Systems, Fairfax, VA, 2007
2. LT Robin Marling, USN, “*A Semantic Differential Approach to Incorporating Qualitative Data into Models*,” article submitted for publication, Quantico, VA (undated)
3. Maj Jon Ault, USA, TRAC Monterey; LT Robin Marling, USN, MCCDC OAD; Mr. Edmund Bitinas, NGMS, “*Pythagoras User’s Group Update*,” (undated)
4. David R. Heise, “*The Semantic Differential and Attitude Research*,” in *Attitude Measurement*, Gene F. Summers, ed., Chicago: Rand McNally, 1970, pp. 235-253
5. Capt Cheryl L. Hetherington, USAF, “*Modeling Transnational Terrorists’ Center of Gravity: An Elements of Influence Approach*,” Master’s Thesis, Air Force Institute of Technology, Dayton, OH, 2005
6. Lucia Falzon and Jayson Priest, “*The Centre of Gravity Network Effects Tool: Probabilistic Modeling for Operational Planning*,” Defense Science and Technology Organization, Australian Department of Defense, Edinburgh South Australia, Australia, 2004

The P-COIN Conceptual Model Validation Report can be found at http://orsagouge.pbwiki.com/ABSVal.

APPENDIX K ABSVAL APPLICATION AUDIT REPORT

The following pages contain the ABSVal Application Audit Report.

SPA Serial 50222-08

ABSVal Application Audit Report

Agent Based Simulation (ABS) Verification and

Validation Phase II Study

31 July 2008

Prepared for:

Northrop Grumman Space and Mission Systems Corporation

Under Subcontract: GVVAIISPA07, P.O. Number: 7500028231

Prepared by:

2001 North Beauregard Street, Suite 100

Alexandria Virginia 22311

1. INTRODUCTION AND BACKGROUND

In March 2007 the U.S. Marine Corps Combat Development Command (MCCDC) Operations Analysis Division (OAD) initiated a study aimed at identifying best practices and advancing the state of practice for the Verification, Validation, and Accreditation (VV&A) of simulation models that focus on Irregular Warfare (IW) and more generally on the simulation of Agent-Based Simulations (ABSs).

MCCDC OAD convened a contractor team led by Northrop Grumman Mission Systems to investigate this problem. The effort had two phases: First adapt and extend (as needed) DoD standard practices for VV&A to accommodate the distinctive challenges posed by ABSs. The resulting body of methodology was labeled "ABSVal." Second "validate" ABSVal itself by applying the methodology to two test cases of ABS Supported studies. At key milestones, workshops were held at which Modeling and Simulation (M&S) practitioners with special interest in VV&A and ABS were invited to participate in framing the problem and evaluating proposed solution elements. These workshops generated a significant portion of ABSVal content and helped to ensure that the resulting methodology would have or be able to secure the buy-in of the M&S community.

Systems Planning and Analysis, Inc. (SPA) was asked to join the team in an auditor role. We were tasked to evaluate and comment on the content and application of the ABSVal process with a focus on its practical utility: how useful did the process appear to be in providing guidance to a Verification and Validation (V&V) team, and how useful are the products of an ABSVal guided V&V team to an accreditation agent or model manager?

We had three tasks in our SOW:

- Task 1: Participate in the workshops
- Task 2: Review "measures of validity" developed early in the project
- Task 3: Audit (review) the application of ABSVal to test cases.

Although we executed all three tasks, the focus of this report is on the third one. In this report, we comment specifically on the ABSVal applications as documented in the V&V reports provided for our review, but also discuss what we observed during the process of inventing ABSVal, and how it might be extended or otherwise modified for improved utility.

The remainder of this report is organized as follows: First we discuss our approach to the audit task. Next, we give an overview of the ABSVal test applications and the V&V reports produced from them. Third, we comment on each of the two applications in turn and provide summary comments that apply to both applications. Finally we offer some comments and recommendations relating to ABSVal itself, as distinct from the two test cases that were identified in the process of reviewing the test cases.

2. APPROACH

SPA analysts attended the workshops and stayed in touch with the validation team throughout the process so that we would know roughly as much as other participants about what ABSVal is, and so that we would have visibility into how it was being applied to the test cases. Although we shared observations during the development and application of ABSVal from time to time, both in the workshops and during the intervening periods of activity, we remained true to our charter. We consciously avoided contributing to the content of ABSVal, or attempting to steer the process of developing that content, in order to prevent a sense of ownership and compromise our position as independent observers and evaluators.

At the March 2008 workshop, the Audit Team presented a plan for reviewing the application of ABSVal to the test cases. The plan focused on the validation reports as a reflection of the process, and specifically on the utility of those reports to a hypothetical Accreditation Agent, whose role we would play. Just as the purpose of applying ABSVal to the two test cases was not to pass judgment on the test cases, but to learn how well ABSVal worked in practice, our review of the resulting validation reports was not to "grade" those reports, but to infer from them how ABSVal adds utility to V&V products and suggest additional utility. In light of this, with rare exceptions we omit findings and recommendations that relate to organizing and formatting the reports' content for ease of reference. The focus of the review is on what content is and is not present, and what this means about the process that was followed to produce it.

We developed and presented a series of review questions that we planned to ask about each report. These were designed to get at how useful the reports (and by extension the process that was followed to develop the information contained therein) would be to the consumers of a V&V effort on an ABS. During and after this presentation, we received feedback from other workshop participants that resulted in modifications to these questions. The key feedback points were:

- The thrust of the questions appeared to reflect too great an emphasis on supporting accreditation decisions and appeared to neglect model improvement programs.
- When conducting V&V specifically in support of an accreditation decision, there is no need to consider the model's ability to produce valid answers from data other than that actually used (or intended for use) in the study in question.

We modified our questions based on these points. Our experience in model development, V&V and the conduct of model-supported studies leads us to concur strongly that V&V, when timely, can add at least as much value to model improvement as it can in support of an accreditation decision. But we found it necessary to add only one review question to account for the use of V&V in support of model improvement. The questions, after revision were:

- *Does the report clearly identify the application (set of study questions) for which the model is being validated, and the model's role in addressing those questions?*

It is important to bound the problem and document understanding (or reveal any disconnect) between the V&V team and the accreditor about what the model needs to do well.

- *Does the report clearly describe the tests that were performed on the model, the possible outcomes for each test, and the criteria for passing?*

  A significant component of the discussion that underlies the inventing of ABSVal had to do with making the V&V process more science-based, and therefore objective and repeatable, than it has traditionally been. Specifically the notion that model validity is a falsifiable proposition subject to challenge by applying tests that, should the model be invalid, would have a reasonable chance to reveal that invalidity. Results were mixed, but this question assessed how far the V&V team was able to take that idea. Also, regardless of any connection to the scientific method, the accreditation agent would need an easily digested summary of the V&V team's investigations that led to the accreditation recommendation.

- *For each test performed, is the result clearly presented in a way that relates directly to the specified acceptance criteria?*

  We did not expect it in the test cases, but we have all seen reports where an investigator, perhaps in response to actual or perceived political pressure, did not seem able to confront the implications of facts uncovered in his or her investigation. This phenomenon typically manifests itself as a redesign of the scorecard after the game has been played, or an expression of the result in terms that don't relate directly any acceptance criteria, predetermined or otherwise. Obscuring the results diminishes the value of the work done to produce them, and this question is one we recommend that any reviewer consider in any situation where it applies.

- *Does the report provide a recommended decision for the accreditation authority?*

  A recommendation with rationale is more useful to an accreditation authority than "here are some facts we found, make of them what you will," even if the rationale is not accepted by the accreditation authority.

- *Does the report make a convincing argument that the tests conducted collectively provide a sufficient basis for the recommended accreditation decision?*

  In other words, does it tell the accreditation agent and authority why they should agree that the investigation was complete enough to make a recommendation that they and the accreditation agent can stand behind?

- *To how broad an audience does the report make its findings accessible?*

  This question was included after having seen an early draft which we found to have a somewhat esoteric style reminiscent of some academic journals. We wanted to encourage the V&V team to target its arguments to the broadest segment of people qualified to evaluate them. (At the same time, we did not want to encourage oversimplification of important concepts, just so nobody would feel left out.)

- *Are recommendations provided that are actionable by a model improvement program?*

  This question was added to get at the value to a model improvement program. Actionable recommendations are the vehicle for delivering that value. The same background information required by an accreditation agent would also provide rationale for these recommendations.

The audit team reviewed the two test applications against these criteria and presented the findings at the July 2008 workshop. Besides producing answers to the above questions, our review of the two test case VV&A reports generated a number of general observations on ABSVal and its application that do not relate directly to any of the above questions. The questions were developed as a framework for reviewing the ABSVal applications and not as a constraint on scope. The observations were also presented at the workshop and are summarized in the final sections of this report.

For a more concise presentation at the expense of pedantry, the reviews of each application presented in subsequent sections will dispense with the question-and-answer format, the audit questions having been described above. The answers to all of the above audit questions are given.

3. OVERVIEW OF ABSVAL APPLICATIONS AND V&V REPORTS

As noted earlier, there were two separate ABS applications (studies supported by an ABS) on which the ABSVal methodology was tested. The audit team reviewed the application of ABSVal to these test cases. In both cases, the supporting ABSs were built on Pythagoras, a Northrop Grumman developed ABS modeling environment. Pythagoras itself was not subjected to V&V in this exercise directly, although an adverse finding in the Results Validation phase of ABSVal could ultimately be traced to a problem with the modeling environment just as well as to the model itself.

The first application was Pythagoras Counterinsurgency (P-COIN), a model of population dynamics that adapted earlier exploratory work done by MCCDC to "make headway in developing a counterinsurgency model." P-COIN was used to answer a question about MAGTF basing in support of a contingency plan. Specifically, if a MAGTF were to be deployed to Colombia for disaster relief, would basing ashore or afloat minimize the number of locals sympathetic to the insurgency or actively participating in it, when observed at some specified time after arrival of the MAGTF?

MCCDC did the original work to approach this problem. That work included developing a simple influence-based model of population dynamics, segmenting the local population, and deriving the values of quantitative model parameters. The model parameters were derived from the answers to a series of questions posed to subject matter experts familiar with the population indigenous to the scenario location. Northrop Grumman adapted the resulting approach in a Pythagoras application that used different representations of population-dynamic phenomena that MCCDC had used. This departure introduced some ambiguity about what should be taken to be the Conceptual Model "of record" for the application of ABSVal.

There were two separate V&V reports on P-COIN. One focused on validation of the Conceptual Model and one on verification of the instantiated model. This split was a consequence of the allocation of tasking among contractors supporting the effort. In most cases, we expect that there would be one V&V report, which would simplify the accreditation agent's task somewhat; we had to come up with a synthesis of the two reports and supply our own "bottom line." It was not difficult to do this, however.

The second application was the Pythagoras Obstacle Clearance Model (P-OCM), a model that was built and used to explore tactics for using bombs to clear surf zone mines and obstacles in support of an amphibious landing in a hostile environment. There was a single V&V report on P-OCM.

An earlier workshop, held in October 2007, had among its objectives to identify candidate applications (other than P-COIN, which had been selected earlier) on which to test the ABSVal methodology. The response was weak, indicating either that there were relatively few "real world" ABS-based models known to workshop participants, or that it was difficult to obtain agreement to have them reviewed using an experimental process, or some combination of these and other conditions. Both of the applications that were made available and selected for review, P-COIN and P-OCM, had been developed by a member of the ABSVal study team using Pythagoras.

4. REVIEW OF P-COIN VERIFICATION AND VALIDATION

The audit team reviewed the two reports for P-COIN. We found that they clearly identified the intended use for which accreditation was being considered. We were able to learn from the reports what tests and related investigations were conducted in order to determine whether P-COIN could be relied upon to support its intended use. We were also able to learn the outcomes of these tests from reading the reports.

In some cases, the tests were described before the results were presented. In other cases, the fact of a test was inferred from the presentation of its result as a finding of the V&V effort. We recommend that ABSVal-based V&V reports consistently describe the tests conducted (or qualitative questions investigated) as a means of introducing the result and giving the reader a better sense of why the test was done, what would constitute a hoped-for or not-hoped-for result, and what the result means to the question of validity. To give an example:

> Ideal: "The methodology includes a coin toss to determine whether some key event X happens. The Conceptual Model specifies that X should be 50% likely to happen. If the simulated probability of X is off by 10%, then reported MOE Y would be off by an order of magnitude, which would invalidate the model for its intended use. This provides the rationale for the use of a coin, but raises the question of whether the coin is fair. To test whether the coin is fair, we flipped it a hundred times and counted the "heads" outcomes. There were 41 heads observed. With a fair coin, there is only a 4% chance of observing 41 heads or fewer, so either this was a very unlucky outcome from tossing a fair coin, or the coin is not fair. Put another way, this outcome shakes our confidence in the fairness of the coin, and by extension, in the model's validity."
>
> Less than ideal, but also acceptable: "The coin used to determine event X came up heads fewer times than expected, and this outcome should be taken as a caveat on the use of the model". Representative of some of the results presented; this implies a test similar to the one described above, reports the result, but is unclear on the result's significance to the validity of the model.

Many tests were more qualitative in nature than was anticipated when we drafted the review question to check whether the reported results could be related to acceptance criteria. Upon applying ABSVal to the representative test cases, it turned out that the most appropriate acceptance criterion, if not the only one available, for many tests was whether the validating analyst felt that the observed outcome was acceptable. In some cases, an assumption testing result was characterized as a failure, but a summary assessment was offered that the failure did not invalidate the Conceptual Model. Although we accept the relevance of those tests to the accreditation decision, as well as the summary assessment that the Conceptual Model had not been invalidated, this result of applying ABSVal to a real world problem suggests that the goal of a purely objective V&V methodology may be illusory. Under a paradigm in which all elements of V&V must be objectively based one would have ask what the passing of a test whose failure does not invalidate the Conceptual Model can do to add confidence to that model.

From reading the reports, we were able to identify a recommendation to bring forward to the accreditation authority and would be able to make that recommendation with confidence in the underlying rationale and our understanding of it. Since there were two separate reports, there was no overarching summary of recommendation; we had to glean the recommendations from both reports and infer the "bottom line" ourselves, which was not difficult. The task was slightly complicated by the reports' disagreement on whether it was safe to rely on the model to inform the analysis question. We took the more favorable and ambiguous assessments in the V&V report focused on verification to mean only that a conclusive determination of invalidity had not been reached. Combining that with the more definitive statement from the assumption testing (validation) report: "…*we have to advise against reporting out any results from the model as actionable in any sense…"* led us to the recommendation not to rely on the results of the model.

Were the accreditation authority to ask probing questions about the model and the validation effort, we would be prepared, having read the reports, to answer some of them, while others would have to be relayed back to the V&V team.

The report that focused on verification of the instantiated model provided an actionable recommendation to a model development team pertaining to improving the model's alignment to the underlying Conceptual Model. However, questions had been raised in the assumption testing report, which focused on validating the Conceptual Model, about its treatment of the phenomenon in question. Therefore the reports are ambiguous, when taken together, on the question of whether acting on that recommendation would produce a net improvement in the model. It may have been a better idea to adjust the Conceptual Model to conform to the instantiated model, but since the V&V effort for P-COIN was split between two performers who worked mostly independently of each other, this question was left unexplored.

5. REVIEW OF P-OCM VERIFICATION AND VALIDATION

The audit team reviewed the validation report for P-OCM. We found that it clearly identified the intended use for which accreditation was being considered. We were able to learn from the report what tests and related investigations were conducted in order to determine whether P-COIN could be relied upon to support this intended use. From reading the report, we were able to identify a recommendation to bring forward to the accreditation authority, and would be able to make that recommendation with confidence in the underlying rationale and our understanding of it. Were the accreditation authority to ask probing questions about the model and the validation effort, we would be prepared, having read the reports, to answer some of them, while others would have to be relayed back to the V&V team.

There were no recommendations for model improvement in the P-OCM report. There was no model verification nor did the investigator have access to the model, so it would be surprising if any such recommendation could have been generated.

The report stated that there was no verification of the instantiated model done, due to lack of investigator access to the model. We can imagine a situation where results

validation alone might be a sufficient path to accreditation (e.g., a weather model that has historically been predictive), but failing that, it is hard to see how a recommendation in favor of accreditation could be reached under ABSVal without the opportunity to do model verification.

There was an interesting twist in the results validation phase for P-OCM, the impact of which was compounded by the investigator's lack of access to the model. Referent data were available, but they related to the neutralization and displacement of mines and obstacles in a surf zone-like environment by bombs, which was an intermediate result of P-OCM. The top-level output of P-OCM was Amphibious Assault Vehicle (AAV) survivability. Lacking access not only to the model but also to intermediate results that built up to this top-level MOE, the investigator was unable to use the referent to validate even those intermediate results that could have been compared to an available referent. In spite of this, the investigator went ahead and addressed the question of referent validity anyway, because the instantiated model was said to be driven in part by data generated by that referent. Accepting this assertion on faith, any deficiency in the referent could be expected to manifest itself as a deficiency in the model. So there was a point, in this case, to validating the referent, even though referent data could not be used for results validation.

6. GENERAL COMMENTS ON THE TWO VALIDATION EXERCISES

While both exercises produced good reasons not to use the respective models "right now, based on the information available," we cannot fail to note that neither test case reached a final determination of the validity of the model as used. Instead, both subject models were left under a cloud of suspicion with the possibility of being exonerated at some later date based upon further investigation. This is not the ideal outcome for demonstrating the power and utility of ABSVal, because it doesn't prove that the process can lead to a final determination, nor demonstrate what it would take to do so. We recommend identifying more explicitly the minimum sets of information and materials sufficient to reach a final validity determination using ABSVal. There might be several alternate paths to accreditation for a given situation, depending on the artifacts and resources available to the V&V team, and the acceptable risk level associated with the intended use of the model.

The assumption testing report for P-COIN and the V&V report for P-OCM laid out some foundational material and described the ABSVal approach in some detail. But in general, it was difficult to infer much about the overall ABSVal framework: rules, stepwise approach, branches, and criteria for determining what information is necessary or sufficient to support a given conclusion. This material was not well enough developed at the time the Phase I report came out, but a practitioner's guide to ABSVal would be a good product to document rules and guidelines that have been adopted as ABSVal canon. Lacking such a guide, we could not make an assessment of how faithfully the ABSVal process appeared to have been followed. The V&V reports themselves were among the best windows into the process that were available, along with the Phase I report, so we structured the audit approach accordingly.

The test cases suggested that subjective elements are difficult to eliminate from V&V altogether, and that some activities to which no objective approach has been identified still add material value to V&V. In particular, the validity implications of an adverse finding within the V&V process were often subjectively assessed, with the underlying test not supporting a more objective finding.

Under the assumption testing approach to Conceptual Model validation there is no way for the accreditation agent to gauge or verify how completely the validation agent has ferreted out the assumptions underlying the Conceptual Model. Fundamentally, the process of identifying the assumptions is simple brainstorming, a process that has no natural completion signal. We believe that the practice of partitioning the "assumption space" into submodels of the Conceptual Model and types of assumptions and then exploring each of these subspaces individually was helpful in giving the accreditation agent confidence that a diligent search for assumptions was done. There may be nothing better within the bounds of practicality that could be done to ensure and demonstrate completeness in enumerating assumptions. However, given the brainstorming nature of the task, we also recommend that two or more people conduct the extraction of implicit assumptions together whenever resources permit.

Based on discussion in the final workshop, we felt it might have been useful to add the following evaluation question to our list: Could another validation agent use the V&V reports to repeat the underlying investigation? The answer for these two test cases is, "probably." But the question of repeatability takes on greater interest for investigations that have reached a final determination of validity or invalidity.

Neither of the two test cases clearly demonstrated emergent behavior, so it was not possible to showcase and test any particular features of ABSVal that may have been designed to respond to that distinctive challenge of validating ABSs.

We believe the question of how to do results validation is a particularly interesting one for simulations displaying emergent behavior as well as more conventional simulations where referent data, if available, may represent just one or a few outcomes that are known only to fall somewhere within the range of possibility. These same outcomes are not necessarily what such a simulation "should" produce in order to be valid; a simulation need only show such outcomes to be within the range of possibility. Perhaps in such cases the focus shifts away from comparing black box output to referent data, and toward explaining observed results: showing that the surprises or other variances from the referent can be explained without requiring the acceptance of any unreasonable intermediate results produced by the model. Such an approach is supportable only when the investigator has good visibility into intermediate steps taken by the subject model.

7. RECOMMENDED ACTIONS TO IMPROVE THE PRACTICAL UTILITY OF ABSVAL

Because the process of inventing ABSVal overlapped the process of applying it, there does not yet appear to be any "rule book" stating what must be done in a V&V study in order to claim that the ABSVal process was followed. We believe it is important to document the rules; particularly what body of evidence is necessary and sufficient to

support a favorable or adverse recommendation. There could be several alternative routes to a favorable or unfavorable recommendation, and if these are codified and become accepted by the community, then the V&V team could just cite one of those logic chains in its report and provide, and clearly identify, the evidence supporting each completed link.

The ABSVal rules should also be helpful in identifying what information and artifacts must be available to a V&V team in order to enable an investigation to reach a final determination about validity under ABSVal. That is, what must be available to the V&V team to enable it to carry out a series of tests sufficient in power and scope to make an assessment that risk associated with the use of the model falls below some identified threshold[24]. There would be different sufficiency criteria for each acceptable risk level. The appropriate risk tolerance level and associated V&V level-of-effort are specific to the application. In some situations, there might be little value in embarking on a V&V effort if it is known in advance that, under ABSVal, a sufficient body of evidence is lacking to support a recommendation in favor of accreditation. In other cases, it might still be deemed worthwhile to embark on the V&V effort knowing that it could only lead to a conclusive finding against accreditation or no conclusive finding. It seems desirable to for the methodology to equip practitioners to make a quick determination as to the sufficiency of the body of evidence and experimental materials available.

We would encourage any V&V team always to offer an assessment of the implications of a finding for model validity, and suggest that doing so be adopted as one of the ABSVal rules. As noted, this was done inconsistently in the test cases. The accreditation agent will be able to form his or her own opinion if sufficient supporting evidence is offered, but should not be asked to make the call without guidance.

If excising the “art,” subjectivity and expert judgment, entirely from the V&V process was an ABSVal objective, this was not achieved. But the team may have discovered reasons to back away from that as an objective. Perhaps the goal, informed by the hindsight developed from the two ABSVal test cases, should have been only look to identify and eliminate unnecessary subjectivity from the V&V process while accepting the necessity of subjectivity in some areas. In the same way that the sum of a precisely known number and a rough estimate is a rough estimate, subjectivity dominates objectivity whenever a combination of the two underlies a determination, but it is likely that there is still value in eliminating unnecessary degrees of subjective freedom.

When exercising ABSVal or any VV&A methodology in a test case, it might be helpful to be more explicit about the “business scenario” under which the exercise, or each component of it, is being carried out. Having a scenario in mind might focus and guide the approach taken somewhat. Three scenarios are:

- The analysis has already been done and an accreditation authority needs to know whether to endorse the process that produced the findings, as a part of accepting the findings. The only data concerns relate to the data that were actually used.

[24] And not only when the study turns out to support the *a priori* planned course of action.

- A model sponsor wants to develop and document some verified, general information about the model's behavior that applies within a broad class of intended uses. This information would not by itself support any accreditation decision, but its availability would streamline the V&V process, thereby making the model a more attractive candidate for employment in any study where accreditation is desired or mandated. Effectively, the model proponent is "building V&V-ability" into the model, but cannot complete the process leading to accreditation, lacking advance knowledge of the particulars of any specific use of the model with regards to its class of intended uses.
- A study director is contemplating the use of an existing model to support a study that is in the planning stages. "Bounds of validity" are of interest since many of the details of the baseline and planned excursions are pending resolution. In this scenario, information from the second scenario might be available to use in conjunction with study specific inputs to expedite the process leading up to the accreditation decision.

It appears that the first listed scenario was the operative one for both of the test cases we reviewed, although it could also have been the third for P-OCM. It would be interesting to test a scenario that combined the second and third of those listed above, so that the line between reusable and one-time V&V information is sharply drawn and the actual value of the reusable information and any associated limitations in streamlining the subsequent "one-time" V&V process for a specific application.

To comment further on the scenario where validity is "built into" the model as an integral part of its development, it has been our experience that the earlier a validation task is done the more value it adds. In practice, the first listed scenario in which V&V is conducted as almost an afterthought, is inherently adversarial in nature, since study findings and the decisions that hinge on them are already on the table, with advocates and detractors not far removed. The approach taken to V&V in such an environment is also typically constrained or compromised by time pressures. Best of all is to develop validation information in conjunction with the model development and update it as needed in conjunction with version releases. As this requires additional up-front investment, it is tempting to omit the parallel model and V&V development, but every party to the process, from developer to user, has an interest in supporting it. The developer benefits from an outside review of the model at a time when the resulting feedback can be used to improve the model before it is exposed to broader scrutiny. Also, when the V&V activity is carefully documented, much of it may be reusable in support of study accreditation, making the model more attractive. The study director has less vetting to do and can select the model with greater confidence in its utility. He or she is also sheltered from the risk of having uncertainties about the model's validity which could taint an ongoing or completed study.

The ABSVal Application Audit Report can be found at
http://orsagouge.pbwiki.com/ABSVal

APPENDIX L SURVEY

The following pages contain the survey that was given during Workshop #4 that solicited feedback on the ABSVal Framework application results.

The Agent-Based Simulation (ABS) Verification, Validation, and Accreditation (VV&A) Framework (ABSVal Framework) Study

Marine Corps Combat Development Command (MCCDC)
Operations and Analysis Division (OAD)

Application of the ABSVal Framework to Model Application Survey

Background: MCCDC OAD has been developing a general, institutionally acceptable processes and criteria for assessing the validity of agent-based simulations used as part of DoD analyses. Phase I of this project developed a framework for validation. Phase II is applying the framework to specified simulations to test and expand the framework.

Purpose: This survey documents feedback from the third and final workshop in Phase II of the ABSVal Framework Study.[1] The intent is to review the validation presentations provided during the workshop and evaluate the usefulness of the material presented to four different potential audiences:

1. Decision-maker (e.g, O-6, O-7, executive) using the results from a simulation; will receive a 20 minute briefing with questions and a recommendation from support staff
2. Recipient of simulation results (e.g., O-3, O-4, O-5, mid-level manager) reviewing them to make a recommendation; may read analysis report but probably won't look at the analysis details
3. Analyst evaluating the simulation and results (e.g., operations analyst, engineer, SME); will look at report and analysis detail
4. Simulation developer; interested in improving the model

Guidelines: From the perspectives of the four potential audiences above, please evaluate the usefulness of the material provided in the presentations to enable that audience to perform his/her job. Identify any glaring omissions or information that would be particularly useful to the audience. Add any other comments on the validation application process that would improve the ABSVal Framework. If a review of the reports developed for these briefings would be useful, please ask for a copy.

Additional assessments: In addition to assessments from the different audiences, we also would like feedback on the maturity of the processes applied. Scott Harmon and Simone Youngblood are developing a process to assess the maturity of the validation process applied to a model's application and use. This Validation Process Maturity Model (VPMM) has six levels from 0 (no validation process applied) to 5 (the application of an automated validation process) moving through the levels with increasing rigor applied to the process and data. Additional questions are included in this survey that are derived from a survey developed by Harmon and Youngblood to apply the VPMM provided to the ABSVal Team from Harmon as a courtesy.

The ABSVal Framework Study also has conducted audits of the validation applications for additional feedback and insight on the process to guide improvements. The auditors took the role of an accreditation authority or his/her advisor. This survey contains questions similar to some of those asked in the audit for additional feedback.

[1] Information from these surveys will be compiled and used in the final report for the ABSVal Framework Study and may be used in support of further research of ABS Validation.

The Agent-Based Simulation (ABS) Verification, Validation, and Accreditation (VV&A) Framework (ABSVal Framework) Study

Application of the ABSVal Framework to Model Application Survey

Validation briefing _____

Briefer _____ Presentation block _____

Name (optional) _____ Contact Info (optional) _____

Usefulness of the validation report to different audiences

1. Decision-maker (e.g, O-6, O-7, executive): will use the results from a simulation to support decision making; will probably receive a 20 minute briefing with questions and a recommendation from support staff

Based on validation analysis presented, is there sufficient information to assess whether the simulation results should be trusted in supporting the purpose to which the model was put? Is there any additional information that would be useful?

2. Recipient of simulation results (e.g., O-3, O-4, O-5, mid-level manager): will review the results from a simulation based analysis and use them to make a recommendation; may read analysis report but probably won't look at the analysis details

Based on validation analysis presented, can a well-informed recommendation be made on whether the results from the simulation analysis should be trusted? Is there any additional information that would be useful?

3. Analyst evaluating the simulation and results (e.g., operations analyst, engineer, SME): will look at report and analysis details

Does the validation report provide sufficient information for the analyst to assess the use of the model and the simulation results? Is there any additional information that would be useful?

4. Simulation developer: interested in improving the model

Does the validation presentation provide sufficient information to the developer to help guide model improvement? Is there any additional information that would be useful?

Maturity of the Validation Process

1. Were validation criteria present? Yes No
   What specific additional criteria would you suggest?

2. Were error/accuracy criteria identified? Yes No
   What specific error/accuracy criteria would you suggest?

3. How would you describe the referent and its use with respect to sources, acceptable errors, ability to use to assess correctness of the model and simulation results? How might deficiencies be addressed?

4. How would you describe the conceptual model and its representational capabilities with respect to sources, acceptable errors, ability to use to assess correctness of the model and simulation results? How might deficiencies be addressed?

5. How would you describe the simulation results and its representational capabilities with respect to sources, acceptable errors, and sampling strategies? How might deficiencies be addressed?

Audit Questions (please answer Yes or No and elaborate as needed)

1. Does the validation report clearly identify the application (set of study questions) for which the model is being validated, and the model's role in addressing those questions?

2. Does the validation report make a convincing argument that the tests conducted collectively provide a sufficient basis for the recommended accreditation decision?

3. Does the validation report provide a recommended decision for the accreditation authority?

Additional Comments (please use additional pages as needed)

The survey can also be found at http://orsagouge.pbwiki.com/ABSVal.

APPENDIX M DESCRIPTION OF ASSUMPTION TESTING

This section details an existing validity assessment process that the Project Team incorporated within the ABSVal framework and applied to two of the test cases. The process was developed by one of the subcontractors on the team, Mr. Robert Eberth of Sanderling Research Corporation. The process has its origins in two model assessments he performed as an Operations Research student at the Naval Postgraduate School in 1972-74, long before the terms "validity" and "validation" were in vogue in the industry. It has evolved continually since then.

The process has gained highly positive notice. One paper reporting the results of the process applied to a commonly-used battle damage assessment algorithm won recognition as "Best Test and Evaluation Paper" of the 58th Military Operations Research Society (MORS) Symposium. A later paper[25] on the use of the process in the invalidation of a mine warfare simulation was nominated for the Rist Prize of the 63rd MORS Symposium.

M.1 THE BASICS

This particular process, alternatively termed the "Technical Validity Assessment" process, is an exercise in the application of modern scientific method. It depends almost exclusively on two techniques: (1) Assumption Testing, and (2) Reverse-Engineering.

Logical Validity Assessment (LVA) is *not* an "end to end" validation process. It purposely ignores the Conceptual Model of the Mission Space (CMMS), the set of documents that describes the real-world counterpart to whatever will be modeled. The CMMS is ignored because it describes what "is" rather than what "is supposed to be built," and thus contains information that is superfluous or even counterproductive to VV&A. LVA instead begins with and concentrates on the "other" Conceptual Model, defined in the ABSVal framework as consisting of:

1) ***Theoretic model.*** The initial expression, usually in textual and/or graphical form, of the context of the model and of the cause-and-effect relationships believed to be operative in the situation of interest and that are intended to be incorporated within the end model.
2) ***Mathematic model.*** The model that captures the specific logical structures and expressions (equations, conditional statements, logic tables, etc). Note that the relationship between the theoretic model and mathematic model is one-to-many; i.e., there are numerous ways a theoretic model could be represented mathematically.

[25] R. Eberth, *"Assessment of the Marine Corps Combat Analysis Model;"* Appendix H to "*Mine, Minelaying, and Mine Countermeasures Final Report,"* PRC, Inc., 1992

3) ***Algorithmic model.*** The coded form of the mathematic model. Again, note that the relationship between the mathematic model and algorithmic model is one-to-many.

Unless LVA were being applied during the development of a model (which has yet to happen over a span of thirty-five years), it does not assess the above three sub-models separately. It normally starts with whatever document explains the logic used within the model (often termed the "Analyst's Manual"), which generally addresses the theoretic model and the mathematic model as a single entity. If no such document exists, LVA can be applied directly to the algorithmic model as expressed by the source code. Even when a thorough Analyst's Manual does exist, however, it still is necessary to do a "logical verification" of the algorithmic model vis-à-vis the mathematic model, and may be necessary to do additional testing of additional, new, assumptions found in that algorithmic model. (Such assumptions typically are structural assumptions that are made for the first time during coding.)

LVA also does not address Results Validation, which should be the most powerful element of a thorough validation process. Unfortunately, Results Validation often is an unattainable goal, usually because there is no referent available for results comparisons. In that sense, LVA can be a gap-filler. Even in the absence of Results Validation and/or a referent, LVA can provide strong evidence informing an accreditation decision.

M.1.1 Assumption Testing

As stated in the mine warfare simulation report,[26] "Model assessments commonly address the hardware and software engineering aspects of a model, and particularly its *usability* characteristics (e.g., user interface, graphics, input data availability and formatting requirements, clarity and completeness of documentation, maintenance support). While those characteristics are unquestionably important, the *analytic* capabilities of any model are not determined by its hardware and software architecture, communications network, or user interface. Nor can they reliably be determined from the "capabilities" statements – often little more than advertising copy – that typically accompany a model. Rather, analytic capabilities are determined by a model's logic and control structures and their underlying assumptions, its computational algorithms and underlying mathematical assumptions, and its data manipulation and transformation algorithms – all of which are rarely seen by the end user. Moreover, a model's *bounds of validity* also are determined by its underlying assumptions, some of which may not be readily visible even to the model's developers." [27]

The potential usefulness of assumption testing as a validity assessment technique may be seen by considering the nature of models in general. George Box famously stated, "All models are wrong, some are useful." Box was absolutely correct in the literal sense

[26] Ibid.

[27] The condition is largely due to the practice of reusing algorithms and/or code from earlier models without first ascertaining the embedded assumptions.

– no model *is* reality. Rather, every model is an abstraction of reality to some extent. That aspect of a model is widely if perhaps not universally recognized. What is more rarely recognized and far more rarely appreciated is that the mechanism of abstraction is the *assumption.* Thus if we could identify every assumption used to create a given model, we would know how it deviates from reality or, in Box's terms, we would know just how "wrong" it is.

As a practical matter, however, we can not *explicitly* identify every assumption in even simple models. The good news is there is no need to identify all of them. We need only to identify, and "test," the assumptions that have significance to the *intended purpose* of the model and especially to the *analytic questions at hand.* That is, of course, more easily said than done. Part of the art, vice science, of assumption testing is to be able to recognize in at least broad terms which assumptions are likely to be significant, given only a description of the model, the context of the study, and the specific analytic questions at hand. Thus validity assessment analysts generally have to cast a wider net than would be necessary if they had full knowledge going in as to which assumptions are significant. Assumptions having little or no apparent significance are set aside. Ones having apparent significance are tested as described below.

M.1.1.1 Types of Assumptions

There are four sets of assumptions of interest: causal, structural, mathematic, and scenario:

1) ***Causal assumptions*** deal with cause-and-effect relationships among agents/objects/entities and with their environment(s) and other stimuli.
2) ***Structural assumptions*** deal primarily with the processing order of stimuli, decisions, and actions within a model, but also may deal with starting, ending, and boundary conditions within a model.
3) ***Mathematic assumptions*** deal with the myriad assumptions made to enable constructing a determinable mathematic abstract of real-world scenarios, processes, behaviors, and events; mathematic assumptions include the choice of algorithms and other logic structures, and thus assumption testing includes an assessment of those algorithms/structures.
4) ***Scenario assumptions*** deal with bounding the real-world environment (which may be behavioral as well as geophysical) to be addressed within the model, with the geophysical features and environmental conditions contained therein, and with the entities and their characteristics to be "in play" in a particular scenario.

M.1.1.2 Assumption Testing Process

Assumption testing is a three-step process:

> ***Step 1: Identify the assumptions.*** Assumptions, particularly causal and mathematic assumptions, are rarely if ever well-documented and may even have to be reverse-engineered from the source code. In some cases, even some of the algorithms may not be documented. This is *by far* the most difficult aspect of assumption testing.

Step 2: Determine the operational implications and bounds of validity of the assumptions. Accomplished as a cooperative effort between M&S and operational subject matter experts (SMEs).

Step 3: Determine the acceptability of the identified operational implications and bounds of validity to the decision-maker. During M&S system development, the decision-maker is the M&S Sponsor. For a particular application of the M&S system, the decision-maker is the Application Sponsor (also designated by DoD policy as the Accreditation Authority for that particular application).

M.1.2 Reverse Engineering

Reverse engineering is a technique applied to the Mathematic and/or the Algorithmic Models within the Conceptual Model. The purpose of reverse engineering is to identify the assumptions embedded within the various forms of logical expression within the model. The assumptions then are subjected to assumption testing, as above. A "full" reverse-engineering approach would first break down a model into its logical components – structures and algorithms – and then derive those same algorithms and develop those same structures from a zero base. As each derivation proceeds, each assumption necessary to the derivation would be identified, interpreted in terms of its operational implications (i.e., what it would mean in the "real world ") and impact on the bounds of validity, and recorded. Similarly, the operational implications of the logic and control structures and their assumptions would be determined and recorded as their development proceeded. Ultimately, as above, the supported decision maker (i.e., the Application Sponsor) is asked to decide whether those operational implications and bounds of validity – and thus the model – are acceptable for the intended purpose.

A shortened form of the approach attempts to identify a very few "most critical" algorithms (ideally, only one or two) and the principal logic and control structures for analysis. It further attempts to start at some level above the "zero base," relying on recognition of common algorithms that have been subjected to a zero-based assessment in the past. For the short-form approach to be feasible, technical documentation of algorithms and structures must be thorough.

M.1.3 Role of Scientific Method

The Logical Validity Assessment process is based in scientific method, with the thrust being to find evidence that would reject (falsify) the null hypothesis that the model or simulation is valid for the intended purpose. Scientific method applies within assumption testing, but in an indirect fashion. A validity assessment attempts to determine whether a model or simulation is “sufficiently accurate,” *vis-à-vis the real world*, for a particular application. But because assumptions represent purposeful *departures* from the real world, and sometimes quite significant departures, a direct application of scientific method – one that directly compared the assumptions to the real world -- could readily hold the model to be invalid without even considering the intended application. (Which is why Box’s famous “All models are wrong” quotation often is cited,

incorrectly, as evidence that validation is a waste of time and money.) The indirect scientific method approach used in assumption testing notes the departures from reality, but then determines their *operational implications and bounds of validity in the context of the application at hand*. The final step of the indirect approach is for the Application Sponsor to decide whether those operational implications and bounds of validity are acceptable for his or her application. If the operational implications or bounds are *not acceptable*,[28] then the null hypothesis is falsified and the model is declared invalid for that particular application.[29] If the operational implications and bounds of validity *are acceptable*, then falsification has failed and the model is *accepted as valid* for that particular application. (Note, however, that it is never *proven* valid.)

M.1.4 Verification vs. Validation

It is surprisingly common to find errors in the Conceptual Model when applying the Logical Validity Assessment process. The most common are equations or algorithms that are simply mathematically incorrect or that fail to do what the specification said they should do – or do something very different altogether. Coding errors also are common. While such errors indicate failures of the verification process, they immediately become validity issues and have to be treated as such (i.e., they *must be* corrected for a model to be accepted as valid and accredited for a particular application).

M.2 LVA PROCESS STEPS

M.2.1 Identify the Application Sponsor

The Application Sponsor, who also by DoD policy is the Accreditation Authority, is absolutely critical to any validation process. The Application Sponsor is the official ultimately responsible to higher authority for the "goodness" of the product, whether that product is a training event, a multi-service live-virtual-constructive experiment, or an analysis. It generally is the official whose signature will go on the report being released to external organizations. The reason for twinning the responsibilities of Application Sponsor and Accreditation Authority is to ensure that the official responsible for the goodness of the product also recognizes his or her responsibility for the choice of the supporting model and dataset. [30]

[28] The bounds of validity are rarely a problem compared to the operational implications. They can become a major problem, though, if the Application Sponsor wants to use a single model for a series of applications and one or more of them will fall outside the bounds.

[29] *Assumptions* that are shown to be invalid for the application at hand are usually extremely difficult to correct within the timeframe of a study because correction usually would mean redesigning significant elements of the model. *Errors*, however, often are discovered during a logical validity assessment, and may be easily corrected.

[30] And possibly in the near future, accreditation will include the team using the model and the analysis or exercise plan itself.

M.2.2 Identify the Application

Accomplished only with the Application Sponsor: What, precisely, is the problem he/she wants addressed? What are the *specific analytic questions* to be answered? What is the actual *system* being addressed? What decision is being supported and what role will the model's results play in that decision? What are the correct metrics for the system, analytic questions, and supported decision?

M.2.3 Identify the Referent

The referent becomes the standard for comparison for the model. Ideally, it will be an empirically-derived results database. This activity s*hould* be accomplished with the Application Sponsor, but may need to be done with the planned analyst team or independently. Generally, the referent is either obvious or non-existent. Unlike for most validation methods, a referent is nice to have but not *necessary* for the Logical Validity Assessment process to succeed. The step is included here only because the referent *will be necessary* for Results Validation, and this is a convenient time to identify it if it exists.

M.2.4 Identify the Accreditation Criteria

Again, accomplished only with the Application Sponsor, but often that individual doesn't even know what the term means, so the validation analyst has to *educate* as much as *ask,* and may need to suggest appropriate criteria: How accurate, vis-à-vis the real world, do you need the model's results to be (assuming Results Validation will be applied at some point)? Are there things that it absolutely must represent? Especially if Results Validation will not be performed, is the Application Sponsor comfortable with the concept of assumption testing as a *primary form* of validity assessment?

M.2.5 Gather Documentation and Schedule Interviews

Documentation on the model and any earlier VV&A actions should be available from the M&S Sponsor and/or user organizations. The Modeling and Simulation Information Analysis Center (MSIAC) at http://www.dod-msiac.org can be a very effective document research asset. Web searches also can turn up pertinent information that even the developer may not have seen.

After absorbing the available documentation, interviews should be scheduled with development personnel if available. Interviews with model marketers are a waste of time unless they can talk knowledgeably and in detail about the internal logic of the model.

M.2.6 Secure Services of Operational SMEs

There is a place for operational SMEs in LVA, but it's *not* for the purpose of "face validating" the model. Their role is to determine the *operational implications* of the myriad assumptions that the process will identify. Such SMEs often are available at no

cost from the Application Sponsor's own organization. In some cases, the validation team may have the necessary operational expertise within its own ranks.

M.2.7 Identify Assumptions

This is the dog work of the process, and it is more art than science. Virtually every paragraph, every logical construct, every algorithm will contain some number of assumptions. The trick is to decide on the fly which ones could be "significant" to the model's validity for the application at hand. Ideally, at least two analysts will work the problem independently and compare notes afterward. As potentially significant assumptions are identified, they should be sorted into the categories mentioned earlier; i.e., causal, structural, mathematic, and scenario.

When the team hits the equations and algorithms, the level of difficulty jumps considerably. It is extremely rare for the model's documentation to provide details on how the equations and/or algorithms were derived, yet those are the details that are needed. Thus often the only available technique for uncovering the embedded assumptions is to reverse-engineer the logical constructs. Ideally, at least one member of the team will be sufficiently well-versed in modeling mathematics that he or she can recognize the roots of even complex expressions. As stated earlier, the goal is to break the expressions down to their most basic components or roots and then re-derive the expressions. As each derivation proceeds, each assumption necessary to the derivation would be identified and recorded.

M.2.8 Identify Bounds of Validity

This step is most easily performed as significant assumptions are identified. Every assumption prescribes some bound on the validity of a model. Even very standard assumptions can have profound and unintended consequences with regard to bounds of validity. (E.g., the virtually standard assumption of statistical independence of events should never be used when it is known that the "things" being analyzed are highly correlated.) The bounds tell the Application Sponsor and his or her analysts when they are operating within the region in which the model should be valid (assuming it has not been falsified) and when they are outside that region and at some risk of getting invalid results. It should not require operational SMEs, which is why it can be performed as the assumptions are being identified in the first place.

M.2.9 Identify Operational Implications

This is the step requiring operational SMEs. Assumptions are rarely stated in operational terms and thus may well be meaningless to the Application Sponsor. This step translates the assumptions into language he or she not only can understand, but also can fully appreciate. Thus a decision to use "DTED Level 0" terrain data in a computerized experiment translated – with a good SME – into "long range airborne shooters will be able to detect and kill all the targets before the ground-based assets can get close enough to see them." Since the real world operates at a considerably

higher “DTED Level” than zero, thus obscuring many targets from airborne targeting, the assumption was unacceptable for computing damage

M.2.10 Obtain Acceptability Decisions

This step presents the significant assumptions and their operational implications and bounds of validity to the Application Sponsor for an acceptability decision. This is a go/no-go decision. If there’s an operational implication or bound of validity he or she simply cannot accept, the model’s validity for that intended application has been falsified. On the other hand, if everything *is* acceptable, then the model itself may be accepted as valid for the application pending falsification by additional verification or validation steps.