United States Military Academy
West Point, New York 10996

A Validation Methodology for Human Behavior Representation Models

OPERATIONS RESEARCH CENTER OF EXCELLENCE
TECHNICAL REPORT DSE-R-0531
DTIC #: ADA433696

Authors
Lieutenant Colonel Simon R. Goerger, Ph.D.
Assistant Professor, Department of Systems Engineering

Colonel Michael L. McGinnis, Ph.D.
Professor and Head, Department of Systems Engineering

Rudolph P. Darken, D.Sc.
Professor and Director of Modeling, Virtual Environments & Simulation
Naval Postgraduate School, Monterey, CA

Approved by
Lieutenant Colonel Michael J. Kwinn, Jr., Ph.D.
Director, Operations Research Center of Excellence

May 2005

The Operations Research Center of Excellence is supported by the
Assistant secretary of the Army (Financial Management & Comptroller)

Distribution A: Approved for public release; distribution is unlimited.

20050606 075

A Validation Methodology for Human Behavior Representation Models

Senior Investigator

Lieutenant Colonel Simon R. Goerger, Ph.D.
Assistant Professor, Department of Systems Engineering

OPERATIONS RESEARCH CENTER OF EXCELLENCE TECHNICAL REPORT DSE-R-0531 DTIC #: ADA433696

Authors

Lieutenant Colonel Simon R. Goerger, Ph.D.
Assistant Professor, Department of Systems Engineering

Colonel Michael L. McGinnis, Ph.D.
Professor and Head, Department of Systems Engineering

Rudolph P. Darken, D.Sc.
Professor and Director of Modeling, Virtual Environments & Simulation
Naval Postgraduate School, Monterey, CA

Approved by

Lieutenant Colonel Michael J. Kwinn, Jr., Ph.D.
Director, Operations Research Center of Excellence

May 2005

The Operations Research Center of Excellence is supported by the Assistant Secretary of the Army (Financial Management & Comptroller)

<u>Distribution A</u>: Approved for public release; distribution is unlimited.

Abstract

The Department of Defense relies heavily on mathematical models and computer simulations to analyze and acquire new weapon systems. Models and simulations help decision-makers understand the differences between systems and provide insights into the implications of weapon system tradeoffs. Given this key role, the credibility of simulations is paramount. For combat models, this is gained through the verification, validation, and accreditation process required of DoD analytical models prior to their use in weapon system acquisition and other studies. The nature of nondeterministic human behavior makes validation of models of human behavior representation contingent on the judgments of subject matter experts that are routinely acquired using a face validation methodology. In an attempt to better understand the strengths and weaknesses of assessing human behavior representation using experts and the face validation methodology, the authors conducted experiments to identify issues critical to utilizing human experts for the purpose of ascertaining ways to enrich the validation process for models relying on human behavior representation. The research was limited to the behaviors of individuals engaged in close combat in an urban environment. This paper presents the study methodology, data analysis, and recommendations for mitigating attendant problems with validation of human behavior representation models.

About the Author(s)

Lieutenant Colonel Simon R. Goerger is an Associate Professor in the Department of Systems Engineering at the United States Military Academy, West Point, New York. He earned his Bachelor of Science from the United States Military Academy in 1988 and his Masters in Computer Science and Doctorate in Modeling and Simulations from the Naval Postgraduate School, Monterey, CA in 1998 and 2004, respectively. His research interests include combat models, agent based modeling, human factors, and training in virtual environments. LTC Goerger has served as an infantry officer with the 6th Infantry Division in Alaska & Sinai, Egypt, as a cavalry officer with the 2d Armored Cavalry Regiment at Fort Polk, LA & Port-a-Prince, Haiti, and as a software engineer for COMBAT^{XXI}, the US Army's future brigade and below analytical model for the 21st Century.

Dr. Rudolph Darken is an Associate Professor of Computer Science and a Technical Director of the Modeling, Virtual Environments, and Simulation (MOVES) Institute at the Naval Postgraduate School in Monterey, California. He is the Chair of the MOVES Curriculum Committee and directs the Laboratory for Human Performance Engineering. His research has been primarily focused on human factors and training in virtual environments with emphasis on navigation and wayfinding in large-scale virtual worlds. He is a Senior Editor of PRESENCE Journal, the MIT Press journal of teleoperators and virtual environments. He received his B.S. in Computer Science Engineering from the University of Illinois at Chicago in 1990 and his M.S. and D.Sc. degrees in Computer Science from The George Washington University in 1993 and 1995, respectively.

Colonel Mike McGinnis, mike.mcginnis@us.army.mil, is Professor of Systems Engineering and Head of the Systems Engineering Department at West Point. A 1977 West Point graduate, Colonel McGinnis earned a doctorate in Systems and Industrial Engineering from Arizona University, two masters degrees from Rensselaer Polytechnic Institute (RPI) in Applied Mathematics and Operations Research, and a masters degree in National Security Decision-Making from the Naval War College. Colonel McGinnis' previous Army-level experience includes Director of the Unit Manning Task Force, OPMS XXI Task Force, Army

Development System XXI Task Force, INTEL XXI Task Force, and Army Chief of Staff Training and Leader Development Panel.

Acknowledgements

This report covers two and a half years of dissertation and follow on research. The individuals to whom we are indebted represent a number of agencies. The Defense Modeling and Simulations Office, under the supervision of Technical Director for Verification, Validation and Accreditation (VV&A) Simone Youngblood, kept me motivated and reviewed the work in progress. Invaluable assistance in documentation and procedures review to ensure soundness in the validation process was provided by Scott Harmon of ZETETIX and the Defense Modeling and Simulation Office (DMSO) VV&A Technical Working Group; Dale Pace from the Johns Hopkins University Applied Physics Laboratory; Susan Solick, the Army M&S VV&A Standards Category coordinator; Marcy Stutzman from Northrop Grumman of the Navy VV&A team; Jennifer Park, the Navy Modeling and Simulation Management Office VV&A leader; and Bill Rostad.

This research could not have been conducted without the scores of subjects who participated in the two studies. Special gratitude is extended to MAJ William Bestermann, MAJ Mark Hollis, and MAJ John Best of the Infantry Officers Carrier Course at Fort Benning, GA, who coordinated for and supervised the 182 students who served as subject-matter experts. Two other men played a major role in data collection: my research assistants, CPT Benjamin Tipton, US Army, and Petty Officer Second Class Brian Wood, US Navy. We also thank the officers and staff of the Training and Doctrine Command Analysis Center (TRAC), Monterey, who provided subject-matter experts for the pilot study.

Finally, we wish to recognize three organizations that supplied vital resources: the Marine Corps Combat Development Command allowed us to use the agent-based model MANA; Natick Soldier Center offered human-performance data of soldiers conducting operations in the McKenna, Military Operations in Urban Terrain (MOUT) Site at Fort Benning; and the Navy Modeling and Simulation Management Office financed equipment and travel in conjunction with research in Fiscal Years 2003 and 2004.

Table of Contents

List of Figures

List of Tables

Chapter 1. Introduction

Representation of human behaviors in computer simulations is a relatively new and complex area of research that lies at the nexus of modeling and simulations, and behavioral and cognitive psychology. Researchers in this area attempt to model human behavior using computer simulations primarily developed for training, analysis, and research. While each community approaches modeling human behavior from different directions, the boundaries of the area shown in Figure 1 forms a new area of research for validating models with embedded human behavior representation.

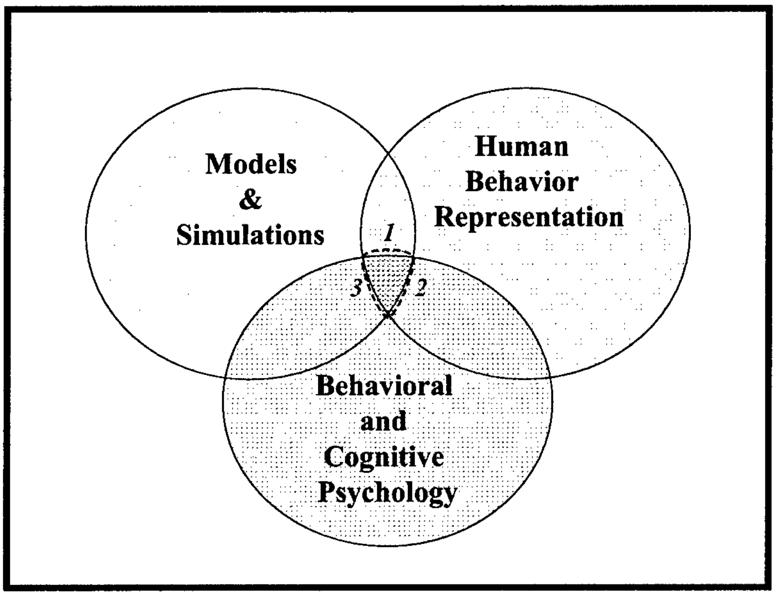


Figure 1. Research Objective: To Define the Common Area

1.1 Problem Statement and Approach

The Department of Defense (DoD) continually pursues new modeling and simulation capabilities to meet the training and analytical needs of America's military establishment. Improvements to the fidelity of physics-based models have raised expectations for modeling human behaviors. However, the lack of verified data has made validating human behavior models difficult.

Although validation of physics-based models is well-defined using long-established standards, the practices are not well suited for validating behavioral models. This is due to several factors:

- The nonlinear nature of human cognitive processes (Department of Defense Directive, 2001);
- The large set of interdependent variables making it impossible to account for all possible interactions (Department of Defense Directive, 2001);
- Inadequate metrics for validating HBR models;
- The lack of a robust set of environmental data to run behavioral models for model validation; and
- No uniform, standard method of validating cognitive models.[1]

This paper contends that subject matter expert (SME) bias demonstrated in the assessment of human behavior representations for human ground combatants can be identified, measured, and mitigated using techniques and standards similar to what is used in assessing the performance of actual soldiers.[2] We tested this hypothesis using a series of studies of company grade Army officers that analyzes their assessment of the performance of soldier tasks derived from *ARTEP 7-8-MTP: Mission Training Plan for the Infantry Rifle Platoon and Squad* (2001). This was done during experimentation sessions were SMEs quantitatively assessed the degree to which computer objects representing soldiers performed tasks to standard.

Human behaviors of interest to the military occur in complex, multi-dimensional environments with an abundance of stimuli. The scenarios developed for studying human behavior models must reflect these complexities. Given this context, two major assumptions bound the research. First, computational requirements of modeling human behavior are beyond the limits of current technology to develop a computable mathematical algorithm or computer program to assess nondeterministic, nonlinear human behavior. Second, fully understanding human behavior requires validating models of human behavior within the context of the decision-making environment where it naturally occurs.[3]

[1] Cognitive models "describe the detection, storage, and use of information" (Solso, 2001). This refers to models that simulate the human thought process to select actions for execution during a simulation.
[2] The term subject matter expert (SME), as used throughout this document, refers to study participants.
[3] Naturalistic decision-making is "the study of how people use their experience to make decisions in field settings" (Klein, 2001).

1.2 Goal

The ultimate outcome of *any* validation process for models of human behavior is to assure *simulated* human behavior is consistent with *actual* human behavior under the constraints and context of a specific domain. This paper presents a methodology for validating HBR model implementations for use in Department of Defense training and research models and simulations. The methodology we identify mitigates issues regarding validation and use of HBR models implemented in legacy and emergent combat simulations.

Chapter 2. Validation Methodology

The methodology for validating human behaviors draws upon three distinct yet related fields: models and simulations; human behavior representation; and behavioral and cognitive psychology. Each discipline has a unique perspective on how it addresses aspects of creating viable HBR models that, until recently, had little in common with the other two disciplines. When considered as a whole, there are key elements from each discipline common to these domains.

The literature contains very few references to formal, statistically based research on creating, implementing, and validating computer-based HBR models. Initially rule-based models of human behavior were integrated into simulations in order to study more advanced concepts and requirements. In doing so, researchers discovered that validation procedures for physics-based models are not adequate for HBR models.

Unlike physics-based models, human behavior models are not mathematically-based making them difficult, if not impossible, to codify. However, human behavior research has collected vast amounts of data that is available to verify and validate HBR models.

Chapter 3. Background

Traditionally, most DoD models and simulations of military forces have focused on replicating armed conflict between two or more sides. This paradigm of physics-based, force-on-force models relies on mathematical algorithms instantiated in computer programs to study battle damage aspects of combat. Metrics, such as the probability of hits and kills, are used to assess the effectiveness of various weapon systems and munitions, fired from various platforms, subject to specific environmental conditions and target types. Over the past decade, however, military operations have placed more emphasis on the actions of the participants rather than on the characteristics of the weapon systems. In response to this new focus, M&S research has shifted to the development of models that represent the human dimensions of operations other than war (OOTW) and combat operations.

As stated in Chapter I, the goal of this research is to integrate into a single framework, a new methodology for validating human behavior models that draws upon three distinct domains: entity-level combat simulations, human behavior representation, and cognitive psychology.[4] The body of behavioral research encompasses many elements of human decision-making to include information gathering, situational awareness, and information processing and communicating. Cognitive models attempt to replicate the human decision-making process through models of human behavior. Cognitive models, linked with physics models, attempt to reproduce human behaviors in a dynamic, simulated environment.

Most behavioral models today deal with a very narrow range of human behaviors that are generally categorized as reactive or procedural. Reactive models follow an input-output, cause and effect protocol where a simulated 'human' agent executes an action that

[4] These domains use numerous terms interchangeably. To reduce confusion and to ensure this research conveys its points, we define some terms in footnotes. The Glossary at the back of the report contains a comprehensive list of terms and definitions for greater clarification.

responds to a stimulus injected into the current situation. Procedural models require simulation agent to follow a prescribed protocol for analyzing a situation, processing information, selecting an appropriate action, and then executing the action. Within the body of research, only procedural models are considered to be cognitive models. Figure 2 shows the relationship between physics-based and behavioral models with respect to the application areas of combat and OOTW. In general, physics-based models perform consistently in either combat or OOTW model applications with no differences in performance characteristics. For example, a model of an assault rifle maintains the integrity of the physical representation and physics of the weapon systems in either domain. Conversely, behavioral models may not perform consistently in either combat or OOTW model applications without noticeable differences in performance characteristics. For example, a model of human behavior for a combat model application cannot be federated with or integrated into another model of an assault rifle in the OOTW model application without altering the context and purpose of the physics model.

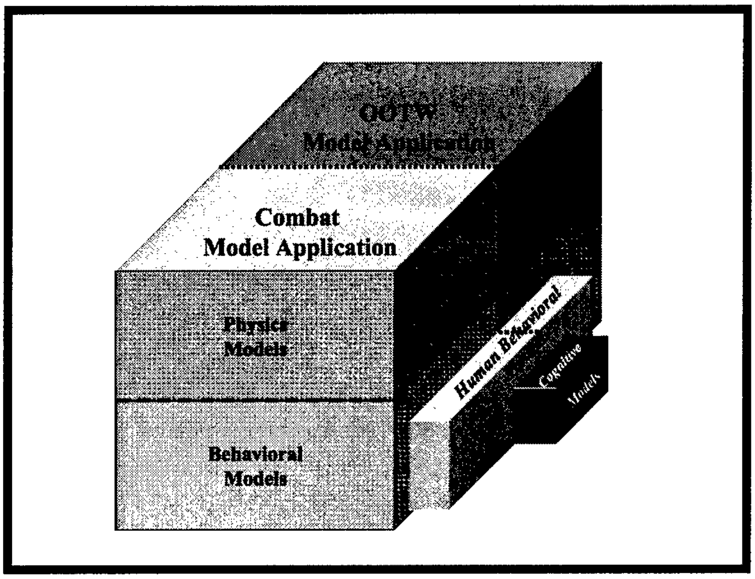


Figure 2. DoD Modeling and Simulation Landscape

3.1 Verification, Validation and Accreditation

Verification, validation, and accreditation (VV&A) are important to ensure that models and simulations are ready for use.[5] Verification and validation are generally conducted concurrently, with accreditation always being the final step in the process (Department of Defense Directive 5000.59, 4 January 1994). Verification ensures model code and algorithms accurately represent the real-world processes or objects modeled (Department of Defense Instruction, 5 October 2001). The Department of Defense Modeling and Simulation Office (DMSO) VV&A Technical Working Group (TWG) defines validation as "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" (Department of Defense Instruction, 5 October 2001). Accreditation is an "official" seal of approval that the designated authority bestows on a model that confirms that the model has been properly verified, validated, and accredited for an intended purpose, application, and scenario.

Figure 3 depicts the iterative sequence of steps involved to VV&A DoD models and simulations. The process begins with identifying, defining, and scoping the problem. Next, an appropriate modeling and simulation method must be selected that is relevant to the purpose of the study and one that generates the right data for the decision-making process. Then a M&S plan is developed for building, verifying, validating, accrediting, and using the model. The model user must decide whether to use a legacy model as is, develop a new model, or federate multiple models together into a family of models.

[5] For reference, key players involved in model VV&A for DoD use are provided in Appendix C.

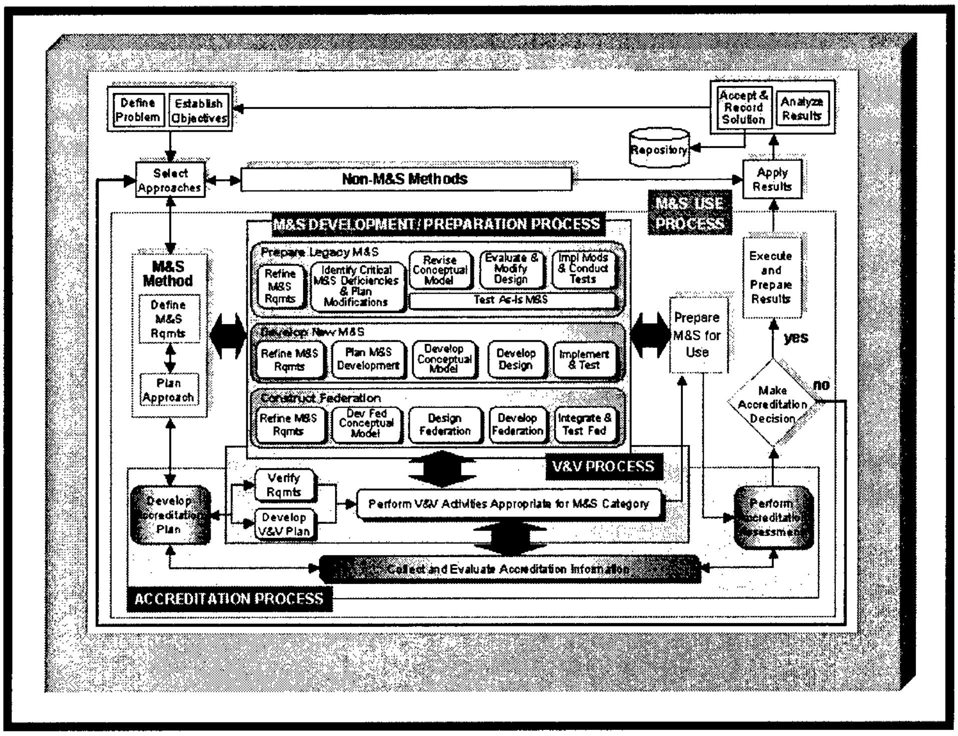


Figure 3. Modeling and Simulation Problem Solving Process From (Department of Defense Modeling and Simulation Office, 15 August 2001c)

The verification and validation (V&V) process begins with a V&V plan that outlines V&V tasks given the type of model under construction. Specific requirements, tasks, and steps depend on the plan for building the model, the type of simulation the model is either integrated into or federated with, and the model's intended use (Department of Defense Modeling and Simulation Office, 15 August 2001c). Although V&V is required for both virtual and constructive simulations, it is understood, and common practice, to tailor V&V tasks to meet the unique needs and limitations of the model. DMSO's "Key Concepts of VV&A" list the following key V&V tasks (Department of Defense Modeling and Simulation Office, 15 August 2001c):

- Verify User Requirements;
- Develop a V&V Plan; and
- Perform the V&V Procedures Suitable for the Model's M&S Category: Validate Conceptual Model; Verify Model Design; Verify Model Implementation; and Validate Model Results.

Although DMSO's key tasks do not address requirements to validate data used to build and to test a model, the Modeling and Simulation (M&S) community recognizes it is not possible to validate a model without test data to produce verifiable simulation results. For this reason, validating agents normally perform validation three times: referent, conceptual model, and model implementation (Department of Defense Modeling and Simulation Office, 15 August 2001a). [6] [7]

Figure 4 illustrates where the validation steps fit into Birta and Özmirak's model validation framework (Birta, et al, January 1996). Birta and Özmirak do not explicitly address the roll of referent in their design of model development and model testing, however, referent is integrated into the diagram to show where it is created, validated, and used in model validation. Model implementation validation is the result of comparing simulation outcomes with real-world results under specific controlled conditions.

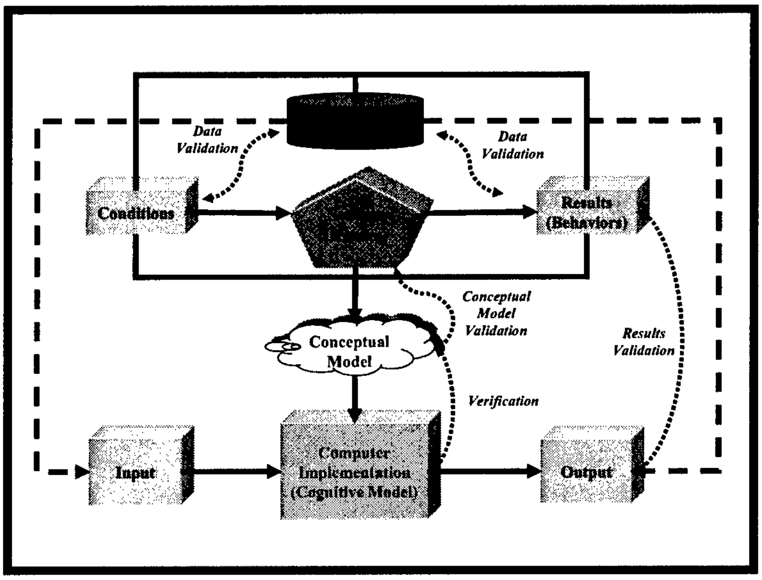


Figure 4. Birta and Özmirak Validation/Verification After (Birta, et al, 1996)[8]

[6] Validation agents are persons or organizations responsible for conducting validation of a model, simulation, or federation and supporting data (Department of Defense Modeling and Simulation Office, 26 July 2002).

[7] Referent is the "codified body of knowledge about a thing being simulated" (Harmon, 16 December 1998) (Department of Defense Modeling and Simulation Office, 27 June 2002c).

[8] The original process proposed by Birta and Özmirak is modified in this document to reflect terms consistent with this research.

Data for the referent comes from many sources. One of these sources is validated models. Examples include models of specific aspects of human behavior, sociological phenomena, and the physiological processes underlying human behavior. Referent is also collected from validated simulations of human behavior (live, virtual, or constructive), empirical observations of actual operations, historical case studies, experimental data, and from SMEs (Department of Defense Modeling and Simulation Office, 25 September 2001). Data also comes in various formats such as narrative, numerical, or tabular. Due to the source and nature of a referent required to build, validate, and operate models, numerous techniques exist for validating the referent.

Table 1 lists five techniques used for validating referents. Validating agents may use combinations of these techniques to provide a more comprehensive validation. Table 1 identifies when it is most appropriate to use each technique from past M&S validation efforts.

Table 1. Steps in Verification and Validation Process Where Comparison Techniques Best Apply After (Department of Defense Modeling and Simulation Office, 30 November 2000b) [9]

Comparison Technique Class	**Validation Process Step**
SME Assessments	Conceptual model, data & face validation[10]
Audits, Inspections & Walkthroughs	Conceptual model & data validation
Visual Comparisons	Data & face validation
Analytical Comparisons	Conceptual model & data validation
Formal Comparisons	Conceptual model, data & face validation

Table 2 presents a list of comparison validation technique limitations identified by DMSO. Limitations of comparison techniques illustrate an important aspect of validation plans and referents. Model requirements and specifications must be detailed and unambiguous. If they are not, the use of SMEs, auditors, and inspectors results in an unfocused validation effort. Comprehensive and explicit requirements and specifications scope the problem making model validation more manageable; however, they also focus

[9] Knowledge base validation and other forms of complex data which the conceptual model may not represent fall under the term **data validation** (Department of Defense Modeling and Simulation Office, 30 November 2000b).

[10] The original table labels face validation as results validation. Face validation is used in the dissertation to maintain consistence in terms.

the validation making it difficult to abstract the results and accredit the model for use in other domains.

Table 2. General Limitations of Different Comparison Techniques From (Department of Defense Modeling and Simulation Office, 30 November 2000b)

Comparison Technique Class	**Limitations**
SME Assessments	• SMEs should be available & properly prepared • All information should be understandable to SMEs
Audits, Inspections & Walkthroughs	• Teams should be properly composed, available, and prepared • Sufficient information should be available for review sessions
Visual Comparisons	• Information should lend itself to meaningful visualization • Visualizations should be scaled correctly
Analytical Comparisons	• Referents and requirements should be described in forms that permit comparison with model or simulation representations (e.g., UML)
Formal Comparisons	• Information should take a formal, usually quantitative, form • Uncertainties may need to be described but should absolutely be understood

Inconsistent or skewed data display can introduce a scaling effect when using *visualization comparison* techniques. This can distort validation results by exposing SMEs to perception bias.[11] Placing the data in proper perspective is often difficult and current technology limits the use of this technique. Therefore, validating agents normally use visualization comparison in conjunction with at least one other method validation technique. The degree of rigor and extensive resources required to use *analytical comparison* techniques make them less attractive than more informal techniques, however, they are excellent for validating conceptual models and knowledge bases due to their ability to investigate the composition and causality of models and simulations. Strictly defined specifications for extracting data used in *formal comparison* techniques make them the preferred means of verifying and validating a physics-based model's knowledge base, conceptual model, and results. However, the rigorous characteristics of this method limit the technique's applicability due to the time and money required to collect large amounts of data.

To assist the military M&S community with VV&A, the DoD has developed a series of instructions, regulations, and publications. Verification and validation

[11] Performance bias is defined in Subsection 3.6.2. Bias.

procedures set forth in DoD and the three Services outline policies, assign responsibilities, prescribe general procedures, and provide a list of standard products required for accrediting a model. The documents do not provide a fixed set of procedures or a set of referent to validate models. The procedures follow the general phases outlined in Figure 5 and listed in detail in Appendix C. (Key Players in Verification, Validation and Accreditation).

In Figure 5, the clouds represent inputs into the system. *User objectives* help model developers characterize the requirements for the model. For example, an artillery battalion needs to have a cognitive model integrated into a new automated call for fire trainer (objective). The automated fire direction center (FDC) would need to interpret verbal calls for fire from forward observers (FO) (requirement). *Requirements* help developers filter through the *available referents* to identify the relevant referent(s) for use in developing algorithms and validating the final model. Developers do not use all referents during the development and initial testing of the model. Developers often place some referents aside for validation runs of the model. Examples of possible referents are the ability to receive calls for fire, how to parse and evaluate call for fire messages, allocation of indirect fires, and time required to process a call for fire. *System information* provides insight to developers about the physical system or processes. Examples are weapon systems utilized, amount and type of ammunition available, ballistics of the ammunitions, and ammo/target pairings. Referent provides inputs for algorithms developed from the system characteristics to produce results. Validating agents compare these system results against the requirements using the validation referent. The final product is a set of documents that describe how well the codified model's results match the selected test referent.

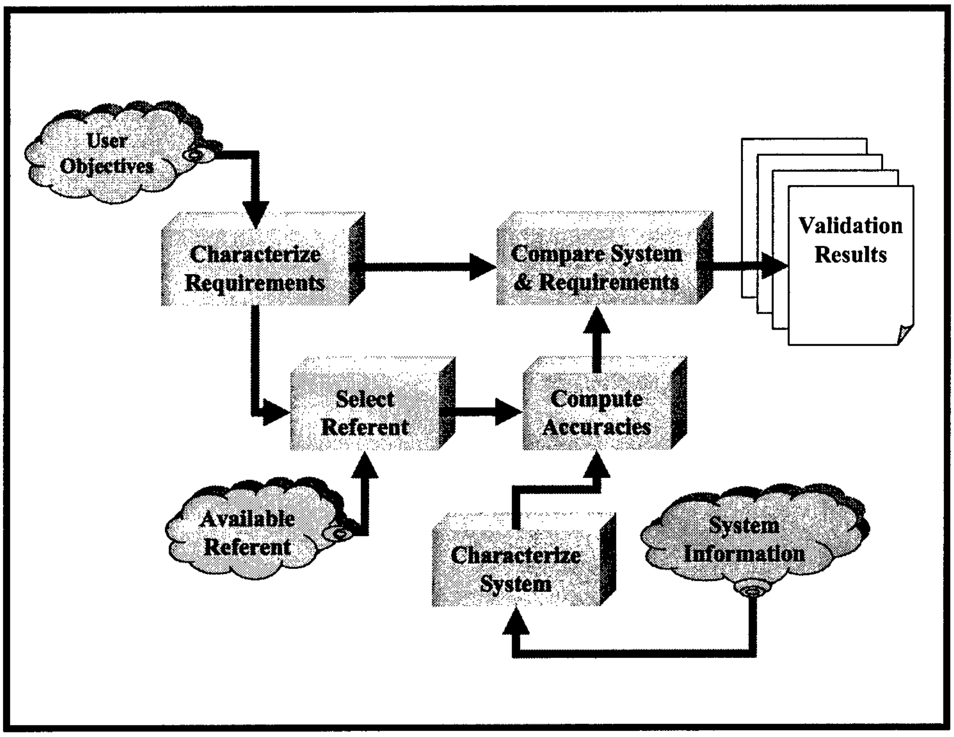


Figure 5. Essential Steps for Validating Models and Simulations From (Department of Defense Modeling and Simulation Office, 30 November 2000b)

One of the most difficult phases of this process is the identification, collection, and selection of suitable referent to develop and validate the model. Table 3 presents four categories of information required for model validation and their associated sources. DMSO identifies SMEs as sources for three of the four categories. One of three is referents (Department of Defense Modeling and Simulation Office, 30 November 2000b).

Table 3. Requirements and Sources of Validation Information From (Department of Defense Modeling and Simulation Office, 30 November 2000b)

Validation Information Requirement	Information Sources
Requirements	SMEs, other user representatives, user documentation (e.g., concepts of operations)
Referents	SMEs, existing system documentation, experimental data, analysis and study reports
Model/Simulation	Conceptual model, design documentation, development team members
Comparison Techniques	Recommended Practices Guide (RPG), technical papers, SMEs

3.2 Psychology

The focus of psychology is the study of the representation and processing of information by complex organisms. It most often deals with species that process information in an intelligent manner. Intelligence implies the ability to obtain and process information in a manner that allows the organism to select behaviors with the best chance of "achieving the fundamental goals of survival and propagation" (Wilson & Keil, 1999). Previously, psychology focused on processing information amid sensory inputs and motor actions. Since psychologists consider humans "capable of the most complex and most domain-general forms of information processing," most psychology research focuses on the nature of human intelligence and information processing (Wilson & Keil, 1999).

One can see the diversity of psychology in its many fields or areas of interest. Behavioral psychology, cognitive psychology, cross-cultural psychology, and ecological psychology are four of these fields.[12]

Behavioral psychology deals with the study of overt responses to stimuli. Its focus is on overt responses to stimuli rather than on the mental processes. This focus failed to provide reasons for diversity in human behavior and neglected to account for elements such as "memory, attention, consciousness, thinking, and imagery" (Solso, 2001). In many cases, behavioral psychology rejected the theories of "mentalistic" (Wilson & Keil, 1999).[13] Previously, behavioralists attempted to operationally define these internal functions of the brain and roll them into a more general study of the mind (Solso, 2001). Although less popular than other areas of psychology, behavioral research continues today using many tools utilized by the natural sciences (Wilson & Keil, 1999).

Cognitive psychology focuses on the scientific study of the human mind (Wilson & Keil, 1999). A cognitive psychologist studies how an individual or a group of individuals reasons through a problem. In doing so, the psychologist is concerned with

[12] Additional fields of psychology include: clinical psychology, comparative psychology, developmental psychology, personality psychology, and social psychology.

[13] Mentalistic refers to processes that are mental in origin (e.g. general knowledge, situational awareness, intent/goal, commitment, etc.) rather than physiological or physical (Shoham, 1993).

perception, thought, and memory. Perception of knowledge deals with how an individual obtains information from the environment. Thought is concerned with how one solves problems and executes thoughts or relays thoughts to others. Memory involves the storage, retrieval, and processing of the information by the human brain. The domain of cognitive psychology is vast, covering as many as twelve principle areas: attention, cognitive neuroscience, consciousness, developmental psychology, human and artificial intelligence, imagery, language, memory, pattern recognition, perception, representation of knowledge, and thinking and concept formation (Solso, 2001). [14]

Ecological psychology research deals with how an organism's behavior is based on its perception of the environment. This includes the shapes of objects, movement and change of objects, the organism's state and movement through the environment, and the organism's ability to influence the environment through effective actions. These perceptions differ for each organism. This is due to the ability of each organism to sense its environment and construct its own mental map of the world (Wilson & Keil, 1999).

This is similar to how situational awareness or mental maps depict an individual's perception of the world. Situational awareness refers to a person's perception of the world based on sensory inputs, memories, and mental possessing. One's situational awareness effects the actions one takes. Because of this, many cognitive models include a situational awareness module. Shattuck and Miller have been conducting research to

[14] *Attention* is concerned with the ability to simulate input and/or process events stored in memory. *Cognitive neuroscience* is a study of how the mind-brain works at the level of the neuron. *Consciousness* deals with one's awareness of his/her internal or external conditions. Often deemed its own domain of psychology, some consider *developmental psychology* a subset of cognitive psychology. As discussed earlier, developmental psychology deals with how human behavior develops/changes over time. *Human and artificial intelligence* deals with recognizing and defining human intelligence so model developers can replicate it using a computer model. *Imagery* focuses on the mind's ability to take physical images to create a mental map from which the individual develops ideas and translates them into meaningful actions. The study of how humans learn and use *language* is often regarded as a subfield of developmental psychology. It concerns itself with the meaning of gestures and body posture as well as the written and spoken word. The field of *memory* research is involved with studying how the mind processes and stores events in short-term, working, and/or long-term memory. *Pattern recognition* is the study of how sensory inputs are grouped together to form recognizable patterns that are interpreted as a meaningful representation of information to be stored or retrieved from memory. *Perception* deals with "the detection and interpretation of sensory stimuli." (Solso, 2001). It attempts to determine how an individual takes sensory input and creates features and objects, categorizes and classifies these features and objects to develop a perception of the world. How information is represented, stored, and processed by the mind is the focus of knowledge *representation*. *Thinking and concept formation* is concerned with how thoughts and concepts are generated, confirmed, and modified.

address the effects of situational awareness on decision makers to determine measures of effectiveness for assessing the impact of systems designed to provide information for commanders to develop their situational understanding of the combat environment. (Miller & Shattuck, 2004)

Cross-cultural psychology "observes human behavior in contrasting cultures" where a culture is widely defined but routinely seen as pertaining to "patterns of behavior, symbols, and values" often transmitted over time (Gale Group, 2001). This field of psychology asserts that the environment where an individual spends a great deal of time plays a dominant role in the behavioral patterns of an individual. These patterns can influence everything from an individual's ability to extract information from symbols to how they perceive technology in general.

Cross-cultural distinctions can be large or small in scope. Psychologists consider global cultural characteristics based on environmental regions, religions, or systems of government as factors for cross-cultural studies; however, cultures can be even smaller. Examples of smaller cultural communities are branch of service (Infantry, Armor, Aviation, etc.) or unit type (light infantry, mechanized infantry, motorized infantry, or special operations). Psychologists may also use technology as a means of distinguishing cross-cultural characteristics. For example, categorizing behavior patterns based on three forms of technology exposure: Those who have never used computer technology, those who recently transitioned to the use of computer technology, and those raised with computer technology integrated into nearly every aspect of their daily lives. Prensky refers to these last two groups as digital immigrants and digital natives, respectively (Prensky, 2001).

Understanding the varied fields of psychology allows us to investigate the impact of the various perspectives offered by the different fields within psychology on HBR models. The procedural aspect of behavioral psychology can be seen in many of the rule-based implementations of modern HBR models where models abstract responses based on stimuli with limited consideration for the thought process behind those decisions. One can also see this abstraction in the use of face validation techniques to validate the overt results of HBR models.

The cognitive psychologist Wilhelm Wundt heavily used *introspection* in the 1880s and 1890s. His method required trained observers to analyze "their own thought processes as they performed various cognitive tasks" (Wilson & Keil, 1999). This self-analysis often lead to biased results, skewed towards how observers were prone to hypothesize. Because of its inconsistencies and apparent lack of objectiveness, many psychologists viewed introspection as "unscientific." Behavioral psychologists were some of the first psychologists to rebuke introspection techniques as a valid means of collecting data (Russell & Norvig, 1995). Since the 1930s, its use in the field of psychology for collecting information has been limited (Wilson & Keil, 1999).

Today, research personnel use a modified version of introspection, cognitive task analysis, to collect information about a specific domain. However, instead of using observers trained in the field of psychology as sources of information, SMEs are the source of information and psychologists collect the data. As with introspection, biases may impact data collected, however, this bias is based on preferred techniques of SMEs, SMEs developing cognitive maps that differ from the facts presented, and the training effect of SMEs reviewing numerous tasks and scenarios.

HBR models, as used by psychologists, are tools to represent observations and assumptions of how the mind works. Psychologists use HBR models to explain a specific theory, further research in cognitive psychology, and study complex concepts of storage, retrieval, and processing of memories. HBR models help to develop hypotheses and make behavioral predictions. One of the most famous and simplistic cognitive models is Waugh and Norman's model of human memory (Figure 6).

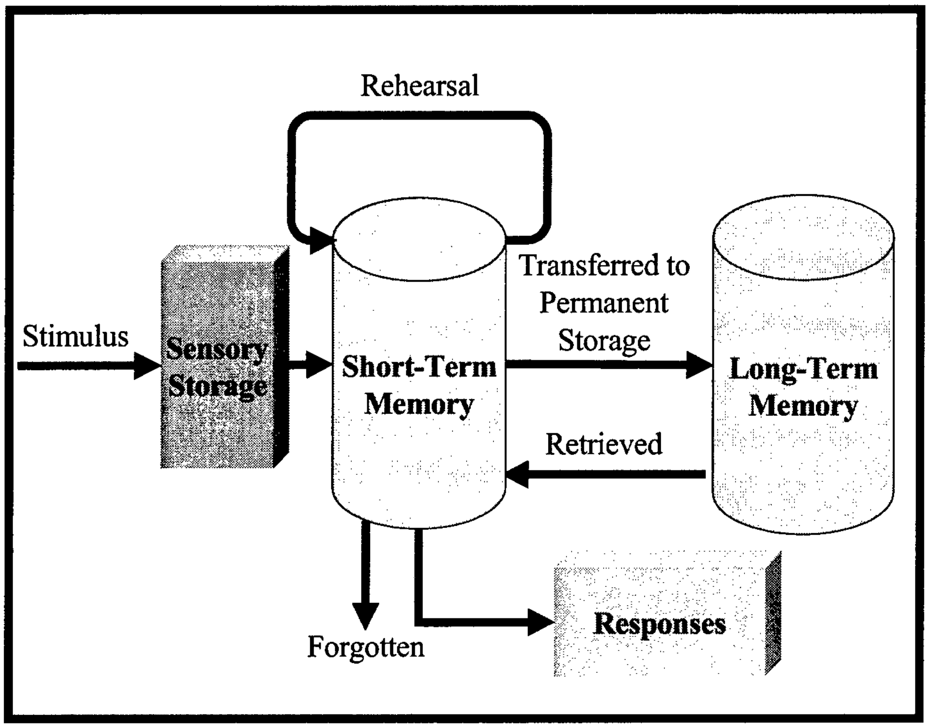


Figure 6. Waugh and Norman's Model of Human Memory From (Solso, 2001)

Many cognitive architectures use variations of this human memory model to represent the storage and retrieval of facts. Understanding this theory may lead to better techniques for validating HBR models as we identify the types of information stored in each section of the model, when a segment of memory is accessed to make decisions, and when memories are lost or are inaccessible. Other constraints may limit the search for an optimal decision where the decision maker abandons or bypasses more formal thought processes to quickly select a plausible solution.

One can see a commonality between ecological psychology and the manner in which military decision makers address situations based on a leader's prior assignments. Lessons learned and techniques used in previous assignments may lead decision makers to recognize certain enemy behavior patterns and select a behavior to address the perceived situation. Examining research techniques used in the field of ecological psychology may provide insight into new methods of identifying ways to represent situational awareness in HBR models, the fusion of information, and presentation of the common operating picture in combat simulations.

Another issue is SME bias based on ecological and cross-cultural influences. This is present in both the development of a HBR model and the collection of referents when using SMEs. This bias discounts possible options based on the way people were raised and trained to think, the region of the world an individual was reared, and other cultural influences which affect an individual's performance. An example of such influence is the value people place on a human life. Cultures who place a relatively higher value on a single life may not consider the option of using suicide bomber(s). On the other hand, a culture which values the well-being of the majority over a single life may see suicide bombers as a viable option to its current dilemma. The reasoning processes of individuals in each culture may lead to seemingly dissimilar behaviors.

3.3 Cognitive Models

As with model taxonomies, cognitive models can be described at three different levels: representations, architectures, and implementations.

3.3.1 Representations and Architectures

As stated earlier, cognitive models deal with the human decision-making process. Cognitive model representations provide a means of describing different methodologies for representing codified cognitive functionality. Codified cognitive modeling has been the focus of two major communities over the past fifty years, artificial intelligence and artificial life.

The artificial intelligence (AI) community has numerous goals but in general, the focus has been on comprehending intelligent computerized entities (Russell & Norvig, 1995). The techniques used by the AI community generally involve a top down approach requiring an attempt to codify all relevant behavioral details (Ralston, Reilly, & Hemmendinger, 2000). These techniques use inductive and deductive reasoning to identify and codify entities to display rational behavior (correct actions) (Russell & Norvig, 1995). The emergent field of artificial life (AL) attempts to model the behavior of biological systems (Freedman, 1999). The AL community uses a bottom-up approach to identify and codify characteristics in computer entities allowing entities to evolve and emerge to perform intelligent actions. The focus of AL is emergent behaviors of entities as they attempt to survive in complex environments (Ralston, Reilly, & Hemmendinger, 2000).

The two communities have developed numerous techniques for implementing their approaches. Some of these techniques fuse the boundaries between AI and AL, such as multi-agent systems, while others are contained primarily in one domain. Examples of cognitive model representations are Agent-Based, Bayesian-Network, Multi-Agent System, Neural-Networks, and Rule-Based.

Agent-Based representations demonstrate intelligence through codified objects that perceive characteristics of the environment and act on those perceptions (Russell & Norvig, 1995). There are several types of agent-based cognitive architectures. Two of these are reactive and rational agents.[15] A reactive agent bases its actions solely on the last set of sensory inputs. Often the approach uses a simple condition-action rule (e.g., this is my perceived state of world; I choose this action). A rational agent uses sensors to perceive its environment and performs actions on the environment using effectors. Rational agents maintain a state of situational awareness based on their past knowledge of the world and current sensory inputs (Russell & Norvig, 1995).

The *Multi-Agent System* (MAS) is a relatively new representation for replicating behaviors based on the Complex Adaptive System (CAS) theory. Developed in the late 1970s, MAS is a system with autonomous or semi-autonomous software agents that produce adaptive and emergent behaviors.[16] The model uses a bottom-up approach where software agents have independent micro-decisions that generate group level macro-behaviors. A MAS can use any form of agent-based software technology (reactive, rational, goal-based, utility-based, etc.) with the agents characterized as possessing intentions that influence their actions. Multi-agent systems are used in large domains were non-linearity is present (Holland, 1995). The MAS, limited only by the physics constraints of the simulation boundaries, uses an indirect approach to search the large

[15] Russell describes agents as three types: reflex agents or reactive agents, goal-based agents that attempt to achieve a specified goal, or utility-based agents that attempt to achieve the best possible state from their point of view (Russell & Norvig, 1995).

[16] Adaptive behavior is the process of fitting oneself to the environment. A MAS generates emergent behavior at a higher cognitive level based on the behaviors and interactions of agents at a lower level. Schelling describes this as micro decisions leading to macro behaviors (Schelling, 1978).

domain for viable results. Another feature of MAS is its ability to allow agents to evolve to create new agents which, in general, are more optimized to survive/thrive in the simulated environment (Ferber, 1999). If coded with a *brain lid*, one can interrogate agents for the reasoning behind their actions as well as view their overt behaviors (Lewis, Zyda, and Hiles, 2002).[17] Examples of MAS are the Irreducible Semi-Autonomous Adaptive Combat (ISAAC), Pythagoras, Socrates, Enhanced ISAAC Neural Simulation Toolkit (EINSTein) and Map Awareness Non-uniform Automata (MANA) (Ilachinski, 1997) (Project Albert Fact Description, 10 December 2002).

Cognitive model *architecture* is the framework for establishing how the components of the cognitive model relate to each other. Cognitive model architectures use one or more cognitive model representations to structure the schema behind a specific cognitive model. An architecture is not a functioning model implementation, but the design for an implementation. Examples of cognitive model architectures are the Adaptive Control of Thought (ACT-R), COGnition as a NETwork of Tasks (COGNET), Connector-Based Multi-Agent System (CMAS), Executive-Process Interaction Control (EPIC), and State, Operator And Result (Soar). Table 4 indicates some of the means by which these architectures can provide information to explain their actions. Each architecture can demonstrate its overt behaviors, but most are limited to their ability to provide information about the specifics behind the cognitive processes they used for their behavior selection.

[17] Programmers code a ***brain lid*** into an agent to allow inspection of the agent to determine its situational awareness and decision processes leading to a specific action (Roddy & Dixon, 2000).

Table 4. Model Architecture Action Information Sources After (Pew & Mavor, 1998) (Osborne, September 2002)

Model	Information for Action Explanation
ACT-R	▪ Overt Behaviors ▪ Encoded Knowledge ▪ Encoded Rules ▪ Decision Stack ▪ Declarative knowledge used ▪ Changes in working Memory ▪ Final Parameters ▪ New Rule & Productions ▪ New Declarative Memory
CMAS	▪ Overt Behaviors (Actions) ▪ Goals ▪ Tickets (Possible Actions to achieve a specific goal) ▪ Outer Environment (State of the model) ▪ Inner Environment (An agents Situational Awareness) ▪ Entity State ▪ Connectors (Possible entity interactions)
COGNET	▪ Overt Behaviors ▪ Conditions/ Rules ▪ Blackboard (Situational Awareness)
EPIC	▪ Overt Behaviors ▪ Encoded Knowledge ▪ Encoded Rules
Soar	▪ Overt Behaviors ▪ Encoded Knowledge ▪ Decision Stack ▪ Knowledge Stack

3.3.2 Implementations

A cognitive model implementation takes a generic cognitive model architecture with its supporting cognitive model representation(s) and provides code and data for each component. An implementation is a functional representation of the architecture.

Ilachinski created the *Irreducible Semi-Autonomous Adaptive Combat* (ISAAC) model in 1997 for the U.S. Marine Corps to investigate the utility of agent-based systems. One of the goals of ISAAC is to show that land combat can be modeled using a CAS. As an implementation of AL, ISAAC introduces dynamic emergent behavior in an attempt to overcome shortcomings of Lanchester-type combat models. (Ilachinski, 1997) As an AL implementation, ISAAC exhibits the effects of a model with no central control; the interaction between autonomous or semi-autonomous entities often produces unpredictable outcomes. The model attempts to fill some of the perceived gaps between the current needs of the M&S community and the shortcomings of previous HBR implementation to represent dynamical human behaviors.

The model uses agents with four properties to generate believable behavior:

- Embedded "doctrine" is a default set of local-rules used to specify how an agent is to act in a generic environment
- A "mission" is a goal directing behavior
- "Situational awareness" results from sensors generating an agent's internal perception of the environment
- Behaviors and/or rules are altered through an internal adaptive mechanism (Ilachinski, 1997)

The system can run in an evolutionary mode utilizing a genetic algorithm to increase an agent's ability to survive.[18] Using the evolutionary mode of operation, ISAAC has shown an impressive catalog of emergent behaviors. This list includes the ability to perform a frontal attack, local clustering, penetration, retreat, containment, flanking maneuvers, and encirclement of the enemy (Ilachinski, 1997).

The *Map Awareness Non-uniform Automata* (MANA) model is another model in the Marine Corps Combat Development Command's (MCCDC) Project Albert. Project Albert is the Marine Corps' research effort to assess the general applicability of the use of CAS to study land warfare. Other HBR models in Project Albert include Pythagoras, Socrates, and ISAAC (Project Albert Fact Description, 2001) (Project Albert Fact Description, 10 December 2002).

The Defence Technology Agency of New Zealand developed MANA to conduct research into implications of chaos and complexity theory for combat and other military operational modeling.[19] MANA is an agent-based representation developed based on Enhanced ISAAC Neural Simulation Toolkit (EINSTein) and its precursor ISAAC.

As with other agent-based models (ABM), MANA consists of entities controlled by decision-making algorithms. The model's developers further classify MANA as a CAS. MANA's entities represent military units which make decisions based on a "memory map" which provides individuals or entities with goals to guide them about the battlefield.

Some of the aspects that allow MANA to be designated as a CAS are:

- MANA has the ability to exhibit "global" behavior, materialized based on local interactions;

[18] A genetic algorithm searches the collection of individual agents to find the agent that maximize the fitness function and then uses the agent(s) to produce new agents. The fitness function takes the agent as an input and delivers a numerical output based on the agent's internal state and resulting performance function. A fitness function can be derived from anything configurable as an optimization problem (Russell & Norvig, 1995).

[19] The following description of MANA is drawn directly from the *MANA, Map Aware Non-uniform Automata, Version 3.0, Users Manual (Draft)* (Galligan, Anderson, & Lauren, 2003).

- MANA uses feedback to update agents regarding changes to the environment;
- MANA cannot be analyzed by decomposing it into simple independent parts; and
- Similar to human behavior, agents "adapt" to their local environment and interact with each other in a non-linear manner.

MANA has the ability to incorporate several additional features which ISAAC did not have when MANA was initially developed. These include:

- Shared memory of enemy contacts provides agents with enhanced situational awareness. MANA uses two mechanisms to provide situational awareness, "squad map" and "inorganic map". The "squad map" maintains group contact data. The "inorganic map" stores contacts based on communications from other units.
- Communications exists between units in order to pass contact information. The model can alter information accuracy based on the influence of unit activities and environmental conditions on communications.
- *Terrain Maps* contain features such as roads which increase agent speed and undergrowth which agents can use for concealment.
- The use of waypoints for routes provide intermediate goals to facilitate coordination of units and achievement of an ultimate goal.
- Agent personalities can be event-driven. Events (e.g., making enemy contact, being shot at, engaging others, reaching a waypoint, etc.) can activate a special personality trait, present for a limited amount of time or until modified by another event. Personality changes can be set individually or for an entire unit.

MANA divides its parameters into four categories: personality weightings, move constraints, basic capabilities, and movement characteristics. *Personality weightings*, determine an automaton's propensity to move towards friendly or enemy units, towards its waypoint, towards easy terrain, and towards a final goal point. Next, *move constraints* act as conditional modifiers. An example of a modifier is the "Combat" parameter, which determines the minimum local numerical advantage a group of agents needs before approaching the enemy. *Basic capabilities* describes an agent based on its use of weapons, its use of sensors, its movement speed, and its tendencies for interaction with other agents. Finally, *movement characteristics* of the agents, include the effects of terrain on agent speed, the degree of random agent movement, and agent's desire to avoid obstacles (Galligan, Anderson, & Lauren, 2003).

3.4 Human Behavior Representation

Human behavior representations (HBR) model human behavior at one of four levels: combined organizations, organizations, individuals, or components of individual performance. They may represent one or more cognitive functions such as perception, inference, planning, or control. HBRs can also portray the effects of behavior modifiers: stress, injury, fatigue, discomfort, motivation, and emotion. They often have human performance restrictions such as decision latencies or bandwidth allocated for sensing (Department of Defense Modeling and Simulation Office, 25 September 2001).

Within DoD M&S, HBRs are referred to as one of the following:

- Automated FORces (AFOR),
- Command FORces (CFOR),
- Computer Generated Forces (CGF),
- Semi-Automated Forces (SAF and SAFOR), or
- Synthetic forces (Department of Defense Modeling and Simulation Office, 25 September 2001).

3.4.1 Human Behavior Representation Verification and Validation Procedures

Although the purpose and implementation of physics-based and HBR models are fundamentally different, the V&V processes are the same. The validating agent must evaluate the capabilities of the physics-based and HBR model at four discrete phases. Figure 7 is a graphical depiction of the four phases of model development and the high-level validation tasks that DMSO defines as necessary for a validation agent to perform a comprehensive validation of a an HBR model: (1) conceptual model design; (2) contents of the knowledge base; (3) implementation of the model and its knowledge base; and (4) integration of the model into the simulation. The degree to which the validating agent can validate a model in each phase is dependent on the model representation. Representations such as neural networks can only undergo face validation due to the complexity of the underlying model, which validating agents often treat as a "black box" (Department of Defense Modeling and Simulation Office, 25 September 2001). Within the four phases, HBR VV&A requires the completion of several high level tasks is essential:

(a) Collecting a complete a set of requirements and acceptability criteria;

(b) Identify referents for in assessing the HBR's validity;

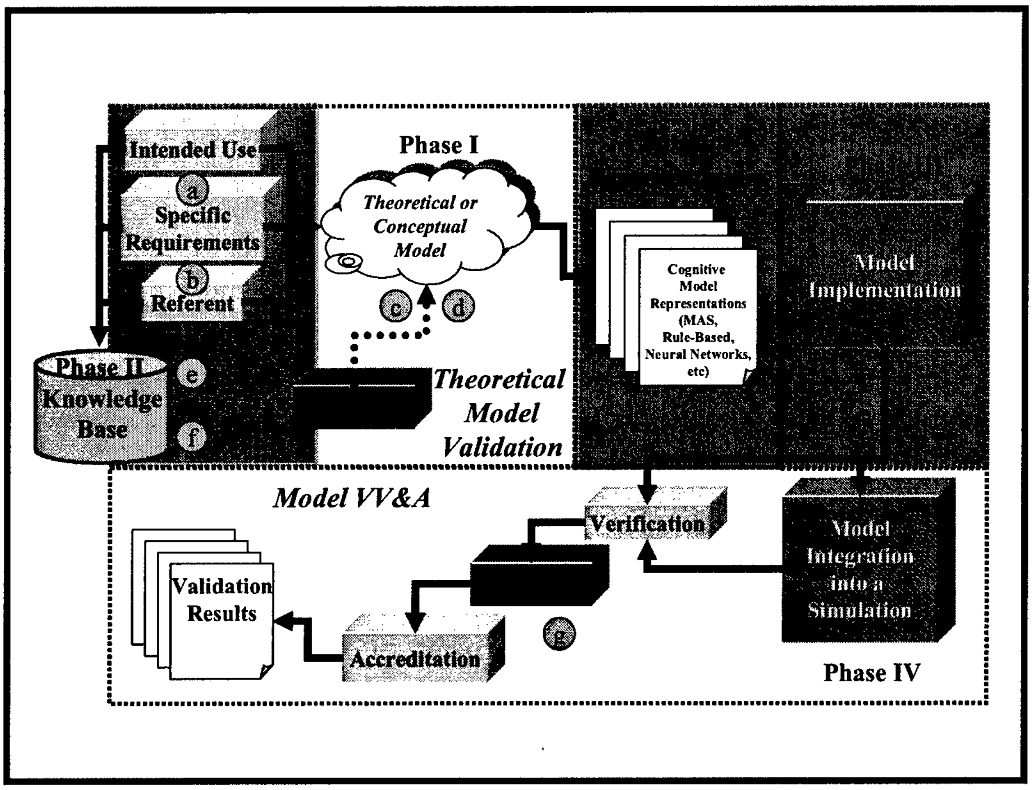
(c) Validate conceptual model against the requirements using the referents;

(d) Analyze conceptual model to identify areas of high complexity to focus model implementation validation efforts;

(e) Validate knowledge base against requirements using referents;

(f) Analyze knowledge base to identify areas of high complexity to focus model implementation validation efforts; and

(g) Validate integrated HBR implementation against requirements using referent and concentrating on key areas identified during the conceptual model and knowledge base analysis (Department of Defense Modeling and Simulation Office, 25 September 2001).



Phase I
Intended Use
a
Specific Requirements
b
Referent
Theoretical or Conceptual Model
c
d
Phase II Knowledge Base
e
f
Theoretical Model Validation
Cognitive Model Representations (MAS, Rule-Based, Neural Networks, etc)
Model Implementation
Model VV&A
Verification
Model Integration into a Simulation
Validation Results
Accreditation
g
Phase IV

Figure 7. Verification, Validation, and Accreditation Tasks for a Human Behavior Representation Model After ([20])

Prior to use, the model must be validated. For physics-based models, this normally includes completing a proof and conducting a numerical validation of the model. For HBR models, SMEs normally perform the less quantifiable and more qualitative method of face validation on the conceptual model to determine if the model has any major theoretical faults based on the current understanding of the human thought process. This research assumes the cognitive architecture is valid, and focuses on face validation of the coded implementation of the HBR model.

[20] See (Department of Defense Modeling and Simulation Office, 2000b) (Department of Defense Modeling and Simulation Office, 25 September 2001b) (Department of Defense Modeling and Simulation Office, 25 September 2000c)

3.4.2 Referent Categories

There are numerous ways to categorize referents. DMSO's "Key Concepts of VV&A" section of its Recommended Practices Guide (RPG) describes six categories of correspondence useful for determining referent for HBR: computational correspondence, domain correspondence, physical correspondence, physiological correspondence, psychological correspondence, and sociological correspondence (Department of Defense Modeling and Simulation Office, 25 September 2001).[21]

Viewing the human mind as a machine made of an immense assortment of computational devices, *computational correspondence* addresses the ability of the human nervous system to take inputs, process the inputs, store information, retrieve stored information, make decisions, and produce outputs. Cognitive psychologists commonly accept that the brain performs these functions, however the physical specifics of how the brain performs these tasks is not well understood. However, psychological studies have identified bandwidth and storage limitations of the human brain for specific tasks. Validating agents have used this referent in conjunction with theories of brain computational performance to conduct limited validations of cognitive models (Department of Defense Modeling and Simulation Office, 25 September 2001).

Domain correspondence addresses the use of SMEs to examine the knowledge base and outcomes of human behavior in their specific area of interest. The data collected is normally qualitative and leads to referent viable for face validation. Researchers often equate this form of validation to a Turing Test (Department of Defense Modeling and Simulation Office, 25 September 2001). This referent is generally gathered from the research of behavioral psychology.

Comparing the results of physics-based models against human physical constraints is known as *physical correspondence* (Department of Defense Modeling and Simulation Office, 25 September 2001). This referent is normally limited to the more obvious physical constraints of the human body (e.g. how fast a human can run, how much a human can carry, etc.).

Physiological correspondence resembles data used to validate physics models. It uses information from neurologists, neurosurgeons, or physiologists to determine if a model's components react similar to the portion of the brain they simulate. This form of validation has become more viable over the last two decades due to advances in understanding of the

[21] Correspondence is the agreement of a model to different levels of abstraction.

physiology of the human nervous system. Physiological correspondence is an immature area of study but it has demonstrated use in validating neural networks (Department of Defense Modeling and Simulation Office, 25 September 2001).

The SME for *psychological correspondence* is the psychology professional. Similar to SMEs and domain correspondence, psychologists provide qualitative analysis of the real-world behavior and model results to determine if the model exhibits human-like behaviors. One can mine data to support psychological correspondence from the numerous volumes of experimental data on human performance in varying real-world scenarios (Department of Defense Modeling and Simulation Office, 25 September 2001).

Validating a model using psychological correspondence has potential issues with the qualitative nature of the referent and unintentional bias of the psychological experts, similar to that identified in introspection. However, psychological correspondence testing has the potential for greater credibility as the M&S and Psychology communities codify and validate more models of emotional phenomena. These validated models may provide baseline data and reduce the need for an exhaustive search of psychological problem space to identify appropriate referent. This shows most promise for models that incorporate aspects of stress and emotion (Department of Defense Modeling and Simulation Office, 25 September 2001).

For cognitive models of group behavior, *sociological correspondence* provides data on the interactions between groups and individuals. It includes groups operating under a unified organizational structure and unordered groups (crowds, mobs, etc.). An extensive body of knowledge exists from simulated and real-world situations from which one can acquire referent on sociological correspondence. The body of knowledge includes interactions between groups, between groups and individuals, and between individuals in groups. Sociological correspondence also has the luxury of well-established experimental protocols of sociological experiments to develop validation tests (Department of Defense Modeling and Simulation Office, 25 September 2001). This form of correspondence is closely related to cross-cultural psychology.

3.4.3 Face Validation

To date, the most common means of validating cognitive models has been through face validation using SMEs (Department of Defense Modeling and Simulation Office, 30 November 2000b). Often this technique uses a SME to exercise the HBR in a scenario where the SME manipulates the model through the simulation space by issuing orders or varying the stimulants, observing the resulting behavior, and determining whether the behavior meets a user's requirements for realism. SMEs often use personal opinions or qualitative referent provided by validating agents for face validation of HBR models (Department of Defense Modeling and Simulation Office, 25 September 2001).

Harmon and Metz propose new criteria for the validation of HBRs. They believe a strict level of validation for HBR models is idealistic. Harmon feels establishing a set of validation levels for the validation of an HBR would provide the M&S Community with a more meaningful and attainable validation process for HBR models (Harmon, 4 August 2003). Goerger, who concurs that a single validation standard for all HBR models is impractical, proposes a sliding scale of validation to indicate the flexibility of an HBR model (Goerger, 2002) (Goerger, 2003).

3.4.4 Subject Matter Experts

The Defense Modeling and Simulation Office VV&A TWG provides a list of general attributes individuals should demonstrate if they are to be used as SMEs (Department of Defense Modeling and Simulation Office, 30 November 2000a). These traits include independence, recognized competence, trust, good judgment, and perspective (Department of Defense Modeling and Simulation Office, 30 November 2000a) .[22] Pace and Sheehan feel these five traits fall short of providing standardization for SME certification. They propose more ridged guidelines for SME certification similar to those used by the judiciary system to classify individuals as expert witnesses. Such standards of excellence could help to ensure the legitimacy of a SME pool (Pace & Sheehan, 22-24 October 2002).

As described earlier, model developers use SMEs throughout the VV&A process to perform tasks such as collecting data, validating the knowledge base, validating the theoretical model, and validating the model implementation. The use of SMEs to perform face validation is analogous to the use of introspection. Despite the limited use of introspection in psychology, validating agents still use "behavior visualization techniques (which are similar to introspection,

[22] Independence suggests that a SME is impartial and can provide an "honest and probing assessment". A SME is one with the level of experience and knowledge of the subject matter and process to perform the task(s) the validating agent is asking him to execute. Trust is the "confidence that an SME has no hidden agenda detrimental to the simulation development." Good judgment indicates a SME can judge when he (or his team) has sufficiently examined the model to provide a proper assessment of its capabilities and limitations. Perspective is a SME's ability to maintain focus on the objective and limitations of the validation effort (Department of Defense Modeling and Simulation Office, 30 November 2000a).

because these techniques) can greatly help SMEs examine simulation results, particularly for simulations with which they (the SMEs) can interact." (Department of Defense Modeling and Simulation Office, 30 November 2000b)

3.4.5 Issues

Although preferred, formal validation is not always attainable. "Current state-of-the-art proof of correctness techniques are simply not capable of being applied to even a reasonably complex simulation model. However, formal techniques serve as the foundation for other V&V techniques." (Balci, 1997) Because multiple V&V agencies with non-standard criteria or non-uniform referent perform validation, validating agents inconsistently apply the validation process (Department of Defense Modeling and Simulation Office, 25 September 2001). This often leads to an invalid comparison of cognitive models due to the non-uniform means of validation and inconsistent validation efforts.

The high-level V&V tasks and issues with referents lead to other innate difficulties in validating human behavior models. DMSO has identified four factors, making validation of HBR models difficult. First is the very large set of possible actions for the simplest human behaviors. This makes it difficult to ensure complete consideration of all viable solutions. Second is the general non-linear characteristic of the constrained space of consideration. The non-linearity of the space prevents a simple causal relationship to be drawn between situational parameters and resulting actions. Next is the tendency of behavioral model developers to use stochastic algorithms in HBR models to demonstrate unpredictability. This 'unpredictable', unless it can be made deterministic, typically makes repeatable runs of the model impossible. Therefore, the model becomes difficult, often impossible, to validate. DMSO's fourth hindrance to validation is the chaotic behavior exhibited by HBR model implementations that are sensitive to initial and boundary conditions. Models with such sensitivity issues are limited to the breadth of their validation to the subset of scenarios where they exhibit stable behavior (Department of Defense Modeling and Simulation Office, 25 September 2001).

3.5 Validation Efforts of Human Behavior Models

Over the years, the M&S and psychology communities have developed numerous HBRs for a variety of purposes. The National Research Council conducted a study in 1988 to review the state of HBR and organizational modeling. One of the products of the study is a survey of validation efforts for many of the HBRs in existence or under development at the time. Table 5 summarizes and compares the different HBR validation approaches discussed in the study (Pew & Mavor, 1998).

Table 5 includes the domain for which each cognitive model was developed, the types of correspondence used for validation, and the sources of referents. Correspondence categories were limited to either domain, physiological, or psychological based on the techniques employed by validating agents at the time of the report. As stated earlier, domain and psychological correspondence gather their referents from SMEs. The use of SME-derived referents makes these two forms of validation subject to bias, frequently limited to qualitative data, and routinely resulting in face validation of the model. Models validated using more than one category of correspondence often focus on domain and psychological correspondence, which are typically limited to face validation of overt behaviors.

Table 5 illustrates the difficulties in comparing models based on their validation efforts since not all models are validated using the same techniques or correspondence. It also expresses the need for developing standardized procedures for the validation of HBR models to ensure model users provided more than a cursory review of the model prior to their use in a simulation. Finally, the table indicates the difficulty in collecting referents for each category of correspondence for use in developing and validating HBR models for different domains. While not the easiest data to collect, human performance data is definitely an area in which the DoD has focused a majority of its referent collection resources.

Table 5. Comparison of the Validation of Different HBRs from (Pew & Mavor, 1998)

Cognitive Model	Domain Types	Correspondences			Validating Data Sources
		Domain	Psychological	Physiological	
ACT-R	submarine TAO & Aegis radar operators	X	X		• human behavior data
COGNET	anti-submarine warfare	X			• human behavior data
EPIC	computer interaction tasks	X	X		• human behavior data
HOS			X		• validated theory
Micro SAINT	helicopter crew, ground vehicle crews, C2 message, tank maintenance & harbor entry operations	X			• human behavior data
MIDAS	757 flight crew	X			• human behavior data
Neural Networks			X	X	• validated theory • human behavior data
OMAR			X		• validated theory • human interaction
SAMPLE			X		• validated theory

Cognitive Model	Domain Types	Correspondences			Validating Data Sources
		Domain	Psychological	Physiological	
Soar	air traffic control, test director, automobile driver, job shop scheduling	X	X		• validated theory • human interaction • human behavior data
ModSAF	ground warfare	X			• human interaction
CCTT SAF	ground warfare	X			• human interaction
MCSF	small unit operations	X			• human behavior data • human interaction
SUTT CCH	small unit operations	X	X		• human behavior data • human interaction
IFOR (see Soar)	fixed & rotary wing air operations	X	X		• validated theory • human interaction • human behavior data

All validation techniques have limitations. The cognitive models listed in Table 5 indicate there are two significant limitations of HBR correspondence used for validation. First is the unrealistic requirement of domain correspondence to search very large and nonlinear behavior spaces. For example, identifying and codifying every factor influencing a soldier's decision on a dismounted route through the woods, swamp, jungle, desert, arctic, or urban terrain includes elements of mission, enemy, terrain, time, troops, weather, equipment, etc. Second concerns testing for psychological and physiological correspondences. These two forms of correspondence usually require the use of extensively validated models of psychological and physiological phenomena to produce referent (Department of Defense Modeling and Simulation Office, 15 August 2001b). In essence, one must find results from other valid HBR models or build and validate another HBR model to provide referents for validation of a new model. This dependence on other models makes validation using psychological and physiological correspondences tenuous at best.

3.6 Human Performance Evaluation

Supervisors evaluate personnel for two reasons. First is to determine who is due just rewards and promotions. Second is to determine what additional training is needed to help develop individuals and teams (Tziner, Joanis, & Murphy, 2000). This process is complex and fraught with potential issues which human resource personnel have established techniques to help resolve. To address some of these issues and techniques, the remainder of this subsection covers the fundamental elements of human performance evaluation, the common problem of evaluator bias, and some of the possible techniques shown to mitigate bias.

3.6.1 Procedural Versus Declarative Knowledge

Knowledge normally used to provide input to human performance evaluation is categorized as either declarative or procedural. Declarative knowledge is facts -- the "what". Examples of declarative knowledge are an M16A2 is a semiautomatic rifle used by the US Army, an M16A2 semiautomatic rifle uses a 5.56mm round, and an M16A2 can fire a using 3-round burst or single shot modes. Procedural knowledge involves comprehension of the process -- the "how". For example, before firing an M16A2, one must load the weapon by inserting a magazine containing one or more rounds of ammunition, allow the bolt to slide forward to chamber a round, and move the shot selection switch from safe to single shot or burst mode.

Procedural knowledge is declarative knowledge interpreted within the context of situational understanding. Without declarative knowledge, procedural knowledge has no foundation. Without procedural knowledge, declarative knowledge is limited to the statement of facts. This difference allows one to look at an incident in two ways. Declarative knowledge allows you to collect the facts of what happened, while procedural knowledge allows you to determine why it happened. This is illustrated by comparing overt behaviors with cognitive processes. Overt behaviors are described as declarative knowledge, while cognitive processes allow the user to understand why a particular behavior was selected. A combination of the two categories permits supervisors to provide a more complete assessment of personnel by demonstrating if the sum of the facts is equal to the whole. This explains why assessment requires context and not just analysis of the raw facts.

3.6.2 Bias

As defined by Webster's Dictionary, bias is "systematic error introduced into sampling or testing by selecting or encouraging one outcome or answer over others." (Merriam-Webster's Collegiate Dictionary, 2003) Bias often occurs in the assessment of human performance. Research literature describes at least five types of bias applicable to SMEs: judgmental, decision, heuristic, informational, and normative.[23] One can further classify judgmental and decision bias into at least twenty subcategories: anchoring, adjusting, association, availability, base rate neglect, belief, certainty effect, central tendency, confirmation, conjunction, conservatism,

[23] The Glossary provides definitions for each bias category.

contrast, framing, halo, hindsight, illusory correlation, insensitivity to the prior probability of outcomes, leniency/severity, overconfidence, regression to the mean, representativeness, response bias, sunk costs, and the Law of Small Numbers (Tversky & Kahneman, 1971) (Tversky & Kahneman, 1974) (Kahneman, et al, 1982) (Cohen, 1993) (Barnett, et al, 1993) (Perrin, et al, 1993) (Cascio, 1998) (Stein & Stein, 1998) (Gilovich, et al, 2002).[24]

Pace and Sheehan categorize bias associated with the use of SMEs into three dimensions: perspective, performance, and perception (Pace & Sheehan, 22-24 October 2002). *Perspective* addresses a SME's ability to maintain focus on the intended purpose of the model. A SME may lose focus as he allows his real-world experiences to cloud his view on what the model should have the capability of doing. *Performance* deals with the SME's ability to execute the validation process. This ability may be hindered by demands on the SME's time, the availability of data, the SME's ability or desire to comply with specified validation procedures, or the ability of the expert to understand the simulation. Finally, p*erception* addresses the bias an expert brings to the process based on his education, training, real-world experiences, exposure to simulations, and organizational loyalties. These factors may unduly focus a SME's attention on certain aspects of a model's performance (Pace & Sheehan, 22-24 October 2002).

Three subcategories of perception bias, which this research addresses, are anchoring, contrast, and confirmation. *Anchoring bias* emerges when an individual embraces an initial hypothesis and maintains this view regardless of incoming facts. This results in overemphasis on the hypothesis and an inappropriately minimal shift from the initial viewpoint (Tversky & Kahneman, 1974) (Kahneman, et al, 1982) (Cohen, 1993) (Duffy, 1993) (Perrin, et al, 1993) (Stein & Stein, 1998). *Contrast bias* materializes when one seeks information to contradict an original hypothesis, ignoring or undervaluing evidence in support of the hypothesis (Tversky & Kahneman, 1974) (Kahneman, et al, 1982) *Confirmation bias* is demonstrated when an individual overvalues select pieces of information relative to consistent evidence indicating an alternate conclusion (Cohen, 1993) (Duffy, 1993) (Perrin, et al, 1993) (Stein & Stein, 1998).

Subject matter experts show bias on many levels. One characteristic of a SME is his ability to quickly develop a solution or response based on his experience. This can manifest itself as perception bias when SMEs use aspects of the Recognition-Primed Decision (RPD) pattern

[24] This work only defines those subcategories specifically addressed in this dissertation: anchoring, contrast, confirmation, and the Law of Small Numbers. The remaining subcategories are listed to provide an indication of the vast number of bias which might effect evaluation results.

matching process (Klein, 2001).[25] Such bias may not be wise to mitigate. However, until one can identify, measure, and mitigate perception bias, we have little understanding of practical bias. *Practical bias* is not a category or subcategory of bias. It is a measure of the magnitude and importance of the impact of participant inconsistency and inaccuracy. In other words, how much does bias skew results.

3.6.3 Performance Appraisal

Supervisors have used many methods to evaluate human performance over the years. Some of these means are purely qualitative in nature. Methods that describe the performance without ranking performance against others are known as absolute rating systems. There are four general methods involving absolute rating systems: behavioral checklists, essays, critical incidents and graphics rating systems. *Behavioral checklists* are similar to declarative knowledge in that they merely state facts regarding the existence or non-existence of a behavioral trait. These checklists are Go/No-Go in nature and fail to indicate a level of performance. *Essays* allow raters to provide a more extensive description of the observed performance without limiting the assessment to a specific list of behaviors. However, essays do not provide standard rater responses and require a great deal of time to complete. *Critical incident reports* provide specific examples of performance, but require raters to witness the act (Cascio, 1998). Thus, essays and critical incident reports typically concentrate on procedural knowledge by allowing the rater to place the facts in context of the situation in which they were performed.

In an attempt to provide a quantitative means of assessing performance, supervisors can use *graphic rating scales*. These scales consist of a series of performance-based questions with standardized scales for evaluators to provide their assessment of subordinate behavior (Cascio, 1998). One example of a graphic rating scale is a Likert Scale. Likert Scales have an odd number of possible responses with one side of the midpoint representing substandard performance and the other side of the midpoint representing above average performance. The midpoint represents average performance. Scale values are general and subjective in nature but provide a means of quantifying subordinate performance. Examples of possible responses equated to a 5-Point Likert Scale are outstanding, above average, average, below average, and poor.

[25] The RPD model is described in subsection 3.7. Naturalistic Decision-Making.

Graphic rating scales provide evaluators with four advantages over using open-ended questionnaires. First, graphic rating scales require less time to complete since they only require evaluators to choose one of the available options. Second, they allow evaluators a means of converting qualitative information into quantitative data. Next, since they are less time consuming, assessment forms can include more questions allowing for a broader assessment of an employee's performance. Finally, quantitative employee performance data allows for comparison across evaluators and evaluates. Thus, graphic rating scales help evaluators capture aspects of procedural knowledge of individual behavior by acquiring more information about the employee while converting qualitative information into declarative knowledge.

Understanding bias is present in the assessment of human performance, Smith and Kendall suggest human resource personnel can assist supervisors in assessment of personnel by providing better assessment worksheets. These researchers developed a rating scale consisting of a series of assessment questions with possible responses which include explicit examples of performance for each response listed (Smith & Kendall, 1963). This scale is often referred to as the *Behavioral Anchored Rating System.*

Creation and validation of such evaluation forms is expensive and time consuming. However, they provide supervisors with a powerful yet relatively simplistic tool to assess the performance of their subordinates. More complex and time-consuming assessment methodologies have been devised to provide a better assessment of personnel performance. According to King et al., over time, the Behavioral Anchored Rating System has proven itself as viable and reliable an assessment process as systems that are more complex (King, et al, 1980).

The *behavior observation scale* is a hybrid version of a graphic rating scale and behavioral check lists. The scale allows the supervisor to track the frequency of specified occupational behaviors (Tziner, Joanis, & Murphy, 2000). Because of this, it provides more information about the kinds of behavior a subordinate is performing, but still fails to address the quality or context of this behavior.

The most often used method of assessment it the graphic rating scale (Cascio, 1998). Each performance appraisal technique is subject to the observation and judgments of the supervisor. As such, they are subject to misinterpretation and bias. Some performance appraisal techniques are better at mitigating misinterpretation and bias than others.

3.7 Naturalistic Decision-Making

Klein characterizes *naturalistic decision-making* (NDM) as a paradigm designed to describe how people perform rather then being a method to improve performance (Klein, 1997). The focus is on how experts use their experience to make decisions when concerned with the execution of tasks in complex environments (Zsambok, 1997). Cognitive psychologists have demonstrated that, for expert decision makers, methods and models associated with NDM more accurately describe the human decision-making process than previous paradigms. This is especially true when the situation involves a "high stakes, dynamically changing environment, time pressure, (with) ambiguous or incomplete goals" (Tolk, 10-11 December 2002). These characteristics typify decisions made by military personnel during times of crises decision-making and execution of military operations.

In the late 1980s, Klein developed a theoretical model of decision-making refered to as the Recognition-Primed Decision (RPD) model. The *RPD* model asserts that expert decision makers use pattern matching to provide viable solutions to a situation. When an expert cannot match the situation to a known pattern, he uses a modified decision-making process to provide a solution until the situation changes. In these situations, the expert may modify his mental model of the world or generate a story to explain the difference in what he is observing and what his mental model tells him should be occurring. Research has validated the RPD theoretical model as a decision model offering merit for military operations. However, as of January 2004, no computational implementation of the RPD model at the operational-level for military decision-making exists (Klein, 2001). RPD was never meant to be a computational model with predictive capabilities. It was developed to help understand how expert decision makers draw conclusions and select a course of action.

As with any model, RPD has its limitations. Due to the Law of Small Numbers, using RPD, or any model, for describing the decision-making process has limited statistical strength if one has a limited number of SMEs.[26] This could lead to an incomplete assessment of the decision-making process. Also, using experts exposes the process to human error. Although less likely than non-experts, SMEs may introduce bias into the decision-making process by negating

[26] The Law of Small Numbers takes effect when a person over infers the likely hood of the frequency of an event based on a limited number of observation (Tversky & Kahneman, 1971).

plausible courses of action due to their incomplete collection of situational patterns. This bias comes in the form of knowledge-based mistakes, decision errors, and judgment errors.[27] Thus, even though "the decision processes typically studied in NDM consist of a series of decisions or a sequence of intermediate outcomes," validating agents must use it with care to limit possible negative effects from potential SME bias (Lipshitz, 1997). Nonetheless, the nature of the validation process for HBR models, where one must take into account the context in which the task is being performed, suggests a fit between the face validation process and the NDM paradigm.

The NDM paradigm is applicable beyond the collection of referents and the face validation of HBR models. Validating agents can also apply its context dependent nature to the training and retraining of SMEs for the validation process (Cohen, et al, 1997) (Lipshitz & Shaul, 1997). Validating agents must train and focus SMEs to ensure SMEs only assess the model for the specific domain. If problems occur with performance of the SME that require retraining, remedial training methods must also be domain specific (Lipshitz & Shaul, 1997).

Since face validation concerns experts making decisions about performance, it is apparent that the NDM paradigm is applicable to the face validation process where an assessment of the model's performance is made for a specific yet still complex environment. Specifically, validating agents may use the RPD conceptual model to validate HBR models and to train SMEs to perform validation for combat tasks through pattern matching.

Methods used by NDM researchers, such as cognitive task analysis (CTA), have been used for the initial stage of simulation design to assist in identifying important aspects of the task to be modeled (Miller & Woods, 1997). This technique has similar requirements to validation techniques which require SMEs to assess a model in a context dependent situation. However, CTA requires one to look deeper then just the overt behaviors of a decisions maker.

Klein defines a task analysis as the direct observation of a person performing an action resulting in a detailed description of the tasks one accomplishes in order to achieve a goal. A *cognitive task analysis* is a more extensive/detailed look at cognitive components of the task. It seeks to describe the cognitive processes underling the performance of tasks and the cognitive

[27] "Decision errors pertain to situational assessment, mental models, and sequential option generation/evaluation rather than concurrent choice" (Lipshitz, 1997a).

skills required to respond appropriately to complex situations (Klein, 2000). Thus, it examines actions and the decisions leading to those actions.

A CTA does not predict actions. Information collected by performing a CTA can be used to produce a descriptive model developed through interviews with SMEs and is qualitative in nature. In the past, CTA studies have been conducted for the design of human-computer interfaces, instruction and training, organizational design, system development, product design and marketing.

Many variations of CTA have been developed. Klein describes CTA as consisting of five steps: identifying sources of expertise, assessing the knowledge, extracting the knowledge, codifying the knowledge, and applying the knowledge (Klein, 2000). Aronson's taxonomy includes four phases: knowledge elicitation, analysis, knowledge representation, and validation (Aronson, September 2002). Finally, Harvey separates the process into four phases: preliminary phase, identifying knowledge representation, knowledge elicitation techniques, and representations (Harvey, 2001).

Using Harvey's phases, the *preliminary phase* requires individual(s) performing CTA to become conversant in the area they wish to study. It may consist of reading relevant professional or training manuals, unstructured interviews with SMEs, and participant questionnaires to collect information about the tasks required to achieve a goal or accomplish a task (Harvey, 2001).

After achieving a sufficient understanding of the basic issues and tasks relevant to the problem domain, the next step is to determine how best to represent knowledge. Two ways of representing the knowledge are procedural and declarative. The factual or conceptual nature of declarative knowledge allows one to use the information in ways not originally foreseen. Since procedural knowledge is a more precise means of describing how an individual accomplishes a task, it is an efficient but less germane means of depicting how to perform a task. When determining which data representation to use, the individual(s) conducting the CTA must consider the nature of the information and processes to be modeled (Harvey, 2001) (Wray, et al, 1992).

With a basic knowledge of the problem space and a decision on how to represent the domain knowledge determined, collection of the detailed knowledge set is undertaken. Data collectors usually conduct this phase using structured interviews of SMEs to gather significant content that researchers will analyze and model developers will codify (Harvey, 2001).

Information representations can take many forms (e.g., flow charts, structured English syntax, entity relationship diagrams, Unified Modeling Language (UML) diagrams, etc.) (Harvey, 2001). There is no prescribed format for representing the information gathered during a CTA. The specific purpose of the CTA and the complexity of the tasks one is modeling will steer the individual(s) conducting the CTA to choose one or more of these methods for representing data. The more complex the task, the more important it is to have a well-understood language or technique for representing the information collected.

3.8 Assessment of Previous Work

Pew et al.'s statement that "few individual combatant or unit-level models in the military context have been validated using statistical comparisons for predication" points to a major issue with emergent military simulations (Pew & Mavor, 1998). Until recently, a limited number of research efforts have attempted to address the issue of validating HBR models. Some of these most prominent have been project Agent-based Modeling and Behavior Representation (AMBR), Birta and Özmirak's automated result validation model, Caughlin's metamodel methodology, Gonzalez and Murillo's validation through automated observations, and current work on alternative scales for face validation results (Air Force Research Laboratory, 1 June 2001) (Birta, et al, January 1996) (Caughlin, 1995) (Gonzalez & Murillo, 1998) (Harmon, 4 August 2003). Additional work such as Tactical Decision-making Under Stress (TADMUS), demonstrated insights to issues such as SME bias (Barnett, et al, 1993) (Hutchins, et al, 1996a) (Hutchins, et al, 1996b).

Project Agent-based Modeling and Behavior Representation (AMBR) is an Air Force Research Laboratory (AFRL) program designed to "advance the state-of-the-art in cognitive and behavioral modeling for military applications" (Air Force Research Laboratory, 1 June 2001). Researchers compared and contrasted HBR architecture implementations as they performed a series of "standard problems" in a simulated environment. During the project's initial phase, program personnel conducted a comparison of the effectiveness of four cognitive architectures: ACT-R, D-COG, EPIC-Soar, and iGEN.

An impartial moderator, BBN Technologies (http://www.bbn.com/), handled the comparison of the models and completed the study in 2000. The focus of the initial phase was multi-tasking. The domain was a simplified version of an enroute air traffic control system. Model developers modified and integrated each cognitive architecture into the virtual air traffic control system and exercised the architectures to determine their ability to simulate the behaviors and perform in a multi-tasking mode. All the models were able to replicate the referent within tolerances. Experimental control personnel noted the differences in how each architecture implemented the multi-tasking requirement.

BBN Technologies' review of the methodology used during the study identified many important issues. Two major criticisms were the limited number of tasks and sparse number of referents used during the comparison. These issues made it difficult to perform an exhaustive comparison of the capabilities of the cognitive models. The referent used in the study also lacked the ability to make a "head-to-head comparison" of the models. Due to limited time for coding modifications, the architecture implementations lacked the capability to represent expert cognitive processes (Gray, 2000).

A summary of the results of the study by BBN Technologies indicates the focus of the project was too vague. Were they to compare the overt behaviors of the models or the cognitive process behind the actions? Were the architectures supposed to simulate behaviors at the performance level or at all levels of interaction (Gray, 2000)? These questions reflect the difficulties of comparing the capabilities of cognitive models. They also identify problems with a lack of consistent validation standards for HBR models.

Although phase one of Project AMBR failed to provide a comprehensive comparison of the four initial cognitive models, it did help to identify some of the fundamental difficulties with such a process. Although its focus was narrow, a specific non-real world task with limited referent, it is a starting point for future work in the development of cognitive model comparisons.

In 1995, Caughlin introduced the idea of using reduced order metamodels to validate models and simulations. He claimed this new method would be a more timely and cost effective means of validation.

The creation of a metamodel requires *a priori* knowledge, data, metamodel structures, and rules to determine which original model will produce the referent (Caughlin, 1995) .[28] Caughlin describes two methods researchers can use to construct metamodels for validation,

[28] *A priori* knowledge is knowledge derived "independent of all particular experiences" (*Encyclopedia Britannica*, 2002).

direct and inverse (Figure 8). The direct method requires creation of a second model, the metamodel, composed of subcomponent models that are lower fidelity replicas of the original components. The issue with the new, lower-fidelity metamodels is the difficulty of ensuring they properly represent the original model and all its functionality. Traceability of the direct method is less of an issue with the inverse method. The inverse method produces a reduced order model using input data and output results from the original model. Although a mathematical approximation of the initial model, the metamodel created using the inverse model, has to deal with issues relating to fidelity, sensitivity, and accuracy of results (Caughlin, 1995) .

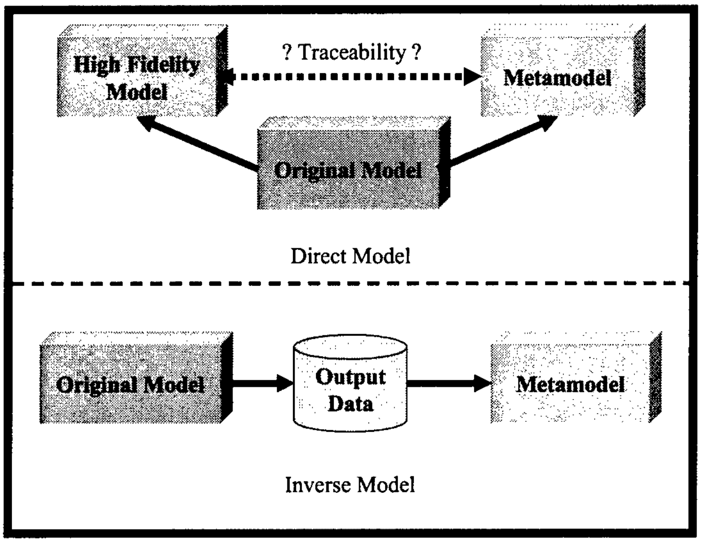


Figure 8. Metamodel Correspondence From (Caughlin, 1995)

Caughlin's metamodel approach to validation holds promise for analytical models that can be reduced to a more simplistic representation. However, this method of validation is not applicable to analytical models that are already in their most simplistic state. Nor has anyone shown the method to be applicable to models whose complexities make it impossible to create metamodels (e.g. cognitive models).

Birta and Özmirak proposed an automatic means to uniformly "validate" discrete, continuous, and combined simulation (Birta, et al, January 1996). Their technique focuses on an automated *face validation* of a model.[29] They felt a single face validation of a model could not perform an "absolute" validation. Instead, an experimental process is required. Figure 9 shows

[29] Birta and Özmirak used the term "*behavioral validation*" in their report. Although not specifically defined the technique is similar to face validation. To reduce confusion the term face validation is used in the section as a replacement for the term behavioral validation. It is NOT restricted to the validation of human behaviors.

the four modules contained in their process: simulation model, validation knowledge base, experiment generator, and evaluator.

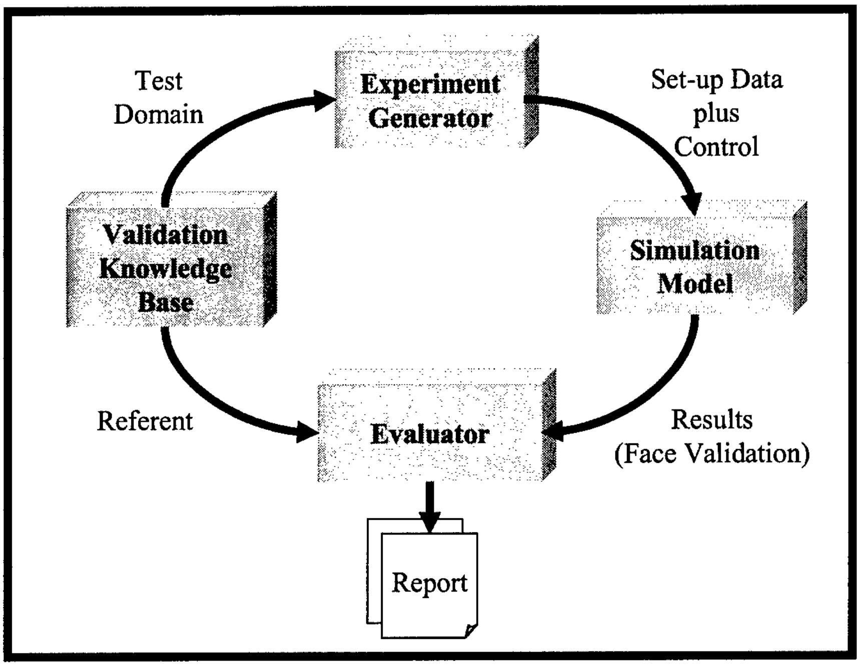


Figure 9. Global Architecture for Birta and Özmirak's Automated Result Validation Model From (Birta, et al, January 1996)[30]

The *simulation model* is the implemented program representing the system the user wishes to simulate. *Validation knowledge base* (VKB), the key component of the model, is the fundamental knowledge of input and associated outputs for the model. It represents the referent required by the model to meet its design specifications and intended use. Researchers use the VKB to develop the experiments used to validate the model's performance and the data to compare with the model's results. The *experiment generator* uses the input values provided by the VKB to design test cases for the simulation. Its goal is to produce the minimum number of test cases required to ensure a comprehensive validation of the model. Finally, *evaluator* takes the results from the simulation runs and compares them with the referent provided by the VKB, conducting a "critical evaluation of the simulation model output" (Birta, et al, January 1996). The results of the comparison are stored in the *report* files.

Birta and Özmirak use dynamic objects to identify the data required by the VKB. The dynamic objects are abstractions of dynamic behaviors represented in the simulation. A dynamic

[30] Birta and Özmirak used the terms Reference Data and Behavior Data. These terms are changed to Referent and Results, respectively, to make them consistent with the terminology of this document.

object, O, is described as an ordered pair of vectors X and Y where O = (X, Y). X is the generalized input and Y is the output of the object. A causal relationship existing between the two vectors infers a change in X results in a change in Y. The fundamental property of all dynamic objects is their "ability to generate (exhibit) behavior over some prescribed time interval" (Birta, et al, January 1996).

The VKB must possess all possible instances of the dynamic object. This means an exhaustive search of the problem space must occur to ensure every possible X, Y combination for the dynamic object is represented in the VKB. These pairings are a set of three disjointed types of specifications: formal, qualitative, and observable.

Formal specifications are X, Y relationships that always hold true (e.g., a 70-ton tank weighs more than a 60-ton tank). A qualitative specification displays the causal relationships between the input and output vectors (e.g. the main gun of a tank stops firing when it is out of ammunition). Finally, an observable specification is a means of ensuring the simulation replicates real-world behaviors when the experimental generator presents similar situations. This data is derived from the observation of previously validated simulations or real-world systems (Birta, et al, January 1996).

Birta and Özmirak's knowledge-base approach to model validation is a means of face validation. It attempts to accomplish validation through an automated system. This can reduce the bias injected into the face validation process by SMEs. The VKB appears to be a set of all available referents, powerful in its content but unlikely to be exhaustive for topics such as human behaviors. The approach also fails to address the non-deterministic nature of human behaviors.

In 1998, Gonzalez and Murillo proposed a method to validate human behavior models by means of automated observation. The technique allows a human behavior model to watch and learn from SMEs performing procedures in a standalone or networked simulation. Computerized agents compare the behaviors of SMEs and simulations performing the same tasks to determine if the model's actions were similar. Later, additional SMEs can analyze the differences noted by the computerized agents to determine if the simulated behaviors were viable (Gonzalez & Murillo, 1998).

Another aspect of this method is its ability to allow models to learn from SMEs as the two execute in parallel environments. As "serious" inconsistencies arise between the actions of SMEs and the simulation, a difference analysis engine (DAE) compares the two actions. If both actions were viable, the DAE would note the differences and allow the simulation to continue. If the computerized agents judge the model's behavior to be inappropriate, the automated system modifies the model's behavior to match the performance of the SME (Gonzalez & Murillo, 1998). This is similar to the training of a neural-network. It is also limited to the extent of modifications it can make based on the type and amount of input data available and the parameters of the algorithms.

Although the methodology may provide a means of training models, it must still address the issue of training behaviors valid for a simulation environment instead of replicating human behaviors in the real world. Developers face the same problem when using the method to validate simulation behaviors. Do these actions/behaviors transfer to the real world? Furthermore, the problem of creating a deterministic program to assess a non-deterministic model of behaviors demonstrating a non-linear nature is NP-complete and thus computationally intractable (Mallery, 28 March - 03 April 1988). The method is another means of conducting a face validation of a simulation; however, as of January 2004, it has not been prototyped and tested.

The Defense Modeling and Simulation Office has determined that the current VV&A process for HBR models is inadequate. Work currently underway by Harmon and Metz seeks to determine if HBR model validation can be broken down into a series of validation levels based on the quantitative nature of the information available to assess them versus the current subjective methods (Harmon, 4 August 2003). Preliminary results from this research are due the summer of 2004.

Goerger presents an alternative methodology, which uses a continuous scale for validating HBR models instead of a binary valid/invalid scale (Goerger, 2002). The scale is anchored on one end by a simple reactive agent HBR model and on the other end by the optimal HBR model, a human being. A model can be placed along the continuum of the validation scale indicating its degree of validity and allowing a relative comparison of similar models. The author's methodology addresses the diversity of HBR models and the varying degrees of information available to validating agents based on the model representation utilized to codify the theoretical model. Goerger argues that a validating agent can provide a more extensive assessment of a model's capabilities if the agent can query the model's cognitive process for information on its situational awareness and the plausible courses of action it is considering.

With this information, the validating agent can assess if there are issues with the development of an adequate situational awareness, the cognitive process, or if the model lacks the diversity of options to address the situation. The methodology fails to address the

The *Tactical Decision Making Under Stress* (TADMUS) program developed a decision support system for enhancing the quality of the air warfare decision-making process. Aegis ship commanding officers and tactical action officers engaged in demanding littoral scenarios using a mock up of their current Aegis displays and performance was recorded. These scenarios were characterized as involving time-sensitive, ambiguous, dynamic situations. Significant improvements in air warfare decision-making performance (i.e., improved situational awareness, more of the correct tactical actions were taken, and decreased levels of communications) resulted when decision makers used the new decision support system (Barnett, et al, 1993) (Perrin, et al, 1993)

(Hutchins, et al, 1996a) (Hutchins, et al, 1996b).

One separate, but related, issue investigated under the TADMUS program was cognitive bias in the decision-making process. Tactical action officers engaged in challenging scenarios and performance was recorded and analyzed. Biases in the air warfare decision-making process were identified; these biases included anchoring, contrast and confirmation (Barnett, et al, 1993) (Perrin, et al, 1993).

Chapter 4. Recommendations

4.1 *Training*

Performance bias affects both accuracy and consistency. One can mitigate a SME's inability to comply with validation procedures through additional training and the use of specific textural and visual examples of poor, fair, and excellent task performance. Training may help the validation agent identify SMEs who possess or develop an uncooperative attitude toward the validation process. Bias can be addressed either through counseling or by removing the SME from the process if necessary. Additional training can allow the SME pool to obtain and maintain a level of proficiency in the validation process. Training and practice sessions help to identify SMEs with the potential for bias and provided an opportunity to mitigate bias through further training or process modifications.

4.2 *Scale*

One method to increase accuracy is to provide SMEs with more precise descriptions for Likert Scale responses. Grounding assessment scales with specific descriptions for each response is a method used by human resource personnel to enhance the evaluation process of employees (Charlton and O'Brien, 2002) (Druckman, and Swets, 1988) (Gawron, 2000) (Stufflebeam, 2002).

There are two means for grounding assessment scales. The first method fixes values for the tails of the scale for each subtask, *general grounding*. The second method is to ground each scale value for each question, *explicit grounding*. General grounding fixes the boundaries of the assessment scale while affording SMEs flexibility to judge questionable actions based on their experiences. Although the process fixes the extremes, it will not preclude imprecise responses about the scale's median score. Explicit grounding fixes the internal scale values as well as the boundary values. The process can make judgment of borderline and boundary behaviors more accurate between SMEs.

Mitigating SME inconsistency can be done by allowing SMEs to place a weighting factor on each sublevel response they feel affects the level assessment to a greater or lesser degree. Weighting factors increase consistency by allowing the mean of the sublevel assessments to correlate more closely with the assessment value of the level. Thus helping ensure the whole is a reflection of the parts.

4.3 Automation

A computerized system for identifying bias and consistency discrepancies during assessment would support SMEs and help improve validation efforts by providing SMEs with quick and accurate feedback. Numerous sublevel questions make it difficult for SMEs to mentally tally and track the numerous sublevel scores. A computerized system to calculate intra-SME consistency and warn the SME of potential inconsistencies could alleviate the need for SMEs to track their sublevel scores. The system could also provide justification for inconsistencies, modify their responses to mitigate inconsistencies, and provide an inter-SME consistency report to the validation agent who can investigate and deconflict any issues.

Chapter 5. Experiment

Studies conducted in support of this research were designed to investigate the aptitude of SMEs to assess the face validity[31] of an HBR model. The experimental design was based on a validation plan utilizing Map Aware Non-uniform Automata (MANA), an agent-based model that consists of entities representing military units that make decisions following a "memory map" which guide them about the battlefield (Galligan, Anderson, and Lauren, 2003). For this research, MANA provided the visual display of simulated human behaviors by individual dismounted soldiers which were assessed by SMEs for validity.

The experiment was conducted at the Infantry Captains Career Course (ICCC), Building #4, Fort Benning, GA. The facilities accommodated groups of 20-30 SMEs. The model user interface was projected on a 5-foot by 5-foot screen at the front of each room allowing all SMEs to view the model as it ran. A total of 182 SMEs were recruited from the Infantry Captains Career Course student body consisting of senior first lieutenants (1LT/02) and junior captains (CPT/O3) who had previous urban warfare experience.

5.1 Simulation Environment

The layout of the McKenna military operations in urban terrain (MOUT) Site, Fort Benning, GA (Figure 10) was modeled in MANA. This environment consisted of 28 buildings and a supporting road network. The environment was selected for two reasons. First, the accessibility to data from past experiments performed at McKenna such as the Natick study by Statkus, Sampson, and Woods in which squad size units were observed performing offensive and defensive tasks in an urban environment (Statkus, 2003). Second, the familiarity of SMEs with the McKenna environment.

[31] Face validation is the use of experts to view a model's performance to determine if it is reasonable under the conditions of the study.

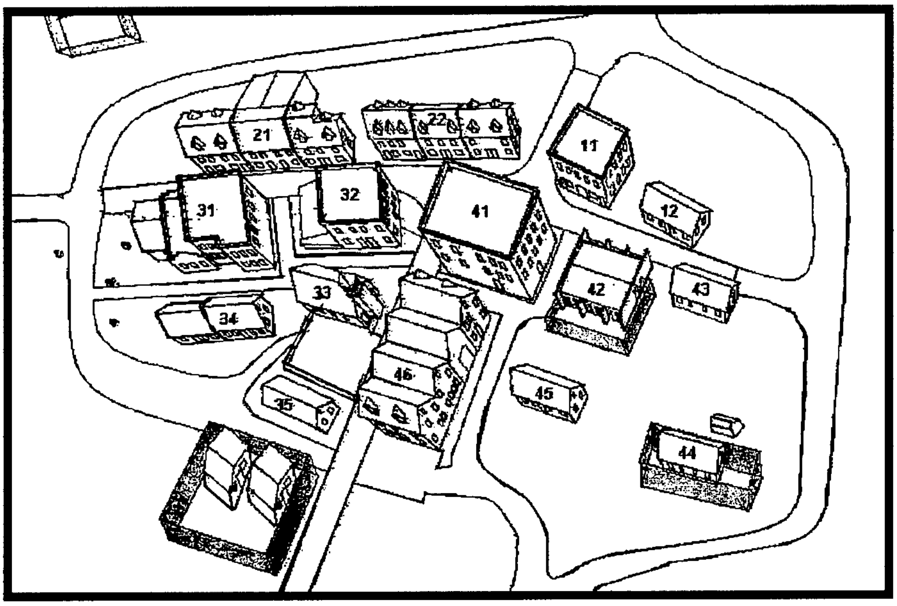


Figure 10. McKenna Test Environment Sketch From (Statkus, 2003)

5.2 Data Collection

Demographic data was collected on the SMEs using the Neuroticism, Extraversion, and Openness Five-Factor Inventory (NEO-FFI). Demographic data included military experience, combat experience, video game and simulation experience, and urban operations training. Data was collected on SME responses to two offensive and one defensive test scenarios involving the McKenna site. While the offensive scenarios use the entire McKenna village and the defensive scenario used only a portion of the south central section of the site.

SME assessment data was collected using worksheets modified from the *ARTEP 7-8-MTP* evaluations forms. Observing behaviors through the MANA interface, SMEs recorded their opinions on the evaluation worksheets using a quantitative scale and provided qualitative comments. Research personnel transferred the quantitative data from the assessment forms to Excel® spreadsheets that were then imported into JMP® for analysis. Information collected from the debriefing questionnaires was used to modify experimental design factors for future experiments and to provide insight into issues.

5.3 *Experimental Design*

The experiment consisted of two studies. Each study was conducted in five phases: In-processing, familiarization, training, data collection, and debriefing. The first study investigated biases by SMEs when responding to scenarios given their belief that they were observing either a live or simulated event using a computerized 2D map or textural display. Confirmation of SME biases when validating CGF performance or evaluating human performance was designed to determine whether or not SMEs apply the same criteria when evaluating either real-world performance or simulated performance under identical conditions. The second study identified and quantified the relative differences in consistency and accuracy of SME assessments of human performance and simulated human behavior.

5.4 *Hypotheses Study #1 - Bias*

The first study assessed whether SMEs demonstrated performance, anchoring, contrast, and confirmation biases when assessing perceived human performance or simulated human behavior. Performance bias occurs when a SME fails to respond to 20% or more of the assessment questions. Anchoring bias measures how far a SME varies from the initial hypothesis of the validity or non-validity of the model regardless of the information presented when a mixture of proper and improper performance is present. Contrast bias exists when a SME rejects the hypothesis regardless of the evidence presented. Confirmation bias measures the extent to which a SME diverged from the hypothesis regardless of the evidence presented. SMEs were categorized into two groups: those who believe they were assessing simulated behaviors and those who believe they were assessing real-world behaviors.

Null Hypothesis H_O^1: The assessment of human performance shows no difference with regards to bias between the two groups of SMEs using conventional validation methods as outlined in the Defense Modeling and Simulations Office (DMSO) Verification, Validation and Accreditation (VV&A) Recommended Practice Guide (RPG) for HBR.

Alternative Hypothesis H_A^1: The assessment of human performance by SMEs shows a difference with regards to bias for the two groups of SMEs.

5.5 Hypotheses Study #2 - Consistency and Accuracy

The second study assessed SMEs levels of consistency and accuracy when evaluating human performance versus simulated human behavior. It identified and quantified the relative difference in inter-SME consistency, intra-SME consistency, intra-SME consistency impact, intra-SME accuracy, and intra-SME accuracy impact for SMEs assessing human performance and simulated human behavior using one of three scales.

Null Hypothesis H_O^2: SMEs demonstrate the same levels of effect on consistency and accuracy during validation of an HBR model implementation using a 7-Point Likert Scale as they do when using a 5-Point Likert Scale or Go/No-Go Scale.

Alternative Hypothesis H_A^2: At least one scale (7-Point Likert, 5-Point Likert, or Go/No-Go) produces different effects on SME consistency and accuracy during validation of an HBR model implementation.

Chapter 6. Analysis

6.1 Bias

Biases generally defined as systematic error introduced into the rating process by a SME who consistently selects one response over another disregarding the actual information presented.

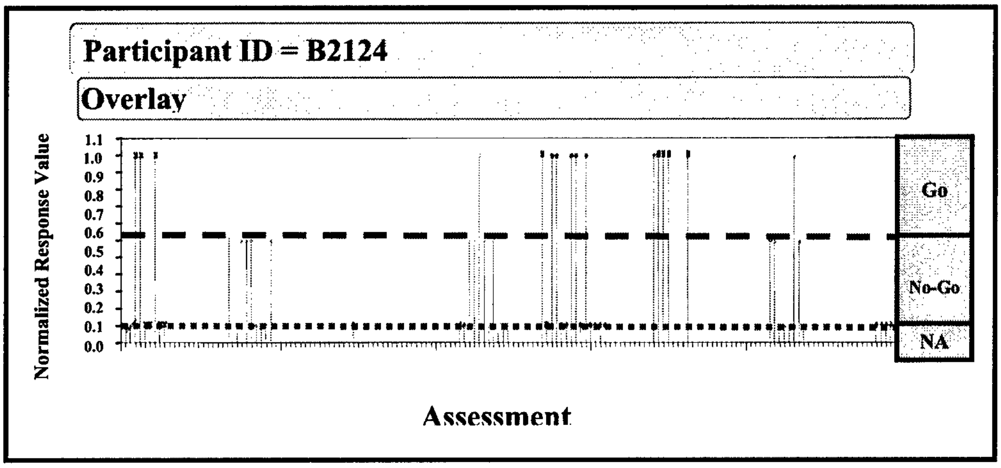


Figure 11. Performance Bias Example

Performance bias deals with the SME's ability to execute the validation process (Pace & Sheehan, 2002). SMEs demonstrate performance bias for two reasons. First, a SME may be unable to make assessments due to the availability of data. Second, a SME lacks the ability or desire to comply with specified validation procedures. For this research, a SME who chooses not to provide definitive responses to 20% or more of the assessment questions is categorized as displaying performance bias.[32] Figure 11 illustrates a performance bias response pattern. The x-axis is the assessment question. The y-axis is the normalized response of the individual to the assessment question. The bar graph indicates the participant's assessment of the specific subtask, task, or scenario. Of 159 questions, SME B2124 only responded to 16 (10%) as indicated by the bars and marks above the dashed Go/No-Go line in the figure. Based on his comments, B2124

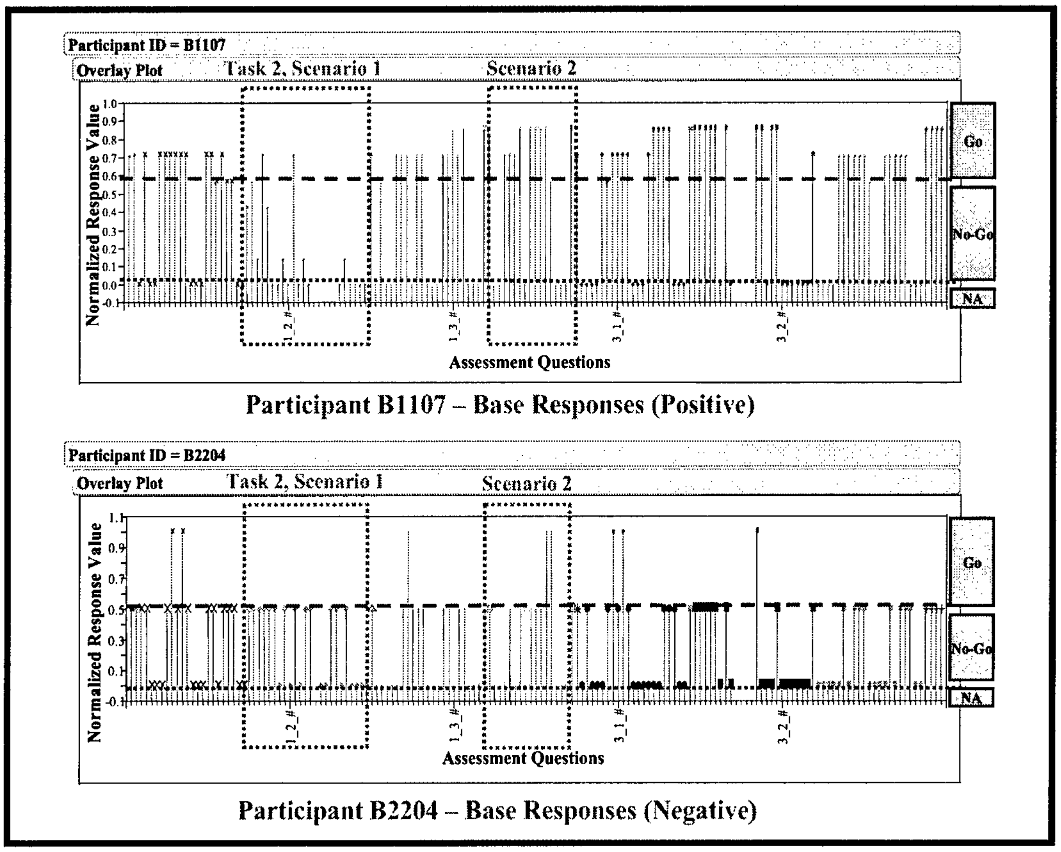
[32] A definitive response to an assessment question is a "Go" response, graphed above the dashed line or "No-Go" response, graphed between the dashed and dotted lines. "Not Applicable", graphed along the dashed line, or "No Opinion", graphed along the dotted line, responses are not definitive responses.

felt the simulation failed to furnish enough information to make an assessment. Of the 182 SMEs, 23 (13 %) displayed performance bias.

Anchoring bias occurs when a SME believes an initial hypothesis and maintains this view regardless of additional facts (Tversky & Kahneman, 1974). Anchoring bias is exhibited in two ways. First, when a SME judges the first task, and associated subtasks, as a "Go", and then, after viewing the second task and associated subtasks, which were not performed correctly, judges the remainder of the model performance as "Go" for more than 90% of the assessment questions. Second, when a SME judges the first scenario, associated tasks and subtasks, as "No-Go", and then after viewing the second scenario and associated subtasks judges the remainder of the model performance as "No-Go" for more than 90% of the assessment questions for which he provides a passing or failing appraisal. Figure 12 illustrates two different anchoring bias response patterns. The x-axis and y-axis are the same as those in Figure 11. The dashed boxes indicate subtasks assessments which relate to Task 2 of Scenario 1[33] and Task 1 of Scenario 2.[34] Participant B1102's responses are an example of positive anchoring bias with only two responses after Task 2 of Scenario 1 being assessed as negative. Participant B2204's responses show an opposite trend as even the obviously proper performance during Task 1 of Scenario 2 was assessed negatively, as indicated by the six bars above the dashed line; an example of negative anchoring bias. Thirty SMEs (16%) displayed anchoring bias.

[33] Task 2 of Scenario 1 is *React to Snipers* where the squad is engaged by an enemy sniper as the squad moves through the town's streets. The sniper kills two of the squad members while the remainder of the squad fails to react to the sniper or the loss of two soldiers. In accordance with doctrine, this results in a majority of the required sub-tasks for *React to Snipers* not being achieved to standard.

[34] Task 1 of Scenario 2 is *Conduct a Strongpoint Defense of a Building* where the squad defends a section of the town killing an entire squad of enemy personnel which attempts to infiltrate its position without the loss of any friendly soldiers. In accordance with doctrine, this results in the successful completion of nearly all the subtasks for this task.



Participant B1107 – Base Responses (Positive)

Participant B2204 – Base Responses (Negative)

Figure 12. Anchoring Bias Examples

Confirmation bias is demonstrated when an individual overvalues select pieces of information relative to consistent evidence indicating an alternate conclusion (Cohen, 1993). When a SME feels certain factors are more important than others, the final assessment may differ from what the supporting assessment factors would suggest is warranted. Confirmation bias manifests itself in two forms. First, when differences between sublevel mean scores and level responses tend toward no difference in response but the overall response differs. Second, when differences between sublevel mean scores and level more lenient but the overall response differs from this trend. Figure 13 illustrates these two different response patterns of confirmation bias. The x-axis is the level, assessment question. The y-axis is the difference between the average sublevel assessment value for the level and the level assessment value.[35] The large dashed ovals are groupings of tasks for a scenario, the smaller dotted circles are the scenario assessments, and

[35] A negative value indicates the level is assessed more harshly than the average sublevel value assessment; a positive value indicates an assessment more favorable than the average sublevel value assessment; and zero means the level assessment and average sublevel assessment are statistically the same.

the small solid ovals are the overall assessments of the three scenarios. Data from 55 SMEs (30%) displays confirmation bias.

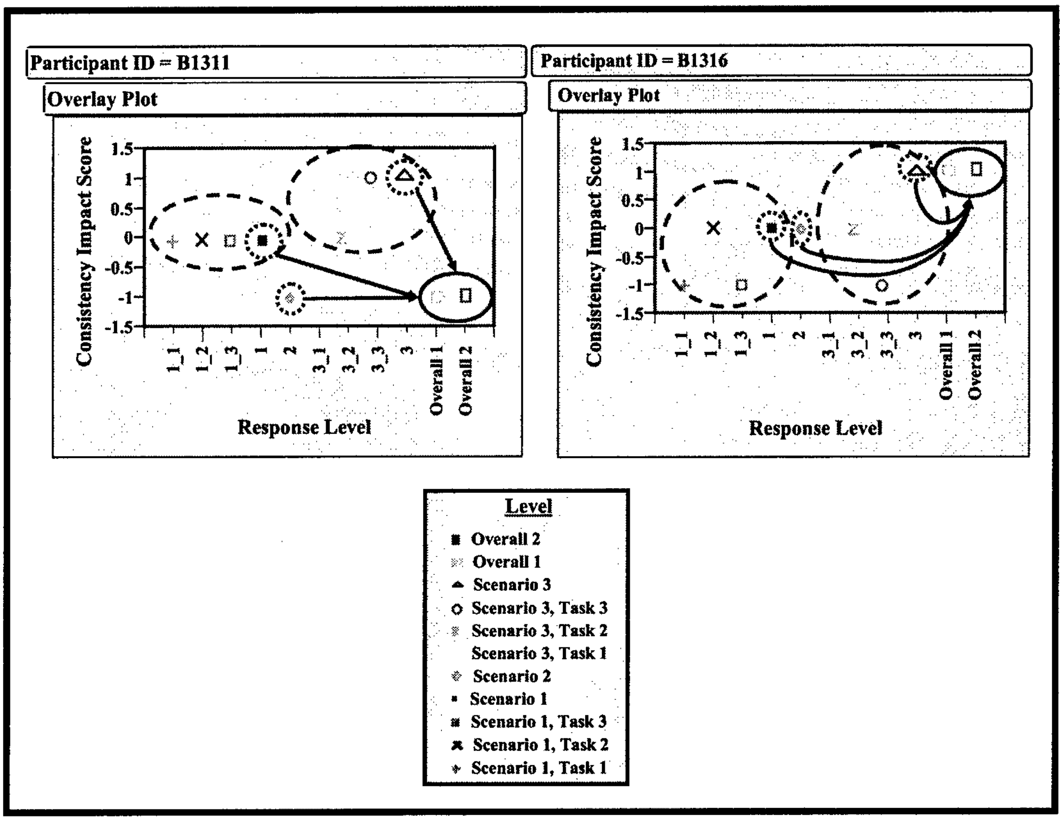
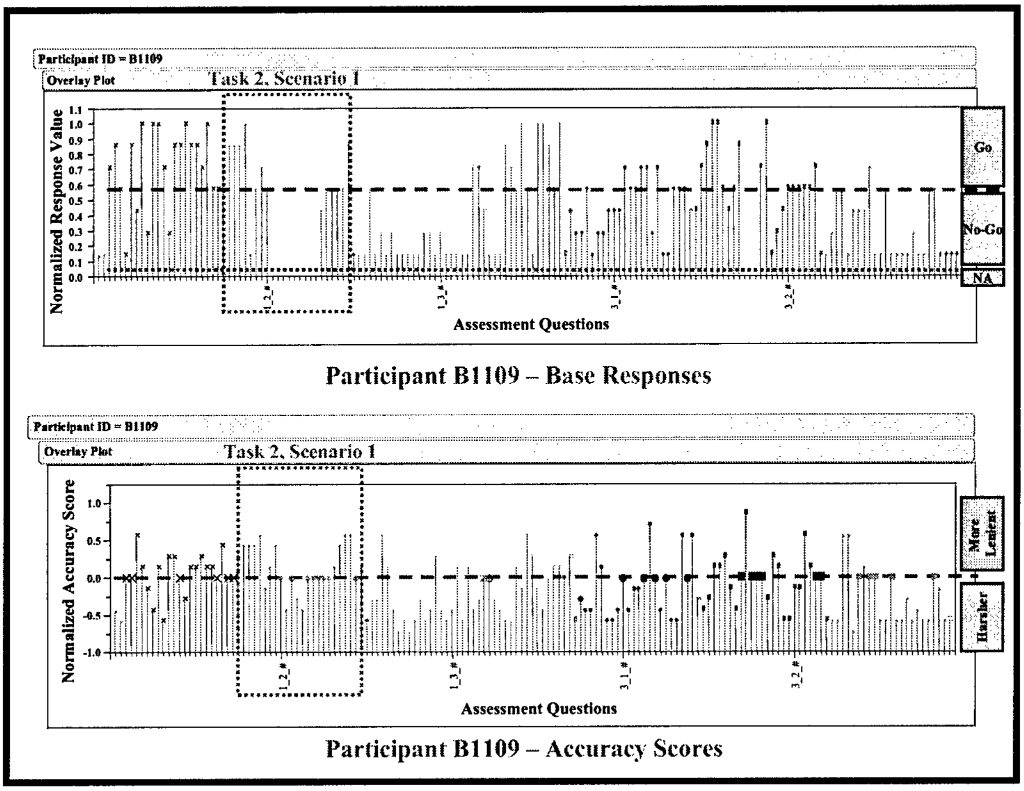


Figure 13. Confirmation Bias Examples

Contrast bias materializes when a SME contradicts an original hypothesis, ignoring or undervaluing evidence in support of the hypothesis (Tversky & Kahneman, 1974). Potential contrast bias occurs when a SME started with either a negative or positive opinion and after viewing data, which differs from this initial opinion, and negates evidence in support of the original hypothesis and assesses the model based on the initial opinion. A source of contrast bias data is a SME's accuracy scores. The accuracy data plot, the top graph, indicates a shift in a SME's accuracy trend, from harsher, below the dashed line, to more lenient, above the dashed line, or from more lenient to harsher, as the assessment process proceeds. Figure 14 combines SME raw data and accuracy plots to demonstrate contrast bias. The SME's accuracy score plot, the bottom graph, illustrates that nine of the first 45 responses (20%) were harsher, below the dashed line, than the key assessment responses. However, after assessing Task 2 of Scenario 1, the SME scored 65 of the remaining 114 responses (57%) harsher. Five SMEs (3%) displayed contrast bias.



Participant B1109 – Base Responses

Participant B1109 – Accuracy Scores

Figure 14. Contrast **Bias Example**

6.2 Consistency and Accuracy

The *overall assessment* combines SME raw scores for each of the four overall assessment questions by calculating the mean score for the normalized (0 to 1) SME responses for each question. Normalized mean scores equal to, or greater than, 0.667 are categorized as "Gos" or valid behaviors. Values above 0.667 fall into the range of responses which are passing scores. Overall 1 is the SMEs' assessment of the performance of individual soldier skills. Overall 2 is the SMEs' assessment of the squad leaders' performance. Overall 3 and Overall 4 are predictive assessments of the quality or realism of the behaviors as SMEs assess the individual soldier skills and squad leaders' performance.

Table 6 displays overall assessment results for the performance of the model based on group mean scores. For overall assessment scores, only the live simulation belief (*0*) and 5-Point Likert Scale (*3*) group rated the model as invalid, scores less than 0.5. Normalized scores less than 0.5 fall into the range of responses SMEs are told are failing scores. The degree of SME variance depicted in Table 6 indicates there is an issue with inter-SME consistency. Inter-SME consistency refers to the agreement between SMEs when they rated each subtask, task, scenario, and overall question rating. This inconsistency is identified by examining the variability in SME responses for each question.

Table 6. Mean Values for Normalized, Overall Assessment Scores

ID			Number of SMEs		Mean (Normalized 0-1 Responses)			
Simulation Belief	**Scale**	**Sim - Scale**	**Overall 1 & Overall 2**	**Overall 3 & Overall 4**	**Question Overall 1**	**Question Overall 2**	**Question Overall 3**	**Question Overall 4**
0	1	0_1	37	36	0.583	0.598	0.54	0.552
0	2	0_2	25	25	0.92	0.92	0.92	0.94
0	3	0_3	24	24	0.483	0.5	0.442	0.433
1	1	1_1	39	39	0.667	0.696	0.593	0.623
1	2	1_2	25	25	0.82	0.82	0.78	0.8
1	3	1_3	25	25	0.616	0.664	0.6	0.632
All Beliefs and Scales			175	174	0.675	0.694	0.636	0.654

Figure 15 illustrates inter-SME consistency between SME responses when observing and assessing the same behavior event via the model interface.[36] The x-axis is the SME reference number and the y-axis is the normalized assessment response to the assessment question. Each plot is a response by a different SME, participant. The plots show inconsistency amongst SME responses. One hundred SMES (55%) believe the overall performance was "Go", 37 SMEs (20%) believe the overall performance was "No-Go", and 45 SMEs (25%) assessed the overall behaviors as "Not Applicable" or had "No Opinion". This inconsistency precludes consistent and accurate assessment of the simulation. Fifty (31.45 %) subtasks, tasks, scenarios, and overall assessment responses plots exhibit inconsistent distributions.

[36] Plots above the dashed line represents "Go" Assessments, plots on the dashed line represent "Undecided", plots between the dashed and dotted lines represent "No-Go" assessments, and plots on or below the dotted line represent the subtask was deemed "Not Applicable" by the SME.

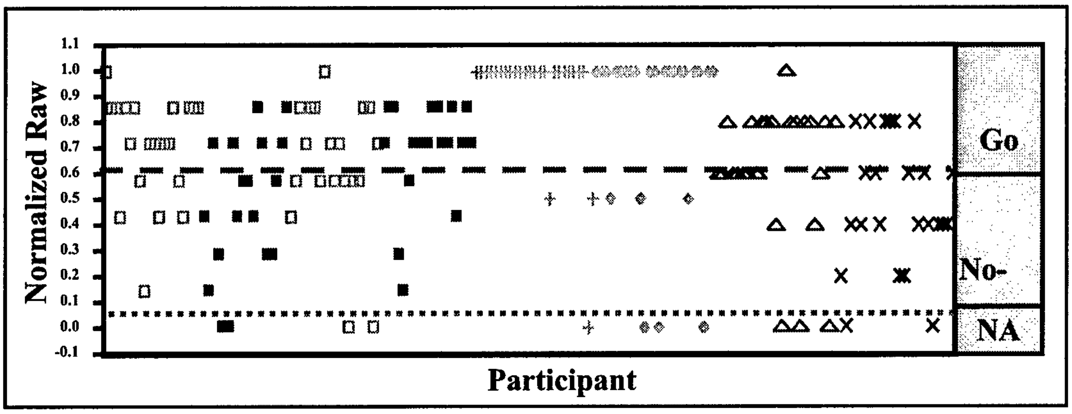


Figure 15. Subject Matter Expert Normalized Responses to Overall 1

Four separate analyses of categorical data (ANOCATs) are performed for each assessment level: Subtask, task, scenario, and overall. In each case, the responses were normalized across levels. Factors considered are the assessment scale used by the SMEs (scale) and whether the SMEs are told the process they are observing is based on live or simulated performance (simulation belief). The model employed for analysis considered the main effects of, scale and simulation belief, and an interaction effect (scale cross simulation belief). With, α = 0.05 and Prob>ChiSq less than 0.05 indicating the factor is statistically significant.[37] Factors are statistically significant at each level of assessment with the Whole Model Test Prob>ChiSq equal to or less than 0.0001. A statistically significant effect for all levels is one with the Effect Likelihood Ratio Test's Prob>ChiSq equal to 0.0000.

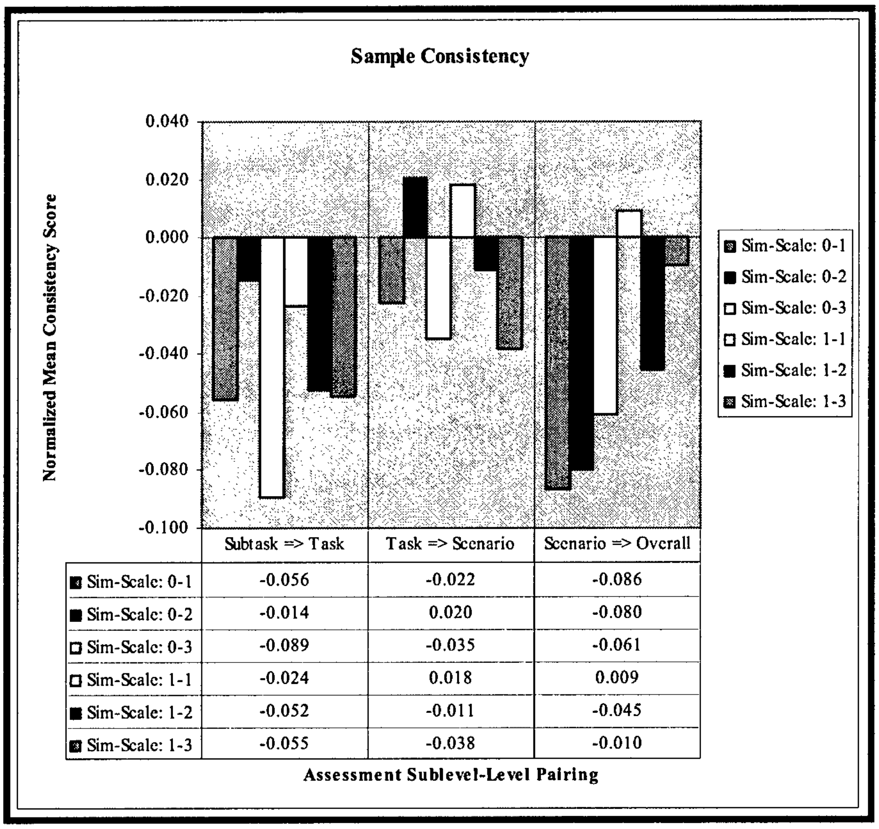
These results indicate the scale used can affect assessments and inter-SME consistency. The type of scale used by the rater also has the potential to mitigate the degree of inconsistency across SMEs and to produce inter-SME results that are both more consistent. Knowing there is inter-SME inconsistency, we sought to determine if SME bias affects inter-SME and intra-SME consistency.

Intra-SME *consistency* is a SME's ability to maintain concurrence between the average of the sublevel response scores and the level score. Analysis shows the statistical likelihood of the factor being significant effect observing an effect based on the factors of scale and simulation belief at each sublevel-level pairing. The data is calculated using the absolute values of

[37] An α = 0.05 and Prob>ChiSq less than 0.05 where chosen as threshold to indicate a 95% confidence the findings would not occur by chance and that less then 5% of the time these factors showed interaction, respectively. These are the thresholds used throughout this paper for the confidence interval and probability of interaction.

consistency score. Values of Prob>ChiSq less than 0.05 indicate a statistically significant effect of the factor. The results show at least one factor is statistically significant for each sublevel-level pairing (Prob>ChiSq = 0.0001). Analyzing effects based on scale, indicates a statistically significant effect on consistency for all pairings (Prob>ChiSq = 0.0000).

Figure 16 shows the Sim-Scale Groups (see Table 6) by sublevel-level groups (x-axis) and the mean values of consistency scores (y-axis). No uniform pattern of increasing, decreasing, or steady assessment was displayed in the general tendencies of assessment based on group, scale, or simulation belief.



	Subtask => Task	Task => Scenario	Scenario => Overall
▨ Sim-Scale: 0-1	-0.056	-0.022	-0.086
■ Sim-Scale: 0-2	-0.014	0.020	-0.080
□ Sim-Scale: 0-3	-0.089	-0.035	-0.061
□ Sim-Scale: 1-1	-0.024	0.018	0.009
■ Sim-Scale: 1-2	-0.052	-0.011	-0.045
▨ Sim-Scale: 1-3	-0.055	-0.038	-0.010

Figure 16. Intra-SME Mean Consistency Scores

Figure 17 graphically displays the correspondence of the normalized, absolute value of the SMEs' mean subtask-to-task scores. The response (y-axis) is the absolute value of consistency scores for subtask and task ratings. The x-axis is the Sim-Scale Group. When grouped by scale, the mean consistency scores for the 5-Point Scale (#-1) are greater than the mean consistency scores for the 7-Point Scale (#-3).

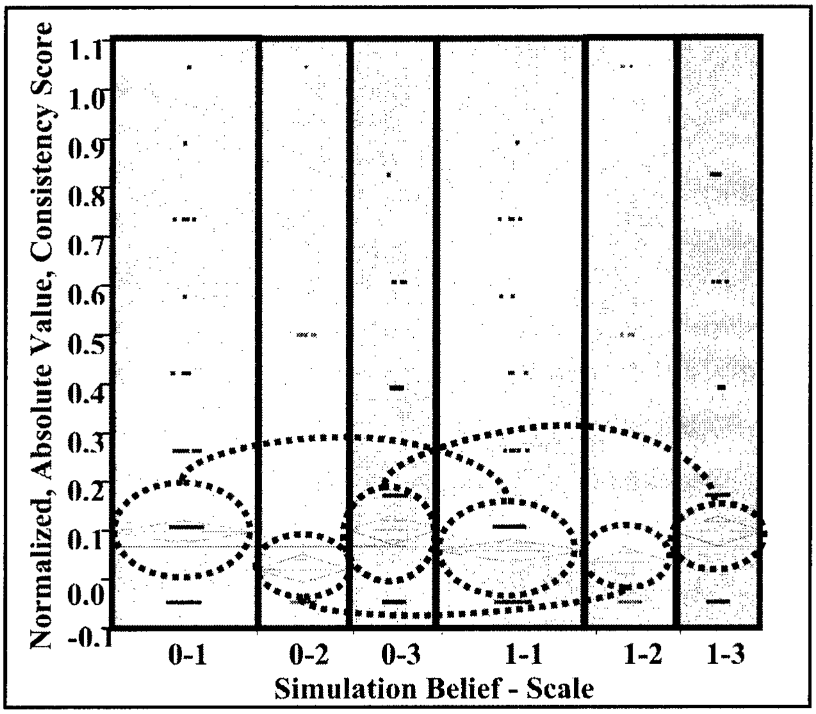


Figure 17. Intra-SME Subtask-to-Task Consistency Scores

Figure 17 illustrates that the 7-Point and 5-Point Likert Scales are less consistent than the Go/No-Go (#-2) Scale. The graphic shows that simulation belief for the subtasks-task pairings are no more or less consistent if SMEs believe they are assessing human performance (1-#) or a constructive simulation (0-#).

Analysis indicates mean SME assessments are inconsistent at each level of interaction (subtask-to-task, task-to-scenario, scenario-to-overall, subtask-to-scenario, etc.) with an effect due to scale. However, the practical effect of inconsistency, *consistency impact*, is the percentage of sublevel-level pairing responses that change their assessment score based on consistency scores, valid versus invalid.

Analysis of consistency impact scores identifies a statically significant effect based on scale for all sublevel-level pairings, Prob>ChiSq is always less than 0.0013. For simulation belief and scale cross simulation belief, no effect is demonstrated, Prob>ChiSq is always greater than 0.4709 or 0.1896 respectively.

Although analyses of mean values for differences between the sublevel-level pairing assessments show no consistent pattern, a question remains regarding process accuracy. For this research, *accuracy* is defined as the rater's ability to maintain relative correctness with respect to a consistent, scale-dependent, assessment key for each subtask, task, scenario, and overall assessment. Accuracy is measured using the normalized (-1 to 1) differences between the base assessment and SME assessments.

Analysis calculates the statistical likelihood of effect on accuracy based on the terms of scale and simulation belief for each level of assessment. Using the absolute values of accuracy scores, a statistically significant effect is found at each level of assessment (Prob>ChiSq < 0.05). Based on scale, the data indicates a statistically significant effect on accuracy for all levels, Prob>ChiSq is always less than 0.05. For simulation belief, no statistically significant effect is present except at the overall assessment level, Prob>ChiSq of 0.0017. Finally, except for the subtask assessment level, Prob>ChiSq of 0.0007, there is no statistically significant effect based on scale cross simulation belief. SMEs using the Go/No-Go Scale rated performance more harshly at the subtask level and more leniently at subsequent levels than the key assessment or SMEs using other scales.

Accuracy impact is the affect inaccuracy has on the general assessment of the subtask, task, scenario, or overall performance. It is the percentage of questions differing in relative value based on differences in accuracy scores, "Go" versus "No-Go". Accuracy impact measures the percentage of level responses that change their overall assessment score based on the response's accuracy score, valid versus invalid.

Analysis of the data denotes an effect at each level of assessment (Prob>ChiSq = 0.0001). Based on scale, there is a statistical effect on consistency for all levels (Prob>ChiSq = 0.0000). For simulation belief, a statistically significant effect is present at the subtask and task level with a Prob>ChiSq of 0.0006 and 0.0024 respectively. Finally, except for the overall assessment level, Prob>ChiSq of 0.1216, there is a statistically significant effect based on scale cross simulation belief.

There are no general trends from assessment level to assessment level based on scale or simulation belief. SMEs who use the Go/No-Go Scale and believe they are assessing human performance demonstrate a trend toward increasingly less accurate responses at each level of assessment. Although the accuracy showed a trend for SMEs using the Go/No-Go Scale to become more lenient in their assessment with each successive level, the impact of the increasing leniency is to keep the assessment slightly negative (between -0.033 and -0.200) for the task, scenario, and overall assessment levels. When SMEs used the 5-Point Likert Scale, scores get progressively harsher from task to scenario to overall assessment level even though the analysis shows accuracy maintaining a relatively constant negative value across all four levels of assessment.

Analysis indicates SMEs using the Go/No-Go Scale were more consistent and accurate at the task, scenario, and overall levels of assessment. However, SMEs using the 7-Point Likert Scale were more accurate and consistent at the subtask to task level of assessment. This means we reject the null hypothesis and accept the alternative hypothesis that scale has an effect on the magnitude of intra-SME consistency, consistency impact, accuracy, and accuracy impact.

Except for groups using the 5-Point Likert Scale, all mean scores for the overall assessment questions increased in value. However, 35 (80%) of the group, overall response, mean scores are more consistent when SMEs with confirmation bias are excluded from the sample data. For those three groups using the 5-Point Likert Scale, all but Sim-Scale 1-1 is more consistent. Figure 18 displays the results of bias identified amongst SME responses from the initial study. SMEs using the 7-Point Likert Scale demonstrated the same number of bias cases whether they believed they were assessing simulated behaviors or human behaviors.

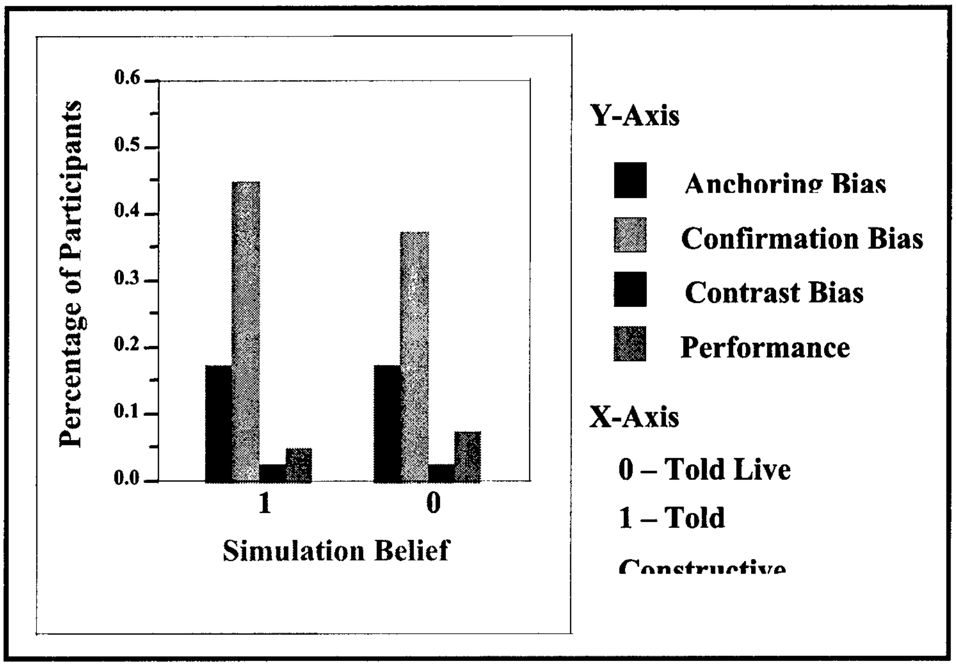


Figure 18. Study #1, Subject Matter Expert Bias for 7-Point Likert Scale

Table 7 shows the overall assessment scores by group after 97 SMEs (53%) demonstrating one or more of the four identified bias are removed. All but one of the twenty-eight cells increased their mean value score. Due to this general increase in the assessment scores, six of the mean scores changed from "No-Go" to "Go". This indicates a decrease in consistency for the mean cell response but results in a higher inter-SME general assessment consistency. Consistency here indicates that normalized mean scores assessed as "Go" in the original sample settings had higher normalized mean assessment scores when SMEs identified as displaying performance bias are excluded from the analysis. Conversely, when SMEs displaying performance bias were excluded normalized overall mean scores assessed as "No-Go" in the original sample settings had lower normalized mean scores and thus were more consistent.

Table 7. Normalized, Mean Overall Assessment Scores - Minus Bias

ID			Number of SMEs	Mean (Normalized 0-1 Responses)			
Simulation Belief	Scale	Sim_Scale		Question Overall 1	Question Overall 2	Question Overall 3	Question Overall 4
0	1	0_1	16	0.589	0.598	0.563	0.58
0	2	0_2	21	1	1	1	1
0	3	0_3	7	0.543	0.543	0.514	0.543
1	1	1_1	16	0.777	0.768	0.696	0.714
1	2	1_2	15	0.967	1	0.9	0.933
1	3	1_3	10	0.7	0.7	0.66	0.66
All Beliefs and Scales			85	0.802	0.808	0.763	0.778

Analysis indicates SMEs using the 7-Point Likert Scale demonstrated the same number of bias cases whether they believed they were assessing simulated behaviors or human behaviors. This means we fail to reject the null hypotheses and conclude that we can use the same MTP evaluation checklist to assess human performance and HBR performance of the same ground combat urban operation tasks.

The general effect on intra-SME accuracy impact when excluding SMEs demonstrating bias indicates, except for Group 1-3, accuracy impact increases for the task, scenario, and overall assessment levels.[38] At the subtask level, those using the 7-Point Likert Scale accuracy impact increased. For groups using the 5-Point Likert or Go/No-Go Scales, the accuracy impact decreased at the subtask level. Accuracy increased by as little as 1% and as much as 100% for 18 of the 24 level and group cells, while decreasing by 2% to 88% for the remaining six cells. The composite mean accuracy score increased from -0.3721 to -0.1882 improving the accuracy score by 49%.[39]

[38] As mean scores approach zero, accuracy impact "increasing". As mean score diverge from zero, accuracy impact "decreases".

[39] This score is calculated using each SME's mean accuracy impact score.

Chapter 7. Significant Contributions

The primary scientific advancement of this research is demonstrating the effects of SME bias and assessment scale on the consistency and accuracy of SME responses during the face validation process for HBR models. The research provides a means of identifying SME bias that can then be mitigated through training or use of human performance evaluation techniques. The results of this research make it possible for the validating agent to deliver a more consistent and accurate assessment of an HBR model to the M&S community than was possible under the legacy face validation process. The result is more realistic models of human behavior for use in training and analysis simulations.

For the Training community, this research can be applied to help ensure reasonable human behavior model responses to soldier inputs, thus providing users with more realistic automated enemy, non-combatant, and friendly entities. The Research and Development community can use these findings to assist in harvesting criteria for the development and validation of new models to enable analysts to better explore, develop, and analysis the possible effects of doctrine, tactics, techniques, and procedures. Finally, the Acquisition community can utilize these results to assist in ensuring its analysis better assesses the potential second and third order effects of developmental equipment on human behavior.

Chapter 8. Future Work

To further investigate the intersection of the overlapping ovals of the methodology, this section outlines additional research areas designed to enhance face validation procedures for human behavior representation models. The fundamental issue is not whether the M&S and Psychology Communities need HBR models or that face validation is necessary. The issues are how to build better HBR models and how to conduct validation in a more consistent, accurate, and cost effective manner.

With respect to using face validation techniques this research demonstrated difficulties with the variability in evaluations based on the consistency and accuracy of SMEs when assessing HBR model implementations. To resolve these difficulties further research is needed to address numerous issues: the appropriateness of assessments criteria, the use of subject matter experts, and the validation procedures.

8.1 Referent

The development of viable referent, assessment worksheets, and examples (for training programs) is a time consuming and costly endeavor. To date, most efforts have focused on the collection of physical data with mixed results in collection of cognitive data for human behavior. Physical and cognitive data are just two categories of referent, each with its own intrinsic costs. Studies must be conduct to demonstrate the trade offs between the cost of collecting, mining, and validating different categories and quantities of human behavior referent. Additionally, the consistency, accuracy, completeness, and usefulness of the ensuing model validation results must be examined.

8.2 Subject Matter Experts

Although there are many issues with the use of SMEs, computability theory indicates we must still use SMEs in order to assess models of human behavior. Since human behavior is non-deterministic, one cannot write an algorithm to assess if a deterministic program, which is replicating non-deterministic behavior, is performing correctly; heuristics apply but are not absolute. Thus, since the use of SMEs is necessary for the validation of HBR models, additional research is required to address issues with categorizing, training, certifying, and supervising SMEs (Goerger, 2004).

8.3 Procedures

Another aspect of the face validation process, requiring further research, is the manner in which the model presents data to SMEs. One might enhance the validation process by modifying the manner in which models display their behaviors. Due to the number of elements and the scope of many analytical models, models routinely present behaviors on a 2D map display or in textural records. Presenting information using 3D models in a stealth view may provide additional information to SMEs. 3D models allow SMEs to observe model behaviors in the same manner that evaluators follow soldiers through the environment in training exercises. Using 3D viewers could potential clarify model behaviors in a manner which 2D displays are incapable. For example, if a SME sees an icon representing a soldier moving through an urban environment stop along the edge of building just short of a window for two to three minutes he may not be able to tell the extent of behaviors the icon is executing. When displayed in a 3D environment, the SME may see a disoriented entity checking its map, an entity stopping to fix his equipment, or an entity attempting to crawl through the wall because it cannot identify the window location. Without the information on the posture and activity of the entity, the SME is left to his own imagination to the status of the entity. There is a need to conduct research in the effectiveness of 2D and 3D displays in providing information to SMEs to determine the level of information the displays provide, their impact on assessment scores, and their cost effectiveness ration.

A corollary effort is the ability to query model implementations for information. This is similar to an after-action review or interview of the model. To enhance a SMEs ability to understand the procedural aspects of the model's overt actions it would be useful to question a model about its situational awareness, possible courses of action, and thought process. A model's ability to provide SMEs with such information would give MSEs a better understanding of why an HBR model implementation performed certain actions. This enhances our ability to make a more comprehensive assessment of the model.

Finally, further research is required to determine the second and third order effects of using grounded and weighted assessment criteria to reduce SME bias and to enhance consistency and accuracy in the validation of HBR models.

Chapter 9. Conclusions

Increasing reliance on virtual and constructive models to provide military leaders with information for the development of new weapon systems, reorganizing force structures, and developing tactics, emphasizes the need for more advanced human behavior representation models. With the increased need for higher-fidelity HBR models comes the matter of validation which has proven to be a difficult and expensive process for the M&S community. This paper provides insights into issues regarding the usage of subject matter experts in the face validation of human behavior representation models via overt behaviors. The results described within this paper are based on data collected as part of an effort to validate a behavioral model utilizing a CGF representation in an entity level, ground combat simulation.

An approved face validation process for HBR models was used and identified issues related to consistency and accuracy, effects based on bias and personality, and a means to mitigate these effects. The validation process required a referent with which to compare the model results, a sequence of military scenarios to exercise the model, and a series of sensitivity tests to indicate variance in SME responses. This paper identified and statistically illustrates three fundamental conclusions with respect to the use of SMEs in the conduct of the model assessment phase of face validation:

(1) There is a statistically significant effect based on the scale used to assess performance that can increase or decrease scores for inter-SME consistency and intra-SME consistency, consistency impact, accuracy, and accuracy impact. ANOCAT results comparing the absolute value of the differences in SME scores for consistency, consistency impact, accuracy, and accuracy impact, based on scale and simulation belief indicate statistically significant effect based on scale. Indicating scale can mitigate effects on these scores.

(2) The use of *Mission Training Plan* assessment worksheets for assessing simulated human behaviors is as valid as using the worksheets for assessing human performance. ANOCAT results indicate simulation belief demonstrates no statistically significant effect on the number of participants displaying performance, anchoring, confirmation, and contrast bias.

(3) The consistency and accuracy of SME assessment responses can be enhanced by controlling SME bias. ANOCAT results indicate SME bias has a statistically significant effect on consistency and accuracy of SME responses.

Bibliography

Air Force Research Laboratory (AFRL). *Agent-based Modeling and Behavior Representation (AMBR) Project.* Retrieved 21 May 2002 from https://www.williams.af.mil/html/ambr.htm, 1 June 2001.

Aronson, W. S. *A Cognitive Task Analysis of Close Quarters Battle.* (Unpublished Master's Thesis). Monterey, CA: Computer Science Department, Naval Postgraduate School, September 2002.

ARTEP 7-8-MTP: Mission Training Plan for the Infantry Rifle Platoon and Squad. (Mission Training Plan). Washington, DC: Headquarters, Department of the Army, 2001.

Balci, O. "Verification, Validation and Accreditation of Simulation Models." In *Proceedings of the 1997 Winter Simulation Conference* (Atlanta, GA, Dec. 7-10). IEEE, Piscataway, NJ, pp 135-141, 1997.

Barnett, B. J., Perrin, B. M., & Walrath, L. D. *Bias in Human Decision-making for Tactical Decision-making Under Stress.* St. Louis, MO: McDonnell Douglas Corporation, 1993.

Birta, L.G. & Özmirak, F.N. A Knowledge-Based Approach for the Validation of Simulation Models: The Foundation. *ACM Transactions on Modeling and Computer Simulation (TOMACS), 6*(1), 76-98. Retrieved 23 August 2002 from http://doi.acm.org/10.1145/229493.229511, January 1996.

Cascio, W. F. Applied Psychology in Human Resource Management (5th ed.). Upper Saddle River, NJ: Prentice Hall, 1998.

Caughlin, D. Verification, Validation and Accreditation (VV&A) of Models and Simulations through Reduced Order Metamodels. Arlington, VA: Proceedings of the 27th Conference on Winter Simulation. pp. 1405-1412. Retrieved 23 August 2002 from http://doi.acm.org/10.1145/ 224401.224832, 1995.

Charlton, S. G., & O'Brien, T. G. (Eds.). *Handbook of Human Factors Testing and Evaluation* (2nd ed.). Mahwah, N.J.: Lawrence Erlbaum Associates, Inc, 2002.

Cohen, M. S. The Naturalistic Basis of Decision Biases. In G. A. Klein, J. Orasanu, R. Calderwood & C. E. Zsambok (Eds.), *Decision Making in Action: Models and Methods* (pp. 51-99). Norwood, NJ: Ablex Publishing, 1993.

Cohen, M. S., Freeman, J. T., & Thompson, B. B. Training the Naturalistic Decision Maker. In C. E. Zsambok & G. A. Klein (Eds.), *Naturalistic Decision-Making: Expertise-Research and Applications* (pp. 257-268). Mahwah, NJ: Lawrence Erlbaum Associates, 1997.

Department of Defense Directive (DoDD) 5000.1: Defense Acquisition System. Alexandria, VA: Department of Defense. Retrieved 2 July 2002 from http://web2.deskbook.osd.mil/htmlfiles/

rlframe/REFLIB_Frame.asp?TOC=/htmlfiles/TOC/061ddtoc.asp?sNode=L46&Exp=N&Doc =/reflib/mdod/061dd/061dddoc.htm&BMK=T16, 2001.

Department of Defense Directive (DoDD) 5000.59: DoD Modeling and Simulation (M&S) Management. Alexandria, VA: Department of Defense. Retrieved 26 June 2002 from http://www.ailtso.com/simval/Documents/5000.59/dod_directive_5000.59.html, 4 January 1994.

Department of Defense Instruction (DoD Instruction) 5000.61: DoD Modeling and Simulation (M&S) Verification, Validation and Accreditation (VV&A). Retrieved 27 June 2002 from https://www.dmso.mil/public/library/projects/vva/products/dodi_5000.61_recoordination_document_5_oct.pdf, 5 October 2001.

Department of Defense Modeling and Simulation Office (DMSO) Verification, Validation, and Accreditation (VV&A) Recommended Practices Guide (RPG) Special Topic - Subject Matter Experts and VV&A. Retrieved 23 January 2003 from http://vva.dmso.mil/ Special_Topics/SME/sme-pr.PDF, 30 November 2000.

Department of Defense Modeling and Simulation Office (DMSO) Verification, Validation, and Accreditation (VV&A) Recommended Practices Guide (RPG) Special Topic – Validation. Retrieved 10 July 2002 from http://www.msiac.dmso.mil/vva/Special_topics/ Validation/Validation.htm, 30 November 2000.

Department of Defense Modeling and Simulation Office (DMSO) Verification, Validation and Accreditation Glossary. Retrieved 26 July, 2002 from https://www.dmso.mil/public/library/projects/ vva/glossary.pdf, 15 October 2001.

Department of Defense Modeling and Simulation Office (DMSO) Verification, Validation, and Accreditation (VV&A) Recommended Practices Guide (RPG) Reference Document - A Practitioner's Perspective on Simulation Validation and Conceptual Model Development and Validation. Retrieved 10 July 2002 from http://www.msiac.dmso.mil/vva/ref_docs/val_lawref/default.htm#toc1, 15 August 2001.

Department of Defense Modeling and Simulation Office (DMSO) Verification, Validation, and Accreditation (VV&A) Recommended Practices Guide (RPG) Reference Document - Human Behavior Representation (HBR) Literature Review. Retrieved 27 June 2002 from http://www.msiac.dmso.mil/vva/Ref_Docs/HBR/beh-ref-pr.PDF, 15 August 2001.

Department of Defense Modeling and Simulation Office (DMSO) Verification, Validation, and Accreditation (VV&A) Recommended Practices Guide (RPG) Reference Document - Key Concepts of VV&A. Retrieved 27 June 2002 from http://www.msiac.dmso.mil/vva/ Key/key.pr.pdf, 15 August 2001.

Department of Defense Modeling and Simulation Office (DMSO) Verification, Validation, and Accreditation (VV&A) Recommended Practices Guide (RPG) Special Topic - Validation of Human Behavior Representations. Retrieved 27 June 2002 from http://www.msiac.dmso.mil/vva/Special_topics/hbr-Validation/default.htm, 25 September 2001.

Department of Defense (DoD) 5000.59-M: Modeling and Simulation (M&S) Glossary, December 1997.

Department of Defense (DoD) 5000.59-P: Modeling and Simulation (M&S) Master Plan, Retrieved 26 June 2002 from http://www.dmso.mil, October 1995.

Druckman, D., & Swets, J. A. (Eds.). *Enhancing Human Performance: Issues, Theories, and Techniques*. Washington, DC: National Academy Press, 1988.

Duffy, L. Team Decision-Making Biases: An Information-Processing Perspective. In G. A. Klein, J. Orasanu, R. Calderwood & C. E. Zsambok (Eds.), *Decision-making in Action: Models and Methods* (pp. 346-359). Norwood, NJ: Ablex Publishing, 1993.

Encyclopedia Britannica. Encyclopedia Britannica, Inc. Retrieved 2 September 2002 from http://www.britannica.com/, 4 February 2002).

Ferber, J. Multi-Agent Systems - An Introduction to Distributed Artificial Intelligence, Reading, MA: Addison-Wesley, 1999.

Freedman, A. *The Computer Desktop Encyclopedia*_(CD Rom - Ver 12.1). Point Pleasant, PA: Computer Language Company Inc, 1999.

Gale Group. The Gale Encyclopedia of Psychology, (2nd Ed). Detroit, MI: Gale Group. Retrieved 12 December 2002 from http://www.findarticles.com/cf_0/g2699/0000/2699000080/p1/article.jhtml, 2001.

Galligan, D. P., Anderson, M. A., & Lauren, M. K. *MANA, Map Aware Non-uniform Automata, Version 3.0, Users Manual (Draft)*. Unpublished manuscript, 2003.

Gawron, V. J. *Human Performance Measures Handbook.* Mahwah, N.J.: Lawrence Erlbaum Associates, 2000.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). *Heuristics and Biases: The Psychology of Intuitive Judgement*. Cambridge, NY: Cambridge University Press, 2002.

Goerger, S. R. *Validating Computational Human Behavior Models: Consistency and Accuracy Issues*. Unpublished Dissertation, Naval Postgraduate School, Monterey, CA, June 2004.

Goerger, S. R. *Validating Human Behavioral Models for Combat Simulations Using Techniques for the Evaluation of Human Performance.* Paper presented at the 2003 Summer Computer Simulation Conference, Montreal, Quebec, Canada, 2003.

Goerger, S. R. *Validation and Evaluation of Cognitive Architectures Using an Emergent Combat Model.* Presentation at the MOVES Open House, Monterey, CA, 22 August 2002.

Gonzalez, A.J. and Murillo, M. *Validation of Human Behavior Models.* (Paper Number *99S-SIW-010)* Simulation Interoperability Standards Organization's (SISO) 1999 Spring Simulation Interoperability Workshop (SIW). Retrieved 24 July 2002 from http://www.sisostds.org/doclib/doclib.cfm?SISO_FID_2442, 1998.

Gray, W.D. *Summary of the AMBR Expert Panel Report.* Retrieved 21 May 2002 from https://www.williams.af.mil/AMBR/ AMBR1_Gray.ppt, July 2000.

Harmon, S. Y. *Levels of VV&A.* Presentation at the Navy Modeling and Simulation Management Office (NAVMSMO) Verification, Validation & Accreditation (VV&A) Technical Working Group (TWG) Workshop #14, Naval Postgraduate School, Monterey, CA, 4 August 2003.

Harmon, S.Y. (Ed.). *Fidelity ISG Glossary.* Ver 3.0. Simulation Interoperability Standards Organization (SISO), Fidelity Implementation Study Group (ISG), Retrieved 2 July 2002 from http://www.sisostds.org/doclib/doclib.cfm?SISO_RID_1000789, 16 December 1998.

Harvey, C.M. *Cognitive Task Analysis.* (briefing slides) Dayton, OH: Dept. of Biomedical, Industrial, and Human Factors Engineering, Wright State University Retrieved 4 September 2002 from http://champ.cs.wright.edu/ai/hfe733-01/cta.pdf, 2001.

Hodges, J. and Dewar, J. Rand Corporation Report R-4114-RC/AF; Is It Your Model Talking? A Framework for Model Validation. Santa Monica, CA: Rand Corporation, 1992.

Holland, J.H. Hidden Order: How applications Build Complexity. Cambridge, MA: Perseus Books, 1995.

Hutchins, S. G., Kelly, R. T., & Morrison, J. G. *Decision Support for Tactical Decision-making Under Stress.* Paper presented at the 40th Human Factors and Ergonomics Society Annual Meeting, Monterey, CA, 25-28 June 1996a.

Hutchins, S. G., Kelly, R. T., & Morrison, J. G. *Principles for Aiding Complex Military Decision-making.* Paper presented at the 40th Human Factors and Ergonomics Society Annual Meeting, Monterey, CA, 25-28 June 1996a.

Ilachinski, A. CRM 97-61.10: Irreducible Semi-Autonomous Adaptive Combat (ISAAC): An Artificial-Life Approach to Land Warfare (U). Alexandria, VA: Center for Naval Analyses, 1997.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). Judgement Under Uncertainty: Heuristics and Biases. Cambridge, NY: Cambridge University Press, 1982.

King, L. M., Hunter, J. E., & Schmidt, F. L. Halo in a Multi-Dimensional Forced-Choice Performance Evaluation Scale. *Journal of Applied Psychology, 65*, 507-516, 1980.

Klein, G. A. An Overview of Naturalistic Decision-Making Applications. In C. E. Zsambok & G. A. Klein (Eds.), Naturalistic Decision-Making: Expertise-Research and Applications (pp. 49-59). Mahwah, NJ: Lawrence Erlbaum Associates, 1997.

Klein, G. A. *Cognitive Task Analysis of Teams.* In J. M. Schraagen, S. F. Chipman & V. L. Shalin (Eds.), *Cognitive Task Analysis* (pp. 417-429). Mahwah, NJ: Lawrence Erlbaum Associates, Inc, 2000.

Klein, G. *Sources of Power; How People Make Decisions.* Cambridge, MA: The MIT Press, 2001.

Lewis, T.G., Zyda, M., and Hiles, J.E *Proposal to Establish a Center for Study of Potential Outcomes.* (Unpublished Technical Report) Monterey, CA: Naval Postgraduate School, Modeling of Virtual Environments and Simulations (MOVES) Institute, 2002.

Lipshitz, R. Naturalistic Decision-Making Perspectives on Decision Errors. In C. E. Zsambok & G. A. Klein (Eds.), *Naturalistic Decision-Making: Expertise-Research and Applications* (pp. 151-162). Mahwah, NJ: Lawrence Erlbaum Associates, 1997a.

Lipshitz, R., & Shaul, O. B. Schemata and Mental Models in Recognized-Primed Decision-making. In C. E. Zsambok & G. A. Klein (Eds.), *Naturalistic Decision-Making: Expertise-Research and Applications* (pp. 293-204). Mahwah, NJ: Lawrence Erlbaum Associates, 1997b.

Mallery, J. C. *Thinking About Foreign Policy: Finding an Appropriate Role for Artificially Intelligent Computers.* Paper presented at the 1988 Annual Meeting of the International Studies Association, Adam's Mark Hotel, St. Louis, MO, 28 March - 03 April 1988.

Merriam-Webster's Collegiate Dictionary. Retrieved 18 July 2002 from http://www.m-w.com/cgi-bin/dictionary, 2002.

Merriam-Webster's Collegiate Dictionary. Retrieved 9 December 2003 from http://www.m-w.com/cgi-bin/dictionary, 2003.

Miller, N. L., & Shattuck, L. G. *Human Performance; Report on DARPA's Future Combat System Command and Control, Experiment 4a, 28 September - 2 October 2003*: Defense Advanced Research Projects Agency (DARPA), 2004.

Miller, T. E., & Woods, D. D. Key Issues for Naturalistic Decision-Making Researchers in System Design. In C. E. Zsambok & G. A. Klein (Eds.), *Naturalistic Decision-Making: Expertise-Research and Applications* (pp. 141-150). Mahwah, NJ: Lawrence Erlbaum Associates, 1997.

Osborne, B.A. *An Agent-Based Architecture for Generating Interactive Stories.* (Unpublished Dissertation). Monterey, CA: Naval Postgraduate School, Computer Science Department, September 2002.

Pace, D. K., & Sheehan, J. *Subject Matter Expert (SME)/Peer Use in M&S V&V.* Paper presented at the Foundations '02, a Workshop on Model and Simulation Verification and Validation for the 21st Century, Kossiakoff Conference & Education Center, Johns Hopkins University Applied Physics Laboratory, Laurel, MD, 22-24 October 2002.

Perrin, B. M., Barnett, B. J., & Walrath, L. D. *Techniques to Reduce Bias in Human Decision-making for Tactical Decision-making Under Stress (Tasks 2 & 3).* St. Louis, MO: McDonnell Douglas Corporation, 1993.

Pew, R. W., & Mavor, A. S. (Eds.). *Modeling Human and Organizational Behavior: Application to Military Simulations.* National Academy Press, Washington, DC, 1998.

Prensky, M. Digital Natives, Digital Immigrants. *On the Horizon, 9*(5), 1-6, 2001.

Project Albert Fact Description. Retrieved 1 April 2004 from http://www.mcwl.quantico.usmc.mil/divisions/albert/index.asp, 2001.

Project Albert Fact Sheet (). Retrieved 1 April 2004 from http://www.mcwl.quantico.usmc.mil/fact_sheets/fs/Pro%20Albert%2007_31_03.pdf, 10 December 2002.

Ralston, A., Reilly, E.D. and Hemmendinger, D. *Encyclopedia of Computer Science* (4th ed). New York, NY: Grove's Dictionaries, Inc, 2000.

Roddy, K. and Dixon, M. *Modeling Human and Organizational Behavior using a Relation-Centric Multi-Agent System Design Paradigm.* (Unpublished Master's Thesis) Monterey, CA: Naval Postgraduate School, MOVES Academic Group, September 2000.

Russell, S. & Norvig, P. *Artificial Intelligence: A Modern Approach.* Upper Saddle River, NJ: Prentice Hall, 1995.

Sage, A. P. Behavioral and Organizational Considerations in the Design of Information Systems and Processes for Planning and Decision Support. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-11*(9), 640-678, 1981.

Schelling, T. C. *Micromotives and Macrobehavior.* New York, NY: W.W. Norton & Co, 1978.

Shoham, Y. *Agent-Oriented Programming.* Artificial Intelligence, 60(1):51--92, 1993.

Smith, P. C., & Kendall, L. M. Retranslation of Expectations: An Approach to the Construction of Unambiguous Anchors for Rating Scales. *Journal of Applied Psychology, 47* (149-155), 1963.

Solso, R. L. *Cognitive Psychology* (6th ed.). Boston, MA: Allyn & Bacon, 2000.

Statkus, M. J., Sampson, J. B., & Woods, R. J. Human Science/Modeling and Analysis Data Project: Situation Awareness Effects on Troop Movement and Decision-making Data Collection Effort, 21 October Through 1 November 2002 (Technical Report No. NATICK/TR-03/033L). Natick, MA: U.S. Army Soldier & Biological Chemical Command, Natick Soldier Center, 2003.

Stein, R., & Stein, M. Sources of Bias and Inaccuracy in the Development of a Best Estimate. Casualty Actuarial Society Forum, 45, 1998.

Stufflebeam, D. L. *Guidance for Choosing and Applying Evaluation Checklists*, Retrieved 12 December 2003 from http://www.wmich.edu/evalctr/checklists/checklistorganizer.htm, 19 November 2002.

Tolk, A. *Human Behaviour Representation - Recent Developments.* Paper presented at the NATO Research & Technology Organization (RTO); Studies, Analyses and Simulation Panel (SAS); Lecture Series on "Simulation of and for Military Decision-making", The Hague, Netherlands, 10-11 December 2002.

Tversky, A., & Kahneman, D. Belief in the Law of Small Numbers. *Psychological Bulletin, 76,* 105-110, 1971.

Tversky, A., & Kahneman, D. Judgment Under Uncertainty: Heuristics and Biases. *Science, 185*, 1124-1130, 1974.

Tziner, A., Joanis, C., & Murphy, K. R. A Comparison of Three Methods of Performance Appraisal with Regard to Goal Properties, Goal Perception and Ratee Satisfaction. *Group and Organization Management, 25*(2), 175-190, 2000.

Wilson, R.A., & Keil F.C. *The MIT Encyclopedia of the Cognitive Sciences* (CD Rom). Cambridge, MA: MIT Press, 1999.

Wray, R., Chong, R. Phillips, J., Rogers, S., and Walsh, B. *A Survey of Cognitive and Agent Architectures.* Ann Arbor, MIP: Department of Electrical Engineering and Computer Science, University of Michigan. Retrieved 6 September 2002 from http://ai.eecs.umich.edu/cogarch0/ index.html, 1992.

Zsambok, C. E. Naturalistic Decision-Making: Where are We Now? In C. E. Zsambok & G. A. Klein (Eds.), *Naturalistic Decision-Making: Expertise-Research and Applications* (pp. 3-16). Mahwah, NJ: Lawrence Erlbaum Associates, 1997.

Appendix A: List of Abbreviations

1	
1LT	First Lieutenant
A	
ABM	Agent-Based Model
ACT-R	Adaptive Control of Thought
AFOR	Automated FORces
AFRL	Air Force Research Laboratory
AI	Artificial Intelligence
AL	Artificial Life
AMBR	Agent-based Modeling and Behavior Representation
AMSO	Army Model and Simulation Office
ANOCAT	Analysis of Categorical Data
ARTEP	Army Training and Evaluation Program
B	
BARS	Battlefield Augmented Reality System
B.S.	Bachelor of Science
C	
CAS	Complex Adaptive System
CFOR	Command FORces
CGFs	Computer Generated Forces
CMAS	Connector-Based Multi-Agent System
COGNET	COGnition as a NETwork of Tasks
COMBAT^{XXI}	Combined Arms Analysis Tool for the XXIst Century
CPT	Captain
CTA	Cognitive Task Analysis
D	
DAE	Difference Analysis Engine
D-COG	Distributed Cognition (AFRL's agent-based modeling architecture)
DMSO	Defense Modeling and Simulation Office
DoD	Department of Defense
DOTSE	Defence Operational Technology Support Establishment (New Zealand)
D.Sc.	Doctorate of Science
DSS	Decision Support System
E	
EINSTein	Enhanced ISAAC Neural Simulation Toolkit
EPIC	Executive-Process Interaction Control
EPIC-Soar	Executive-Process Interaction Control - State, Operator and Result
F	
FDC	Fire Direction Center
FO	Forward Observer
H	
HBR	Human Behavior Representation

HBTWG	Human Behavior Technology Working Group
I	
ICCC	Infantry Captains Career Course
iGEN	
ISAAC	Irreducible Semi-Autonomous Adaptive Combat
L	
LTC	Lieutenant Colonel
M	
M&S	Modeling and Simulation
M16A2	Assault Rifle
MAJ	Major
MANA	Map Aware Non-uniform Automata
MAS	Multi-Agent System
MCCDC	Marine Corps Combat Development Command
MIT	Massachusetts Institute of Technology
MOUT	Military Operations in Urban Terrain
MOVES	Modeling, Virtual Environment, and Simulation
M.S.	Masters of Science
MTP	Mission Training Plan
N	
NAVMSMO	Navy Modeling & Simulation Management Office
NDM	Naturalistic Decision-Making
NEO-FFI	Neuroticism, Extraversion, and Openness Five-Factor Inventory
NPS	Naval Postgraduate School
O	
OOTW	Operations Other Than War
OPMS XXI	Officer Personnel Management System XXI
P	
Ph.D.	Doctorate of Philosophy
R	
RPD	Recognition-Primed Decision model
RPG	Recommended Practices Guide (DMSO V&V TWG)
RPI	Rensselaer Polytechnic Institute
S	
SAF (SAFOR)	Semi-Automated Forces
SME	Subject Matter Expert
Soar	State, Operator and Result (Model)
T	
TADMUS	Tactical Decision-making Under Stress
TRAC	Training and Doctrine Command Analysis Center
TWG	Technical Working Group
U	
UML	Unified Modeling Language
USMA	United States Military Academy

V	
V&V	Verification and Validation
VKB	Validation Knowledge Base
VV&A	Verification, Validation and Accreditation

*This table is sorted alphabetically

Appendix B: Glossary

The following definitions for terms used in this report are excerpted from Department of Defense Directive 5000.59, *DoD Modeling and Simulation (M&S) Management*; DMSO's *VV&A Recommended Practices Guide,* "Key Concepts;" Gary Klein's *Sources of Power; How People Make Decisions*; DMSO's *Human Behavior Representation (HBR) Literature Review*; and other DoD and professional publications.

1. Accreditation

"The official certification that a model, simulation, or federation of models and simulations and its associated data are acceptable for use for a specific purpose." (Department of Defense Directive 5000.59, 4 January 1994) This is the final stage of the verification, validation and accreditation (VV&A) process. ***Accreditation*** is the "official" seal of approval by the designated authority that the model is verified and valid for its intended purpose.

2. Accuracy

For this report, ***accuracy*** is defined as the SME's average difference between the assessment key and the SME's assessment of each observation, where a difference is the assessment value from the key minus the assessment value of the SME for a given subtask, task, scenario, or overall question.

3. Accuracy Impact

For this report, ***accuracy impact*** is defined as the SME's average difference between the assessment key and the SME's assessment of each observation, where a difference refers to a change from Go to No-Go, Go to Unknown, No-Go to Go, No-Go to Unknown, Unknown to Go, or Unknown to No-Go.

4. Anchoring Bias

Anchoring bias emerges when an individual embraces an initial hypothesis and maintains this view regardless of incoming facts. This results in overemphasis on the hypothesis and an inappropriately minimal shift from the initial viewpoint (Tversky & Kahneman, 1974) (Kahneman, et al, 1982) (Cohen, 1993) (Duffy, 1993) (Stein & Stein, 1998) (Perrin, et al, 1993).

5. Assessment

An ***assessment*** or ***rating*** is the value (based on scale) an individual SME gives an observed model or human behavior.

6. Assessment Key

The ***assessment key*** is a set of subtask assessments tallied and averaged to produce tasks assessments, which when tallied and averaged produce scenario responses. The average value for the scenario responses determines the overall assessment of the behaviors. Each scale has its own assessment key and all assessment keys are consistent with each other.

7. Bias

As defined by *Webster's Dictionary*, ***bias*** is the "systematic error introduced into sampling or testing by selecting or encouraging one outcome or answer over others" (Merriam-Webster's Collegiate Dictionary, 2003).

8. Cognitive Task Analysis

A "***cognitive task analysis*** is a method for capturing expertise and making it accessible for training and system design." It results in a "... description of the expertise needed to perform complex tasks." It consists of five steps: (1) identifying sources of expertise; (2) assaying the knowledge; (3) extracting the knowledge; (4) codifying the knowledge; and (5) applying the knowledge (Klein, 2001).

9. Confirmation Bias

Confirmation bias is demonstrated when a SME overvalues select pieces of information relative to consistent evidence indicating an alternate conclusion (Cohen, 1993) (Duffy, 1993) (Stein & Stein, 1998) (Perrin, et al, 1993).

10. Consistency

For this research, a SME's ability to maintain logical correspondence between the average sublevel response score and the level score is ***consistency***. In other words, SMEs derive level responses logically/directly from sublevel responses.

11. Consistency Impact

For this research, the degree to which a SME's consistency/inconsistency influences the assessment of the model by changing a SME's results between sublevel and level from Go to No-Go, Go to Unknown, No-Go to Go, No-Go to Unknown, Unknown to Go, or Unknown to

No-Go is ***consistency impact***. In other words, does the inconsistency, when present, make a practical difference in the outcome of the assessment.

12. Contrast Bias

Contrast bias materializes when one seeks information to contradict an original hypothesis, ignoring or undervaluing evidence in support of the hypothesis (Tversky & Kahneman, 1974) (Kahneman, et al, 1982) (Perrin, et al, 1993).

13. Correspondence

Correspondence is "the agreement of things with one another" (Merriam-Webster's Collegiate Dictionary, 2002). In the validation domain, this term is used to describe the agreement of a model to different levels of abstraction. There are at least six levels of correspondence used in HBR validation: computational, domain, physical, physiological, psychological, and sociological (Department of Defense Modeling and Simulation Office, 15 August 2001b) .

14. Credibility

Credibility is "the relevance that the user sees in a model and the confidence that the user has that a model or simulation can serve his purpose" (Department of Defense Modeling and Simulation Office, 15 October 2001).

15. Decision Bias

According to Cohen, ***decision bias*** is "a systemic flaw in the internal relationships among a person's judgments, desires, and /or choices" (Cohen, 1993).

16. Evaluation

Evaluation is a means of determining how well a model agrees with the portion of the real world it is simulating. It is a less stringent means of agreement then validation and is usually based on qualitative versus quantitative data. It is used to assess the model's quality when a model is non-predictive or incapable of validation (Hodges & Dewar, 1992).

17. Fidelity

"The degree to which a model or simulation reproduces the state and behavior of a real-world object or the perception of a real-world object, feature, condition, or chosen standard in a measurable or perceivable manner; a measure of the realism of a model or simulation; faithfulness. ***Fidelity*** should generally be described with respect to the measures, standards, or perceptions used in assessing or stating it" (Harmon, 16 December 1998). The higher the model's ***fidelity***, the more it corresponds to the complexities and represents the real-world element it is simulating. This term is qualitative in nature and is based on a sliding scale. It is best used to distinguish the relative placement of two or more models with respect to each other.

18. Heuristic Bias

Heuristic bias is based on the belief that humans use "mental short-cuts" for quick assessment and decision making. Through the use of heuristics, experts make decisions without detailed exploration and analysis of the problems space and all possible solutions. This allows for an acceptable although not necessarily optimal assessment of the situation or solution to an issue (Stein & Stein, 1998).

19. Human-Behavior Representation

A ***human-behavior representation*** (HBR) is "a model or simulation of any human function, any individual human, or any group or organization of humans." (Department of Defense Modeling and Simulation Office, 15 October 2001) In this research, HBR will refer to the human cognitive process.

20. Human-Behavior Representation Knowledge Base

"The ***HBR's knowledge base*** contains the computer program that determines the HBR's response to the stimuli it receives from the simulated world. At a minimum, the knowledge base largely determines the HBR's cognitive behavior. It may also contribute to the manifestations of emotion upon behavior" (Department of Defense Modeling and Simulation Office, 25 September 2001).

21. Informational Bias

Informational or cognitive bias occurs when individuals use "intuitive strategies" to acquire and analysis information rather than using proven "optimal" methodologies. This results in the improper interpretation and presentation of data leading to non optimal solutions or improper conclusions. Sage describes twenty seven types of cognitive bias (Sage, 1981).

22. Level

The assessment of behaviors is broken into three separate ***levels*** (e.g. task, scenario, and overall) which consist of sublevel assessments (e.g. subtasks, tasks, and scenarios, respectively). These create ***level and sublevel pairings*** (e.g. subtask to task, task to scenario, and scenario to overall).

23. Model

A ***model*** is "a physical, mathematical, or otherwise logical representation of a system, entity, phenomenon, or process" (Department of Defense Modeling and Simulation Office, 15 October 2001).

24. Naturalistic Decision-Making

Naturalistic decision-making (NDM) is the study of how people use their experiences to make decisions in real-world situations. Its focus is on time-pressured decision-making processes used by experts when information is missing or ambiguous, goals are vague, and conditions are changing (Klein, 2001).

25. Normative Bias

Normative bias is concerned with the interaction between individuals who provide information or skills to the community in order to resolve an issue or cultivate a conclusion (Duffy, 1993) .

26. Overall

The ***overall*** assessment is the final judgment of the model/individual performance derived from a collection of scenarios. For this research, the overall assessment is how well the SME feels the individuals and leader performed their roles.

27. Participant/Rater

A ***participant*** or ***rater*** is an individual taking part in the experiments who performs an assessment of observed model/human behaviors. The participants in this research come from a pool of 182 Army and USMC officers enrolled in the Infantry Captains Carrier Course at Fort Benning, GA. This document refers to these individuals as subject-matter experts (SMEs).

28. Perception Bias

Perception bias is that which an expert brings to the process based on his education, training, real-world experiences, exposure to simulations, and organizational loyalties. These factors color the lenses of the SME's microscope or unduly focus the search area on certain aspects of a model's performance (Pace & Sheehan, 22-24 October 2002).

29. Performance Bias

Performance bias deals with the SME's ability to execute the validation process. This ability may be hampered by other demands on the SME's time, the inavailability of data, a low ability or desire to comply with specified validation procedures, or the expert's failure to understand the simulation (Pace & Sheehan, 22-24 October 2002).

30. Perspective Bias

Perspective bias occurs when a SME's fails to maintain focus on the intended purpose of the model. A SME may lose focus as he allows his real-world experiences to cloud his view on what the model should have the capability of doing (Pace & Sheehan, 22-24 October 2002).

31. Rating

An ***assessment*** or ***rating*** is the value (based on scale) an individual SME gives an observed model or human behavior.

32. Referent

"A codified body of knowledge about a thing being simulated" (Harmon, 16 December 1998). In the case of HBR and this research, this would consist of at least one of the six levels of correspondence. ***Referent*** is the best information we have about the simulated object's functionality and performance. The referent provides the standards against which the results of models and simulations are compared, to assess the level of fidelity they are able to replicate (Department of Defense Modeling and Simulation Office, 30 November 2000b) (Department of Defense Modeling and Simulation Office, 25 September 2001).

33. Resolution

Different from fidelity, ***resolution*** is "the degree of detail used to represent aspects of the real world or a specified standard or referent by a model or simulation" (Department of Defense Modeling and Simulation Office, 15 October 2001). Resolution often refers to the visual characteristics of a model.

34. Scale

A ***scale*** is a set of possible assessment responses SMEs can use to quantify the level of performance of the observed behavior. Three scales are used in this research. Scale 1 is a seven-point Likert scale, where a seven represents the SME's highest confidence the model or individual performed to standard and one indicates the SME's certainty that the model or individual failed to perform to standard. Scale 2 is a Go/No-Go scale where a Go indicates the SME's belief the model or individual performed to standard and No-Go indicates the belief that the model or individual failed to perform to standard. Scale 3 is a five-point Likert scale where five represents the SME's highest confidence the model or individual performed to standard and one indicates the SME's utmost confidence the model or individual failed to perform the behavior to standard.

Appendix C. Key Players in Verification, Validation and Accreditation

Table 8. outlines the roles of key players in the DoD modeling and simulation VV&A process. This table is excerpted from DMSO's VV&A Recommended Practices Guide Reference Document, "Key Concepts of VV&A".

Table 8. Typical Roles and Responsibilities Associated with Modeling and Simulation Verification, Validation and Accreditation From (Department of Defense Modeling and Simulation Office, 15 August 2001c)

Role Activity	User	M&S PM	Developer	V&V Agent	Accreditation Agent	SME
Define Requirements	Lead Approve	Monitor	Assist	Review	Review	Assist
Define Measures	Lead Approve	Monitor	Assist	Assist	Assist	Assist
Define Acceptability Criteria	Assist Approve	Monitor	Assist	Assist	Lead	Assist
Plan M&S Development or Modification*	Assist	Lead Approve	Assist	Assist		
Develop V&V Plans	Review	Assist Approve	Review	Lead	Assist	
Develop Accreditation Plan	Review Approve	Assist		Assist	Lead	
Verify Requirements	Lead-alt Approve	Monitor	Assist	Lead	Assist	Assist
Develop Conceptual Model**	Assist Approve	Monitor	Lead			Assist
Validate Conceptual Model	Assist Approve	Monitor	Assist	Lead		Assist
Develop Design***		Monitor Approve	Perform			
Verify Design	Approve	Monitor	Assist	Lead		Assist
Implement Design		Monitor Approve	Perform			
V&V Data	Approve	Monitor	Assist	Lead		Perform
Verify Implementation	Approve	Monitor	Assist	Lead		Assist
Test Implementation	Approve	Monitor	Lead	Assist		Assist

Role Activity	User	M&S PM	Developer	V&V Agent	Accreditation Agent	SME
Validate Results	Assist Approve	Monitor	Assist	Lead		Assist
Prepare V&V Report				Perform		
Configure for Use	Assist	Lead Approve	Assist			
Gather Additional Accreditation Info	Monitor	Assist		Assist	Lead	Assist
Conduct Accreditation Assessment	Monitor				Perform	Assist
Prepare Accreditation Assessment Report					Perform	
Determine Accreditation	Perform					
Prepare Accreditation Report					Perform	

Lead Leads the task. Normally involves active participation from others

Perform Actually does the task. Normally involves little active participation from others

Assist Actively participates in task (e.g., conducting tests, providing information)

Review Participation normally limited to reviewing results of task and providing recommendations

Monitor Oversees task to ensure it is done appropriately but does not normally participate

Approve Determines when an activity is satisfactorily completed and another can begin. Determines what activity should be pursued next (e.g., whether to continue on to the next scheduled activity or to return to a previous activity).

*This activity refers to planning and scheduling of any M&S development, modification, or preparation

**This activity refers to development of new as well as modification of existing conceptual models

***This activity refers to development of new M&S designs as well as modification of existing M&S designs

Distribution List

The list indicates the complete mailing address of the individuals and organizations receiving copies of the report and the number of copies received. Due to the Privacy Act, only use business addresses; no personal home addresses. Distribution lists provide a permanent record of initial distribution. The distribution information will include the following entries:

NAME/AGENCY	ADDRESS	COPIES
Author(s)	Department of Systems Engineering Mahan Hall West Point, NY 10996	2
Dean, USMA	Office of the Dean Building 600 West Point, NY 10996	1
Defense Technical Information Center (DTIC)	ATTN: DTIC-O Defense Technical Information Center 8725 John J. Kingman Rd, Suite 0944 Fort Belvoir, VA 22060-6218	1
Department Head-DSE	Department of Systems Engineering Mahan Hall West Point, NY 10996	1
ORCEN	Department of Systems Engineering Mahan Hall West Point, NY 10996	5
ORCEN Director	Department of Systems Engineering Mahan Hall West Point, NY 10996	1
USMA Library	USMA Library Bldg 757 West Point, NY 10996	1

REPORT DOCUMENTATION PAGE – SF298

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

1. REPORT DATE *(DD-MM-YYYY)* 24-05-2005	**2. REPORT TYPE** Technical Report	**3. DATES COVERED** *(From - To)* 01 January 2003-24 May 2005

4. TITLE AND SUBTITLE

A Validation Methodology for Human Behavior Representation Models

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Lieutenant Colonel Simon R. Goerger

Colonel Michael L. McGinnis

Dr Rudolph P. Darken

5d. PROJECT NUMBER
DSE-R-0530

5e. TASK NUMBER

5f. WORK UNIT NUMBER
845-938-5661

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Department of Systems Engineering
US Military Academy
West Point, NY 10996

8. PERFORMING ORGANIZATION REPORT NUMBER

DSE-R-0530

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Navy Modeling and Simulation
2000 Navy Pentagon
Washington, DC

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Distribution A: Approved for public release; distribution is unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The Department of Defense relies heavily on mathematical models and computer simulations to analyze and acquire new weapon systems. Models and simulations help decision-makers understand the differences between systems and provide insights into the implications of weapon system tradeoffs. Given this key role, the credibility of simulations is paramount. For combat models, this is gained through the verification, validation, and accreditation process required of DoD analytical models prior to their use in weapon system acquisition and other studies. The nature of nondeterministic human behavior makes validation of models of human behavior representation contingent on the judgments of subject matter experts that are routinely acquired using a face validation methodology. In an attempt to better understand the strengths and weaknesses of assessing human behavior representation using experts and the face validation methodology, the authors conducted experiments to identify issues critical to utilizing human experts for the purpose of ascertaining ways to enrich the validation process for models relying on human behavior representation. The research was limited to the behaviors of individuals engaged in close combat in an urban environment. This paper presents the study methodology, data analysis, and recommendations for mitigating attendant problems with validation of human behavior representation models.

15. SUBJECT TERMS

Validation, Cognitive Model, Modeling and Simulations, Human Behavior Representation, Bias, Multi-Agent Systems, Behavioral Psychology, Cognitive Psychology, VV&A, Human Performance Evaluation

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** LTC Simon R. Goerger
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	None	99	**19b. TELEPHONE NUMBER** *(include area code)* 845-938-5535