A Validation Methodology for Human Behavior Representation Models

Lieutenant Colonel Simon R. Goerger, Ph.D.
Colonel Michael L. McGinnis, Ph.D.
United States Military Academy
Department of Systems Engineering
West Point, NY
simon.goerger@usma.edu

Rudolph P. Darken, D.Sc.
Naval Postgraduate School
Modeling, Virtual Environments & Simulation
Monterey, CA
darken@nps.edu

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 AUTHORS

Lieutenant Colonel Simon R. Goerger currently serves 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.

A Validation Methodology for Human Behavior Representation Models

Lieutenant Colonel Simon R. Goerger, Ph.D.
Colonel Michael L. McGinnis, Ph.D.
United States Military Academy
Department of Systems Engineering
West Point, NY
simon.goerger@usma.edu

Rudolph P. Darken, D.Sc.
Naval Postgraduate School
Modeling, Virtual Environments & Simulation
Monterey, CA
darken@nps.edu

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.

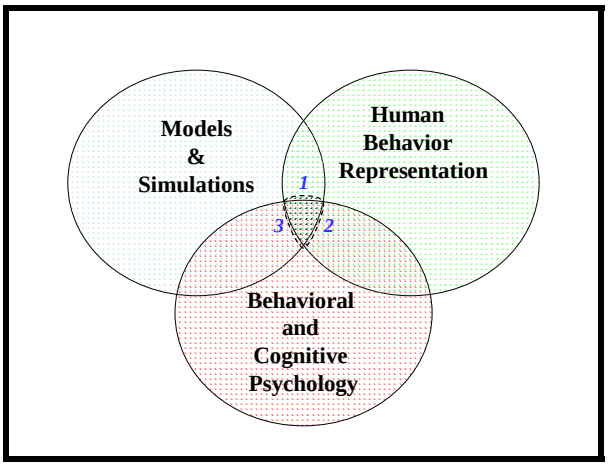


Figure 1. Research Objective: To Define the Common Area

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

[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.

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]

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.

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 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.

EXPERIMENT

Studies conducted in support of this research were designed to investigate the aptitude of SMEs to assess the face validity[4] 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.

Simulation Environment

The layout of the McKenna military operations in urban terrain (MOUT) Site, Fort Benning, GA (Figure 2) 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.

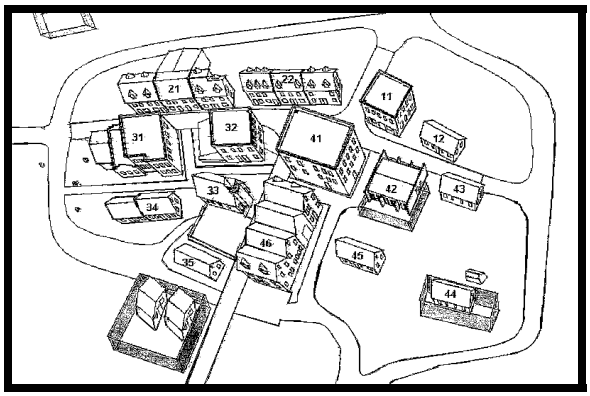


Figure 2. McKenna Test Environment Sketch From (Statkus, 2003)

[3] Naturalistic decision-making is "the study of how people use their experience to make decisions in field settings" (Klein, 2001).

[4] 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.

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.

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.

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) 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.

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.

RESULTS

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.

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.[5] Figure 3 illustrates a performance bias response pattern. Of 159 questions, SME B2124 only responded to 16 (10%). Based on his comments, B2124 felt the simulation failed to furnish enough information to make an assessment. Of the 182 SMEs, 23 (13 %) displayed performance bias.

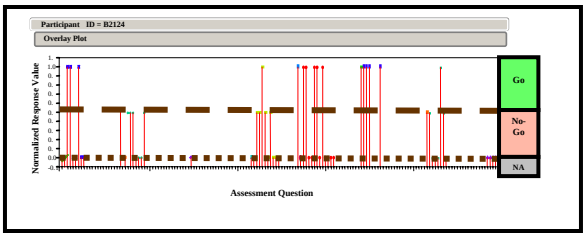


Figure 3. Performance Bias Example

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 4 illustrates two different anchoring bias response patterns. Participant B1102's responses are an example of positive anchoring bias with only two responses after the second task being assessed as negative. Participant B2204's responses show an opposite trend and are an example of negative anchoring bias. Thirty SMEs (16%) displayed anchoring bias.

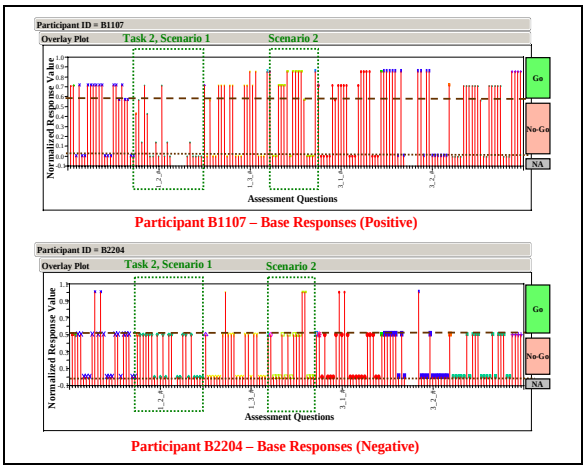


Figure 4. Anchoring Bias Examples

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 responses show a general trend of either harsher or more lenient but the overall response differs from this trend. Figure 5 illustrates these two different response patterns of confirmation bias. Data from 55 SMEs (30%) displays confirmation bias.

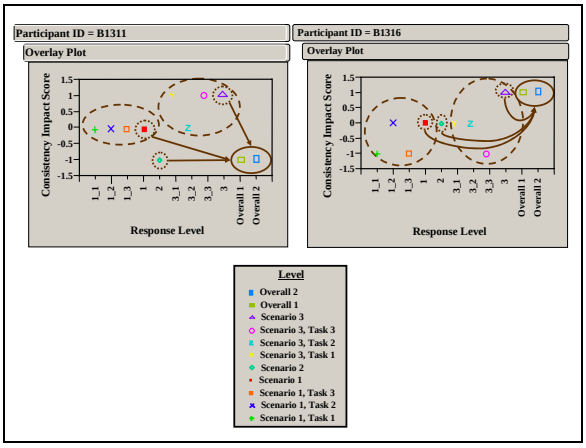


Figure 5. Confirmation Bias Examples

Contrast bias materializes when a SME contradicts an original hypothesis, ignoring or undervaluing evidence

[5] A definitive response is a "Go" or "No-Go" response to an assessment question. "Not Applicable" or "No Opinion" responses are not definitive responses.

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 indicates a shift in a SME's accuracy trend, from harsher to more lenient or from more lenient to harsher, as the assessment process proceeds. Figure 6 combines SME raw data and accuracy plots to demonstrate contrast bias. The SME's accuracy score plot illustrates that nine of the first 45 responses (20%) were harsher than the key assessment responses. However, after assessing task number two, the SME scored 65 of the remaining 114 responses (57%) harsher. Five SMEs (3%) displayed contrast bias.

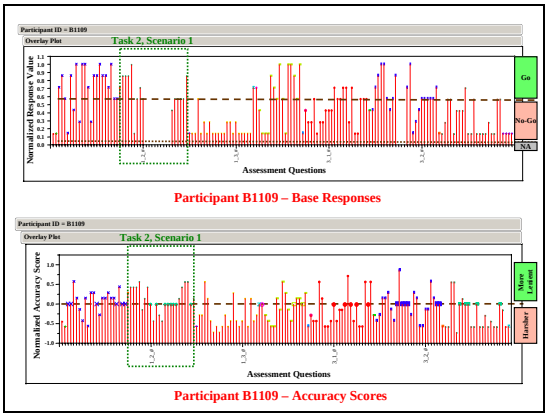


Figure 6. Contrast Bias Example

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 1 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 1 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 1. Mean Values for Normalized, Overall Assessment Scores

ID		Number of SMEs		Mean (Normalized 0-1 Responses)			
Simulation Belief	**Scale**	**Overall 1 & Overall 2**	**Overall 3 & Overall 4**	**Question Overall 1**	**Question Overall 2**	**Question Overall 3**	**Question Overall 4**
0	1	37	36	0.583	0.598	0.540	0.552
0	2	25	25	0.920	0.920	0.920	0.940
0	3	24	24	0.483	0.500	0.442	0.433
1	1	39	39	0.667	0.696	0.593	0.623
1	2	25	25	0.820	0.820	0.780	0.800
1	3	25	25	0.616	0.664	0.600	0.632
All Beliefs and Scales		175	174	0.675	0.694	0.636	0.654

Figure 7 illustrates inter-SME consistency between SME responses when observing and assessing the same behavior event via the model interface. This inconsistency precludes consistent and accurate assessment of the simulation. Fifty (31.45 %) subtasks, tasks, scenarios, and overall assessment responses plots exhibit inconsistent distributions.

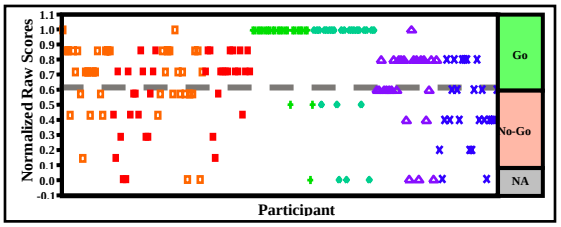


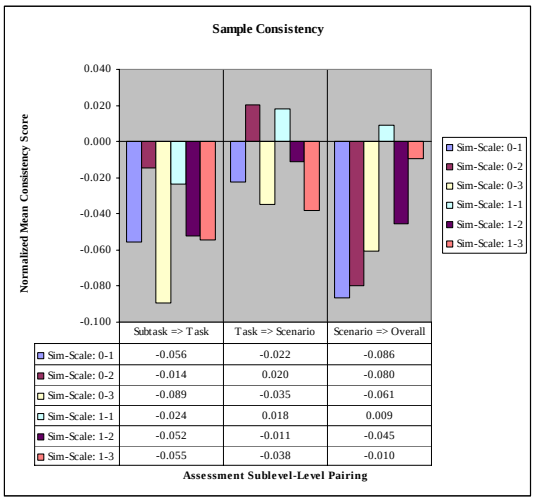
Figure 7. Subject Matter Expert Normalized Responses to Overall 1

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, $\alpha = 0.05$ and Prob>ChiSq less than 0.05 indicating the factor is statistically significant. 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 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 8 shows the Sim-Scale Groups 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 8. Intra-SME Mean Consistency Scores

Figure 9 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).

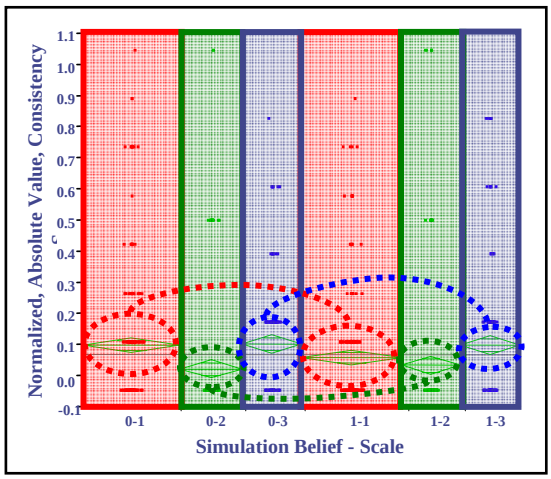


Figure 9. Intra-SME Subtask-to-Task Consistency Scores

Figure 9 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 10 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.

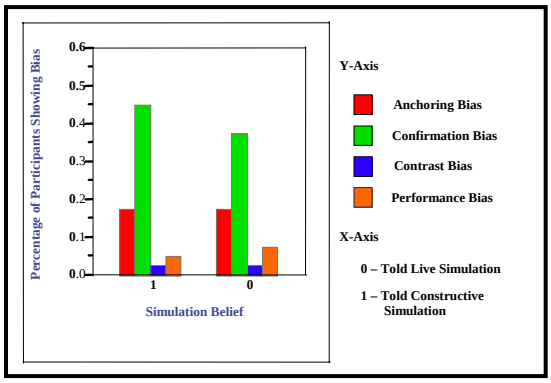


Figure 10. Study #1, Subject Matter Expert Bias for 7-Point Likert Scale

Table 2 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 2. Normalized, Mean Overall Assessment Scores - Minus Bias

ID		Number of SMEs	Mean (Normalized 0-1 Responses)			
Simulation Belief	Scale		Question Overall 1	Question Overall 2	Question Overall 3	Question Overall 4
0	1	16	0.589	0.598	0.563	0.580
0	2	21	1.000	1.000	1.000	1.000
0	3	7	0.543	0.543	0.514	0.543
1	1	16	0.777	0.768	0.696	0.714
1	2	15	0.967	1.000	0.900	0.933
1	3	10	0.700	0.700	0.660	0.660
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.[6] 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%.[7]

RECOMMENDATIONS

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

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

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

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.

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.

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.

SIGNIFICANT CONTRIBUTION

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.

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.

REFERENCES

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

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

Cohen, M. S. (1993). 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.

Department of Defense Directive (DoDD) 5000.1: Defense Acquisition System. (2001). Alexandria, VA: Department of Defense. Retrieved July 2, 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.

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

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

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

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

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

Pace, D. K., & Sheehan, J. (2002, 22-24 October). *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.

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

Statkus, M. J., Sampson, J. B., & Woods, R. J. (2003). *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.

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

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