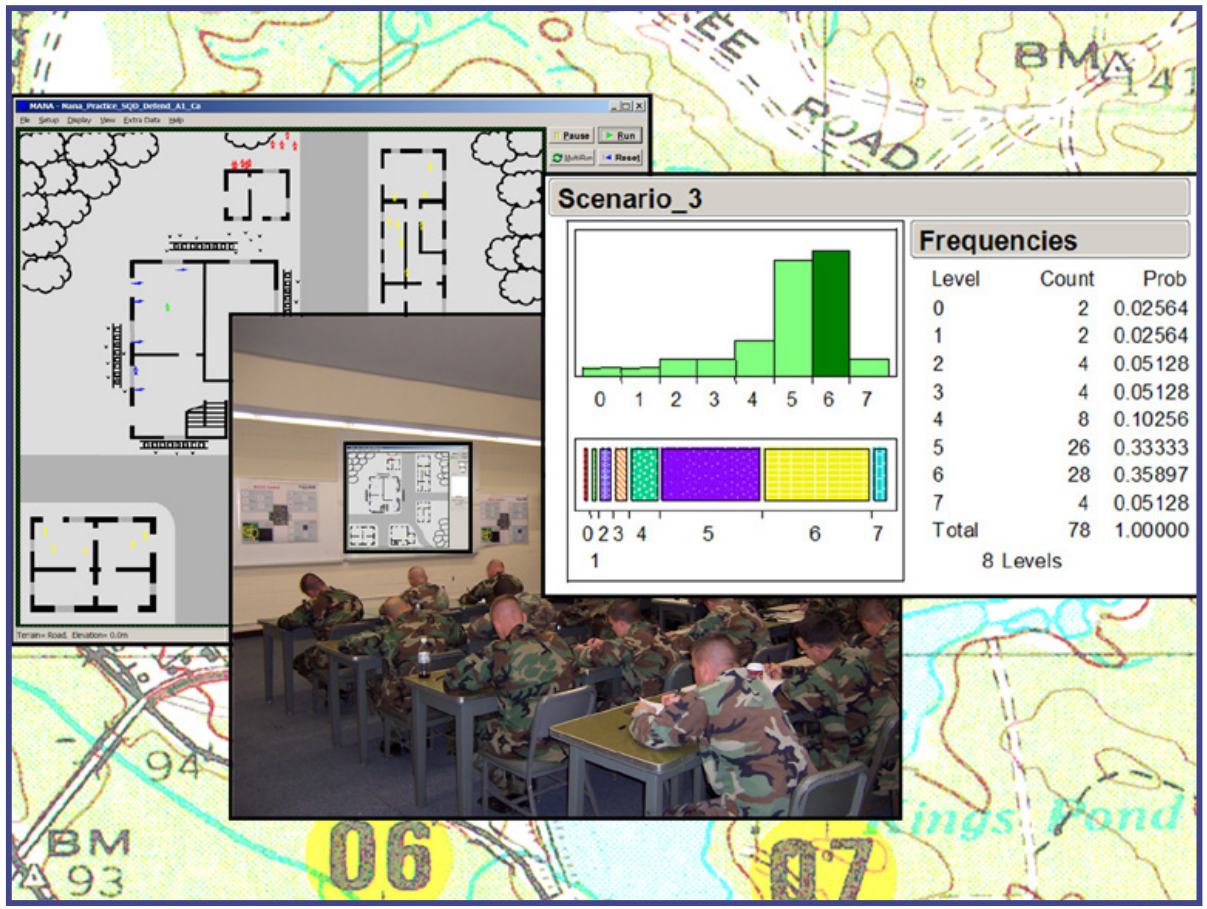
Validating Computational Human Behavior Models: Consistency and Accuracy Issues

Dissertation Defense
22 April 2004
Monterey, CA

MAJ Simon R. Goerger, USA
srgoerge@nps.navy.mil
(831) 656 - 3733

Outline

- Introduction
- Methodology
- Analysis
- Conclusions
- Contributions
- Summary



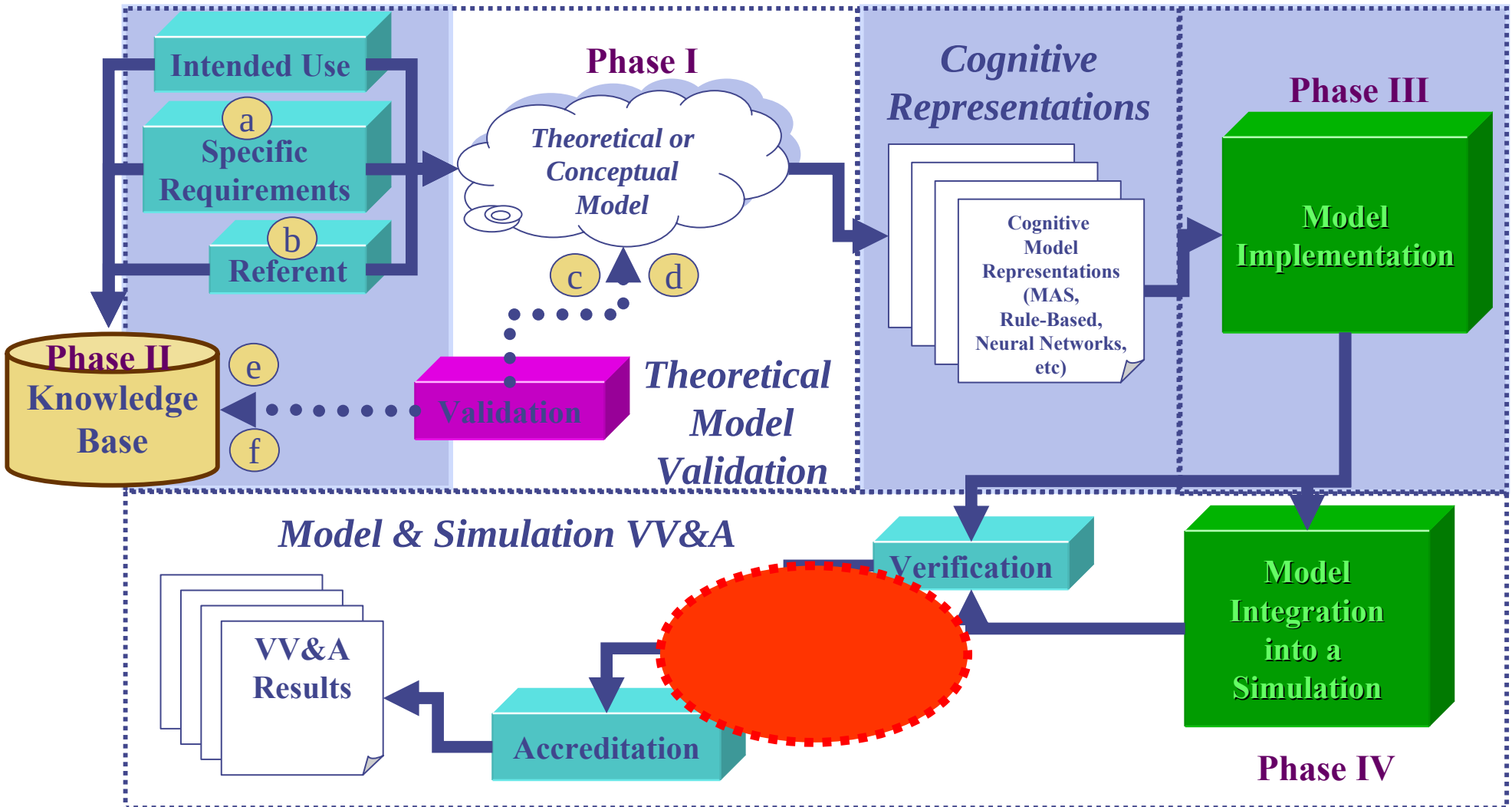
Introduction: Motivation

- The world threat is changing and we can't focus solely on Soviet Tactics
- Viable human behavioral representation (HBR) models are needed for combat simulations
  - "Garbage in equals garbage out"
- Validation is hard for physically based models
  - The "how to" is not well defined
- Validation of HBR models is even more difficult
  - The nondeterministic nature of human cognitive processes
  - Inadequate quantitative measures for validating human behavior representation models
  - Subject Matter Expert (SME) bias

Introduction: Goal

To develop and validate HBR model implementations for use in Department of Defense training and research models and simulations

Introduction: Validation

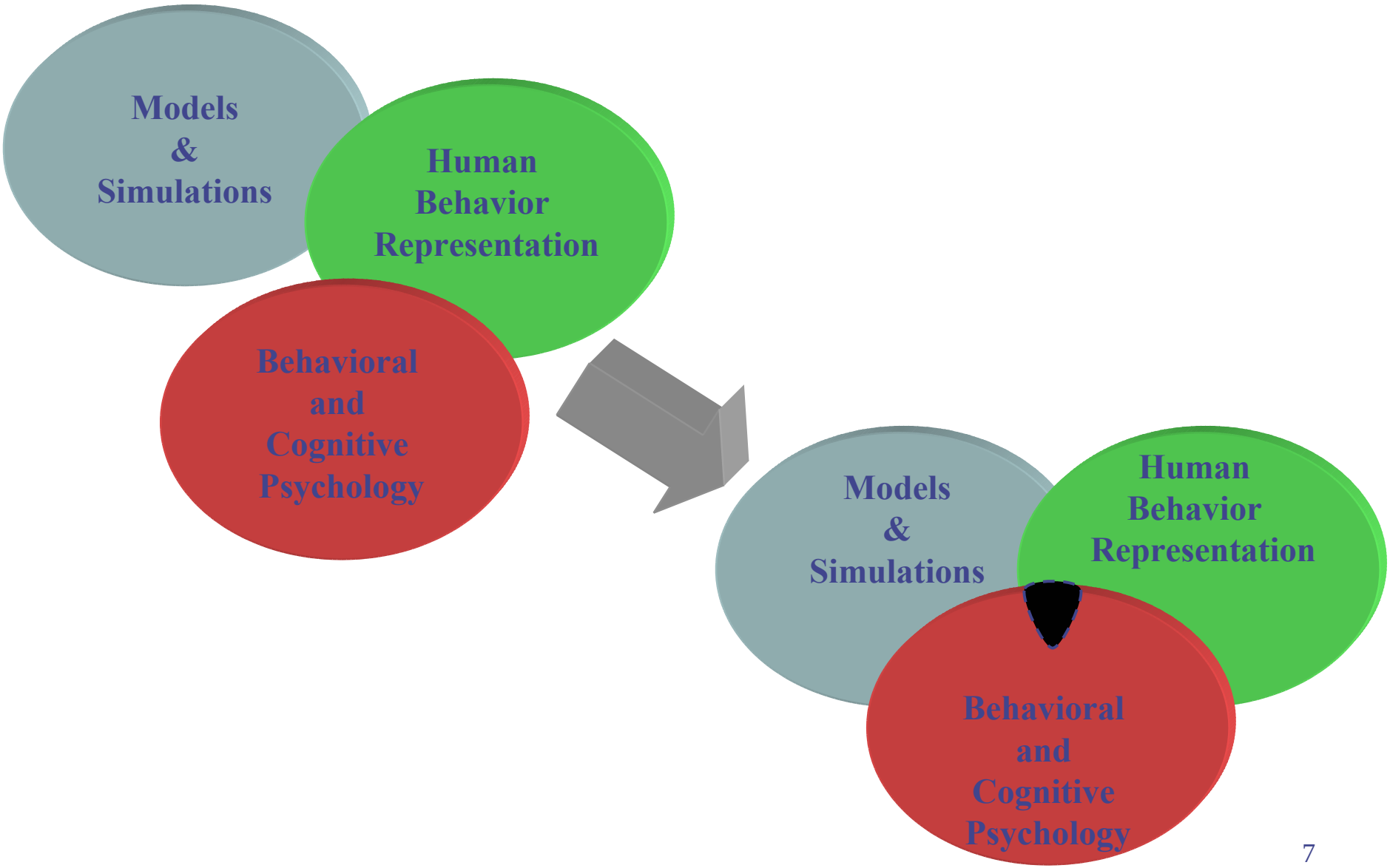


Adapted from DSMO VV&A TWG textural description of the VVA process.

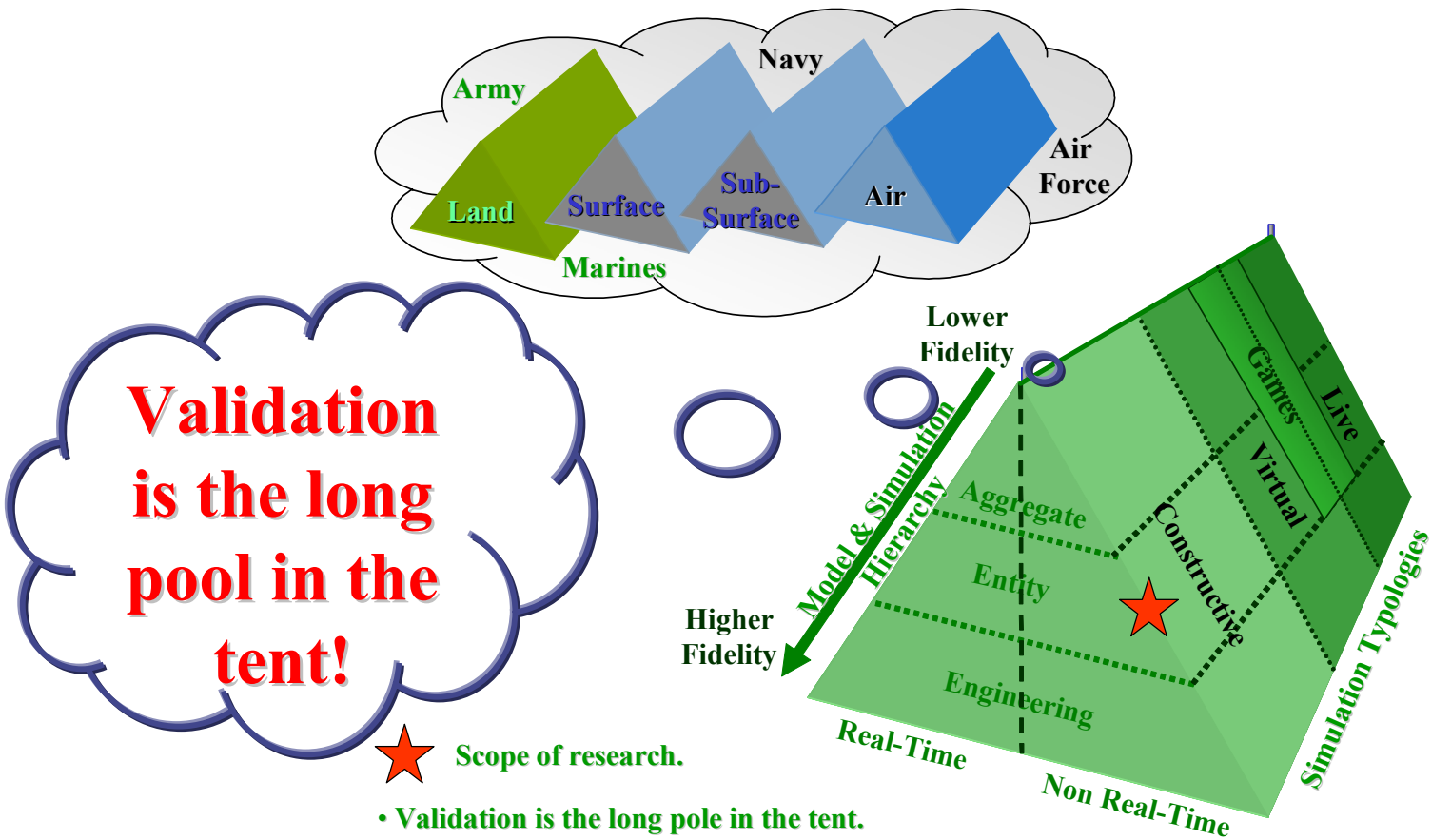
Methodology

- Scope
- Hypotheses
- Experimental Design
- Validation Plan
- Experimental Set-Up
- Experiment Phases
- Assessment Procedures
- Scenario

Methodology



Methodology: Scope



• Validation is the long pole in the tent.

Methodology: Hypotheses

- $\mathbf{H_{o'}}$: The assessment of human performance by SMEs shows different bias as compared to bias shown by SMEs assessing simulated human performance.

- $\mathbf{H_{o''}}$: SMEs demonstrate the same levels of effect on consistency and accuracy during validation of an HBR model implementation.

Methodology: Experimental Design

Participant Group	Offensive Scenario	Defensive Scenario	Offensive Scenario
1	7 Pt Simulated Performance	7 Pt Simulated Performance	7 Pt Simulated Performance
2	*7 Pt Human Performance*	*7 Pt Human Performance*	*7 Pt Human Performance*
3	Go-No Go Simulated Performance	Go-No Go Simulated Performance	Go-No Go Simulated Performance
4	*Go-No Go Human Performance*	*Go-No Go Human Performance*	*Go-No Go Human Performance*
5	*5 Pt Human Performance*	*5 Pt Human Performance*	*5 Pt Human Performance*
6	5 Pt Simulated Performance	5 Pt Simulated Performance	5 Pt Simulated Performance

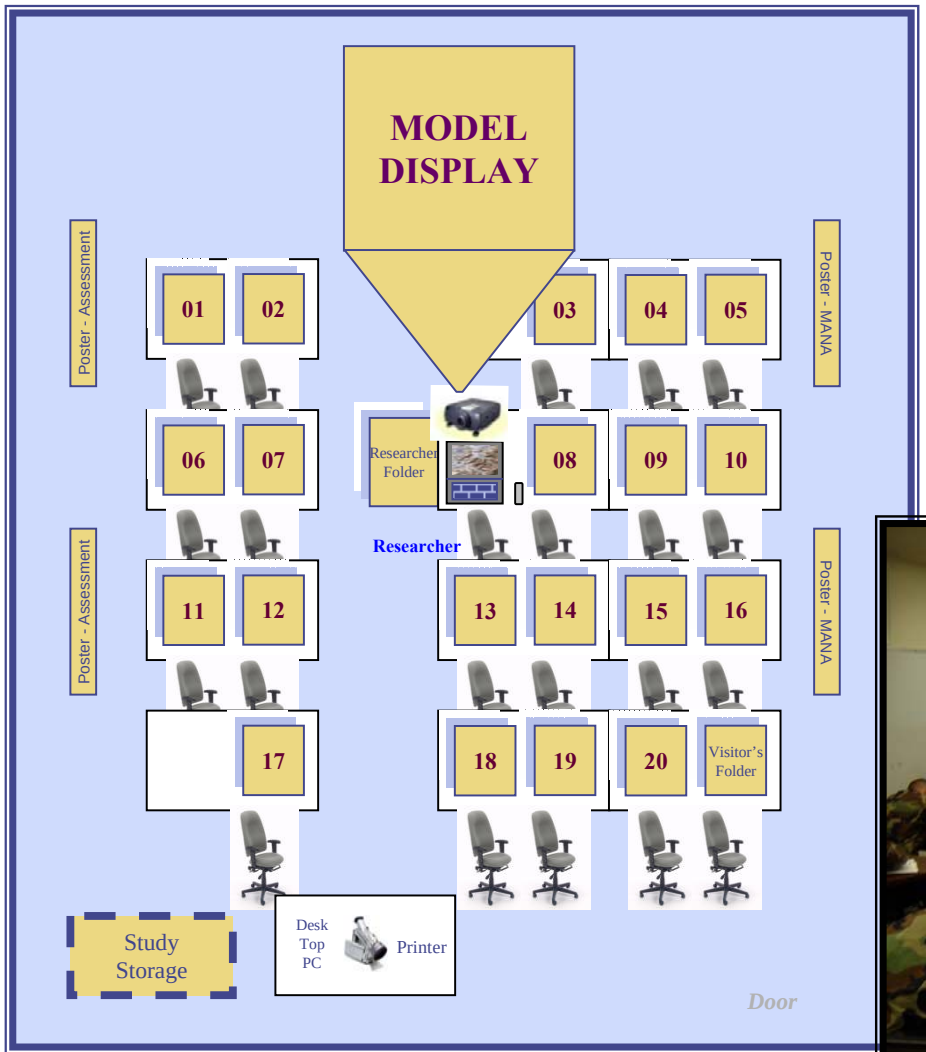
Study #1

Study #2

Methodology: Validation Plan

- Purpose:
  - To ensure the process methodology for this research is similar to the one prescribed by the DoD M&S community to validate HBR models
- Reviewed by:
  - Dr. Dale Pace (Johns Hopkins University Applied Physics Laboratory)
  - Simone Youngblood (DMSO VV&A Technical Director)
  - Scott Harmon (DMSO VV&A Technical Working Group)
  - Susan Solick (Army M&S VV&A Standards Category Coordinator (SCC))
  - Marcy Stutzman (Navy VV&A Team)

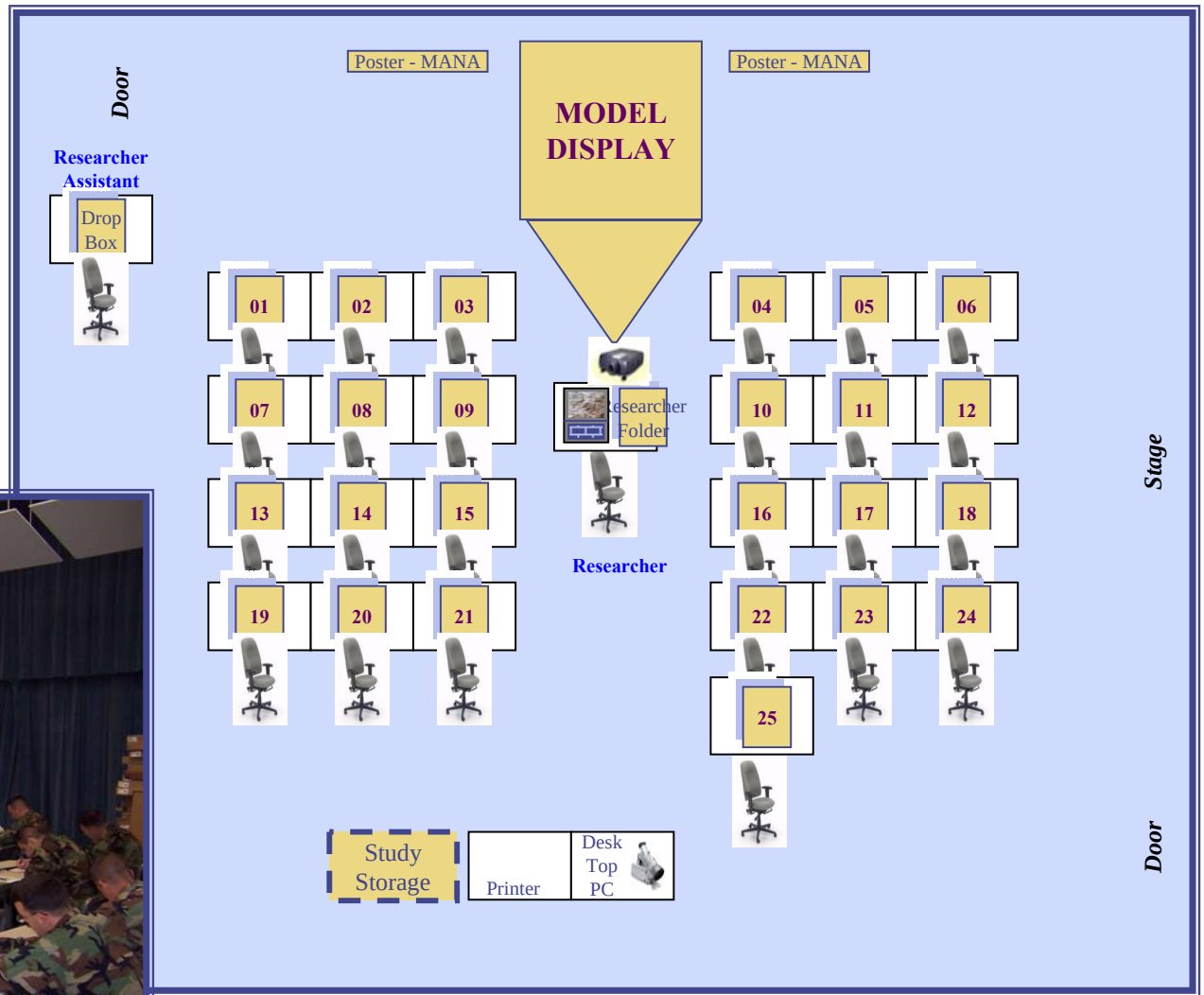
Methodology: Experimental Set-Up



- Study #1
- 80 Participants
  - Rank: 1LT/CPT
  - Branch: Infantry
  - Service: Army/Marine

Methodology: Experimental Set-Up

- Study #2
- 102 Participants
  - Rank: 1LT/CPT
  - Branch: Infantry/Other
  - Service: Army/Marine



Methodology: Experiment Phases

- *Day #1*
  - In-Processing Phase (*15 Minutes*)
  - Training Phase (*25 Minutes*)
  - Participant Practice Phase (*20 Minutes*)
- *Day #2*
  - Review Phase (*10 Minutes*)
  - Assessment Phase (*45 Minutes*)
  - Debriefing (*5 Minutes*)

Likert Seven Factor Evaluation Scale

- Low score, means low or sub par performance
- The higher the score, the better the rated performance

ASSESSMENT SCALE: Use the following scale to assess performance of the squad as it performs this task.

1 – Strongly agree the task, step, or performance measure was *improperly* performed
2 – Agree the task, step, or performance measure was *improperly* performed
3 – Not sure but tend to agree the task, step, or performance measure was *improperly* performed
4 – Undecided
5 – Not sure but tend to agree the task, step, or performance measure was *properly* performed
6 – Agree the task, step, or performance measure was *properly* performed
7 – Strongly agree the task, step, or performance measure was *properly* performed
NA - Not applicable or no means of determining

Likert Seven Factor Evaluation Scale

- Example: with not applicable (NA) and comments section

OFFENSIVE SCENARIO, TASK: Conduct Tactical Movement in a Built-up Area (Infantry Squad) (07-3-1279)

TASK(S), STEP(S) and PERFORMANCE MEASURE(S)	ASSESSMENT
2. Squad moves only after defenders' fires have been suppressed or obscured, if applicable. ***Comments***:	1-2-3-4-5-6-7 NA
3. Squad moves at night or during other periods of reduced visibility using night vision devices (NVDs). ***Comments***: *Due to slow movement to the town, forces were required to attack the town in daylight.*	1-2-3-4-5-6-7 NA
4. Squad moves using concealment of smoke provided by supporting vehicles or assets. ***Comments***: *The squad did not use smoke, but I am not sure any was available from supporting assets.*	1-2-3-4-5-6-7 NA

Subtask Summary and Task Assessment

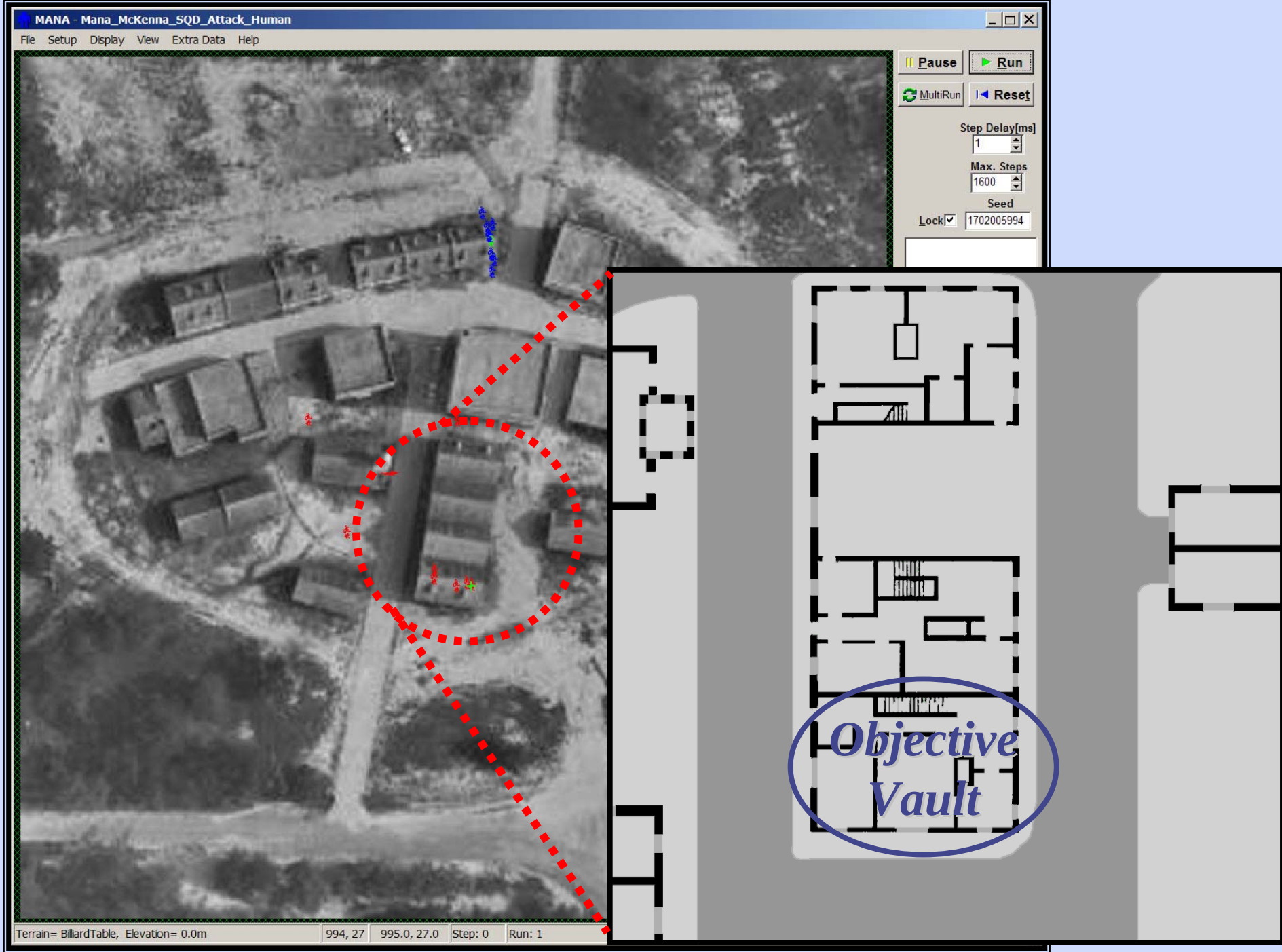
- Example

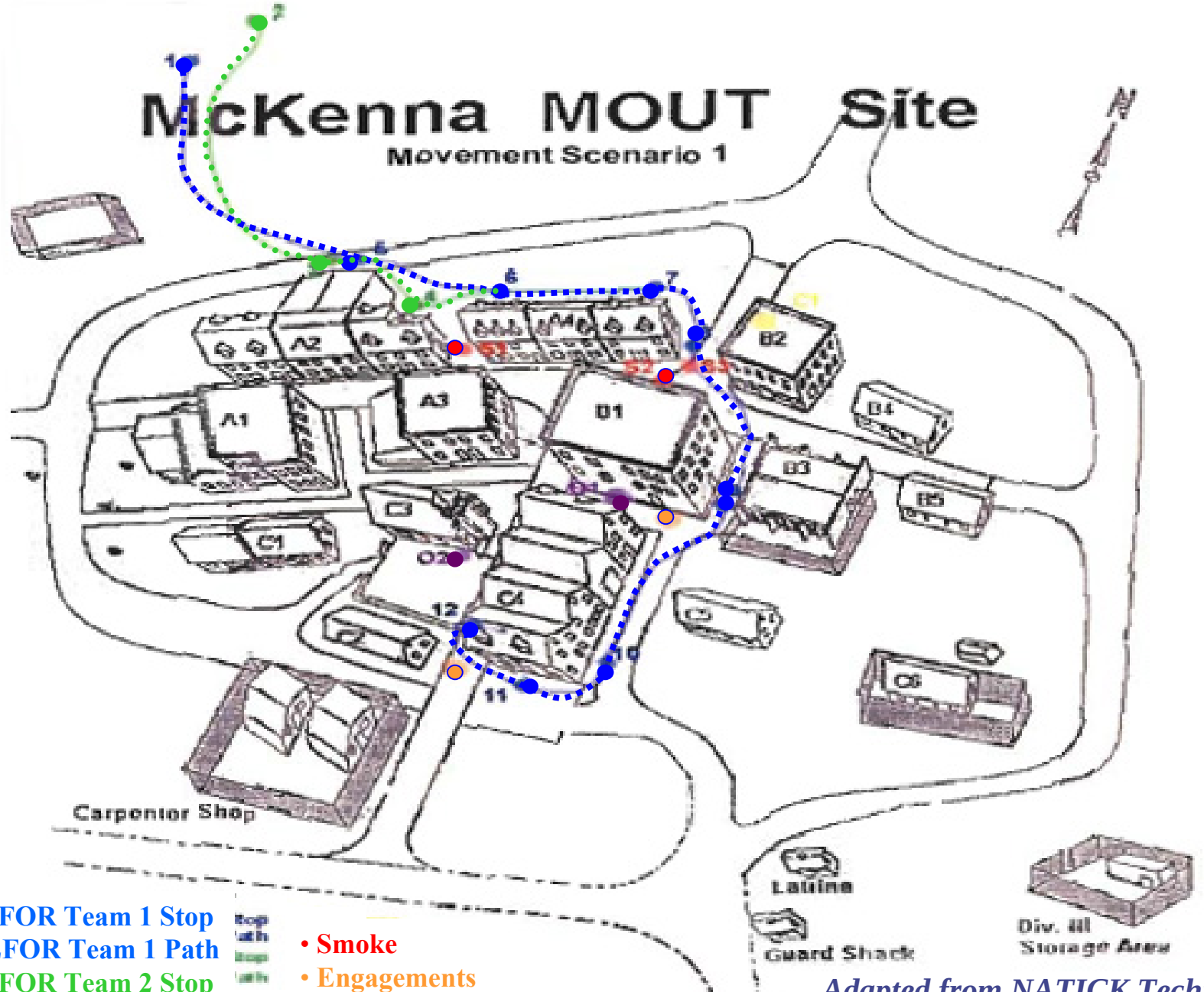
TASK PERFORMANCE SUMMARY BLOCK								
ASSESSMENT CATEGORY	1	2	3	4	5	6	7	NA
Total tasks, steps, and performance measures evaluated at this level.	2	4	4	3	4	3	1	5

TASK: Conduct Tactical Movement in a Built-up Area (Infantry Squad) (07-3-1279)	ASSESSMENT
Squad performances this task to standard. *Comments*: ***Although the squad failed to perform many of the subtasks because it never reached the OBJ, I feel it could perform the subtasks to acceptable standards.***	1-2-3-4-5-6-7 NA

Methodology: Assessment Levels

- Scenario #1 – Offensive Operation
  - Conduct Tactical Movement in a Built-up Area
  - React to Snipers
  - Conduct Tactical Movement in a Built-up Area
  - Scenario Assessment
- Scenario #2 – Defensive Operation
  - Conduct a Strongpoint Defense of a Building
  - Scenario Assessment
- Scenario #3 – Offensive Operation
  - Conduct Tactical Movement in a Built-up Area
  - React to Snipers
  - Conduct Tactical Movement in a Built-up Area
  - Scenario Assessment
- Overall Performance Assessment

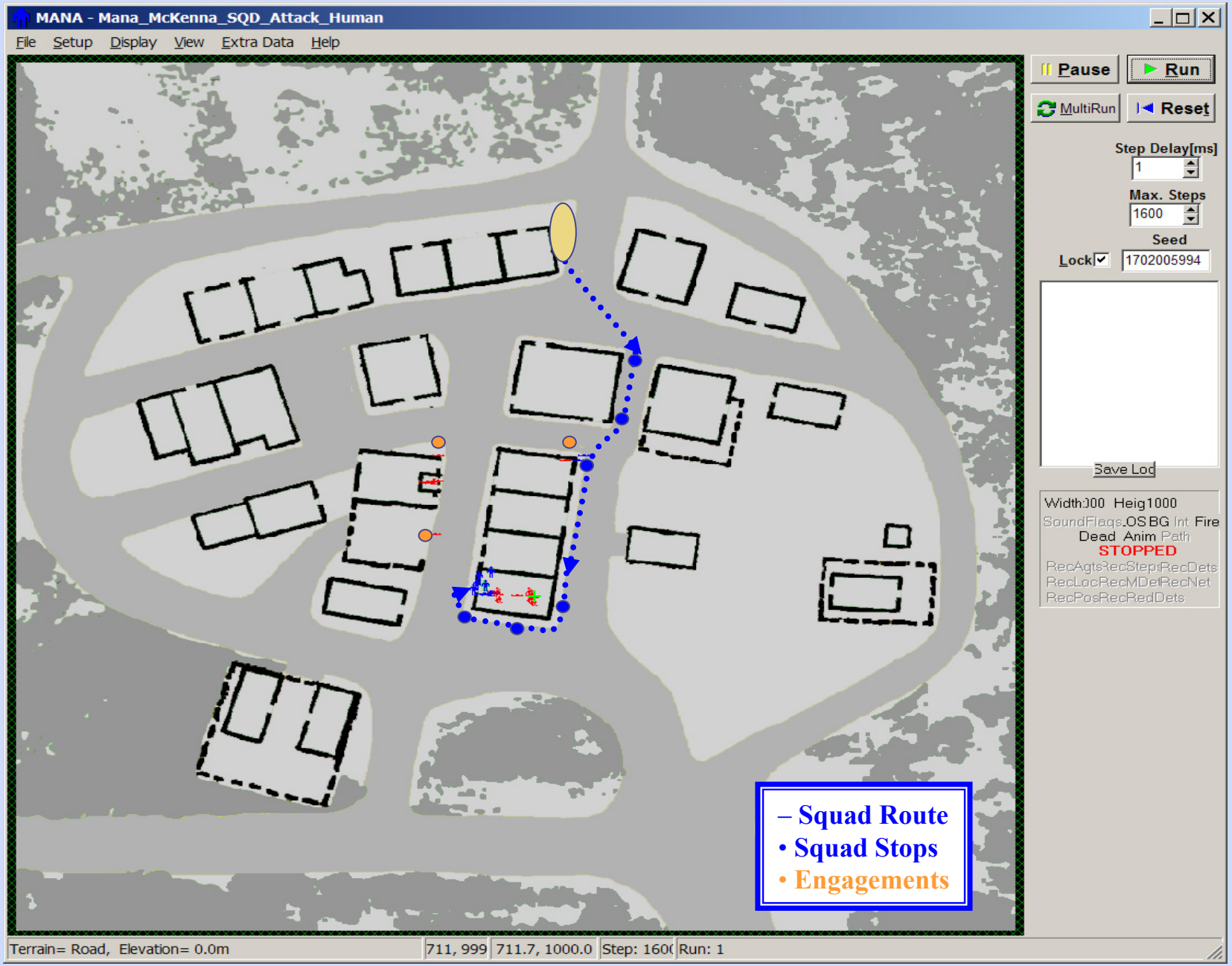

MANA - Mana_McKenna_SQD_Attack_Human
File
Setup
Display
View
Extra Data
Help
Pause
Run
MultiRun
Reset
Step Delay[ms]
1
Max. Steps
1600
Seed
Lock
1702005994
Terrain= BilliardTable, Elevation= 0.0m
994, 27
995.0, 27.0
Step: 0
Run: 1
Objective
Vault



- BLUEFOR Team 1 Stop
– BLUEFOR Team 1 Path
- BLUEFOR Team 2 Stop
– BLUEFOR Team 2 Path

- Smoke
- Engagements
- OPFOR

Adapted from NATICK Technical Report No. NATICK/TR-03/033L .


MANA - Mana_McKenna_SQD_Attack_Human
File Setup Display View Extra Data Help
Pause
Run
MultiRun
Reset
Step Delay[ms]
1
Max. Steps
1600
Seed
Lock
1702005994
Save Log
Width:000 Heig1000
SoundFlags OS BG Int Fire
Dead Anim Path
STOPPED
RecAgtsRecStepsRecDets
RecLocRecMDetRecNet
RecPosRecRedDets
– Squad Route
• Squad Stops
• Engagements
Terrain= Road, Elevation= 0.0m
711, 999
711.7, 1000.0
Step: 160
Run: 1

Analysis

- Primary Hypothesis Study #1: $H_{o'}$
- Primary Hypothesis Study #2: $H_{o''}$
- Effect of SME Bias
- Stepwise Logistical Regression
  - Scale
  - Personality
  - Bias
  - Prior Service
  - Duty Positions
  - Video Games

Analysis: Primary Hypothesis Study #1

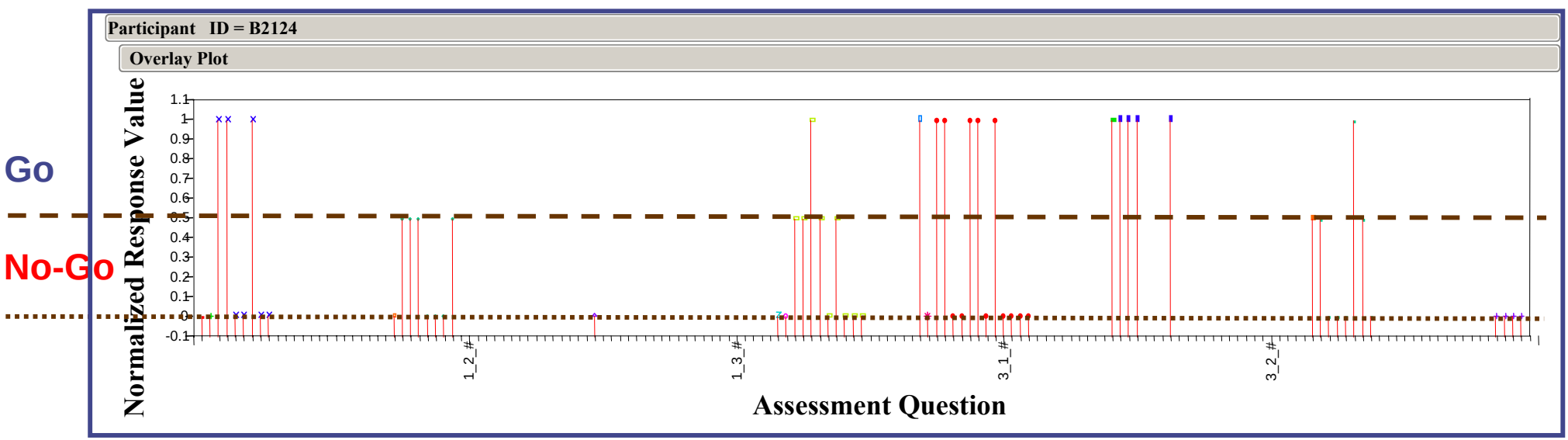
$H_{o'}$: The assessment of human performance by SMEs shows different bias as compared to bias shown by SMEs assessing simulated human performance using conventional methods as outlined in the Defense Modeling and Simulations Office (DMSO) Verification, Validation and Accreditation (VV&A) Recommended Practice Guide (RPG) for HBR.

Analysis: SME Bias

- "Systematic error introduced into sampling or testing by selecting or encouraging one outcome or answer over others" - Merriam-Webster's Collegiate Dictionary
- Types
  - Performance
  - Anchoring
  - Contrast
  - Confirmation

Analysis: SME Performance Bias

- Participant chooses not to respond to 20% or more of the assessment questions
- Pattern



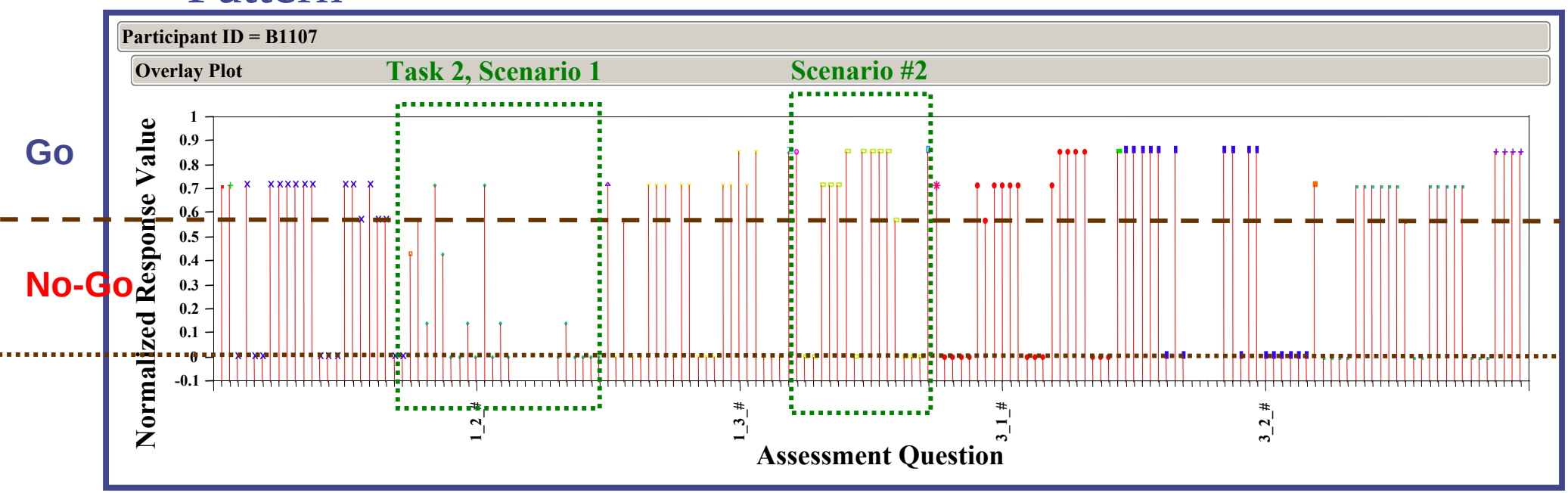
Participant B2124 – Responses

- ***23*** participants (***12.63 %***) displayed performance bias

Undecided Line
NA Response

Analysis: Anchoring Bias

- Participant assesses 90% or more of the questions after the primary task the same as they rated the initial task
- Pattern



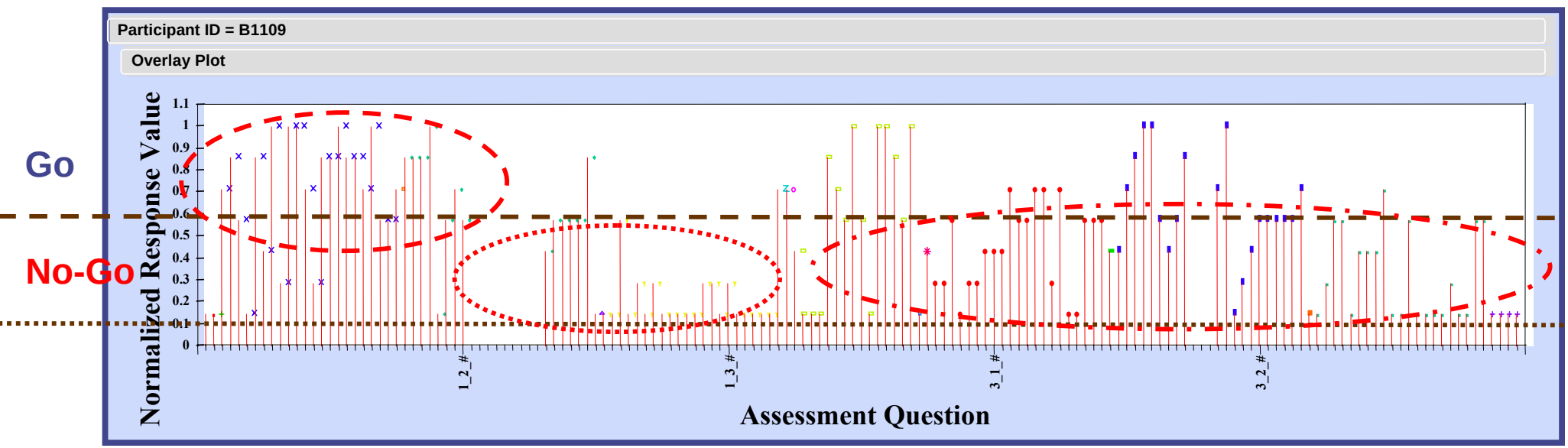
Participant B1107 – Base Responses

- ***30*** participants (***16.48 %***) displayed anchoring bias

Undecided Line
NA Response

Analysis: Contrast Bias

- Step or binary rating
  - Starts with initial hypothesis/rating and after the first sighting of contradictory information, assesses 80% or more of the remain tasks the same way with no apparent regard for contrary evidence
- Pattern



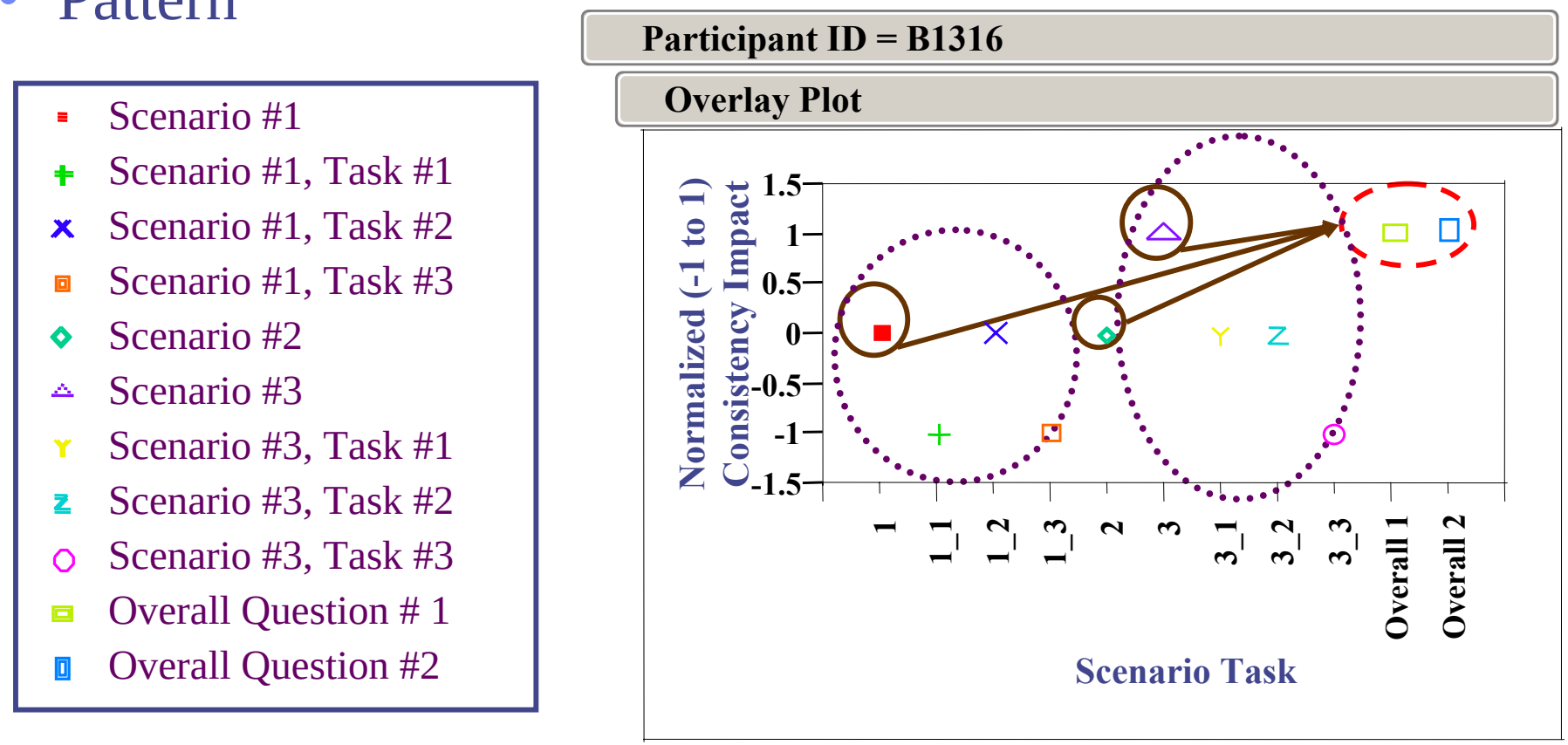
Participant B1109 – Base Responses

- *5* participants (*2.75 %*) displayed contrast bias

Undecided Line
NA Response

Analysis: Confirmation Bias

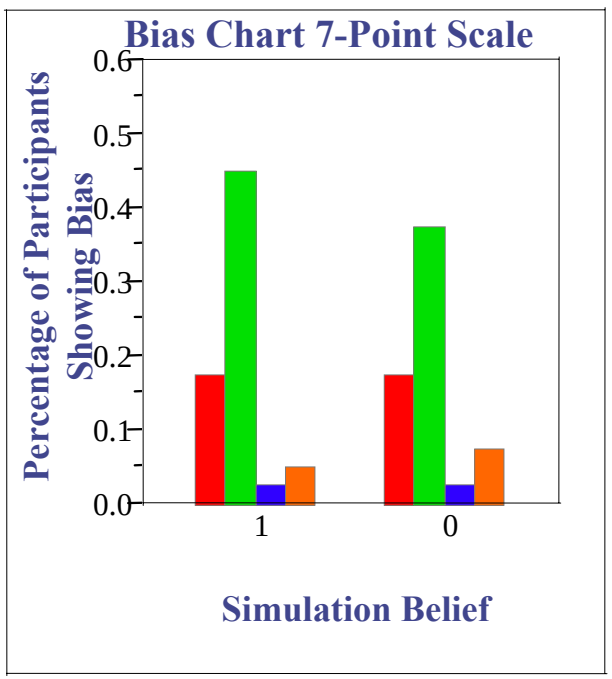
- Weights certain factors more heavily than others
- Pattern



- ***55*** participants (***30.22 %***) displayed confirmation bias

Analysis: Primary Hypothesis Study #1

- Reject the null hypothesis and conclude SMEs demonstrate the same amount of performance, anchoring, contrast, and confirmation biases during assessment of computer simulated and human behaviors.



Y-Axis

- Anchoring Bias
- Confirmation Bias
- Contrast Bias
- Performance Bias

X-Axis

- 0 – Told Live Simulation
- 1 – Told Constructive Simulation

The same amount of bias occurs regardless of simulation belief.

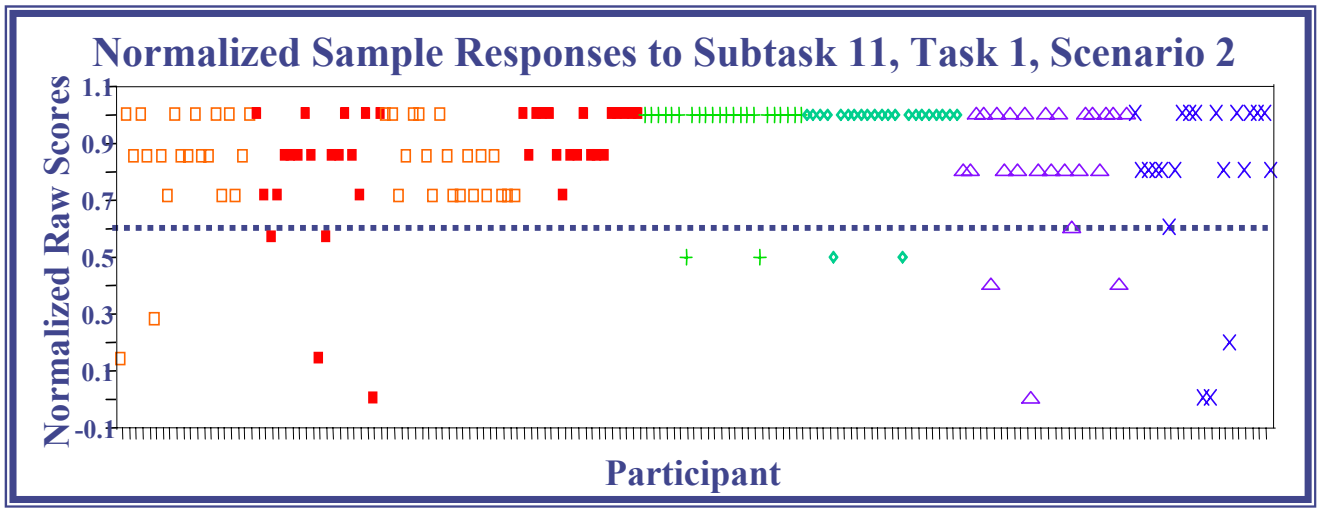
Analysis: Primary Hypothesis Study #2

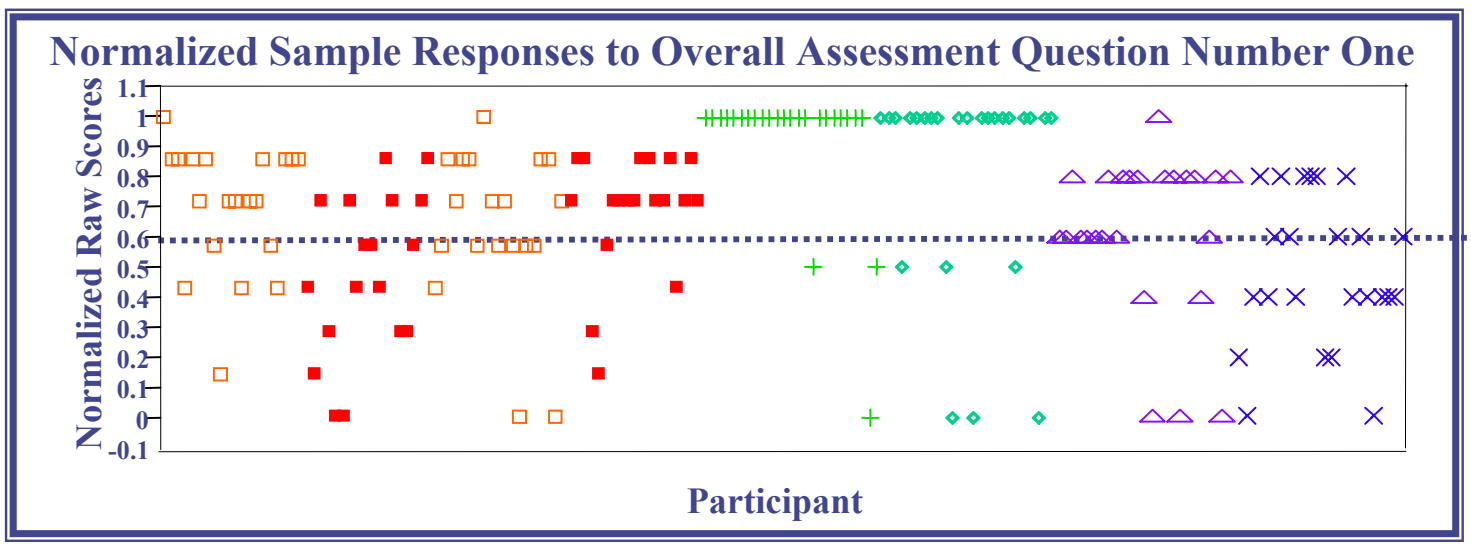
$H_{o''}$: SMEs demonstrate the same levels of effect on consistency and accuracy during validation of an HBR model implementation using a 7-Point Likert Scale as they do when using a 5-Point Likert Scale or Go/No-Go Scale.

Analysis: Inter-SME Consistency

- 67% or more of the participants agree in their assessment of the observed behavior.
- Examples

Consistent =>



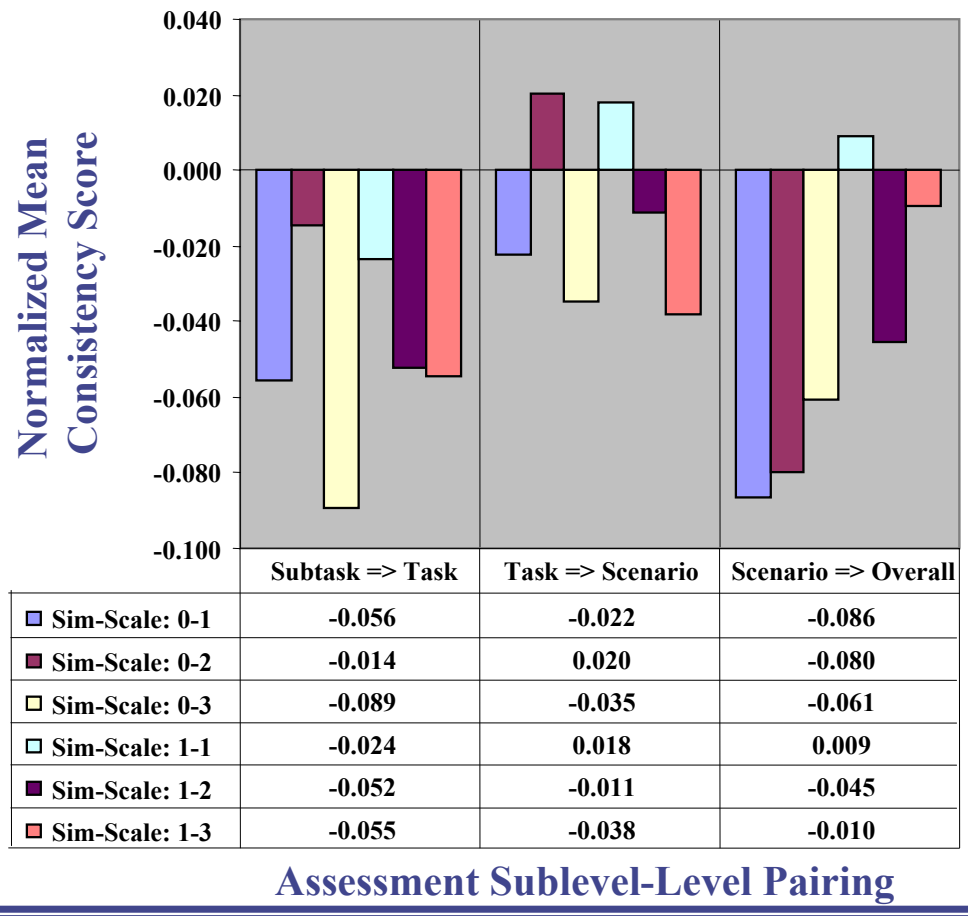


Go/No-Go Line

<= Inconsistent

Analysis: Intra-SME Consistency

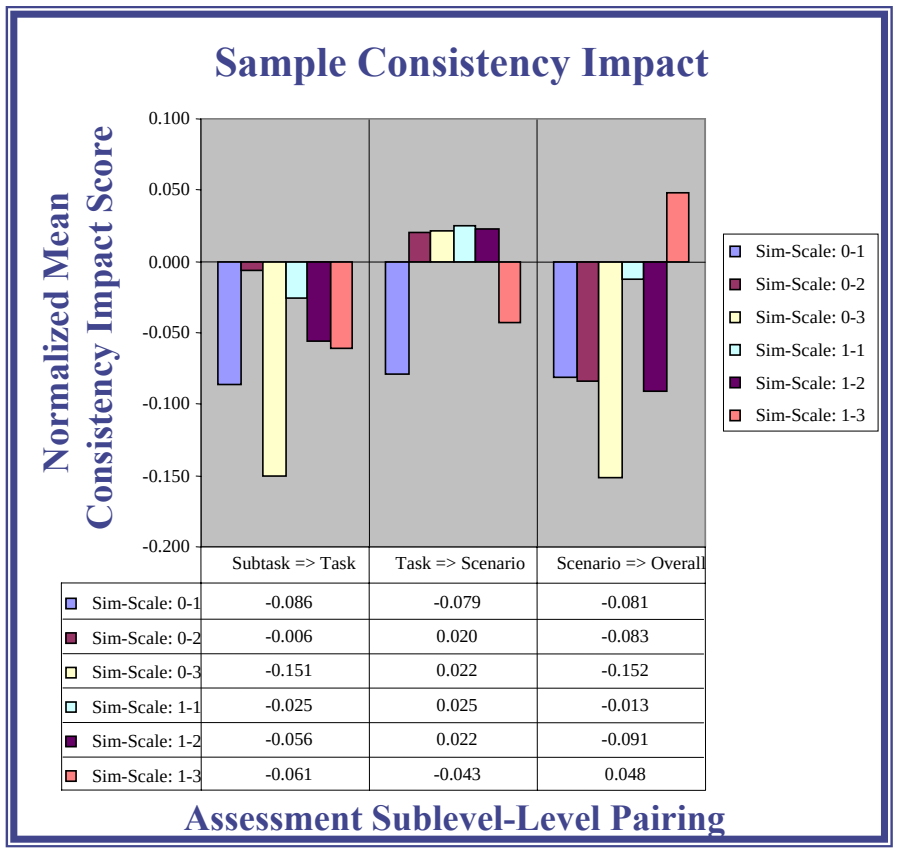
- Participants are internally consistent in their assessments when their level score is the same as the mean score of their sublevel scores.
- **Results**: 150 (82.4 %) SMEs were internally inconsistent in their assessment.



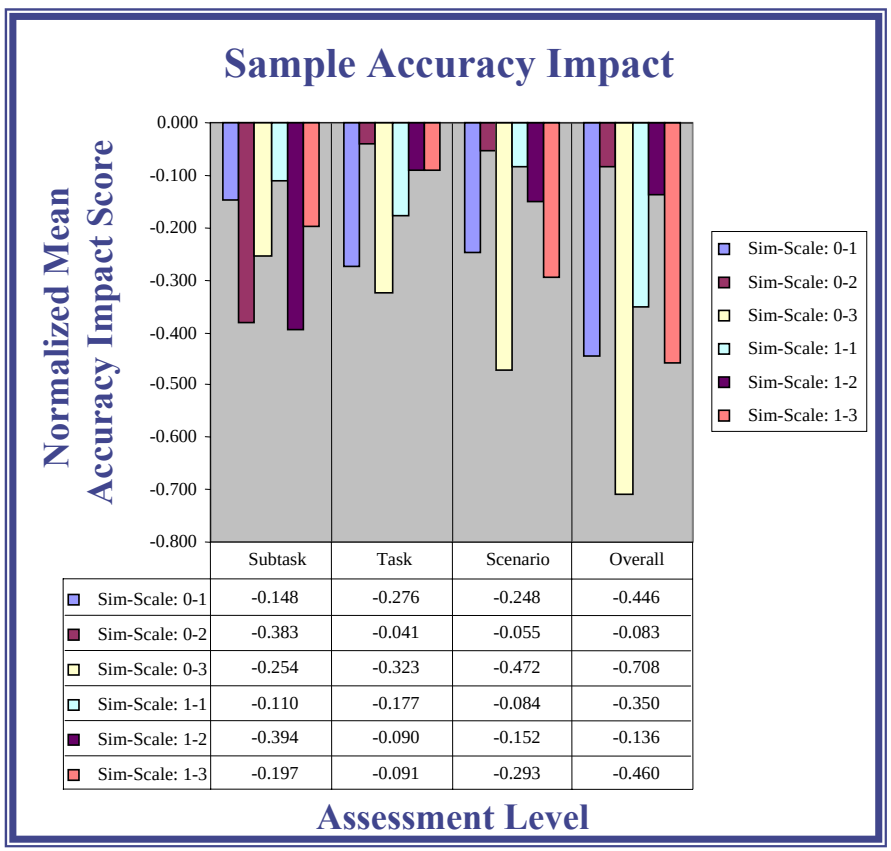
	Subtask => Task	Task => Scenario	Scenario => Overall
Sim-Scale: 0-1	-0.056	-0.022	-0.086
Sim-Scale: 0-2	-0.014	0.020	-0.080
Sim-Scale: 0-3	-0.089	-0.035	-0.061
Sim-Scale: 1-1	-0.024	0.018	0.009
Sim-Scale: 1-2	-0.052	-0.011	-0.045
Sim-Scale: 1-3	-0.055	-0.038	-0.010

Analysis: Other

- Intra-SME Consistency Impact
- Intra-SME Accuracy
- Intra-SME Accuracy Impact



	Subtask => Task	Task => Scenario	Scenario => Overall
Sim-Scale: 0-1	-0.086	-0.079	-0.081
Sim-Scale: 0-2	-0.006	0.020	-0.083
Sim-Scale: 0-3	-0.151	0.022	-0.152
Sim-Scale: 1-1	-0.025	0.025	-0.013
Sim-Scale: 1-2	-0.056	0.022	-0.091
Sim-Scale: 1-3	-0.061	-0.043	0.048

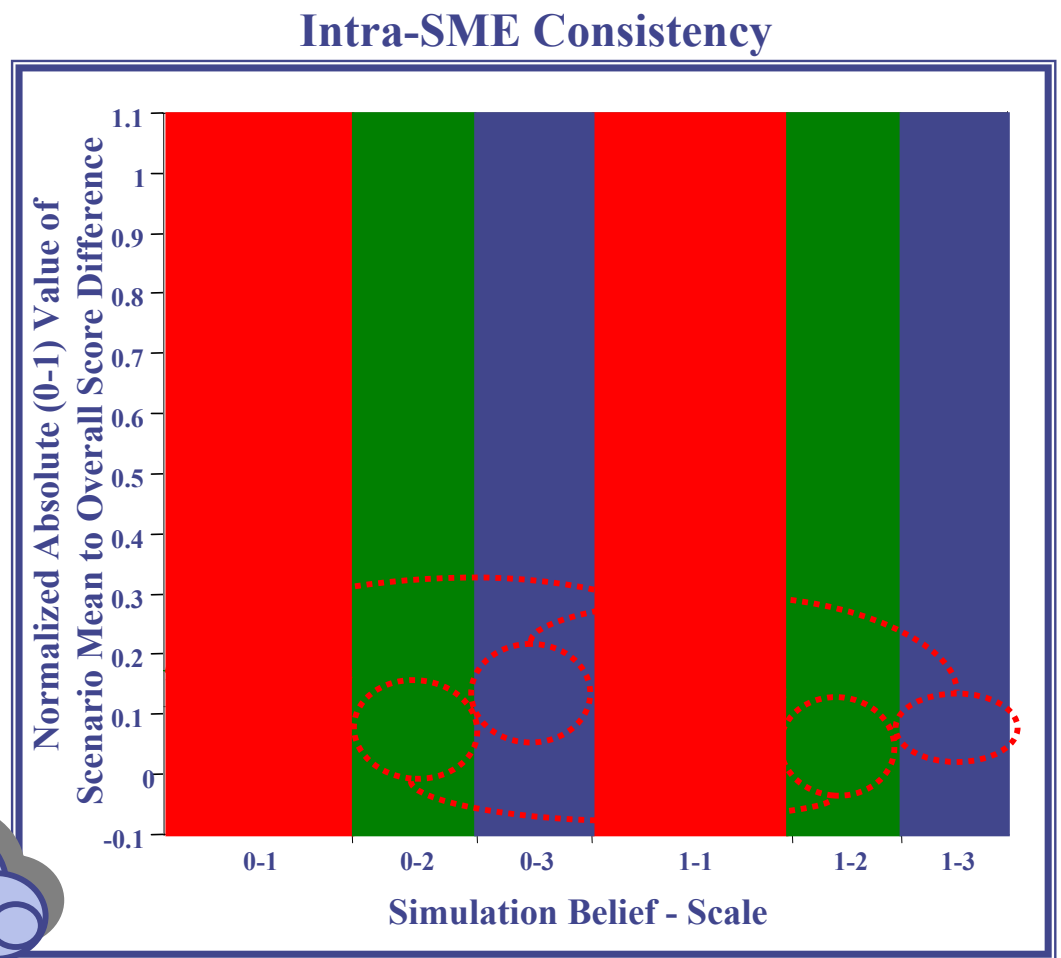


	Subtask	Task	Scenario	Overall
Sim-Scale: 0-1	-0.148	-0.276	-0.248	-0.446
Sim-Scale: 0-2	-0.383	-0.041	-0.055	-0.083
Sim-Scale: 0-3	-0.254	-0.323	-0.472	-0.708
Sim-Scale: 1-1	-0.110	-0.177	-0.084	-0.350
Sim-Scale: 1-2	-0.394	-0.090	-0.152	-0.136
Sim-Scale: 1-3	-0.197	-0.091	-0.293	-0.460

Analysis: Primary Hypothesis Study #2

- Reject the null hypothesis and conclude there is effect on intra-SME consistency due to scale.

The Go/No-Go Scale is the most consistent and the 7-Point Likert Scale is the least consistent.



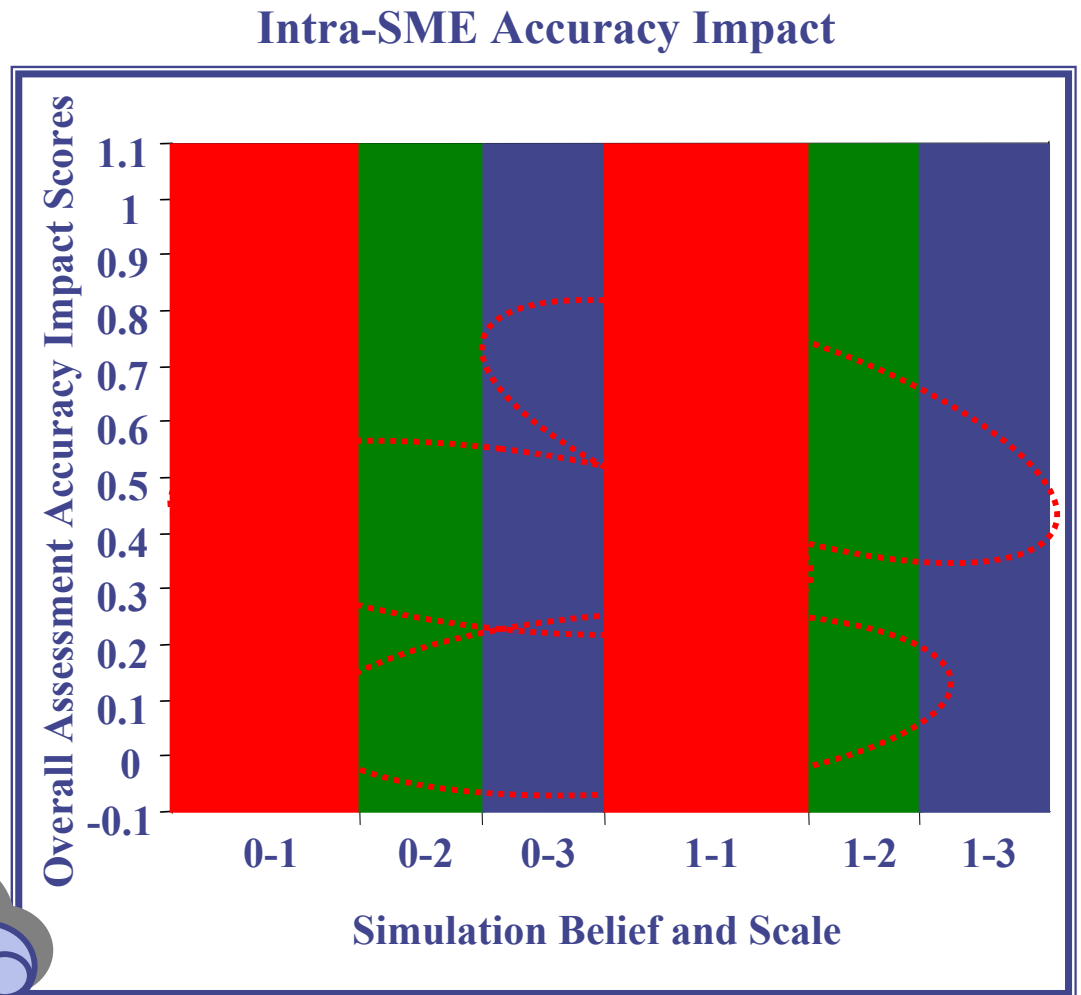
Simulation Belief: ***0 – Live; 1 - Constructive***

Scale: ***1 – 7-Pt; 2 – Go/No-Go; 3 – 5-Pt***

Analysis: Primary Hypothesis Study #2

Reject the null hypothesis and conclude there is effect on inter-SME accuracy impact due to scale.

The Go/No-Go Scale is the most accurate and the 5-Point Likert Scale is the least accurate.



Simulation Belief: ***0 – Live; 1 - Constructive***

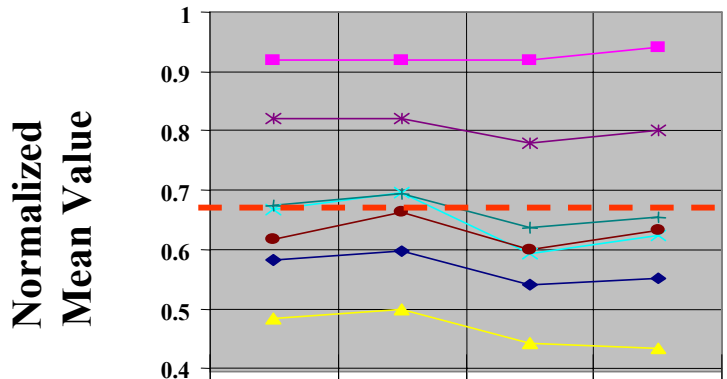
Scale: ***1 – 7-Pt; 2 – Go/No-Go; 3 – 5-Pt***

Analysis: SME Bias

Scores increase in all cases but one, when bias is removed.

- Effect

SME Normalized Responses



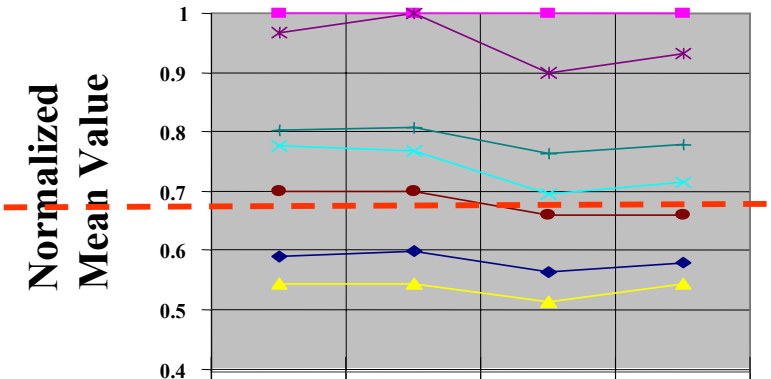
	Overall 1	Overall 2	Overall 3	Overall 4
Sim Scale 0-1	0.583	0.598	0.540	0.552
Sim Scale 0-2	0.920	0.920	0.920	0.940
Sim Scale 0-3	0.483	0.500	0.442	0.433
Sim Scale 1-1	0.667	0.696	0.593	0.623
Sim Scale 1-2	0.820	0.820	0.780	0.800
Sim Scale 1-3	0.616	0.664	0.600	0.632
All	0.675	0.694	0.636	0.654

Overall Assessment Question

Go

No-Go

SME Normalized Responses w/o Bias



	Overall 1	Overall 2	Overall 3	Overall 4
Sim Scale 0-1	0.589	0.598	0.563	0.580
Sim Scale 0-2	1.000	1.000	1.000	1.000
Sim Scale 0-3	0.543	0.543	0.514	0.543
Sim Scale 1-1	0.777	0.768	0.696	0.714
Sim Scale 1-2	0.967	1.000	0.900	0.933
Sim Scale 1-3	0.700	0.700	0.660	0.660
All	0.802	0.808	0.763	0.778

Overall Assessment Question

Go

No-Go

Analysis: Stepwise Logistical Regression

- Scale
  - 7-Pt Scale (Most Accurate)
  - *Go/No-Go Scale (Least Accurate)*
- Personality
  - *Introspectors (Less Accurate)*
  - *"Leaders" or "Competitors" (Less Accurate)*
  - *Higher Neuroticism (Less Accurate)*
- Bias
  - *Performance (Less Accurate)*
- *Infantry (Less Accurate)*
- Time Since Last with Troops
  - > 6 Months (More Accurate)
- Duty Positions
  - *Executive Officer (Less Accurate)*
  - *Squad Leader (Less Accurate)*
- Video Games
  - Experts (Most Accurate)
  - *Novice (Least Accurate)*

Ordinal Logistic Fit for Accuracy Impact - Absolute Values

Term	Prob>ChiSq
Scale {5Pt&7Pt – Go/NoGo}	<.0001
Scale {5Pt - 7Pt}	<.0001
Infantry [-1]	<.0001
Interests: Introspectors [-1]	<.0001
Bias: Performance [1-0]	<.0001
Time Since in Last Unit (>6 Months) [-1]	<.0001
First Shooter Video Experience {Expert - Average & None & Novice}	0.0001
First Shooter Video Experience {Average & None - Novice}	0.0015
Duty Position: XO [-1]	0.0037
Interactions Quads: {"Leaders" or "Competitors"} [-1]	0.0086
Duty Position: SL [-1]	0.0111
NEO-FFI: "Neuroticism" (Average, High, and Very High – Very Low and Low) [-1]	0.0207

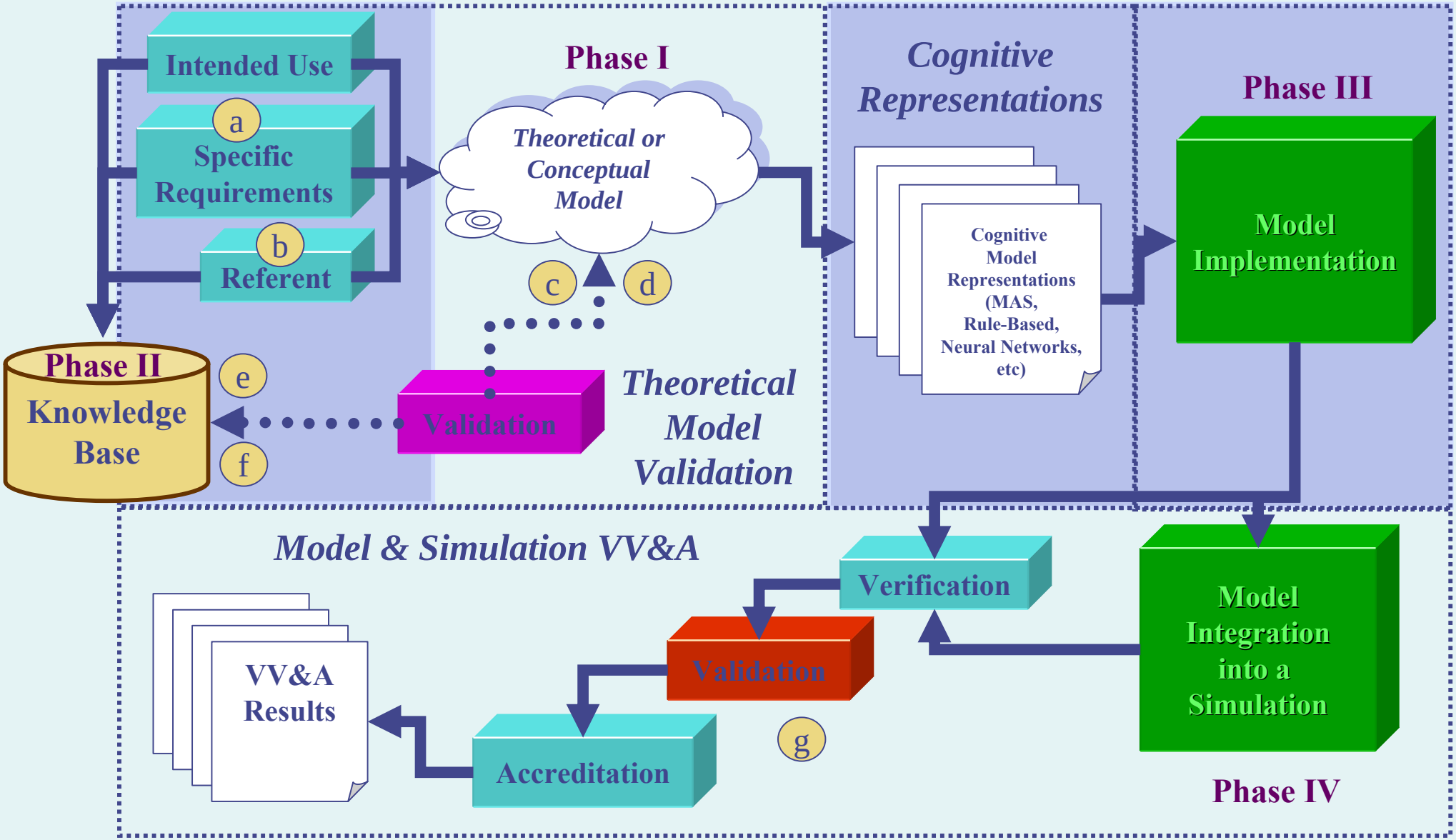
$\alpha = 0.05$

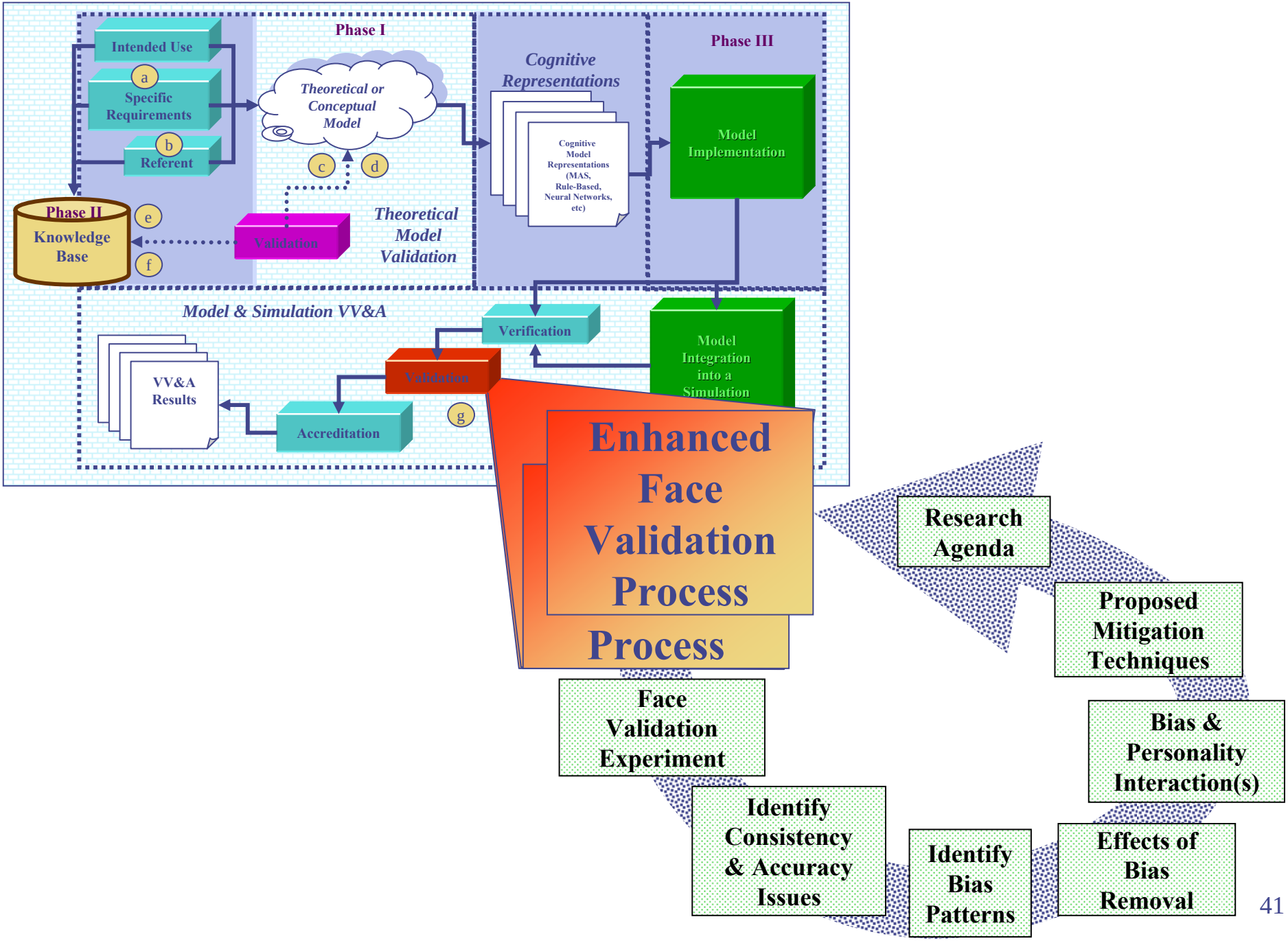
Conclusions

- Validating HBR models is difficult but must be done
- Human Performance Evaluation techniques provide statistically similar results when comparing human and simulated behavior; no difference in bias given simulation belief
- Assessment scales used shows effects on consistency and accuracy
- Bias shows effects on consistency and accuracy
  - Participants showing bias demonstrate similar personality traits
  - Research agenda outlines the need for further research in the area of face validation of HBR models

Contributions

- Lessons learned from the use of human behavior evaluation techniques, such as naturalistic decision-making, in the assessment of human behavior models
- Identified means to increase the consistency and accuracy of ‘face validation’ procedures for HBR models
- Formulation of new techniques for identifying and measuring the presence and impact of participant consistency and accuracy
- Identified quantitative patterns of bias based on SME responses to assessment questions
- Identified methods for removal of participant bias to mitigate participant inconsistencies and inaccuracies
- Establish a statistically significant relationship between bias and Neuroticism, Extraversion, and Openness Five-Factor Inventory personality styles
- Proposed a research agenda for the future enhancement of human behavior representation model validation procedures


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


Phase I
Intended Use
a
Specific Requirements
b
Referent
Theoretical or Conceptual Model
c
d
Phase II
Knowledge Base
e
f
Validation
Theoretical Model Validation
Cognitive Representations
Cognitive Model Representations (MAS, Rule-Based, Neural Networks, etc)
Phase III
Model Implementation
Model & Simulation VV&A
Verification
Validation
Model Integration into a Simulation
VV&A Results
Accreditation
g
Enhanced Face Validation Process
Process
Research Agenda
Proposed Mitigation Techniques
Bias & Personality Interaction(s)
Effects of Bias Removal
Identify Bias Patterns
Identify Consistency & Accuracy Issues
Face Validation Experiment

Summary

The DoD needs viable human behavior representation models for training and analysis. Validation of human behavior representation models is a difficult process which is not well defined. Issues such as subject matter expert bias affect consistency and accuracy of results. Assessment scales can mitigate inconsistency and inaccuracy.

Acknowledgments and Sponsors

- Navy Modeling & Simulation Management Office (NAVMSMO)
- Defense Modeling and Simulation Office (DMSO)
- Marine Corps Combat Development Command (MCCDC)
- US Army Natick Soldier Center
- Infantry Captains Carrier Course (ICCC)

Questions?

Validating Computational Human Behavior Models: Consistency and Accuracy Issues

Dissertation Defense
22 April 2004
Monterey, CA

MAJ Simon R. Goerger, USA
srgoerge@nps.navy.mil
(831) 656 - 3733