Technical Report 97-1

Training in Information Management for Army Brigade and Battalion Staff: Methods and Preliminary Findings

Jared T. Freeman, Marvin S. Cohen, Daniel Serfaty,
Bryan B. Thompson, and Terry A. Bresnick

Cognitive Technologies, Inc.
4200 Lorcom Lane
Arlington, VA 22207

October 1997

United States Army Research Institute
for the Behavioral and Social Sciences

SF298

Technical Report 97-1

Training in Information Management for Army Brigade and Battalion Staff: Methods and Preliminary Findings

Jared T. Freeman, Marvin S. Cohen, Daniel Serfaty,
Bryan B. Thompson, and Terry A. Bresnick

Cognitive Technologies, Inc.
4200 Lorcom Lane
Arlington, VA 22207

U.S. Army Research Institute for the Behavioral and Social Sciences
Armored Forces Research Unit, Ft. Knox, Kentucky 40121

October 1997

This research was performed under contract to U.S. Army Research Institute for the Behavioral and Social Sciences (ARI), Armored Forces Research Unit at Ft. Knox, KY, Contract Number DASW01-97-C-0015. The opinions expressed in this report are solely those of the authors and do not necessarily reflect the views of the Department of the Army.

FOREWORD

The introduction of advanced information technology to Army staff raises the dual prospects of increased access to vital information and information overload. One method of alleviating the problem of information overload is to train staff officers in information management skills.

This report (produced under a Phase I Small Business Innovative Research contract) describes a theoretical framework for developing training in information management, a specific training implementation, technology to support that training, and methods of measuring information management skills. In addition, it presents the results of a pilot study selected elements of the training and networked, training support technology. The experiment compares the performance of former staff officers in control and trained conditions. The findings should be considered preliminary, given the small sample of participants available for the pilot study. However the data generally supported theoretically grounded predictions of positive effects of the training on tactical decision making outcomes and processes, and on communication and coordination in a networked messaging environment.

ZITA M. SIMUTIS
Deputy Director
(Science and Technology)

EDGAR M. JOHNSON
Director

ACKNOWLEDGMENTS

We thank Dr. Bruce Sterling of the Army Research Institute, Armored Forces Research Unit at Ft. Knox, KY for his thoughtful critiques of this work and for providing us with participants and facilities for pilot testing STIM. Dr. Sterling served as a reviewer of this report, as did Dr. Carl Lickteig, whose comments we appreciate. We are grateful to other ARI staff who have also commented on and supported this work, including Dr. Barbara Black, Dr. Kathy Quinkert and Dr. Joe Psotka.

We are indebted to Sue Quensel and her associates at BDM, especially Ken Fergus, for providing us with the Defense-in-Sector (DIS) scenario, which we modified for use in our pilot test. BDM developed this DIS scenario from existing materials specifically to test battalion staff performance, under contract to the U.S. Army Force XXI Training Program and under the direction of the Army Research Institute.

Finally, thanks are due to the consultants and contractors who participated in our initial study of STIM.

TRAINING IN INFORMATION MANAGEMENT FOR ARMY BRIGADE AND BATTALION STAFF: METHODS AND PRELIMINARY FINDINGS

EXECUTIVE SUMMARY

Research Requirements:

The Task Force XXI AWE demonstrated the potential benefits and costs of digital information technology for staff at the brigade level and below (Bruce Sterling, personal communication, April, 1997; Naylor, 1997; Wilson, G., 1997). The new technology opens massive conduits for tactical data. This can be a great resource to staff, but it also increases the burden of filtering data and magnifies the challenge of fusing and interpreting it.

In the research described here[1], we conceptualized and pilot tested components of a networked training system designed to teach staff to filter large data streams, interpret data, and communicate more efficiently. The approach was intentionally generic in character: the instruction was designed to benefit virtually any staff position and the testing interface, while digital, did not resemble any specific, Force XXI technology. The effects of training on tactical decision accuracy, decision processes, and communications strategies were beneficial and large at the mean (though variance was high within the small sample of participants). Furthermore, these effects were measured using instruments that can be implemented in software, where they could drive feedback and adapt training and testing in real time. The system is called STIM, for Staff Training in Information Management.

Two related models, developed in prior research, were adapted to this project and used to focus training development. The first model describes team performance under stress as a function of environmental factors and team process variables. Teams adapt to stress by altering their strategies for decision making, team coordination, team organization (i.e., team structure), and tool selection or parameterization. From this model, we predict that training which addresses only two of these factors--coordination and decision-making strategies--should benefit overall team performance under stress in the digital environment.

[1] This research was conducted with Phase I funding under the Small Business Innovative Research (SBIR) program for U.S. Army Research Institute for the Behavioral and Social Sciences (ARI), Armored Forces Research Unit at Ft. Knox, KY, Contract Number DASW01-97-C-0015.

The second model defines decision making at the individual level as a product of the ability to 1) accurately model the tactical situation and team competencies, 2) set appropriate information goals based on inferences from these models and explicit requests, 3) filter a data stream for material that addresses these information goals, 4) test for and 5) exploit opportunities to use new information as a tool to critique the situation model (which represents situation assessments and plans), and 6) take actions (such as information gathering) that may improve situational awareness, assessments and plans.

From this theoretical base, we predict that staff decision making under high information load should improve if training helps officers to maintain clear and current situation assessments, enhances skill at critiquing situation awareness, and helps officers to test for opportunities to apply these skills. Training in these skills was developed, along with measures of the effects of that training on communications strategies, decision accuracy, and decision-making processes.

Procedure:

A selected, core set of the training concepts and measures was evaluated in a small-scale pilot test at ARI, Ft. Knox. Seven former staff officers served in the training condition; four served as controls. Participants in the training condition received scenario-based STIM training. Controls studied the same practice scenarios as the trained participants and engaged in discussion of the potential challenges of information management in Force XXI, but did not receive STIM instruction. Participants were tested on a Defense-in-Sector vignette derived from the Staff Group Trainer (previously referred to as Commander Staff Trainer) (BDM Federal, 1996), and other Army simulators. All participants played the role of a battalion operations officer (S3) and independently executed the scenario. They responded to a stream of scenario messages using a simple email application. During breaks in the message stream, they responded to a single question from the commanding officer under instructions to answer the question, defend the answer, and indicate actions they would take. Trained participants did this by formulating a recommendation and presenting its defense in the form of a node-link graph in which nodes represented supporting evidence, conflicting evidence, assumptions, other argument components, and actions. Controls responded entirely in email. Researchers reduced the responses of both groups to a common form of text phrases. A subject matter expert (SME) rated these responses blind to experimental condition.

Findings:

Our findings must be interpreted cautiously given the small scale of this study, use of a single test scenario, and the nascent state of the training material. However, the trends in these data are in line with theoretically grounded predictions, and they indicate that STIM may improve staff performance in information management. Specifically, STIM improved the accuracy of tactical decisions by 34% ($p < 0.20$). The persuasiveness of arguments offered in defense of those decisions was 93% higher among trained participants than controls ($p < 0.15$). In contrast with controls, trained participants made more use of evidence supporting their decisions, recognized and attempted to explain apparently conflicting evidence, and more often identified assumptions and gaps in their knowledge. Trained participants specified fewer actions, but their actions were "reasonable" 71% more often than those of controls. STIM also improved communications behaviors. Compared with controls, trained participants filtered out 32% more low-priority messages ($p < 0.05$), were less influenced by the rank of the message sender than, we presume, by the content of incoming messages ($p < 0.05$), were more proactive in their communications ($p < 0.10$), more often issued processed data than simply forwarded it unchanged ($p < 0.05$), and maintained a quieter network, reflecting greater net discipline ($p < 0.05$).

The magnitude of the differences between groups and the size of some of the statistical effects is impressive given that the number of participants in the experiment was very small; participants were expert in staff duties and familiar with the scenario modified for testing (a condition that might have limited the opportunities to improve performance); and training was short, lasting less than two hours.

In sum, results of this pilot study suggest that STIM training may improve staff decisions, decision-making processes, information filtering skill, and information production strategies. Measures of these skills were sensitive to the training manipulation, indicating construct validity. Concepts for automating these measures, producing feedback, and adapting the practice and testing to the individual or the team are developed and presented in the report.

Utilization of Findings:

Opportunities for future research and development include targeting future work on challenges presented by specific Force XXI technology, automating performance assessment using the

measures presented here, and using those measures to drive feedback and adapt training.

We are encouraged by the positive (if preliminary) results of this generic training and training technology that that the performance effects may be strengthened by customizing the instruction to meet specific needs of Force XXI staff and changing the interface to emulate selected Force XXI technology such as the All Source Analysis System (ASAS) Remote Workstation, Maneuver Control System (MCS), Applique, or their successors.

The measures described in this report were largely designed to be taken by a computer-based training system in real time, and to be interpreted by a performance assessment engine that we have conceptualized. One part of the engine would compute measures of communications behaviors (such as proactive information handling and information filtering skill) by applying simple formulas to data concerning the routing and prioritization of messages. A second sub-engine would score decision accuracy on responses to multiple-choice questions. A third sub-engine would evaluate decision-making processes by grading the structure and textual content of responses to test questions. The hybrid architecture of the third engine would mate statistical methods of text encoding with an inferential neural net capable of matching encoded student responses to SME-graded responses. The technology involved is not exotic. Development that integrates it into STIM has a high likelihood of success. Furthermore, successful development of the assessment engine could drive feedback in a highly automated version of STIM. The feedback engine would display performance scores on communications, decision accuracy, and decision-making skills; present model responses; and offer strategic advice. The products of Phase I provide a solid foundation for future research and development training and training systems to combat information overload.

TRAINING IN INFORMATION MANAGEMENT FOR ARMY BRIGADE AND BATTALION STAFF: METHODS AND PRELIMINARY FINDINGS

CONTENTS

Page

INTRODUCTION 1

STIM: STAFF TRAINING FOR INFORMATION OVERLOAD 2

Theoretical Foundations 2
Empirical Foundations 7
STIM Training Content 10
Measures 15

A PILOT TEST OF STIM CONCEPTS 26

Hypotheses and Research Questions 26
Experimental Design 27
Subjects 28
Materials 28
Procedure 29
Apparatus 32
Results 33
Potential Audiences for STIM Training 46
Discussion 47

CONCEPTS FOR FURTHER DEVELOPMENT OF STIM 52

Training 52
Instructional Strategy 54
System Design Concepts 63

CONCLUSIONS 65

REFERENCES 68

APPENDIX A. BRIGADE COMMANDER'S GUIDANCE A-1

B. BATTALION TASK FORCE COMMANDER'S GUIDANCE B-1

C. TASK FORCE EXECUTION MATRIX C-1

D. SYNCHRONIZATION MATRIX D-1

E. TASK FORCE DECISION SUPPORT TEMPLATE E-1

F. ANNOTATED DIS TEST SCENARIO MESSAGE STREAM F-1

G. DEBRIEFING FORM G-1

H. TLX WORKLOAD QUESTIONNAIRE H-1

I. BIOGRAPHICAL SURVEY I-1

J. TRAINING MATERIALS J-1

K. PRACTICE MATERIALS FOR CONTROLS K-1

LIST OF TABLES

Table 1. Methodology Used in Prior Studies of Training in Critical Thinking Skills. 9

2. Mean Persuasiveness of Participants' Arguments. 36

3. Mean Persuasiveness of Arguments Given Additional, Compensatory Response Time. 37

4. Accuracy of Trained Participants at Classifying the Components of Arguments. 38

5. Summary of STIM Training Effects. 48

LIST OF FIGURES

Figure 1. The adaptive team performance model. 4

2. The model of adaptive decision making. 6

3. Toulmin's representation of argument. 18

4. Arguments are represented in STIM as node-link graphs. 19

5. The STIM interface. 32

6. The accuracy of tactical conclusions. 34

7. The persuasiveness of arguments. 35

8. The variety of argument components used. 39

9. The informed rejection ratio (IF-R2). 41

10. The influence of rank on subjective ratings of message. 41

11. The number of messages generated and the compression ratio (IP-R1). 42

12. The ratio of information or status messages to action or planning messages (IP-R2). 43

13. Proactive communications (IP-R5). 43

14. The ratio of processed to forwarded data (IP-R6). ... 43

15. TLX ratings of workload. 44

16. The hybrid argument assessment engine. 57

INTRODUCTION

The digitization of the Army is promoted with the vision that soldiers in the battlefield will become messengers of opportunity, reporting quickly and precisely the important events they perceive. This information stream will make commanders and their staff knowledge rich, allowing them to achieve dominant battlefield awareness and to project force at a rapid tempo wherever and whenever it is needed (CECOM, 1997; Wilson, J., 1977; Terino, 1997).

Increasing information flow may be necessary to ensure victory in future battles, but it is not sufficient, nor is it risk-free. One analyst states the problem in this way:

> While up-to-date technical means of communication and data processing are absolutely vital to the conduct of modern war in all its forms, they will not in themselves suffice for the creation of a functioning command system, and they may, if understanding and proper usage are not achieved, constitute part of the disease they are supposed to cure. (van Creveld, 1985)

Recent interviews with Army officers illustrate the severity of the problem. In one interview, it was revealed that a general in the Desert Storm operation received over one million messages in a single 30-hour period. In another, a Marine officer described waking up from two hours of sleep to find 218 new e-mail messages in his in-box, of which four were relevant to his concerns (Gary Klein, Klein Associates, personal communication, September 25, 1997). Results of the Advanced Warfighting Experiment (AWE) this summer at the National Training Center (NTC) also indicate the emergence of information overload problems. A team sponsored by the director of Operational Test and Evaluation for the Secretary of Defense concluded that, although intelligence gathering by the experimental, digitized brigade was “excellent,” the brigade “failed to act quickly on intelligence most of the time...Information overload was real” (Wilson, 1997).

As the flow of information grows, human ability to manage it may quickly be overwhelmed, threatening accurate, timely decision making. Good software tools--such as automated filters, data fusion systems, and decision aids--can help alleviate the problem, but they are not enough, particularly given the current state of technology. It is necessary also to train staff officers to filter the data and to interpret it well.

In the research described here, we conceptualized and pilot-tested components of training and an inter-networked training

support system designed to help brigade or battalion staff filter and interpret data, that is to prevent information overload and improve tactical judgements. The combined training and software system are called STIM, for Staff Training in Information Management. The initial approach was intentionally generic in character: the instruction might apply to any staff position and the testing interface did not mimic the specific, Force XXI technology that any one staff member uses. In a pilot test, the effects of STIM on tactical decision accuracy, decision processes, information filtering, and information production were beneficial and conformed to predictions based in theory, though they must be interpreted with caution, given the small size of the sample and the formative state of the product. These effects were measured using instruments that can, with few exceptions, be implemented in software and used to drive feedback and adapt training and testing in real time in a more automated training system.

In this report, we first describe theoretical and empirical foundations of the training. Then, the content of STIM training is described and we define an array of measures of information management and decision making, developed in this research project or adapted for it. A pilot test of key STIM instruction, software, and measures is reported. We then present the remaining research products that were conceptual in nature and that were not part of the pilot test. Specifically, these are ideas for automating assessment, feedback, and adaptation of training content in an intelligent tutor based on STIM. We conclude with an overview of the design of a system that integrates these concepts[2].

STIM: STAFF TRAINING FOR INFORMATION OVERLOAD

Theoretical Foundations

Two conceptual models were used to focus the development of STIM. The first model represents the factors that influence team performance under stress, such as that imposed by information overload. The second describes crucial aspects of tactical decision making under conditions of uncertainty and time-stress. In this section, we describe these models and derive predictions

[2]The report addresses each of the four tasks initially proposed for Phase I of this SBIR project: the development of training (Task 1), the definition of measurement instruments and instructional strategy (Task 2; these products are described early in this paper and in the section concerning future directions for STIM), a pilot study of key training components and measurement instruments (Task 3) and concepts for the future design and development of STIM (Task 4).

concerning the type of training that should benefit staff in environments with heavy message loads, such as the digital Tactical Operations Center (TOC).

A Model of Adaptive Team Performance

Dynamic, data-rich work environments are changing our concept of human performance, and particularly our notions of human error. Traditional working conditions are characterized by relative stability, under which people rapidly develop the skills to execute standard operating procedures and to adapt to small or rare perturbations in the environment. Rasmussen (1990) argues, however, that complex human and man-machine systems are designed to address problems that have multiple degrees of freedom for action and many possible “right” answers. These situations require continuous problem-solving and choice among alternatives. Human errors are inevitable under these conditions, as is variance in workload attributable to external forces. “The trick in design of reliable systems" for these environments, claims Rasmussen, "is to make sure that human actors maintain sufficient flexibility to cope with system aberrations...Dynamic shifting among alternative strategies is very important."

The adaptive team performance model specifies strategies by which teams adapt to varying information loads and other stressors. Specifically, a high-performing team adapts its 1) decision making strategy, 2) coordination strategy, 3) organizational structure, and 4) selection and parameterization of tools in order to maintain team performance at acceptable levels (see Figure 1).

From this model, we predict that training teams in any of the four core skills (coordination, decision making, team restructuring, and tool modification) should improve overall team performance under variable information loads. In the Phase I effort, we focused on improving strategies for decision making, or critical thinking, and team coordination, operationalized as routine communication of important tactical information.

A Model of Adaptive Decision Making

The model of adaptive decision making represents the role of tactical knowledge embodied in an individual's situation model or mental model (Cannon-Bowers, Salas, & Converse, 1990) and the decision-making processes with which officers refine this knowledge. Consider this scenario:

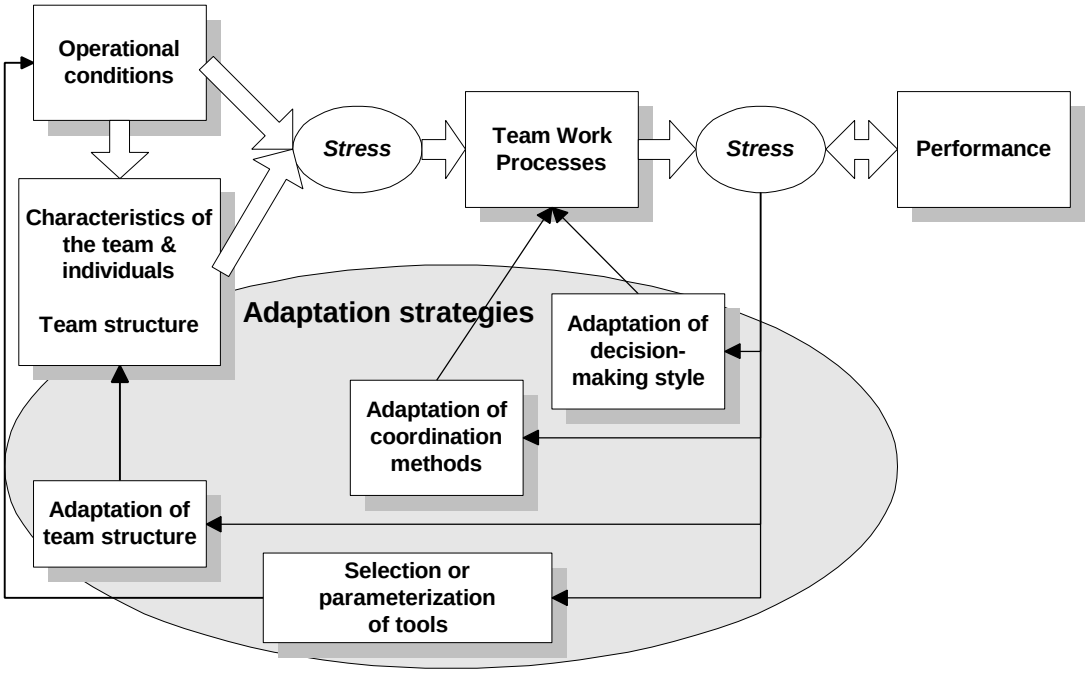

Figure 1. The adaptive team performance model.

In the heat of battle, a battalion S3 is attempting to locate a friendly recon unit at the request of an Army attack helicopter troop. The helicopters, moving against known enemy troop positions, wish to deconflict the friendly unit from enemy targets. The information streaming to the S3's workstation is voluminous and rich. The officer receives messages from personnel and systems in the field, the brigade and division above, the S2, and other members of the staff. From this mass of information the S3 must extract messages from or pertaining to the endangered friendly unit. This is a problem in information filtering, and it is mitigated largely by the clarity of the S3's information retrieval goals (or information goals): the officer knows what information is needed. As the S3 works, the officer notices messages from a marginally reliable and poorly positioned scout asserting that enemy wheeled vehicles have just entered the target zone armed with Stinger-like air defense (AD) weapons. Thus, the S3 is also engaged in opportunistic search through the data stream for surprising events, those that violate the current assessment of the situation, and the S3's predictions concerning the course of battle (see 1, below). Realizing that the potential threat posed by the AD weapons is immediate, the S3 quickly relays coordinates of the friendly unit and the enemy AD to the helicopters (in response to their request) and transmits information concerning the AD to friendly

artillery units (in anticipation of their requests for help coordinating targeting with the helicopters) (2, below). The S3 also senses that the sightings may have larger tactical implications, and that there are a few moments to investigate these. In essence, the S3 critiques the assessment of the situation and modifies it to account for the possibility that the AD unit is part of a deliberate defense of a vital enemy point asset, possibly a command, control, communications, and intelligence (C3I) center concealed near the target zone (3, below). After issuing a call to confirm the sightings, the officer queries intelligence assets and staff concerning enemy communications, radar emissions, and troop movements that might support the suspicion that a C3I center is near the area. Finally, the S3 advises the commanding officer to issue a warning order for tank units to prepare to maneuver towards possible vital enemy assets near the observed AD (4, below).

We can describe this scenario at a more abstract level by applying the model of adaptive decision making (see Figure 2) in the following manner.

1. The officer selects or filters incoming data using either bottom-up, recognition-based faculties or top-down, goal-driven selection criteria that we call information goals. The officer acquires information goals either by inferring them from the interests or responsibilities of others (represented by a mental model of the team), by thinking critically about what information is needed to improve the current assessment of the tactical situation (represented by a mental model of the situation), or by directly acquiring information goals in the form of explicit requests from others.

2. Having selected data to which to attend, the officer rapidly evaluates whether there is time and a need to reason deeply about that data. If there is not, the officer executes a well-practiced response and returns his or her attention to the data stream. If there is, the officer proceeds as follows.

3. The officer engages critical thinking skills to interpret the new data and its implications for the situation model. The officer first attempts to formulate arguments with the new data that bear on specific conclusions derived from the situation model. Then the officer critiques the arguments by ferreting out their weaknesses. Three types of weaknesses can be pursued. The first is a gap caused by failing to formulate some key argument or a lack of data on which to base an argument.

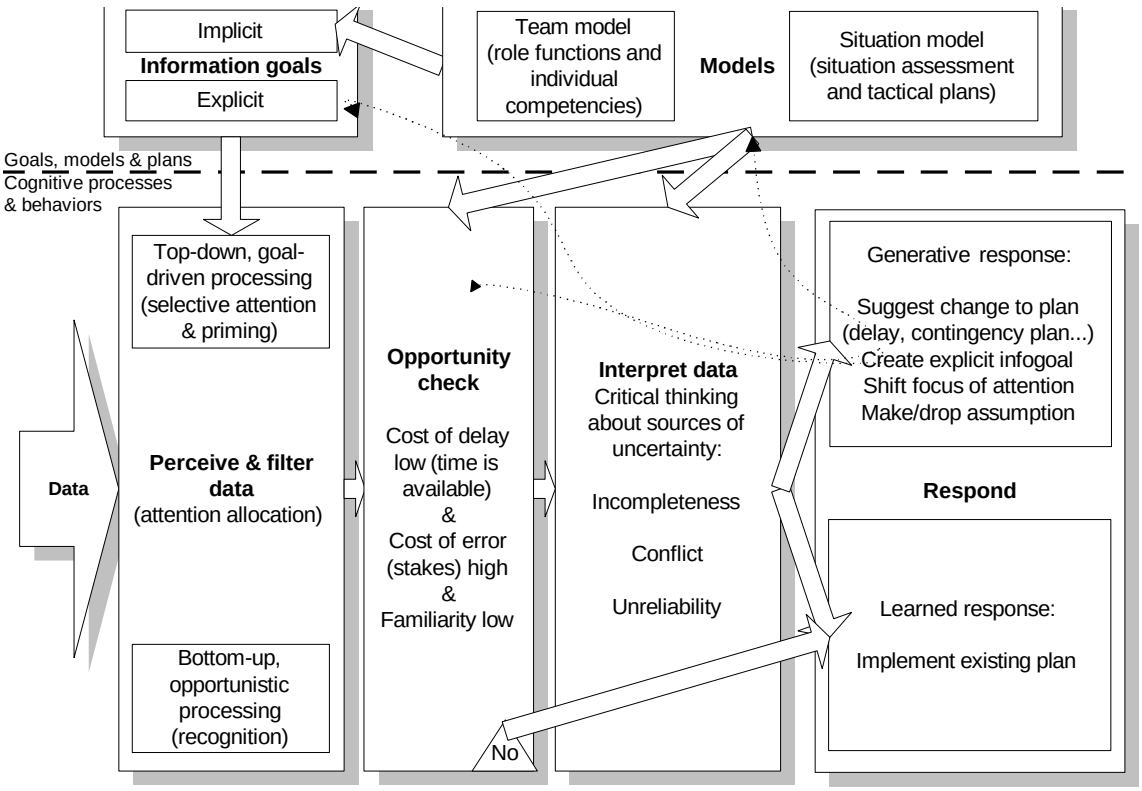

Figure 2. The model of adaptive decision making.

The second is conflict in the conclusions that can be drawn from the available evidence (e.g., several events may point to the conclusion that the enemy will attack at point A; other evidence may suggest the enemy will attack at point B). The third source of weakness is an unreliable argument, which may be based on inaccurate or unreliable data or on faulty inference. In sum, the officer uses the data as a lever to pry at weaknesses in the situation model, and uses the situation model to frame the interpretation of the data. We call this process critical thinking.

4. The officer then acts on the interpreted data by relaying requests, information, or recommendations to other officers or by setting new information goals that shape the officer's own information filtering. The better the officer's understanding of the competencies and responsibilities of team members, the better the officer can express and route information, recommendations and requests for information, and the more proactive these communications will be.

This model has several implications for training officers to manage large volumes of information in the digital environment, that is, to prevent information overload.

1. The more accurate is a staff officer's model of the situation, the more appropriate will be the officer's information goals and the better will be the officer's ability to select useful material from the data stream. If we assume that the staff's commander is most competent to form an accurate assessment, then the commander who communicates his or her assessment and revisions of it to the staff will indirectly improve the staff's filtering ability.

2. Staff who are trained in methods of detecting and handling gaps, conflict and unreliability in their situation models will detect more, or more crucial, weaknesses in the current tactical assessment and plan. They will set better information goals and select data that is more relevant to current tactical concerns.

3. Staff who are more sensitive to time constraints, the potential cost of errors and the accuracy with which they recognize a given problem are more likely to correctly decide when to implement a practiced response and when to engage in critical thinking before taking irreversible actions.

In short, staff should serve their brigade or battalion better when they are trained in several decision making and coordination skills. The Phase I research effort was focused on attaining these effects through instruction and measuring them, using techniques that can be automated in a staff training simulator.

Empirical Foundations

In previous research, members of the project team have successfully tested the effects of training officers in critical thinking and coordination skills. We review this research, below.

Effects of Assessment Updates

In field studies, Serfaty and colleagues (Serfaty, Entin, & Deckert, 1993; Serfaty, Entin, & Volpe, 1993) noted an information management strategy that boosted the performance of staff in Naval Combat Information Centers (CIC). The most effective commanding officers periodically alerted their staffs to the most pressing of their tactical concerns.

Such assessment updates may help staff in several ways. Because the updates may produce a current and common tactical

picture, staff are more likely to accurately infer what information and critiques will help the commander most. This knowledge of information goals in turn should support staff in goal-driven (top-down) filtering for useful data. More subtly, assessment updates set staff up to be surprised by (i.e., to recognize, bottom-up) events that conflict with key predictions of the current assessment.

Serfaty and colleagues (Entin, Serfaty, & Deckert, 1994) tested the effects of teaching staff about assessment updates. Four teams of five Naval CIC officers received training in making and interpreting assessment updates, as well as structured training in six information management skills: preplanning, capitalizing on idle periods, adapting the ratio of informative to administrative communications, pushing information to teammates, balancing the workload among team members, and recognizing the symptoms of information overload. Four teams received training in the six information management skills only, and four additional teams served as controls. Tests of training effects were conducted using high-fidelity, CIC simulators. Participants who received training in assessment updates and other skills performed 28% better on a composite performance index than those who received the reduced training. Staffs who received any experimental training at all performed an average of 21% better than controls, were far less sensitive to changes in workload than were controls, and performed better under high workload than controls did under low workload.

Serfaty's work established the potential value of training commanders and staff to use assessment updates to combat information overload.

Effects of Critical Thinking

Cohen, Freeman, and colleagues have examined the effects of training critical thinking skills for the U.S. Army Research Institute (Cohen, Freeman, et al., 1995) and the Navy Training Systems Division (Cohen, Freeman, & Thompson, in press; Cohen, Freeman, & Wolf, 1996; Freeman & Cohen, 1996; Cohen, Freeman, & Thompson, 1997). The most recent version of that training consists of four lessons. In the first lesson, officers study a simple procedure for building situation models (which we simply call stories, during training). The STEP (Story, Test, Evaluate and Plan) procedure consists of building a story with the evidence on hand; testing the story to identify conflicting interpretations of the evidence and resolving the conflicts, if possible; evaluating the assumptions on which the story is based; and formulating contingency plans to protect against assumptions that cannot be tested. In the second lesson, officers study the issues common to tactical stories, such as enemy goals, opportunities for attack, enemy capabilities, enemy intent,

actions, and outcomes. They practice generating assessments using stories. The third lesson presents a variant of the devil's advocate technique that is particularly useful for reinterpreting apparently conflicting evidence within a story, identifying assumptions, and generating alternative assessments. The fourth and final unit describes how experienced officers apply criteria concerning time, stakes, and familiarity to shift between tasks and between critical thinking and rapid, recognitional implementation of plans and procedures.

Cohen, Freeman, et al. (1995) tested this training in the three experimental studies (see Table 1). The participants were active-duty Army or Navy staff officers with an average of ten years military experience. These officers executed a pretest and posttest, each of which were complex, dynamic scenarios. (In the two Navy studies, high-fidelity computer simulators were used for testing.) Students were asked to monitor the scenarios, formulate assessments and plans, and make arguments in defense of their assessments. Expert judges scored responses for quality.

Table 1. Methodology Used in Prior Studies of Training in Critical Thinking Skills.

Feature	Study 1	Study 2	Study 3
Location	Army Research Institute, Ft. Lewis, WA and Ft. Carson, CO	Surface Warfare Officers School, Newport, RI	Naval Postgraduate School, Monterey, CA
Participants	37 officers ranking from 1st Lt. to Lt. Col.	60 officers, many with CIC experience	35 officers with highly varied expertise
Design	Training (29) vs. control (8), pretest vs. posttest	Training (40) vs. control (20) x pretest vs. posttest	Pretest vs. posttest
Duration of experimental session	One-half day	One day	Five days
Duration of training	90 minutes	90 minutes	4 hours over two days
Training tools for executing practice scenarios	Pencil and paper	Pencil and paper	Computer: DEFTT high fidelity CIC simulator
Test tools	Pencil and paper	Computer: DEFTT high fidelity CIC simulator	Computer: DEFTT high fidelity CIC simulator

The researchers evaluated indices of critical thinking skill, such as the number of arguments made, the number of pieces of evidence cited, and the number of assessments generated. The training improved staff performance. Trained officers made better assessments. The assessments of trained officers conformed more closely to assessments of senior military officers than did those of untrained officers. Furthermore, the plans that trained officers made were congruent with their assessments. The training reliably boosted indices of critical thinking processes by 20% to 60%. These indices concerned the accuracy of assessments, the use of supporting arguments, the identification and handling of conflicting evidence, and the identification of alternative assessments. Interestingly, even though training enhanced the ability of officers to find flaws in their own assessments, trained officers were at least as confident in their assessments as untrained officers and were more decisive in their actions. Finally, officers rated the training positively, and were more likely to do so the greater their tactical experience. These results demonstrated that training in critical thinking skills can improve officers' decisions and actions.

STIM Training Content

In the Phase I research effort, we adapted the training described above. Weak aspects of the previous training were pruned away, presentation concepts were developed with an eye towards implementing them in a multi-media system, and Army scenarios were developed for demonstration, practice and testing. The training concepts were not highly customized to specific battalion staff positions (such as the S2, S3, or Battle Captain). Rather, the instruction was somewhat generic, in that it could benefit staff in most positions. Prior research indicated that such training can have very large effects by helping staff to leverage their own domain-specific knowledge, and thus help them manage information better under stress (Cohen, Freeman, & Thompson, in press; Cohen, Freeman, & Thompson, 1997).

In this section, we describe the training (reproduced in Appendix J) and provide an example of its application to problems in a brief tactical scenario.

STIM training addresses three topics:

- Making and interpreting assessment updates (brief alerts concerning tactical priorities);
- Applying critical thinking skills; and
- Discerning when to exercise critical thinking skills, and when to apply rapid recognitional responses.

The training begins with a brief, motivational unit describing the problem of information overload. It then introduces the notion of using assessment updates to maintain a current and accurate situation model. The utility of assessment updates is established with references to field studies and training experiments in which updates have proved beneficial. Assessment updates are then defined as periodic statements concerning immediate and potential threats. These updates can be made by the senior decision-maker to subordinates, staff to line officers, or, potentially, by subordinates to superiors. Updates concerning immediate threats may have a familiar format: the threat is identified ("enemy APCs at coordinates NK2018") and an action is stated ("move recon unit Charlie to that area"). Updates concerning potential threats are less familiar in form to staff. These updates concern events that do not readily fit a known pattern, such as wheeled vehicles whose origin and intent are difficult to discern. These threats are addressed with a brief story that may account for the observations to date; predict future events; and highlight gaps, conflicts, and weak assumptions underlying the assessment. The STIM training in assessment updates continues with examples of assessment updates in a vignette concerning the actions and intent of two enemy forces poised to attack an American contingency force. (This vignette, presented in installments throughout the training, is called the Frankfurt scenario, and is used in an illustration of critical thinking, below.) The unit concludes with structured practice in interpreting and generating assessment updates, using the Frankfurt scenario.

The next unit opens by motivating the use of critical thinking skills for tactical decision making. It validates the training with a reference to field studies and training experiments. We then describe the goals of critical thinking as finding and handling weaknesses in assessments and plans. Weaknesses are of three types. Gaps are issues that are not addressed in prior planning or current data. Conflict denotes events whose most obvious interpretations appear to discredit the current assessment or plan. Unreliable assumptions are those that have not been carefully examined.

We then present instruction in detecting and handling these sources of uncertainty. The method, called IDEA, consists of five processes. These are not rigidly ordered steps for decision making, but simply tactics for better decision making under stress.

- I = Identify gaps in your knowledge;
- D = Deconflict your understanding of the situation by tentatively explaining the conflict. Look for exception conditions or make assumptions that nullify the conflict;

- E = Evaluate assumptions. Assess the plausibility of assumptions and hunt for other, still-hidden assumptions;
- A = Act on the ideas generated with IDEA. For example, request information, search online for data, recommend new contingency plans, or suggest improvements to the commander's assessment of the situation.

We then introduce a tool with which to identify gaps, deconflict understanding, and ferret out assumptions. The "Crystal Ball" is a variant of the devil's advocate that consists of a few, simple questions.

To help identify gaps, the crystal ball poses this challenge: "Your understanding of the situation hinges on an issue that is not addressed in any message or estimate so far. What is it?" Responses to this question point to gaps in an officer's knowledge and understanding. We find that lists of issues often help officers in their search. For example, officers may use a list of METT-T issues (Mission, Enemy, Terrain, Troops, and Time available) to help them find gaps regarding mission goals, enemy intent, terrain, weather, etc. They may also benefit from considering a list of story elements. A story concerning an attack, for example, might describe why an enemy would attack with the specific assets in question (given the other assets available to it) at a specific location (given other potential targets) and how it would execute the attack, that is how it would localize the target, approach it, strike and hold ground, or escape. Attempting to flesh out a story often makes gaps in knowledge obvious.

To help deconflict situation understanding, the crystal ball says: "You may think this information conflicts with your assessment (or plan) but it does not. Why not?" Answers to this question are exception conditions under which seemingly conflicting evidence can be interpreted as a natural outgrowth of a very specific causal process.

To help officers uncover hidden assumptions, the crystal ball says: "There is another way to interpret this data overall. What is it?" For example, the crystal ball insists that there is a way to interpret the enemy's radio silence other than as a sign of impending attack. The silence could be a result of systems failure, a product of a successful interdiction by friendly forces, or a feint. Responses to this question can be viewed as alternative assessments. A more interesting interpretation is that the responses, once negated, are assumptions of the current assessment. For example, the enemy must not have lost radio communications as a result of friendly interdiction if the radio silence is in fact a sign of impending attack. Assumptions such

as this can be tested, for example, by seeking out battle damage assessments.

It may seem that the crystal ball is poorly named because it asks questions, rather than answers them. What we mean to connote with the name, however, are the omniscience, indefatigability, and simplicity of the crystal ball. The crystal ball always claims to have a better answer than the user. It never tires, but continues to repeat its question until the user believes that a broad and useful set of answers has been generated. Finally, the crystal ball is capable of uttering only the simple, specific questions, above.

The third unit of training describes several criteria used by experienced tactical decision makers concerning when to apply critical thinking and when to suffice with rapid recognitional processing. The criteria, in short, are that there must be time for critical thinking, given other priorities; the stakes (that is the range in the value of possible outcomes) must be high enough to warrant investing the time in the present problem; and there must be sufficient novelty to the situation to throw into question the accuracy of a rapid recognitional response. We then present a demonstration scenario and a practice opportunity.

The training closes with a summary of the lessons concerning assessment updates, critical thinking, and opportunity testing.

Scenario-Based Practice

The training provides scenario-based demonstrations and practice in using IDEA and the crystal ball. Each scenario consists of a background briefing (presented on slides) followed by a series of messages that describe scenario events (delivered on a simple e-mail system at the workstation provided to each participant and also presented to the group on slides). Each scenario is designed to exercise staff's skills in an instructional topic. The principle vignette, called the Frankfurt scenario, opens in the first unit of training and is elaborated in every subsequent lesson. Here, we combine several segments of that scenario into a demonstration of some critical thinking skills. The scenario briefing and several messages appear, below.

Background briefing

> You are a contingency force battalion S3 during a major regional conflict. Your battalion and the brigade of which it is a part are defending a sector with a port through which Allied reinforcements are arriving. Your task force assets are mixed armor and Bradleys. There has been no contact with the enemy for 48 hours, while political negotiators are busy trying to end the conflict. Enemy forces are arrayed about 30km away to

your northwest and to your southwest. Both units are Motorized Rifle Regiments. From either location, the enemy must traverse several rivers to reach the port, which appears to be their objective. The northern enemy force is better equipped for these river crossings than the southern force, and its commander is more experienced than the southern commander. However, poor roads and rough terrain in the north make armor movement there difficult. The southern terrain and roadways support rapid armor movement, and the southern enemy force has a more direct path to the port. The port is in the southern part of your sector, and it poses an attractive target to the enemy. Furthermore, the enemy has had marked success attacking your southern sector (but not the northern sector). Soviet doctrine, on which the enemy relies, is to exploit success.

Incoming messages

- Small contingent of southern enemy forces moving toward bridge Alpha on apparent approach to port.
- Enemy forces near bridge Alpha firing at US recon.
- Allied air interdiction campaign and indirect fire begin in southern section.
- POW reports lots of preparatory activity in main camp of southern enemy.
- BDA reports from Air Force indicate multiple southern enemy units struck.
- Intel reports that the southern enemy forces appear to have destroyed bridge Alpha to their front.
- Intel reports that enemy radio activity has ceased in south and north.

The battalion commander's initial prediction was that the enemy would conduct its main attack from the south. The commander now issues an update to that assessment, consisting of a story-like account of recent activity and a pointer to the story's main weakness.

Assessment update: Southern enemy is moving to evade interdiction. It has initiated radio silence, which is SOP for attack approach. Southern enemy will use its current movement to begin attack approach, but why is he shifting some forces to the north?

We ask staff to interpret this update and to advise the commander. By applying the IDEA method and the crystal ball, they

can formulate an argument in support of the current assessment and specify actions they can take to confirm their reasoning.

The officer takes as the tentative conclusion that the enemy will attack from the south. Several pieces of evidence support this claim, among them the doctrinal pursuit of success by the southern brigade and its advantageous routes to the port. However, the message that the southern enemy has destroyed a bridge in its own line of advance seems to conflict with the conclusion. Staff might use the crystal ball to address this conflict, as follows: "You may think that the reported bridge destruction conflicts with your assessment that the southern enemy will attack but it does not. Why not?" One of many possible responses neatly nullifies the conflict. It is to assume that the enemy does not plan to cross the river at the bridge, but elsewhere, and that its strike on the bridge is a proactive move to hinder friendly reconnaissance and defensive forces.

The crystal ball aids in identifying a gap in the available data regarding this assumption. The crystal ball says: "Your understanding of the situation hinges on an issue that is not addressed in any message or estimate so far. What is it?" One response is that the information concerning alternative crossing sites is sparse. This should cue staff to reevaluate the known river crossings.

Assumptions are lurking below the blue commander's assessment that the southern enemy will attack. The crystal ball helps to reveal some of them with this query: "There is another way to interpret this data overall. What is it?" One response is that the enemy does not in fact plan to attack, but merely to gain ground it wishes to claim during the current negotiations. When negated, this alternative assessment constitutes an assumption underlying the current assessment that the enemy will attack. The assumption can be tested, if only weakly, by studying whether the sites the enemy currently occupies have long-term strategic value or short-term tactical weaknesses. Staff can take this very action.

Measures

To measure the effects of the training described above, we developed instruments to assess three aspects of information management. These were the accuracy of decisions, by which we mean responses to command requests for tactical judgments; the quality of decision-making processes, by which we mean the manner in which staff assemble their knowledge to defend their decisions; and communications behaviors, particularly information filtering and production. In the following sections, we describe each of the measurement instruments in detail.

Accuracy of Decisions

STIM's measures of the accuracy of decisions are operationalized as the accuracy of responses to multiple choice questions requiring situational awareness and tactical judgment. For example, the questions used in the pilot experiment described below required the participant to state whether (1) the enemy force encountered thus far was the main force or the forward support element, (2) whether to continue fire missions despite the potential for fratricide, and (3) whether and how to displace forward units. The first two are close-ended questions multiple-choice questions. The third is partly closed (whether to displace) and partly open (how to do so). The number of reasonable responses to this question is large but limited, indicating that it can be presented in multiple choice format, but that responses may be rank ordered on accuracy. This offers interesting opportunities for assessing not only mean accuracy, but also variance in accuracy.

The challenge of constructing multiple choice problems is to complicate the process of choosing between the few alternatives available. To accomplish this, scenario authors can manipulate the level of uncertainty in a scenario. The principles we have applied to accomplish this in the sample STIM scenario are to ensure that (1) critical information is missing, (2) some events conflict with reasonable assessments or plans, and (3) events elicit highly unreliable assumptions. For example, in the Frankfurt scenario used in the pilot training material, reports that the enemy has destroyed a bridge to its front conflict with the assessment that the enemy planned an approach across that bridge. The same reports commonly elicit the marginally unreliable assumptions that the report is based on accurate observations--that the enemy has destroyed the bridge, and that it has done so intentionally, when in fact friendly units may be responsible or the enemy may have hit the bridge by accident.

In sum, measurement of decision accuracy in STIM is operationalized as the accuracy of responses to multiple choice questions requiring tactical judgement in complex scenarios laced with uncertainty.

Structural Measurement of Decision-Making Processes

An irony of testing decision making is that weak decision-making skills and good luck can combine to produce the same highly accurate decisions as strong decision-making skills alone. It is important to distinguish between these cases. Good decision-making skills are necessary for sound reasoning, and the most explicit, and therefore measurable form of reasoning is argument. We take as an index of decision-making skill the

structure and content of arguments that staff make in defense of their decisions.

This approach to measuring decision-making skill requires a well-defined notion of argument. A number of researchers have attempted to specify the structure of argument. Kuhn (1991, 1992) has developed a model of argument in which evidence bears on the validity of a hypothesis, and counterarguments potentially invalidate the hypothesis (by demonstrating that it postulates unnecessary or insufficient causal mechanisms) unless rebuttals are offered to neutralize them. Kuhn has demonstrated that higher education (college training) is correlated with successful argument, while age and domain expertise are not. Specifically, she has demonstrated that people without college education have difficulty distinguishing the causal model that constitutes a hypothesis from the evidence used to validate it, and that they often cannot conceive of alternative hypotheses, identify evidence that conflicts with their own hypotheses or rebut that evidence[3].

Toulmin's studies of business, law, management, the arts and ethics also focus on the nature and use of argument (Toulmin, Rieke, & Janik, 1984). They provide another definition of argument and a graphical representation (see Figure 3). Toulmin conceives of arguments as a linked structure of claims (or conclusions) based on grounds (facts or assumptions used as evidence) whose relevance and strength are a function of warrants (domain-specific rules for drawing conclusions) supported by backing (evidence for warrants). Rebuttals specify conditions in which a claim may be unjustified, thus breaking the link between grounds and claim. The existence of a rebuttal leads one to qualify one’s conclusions. However, rebuttals can themselves be rebutted, thus revitalizing a claim. Toulmin’s is a generalizable and representation of argument.

[3]There are circumstances in which these seeming deficiencies in arguments may be intentional and advantageous. For example, a prosecuting attorney may avoid introducing hypotheses that posit the defendant to be innocent, a defense attorney may intentionally conflate evidence with hypotheses in order to confuse the jury concerning what is fact and what is conjecture, and neither side is likely to raise evidence that conflicts with their hypotheses. However, these are cases in which the goal is to persuade, and not to pursue the truth. Staff officers are tasked to discover ground truth.

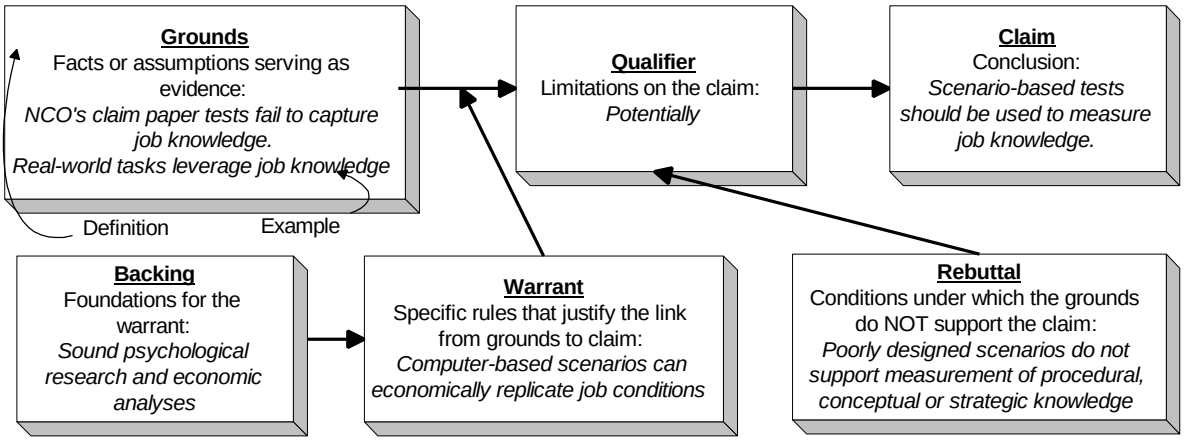

Figure 3. Toulmin's representation of argument.

We have taken prior work by Kuhn and Toulmin as a starting point to represent the structure of argument, but we have attempted to map our representation directly to the trained critical thinking skills. In our framework, a robust argument consists of a ***conclusion*** backed by ***supporting evidence.*** The conclusion may be weakened by ***conflicting evidence*** unless ***deconflicting assumptions*** or assertions are made that, like Toulmin's rebuttals of rebuttals, neutralize the conflict. We also expect a strong argument to recognize other sources of weakness or uncertainty, as well, namely ***gaps*** (or missing information) and ***assumptions.*** Finally, we extend the notion of argument somewhat to make it more relevant to the action-oriented domain of brigade or battalion TOC. We assert that a good argument suggests ***actions*** that can resolve uncertainty, such as requesting data, forwarding data, making recommendations, and formulating contingency plans.

There is a natural graphical format for this notation. The format employs nodes representing a conclusion, supporting evidence, conflicting evidence, deconflicting assumptions, gaps, and the assumptions covering them, other ("evaluated") assumptions and actions[4]. In Figure 4, we use this notation to illustrate an argument that the enemy in the Frankfurt scenario will attack from the south.

[4]In a training project begun since the completion of this Phase I contract, the argument syntax and its graphical representation have been revised to represent only the following. (The acronym for the approach, IDEAS, is formed from the capitalized letter in each component name.) Identified gaps; Deconflicted evidence; Evaluated conclusion; Action; and Supporting evidence.

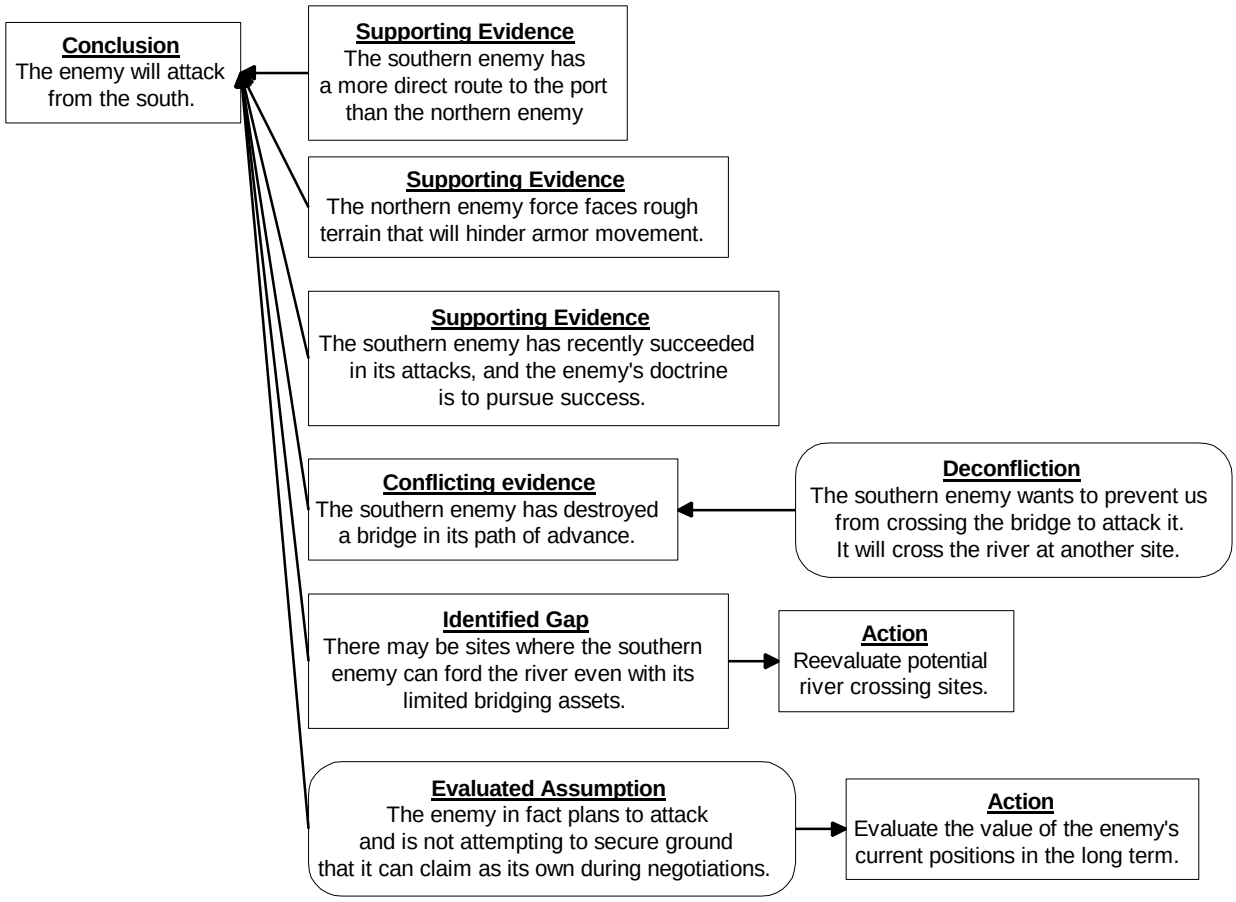

Figure 4. Arguments are represented in STIM as node-link graphs.

If officers can reliably represent their arguments with this graphical notation (or some variant of it), then there are intriguing opportunities for manually or automatically assessing arguments based purely on their syntactic or structural characteristics (in addition to qualitatively evaluating the accuracy of argument conclusions, discussed above, and the persuasive impact of arguments overall, as addressed below.) For example, one might award higher score for argument graphs that have a 1) greater variety of these components and 2) more instances of specific components. Greater variety is an indicator of broader competency in critical thinking skills. Greater number of components (assuming incomplete variety) may be an index of limited expertise in critical thinking but deeper domain knowledge. In evaluating arguments by their structural characteristics, we are inclined to give more weight to the variety than to the number of components, particularly for officers who have less domain expertise or no former experience with these critical thinking skills. However, it is an empirical question (not addressed in this study) whether breadth, depth, or their interaction with each other or other factors (such as problem type) best predict the persuasiveness of arguments, the

quality of actions intended to test assumptions, or the overall accuracy of conclusions.

Measuring Argument Persuasiveness

Good structure is a necessary component of a persuasive argument, but it is not sufficient. To measure the persuasiveness of an argument requires a metric of content; a measure of the evidence that is brought to bear on a conclusion and that which is omitted; the specific gaps and assumptions that are recognized or missed; how conflicting evidence is handled; and what actions in particular are proposed. The measure recommended for STIM is an SME rating of the persuasiveness of an argument over all of the evidence and reasoning presented. This can be supplemented by ratings and critiques of argument substructures (chains of components such as conflicting evidence, deconflicting assumption and actions to test the assumption), and individual argument components.

Communications Patterns

Communication behaviors--who consumes and produces what messages--are indicators of how the team processes data and how it adapts to changes in information load. To guide our approach to measuring communications behaviors, we used a process model. The model describes a two-phase information management process:

$$IM = IF + IP$$

where IM = information management, IF = information filtering, and IP = information production.

The IF process reveals the subject’s ability to filter out irrelevant incoming information and filter in relevant and critical information. IF is operationalized of as a categorical rating of incoming messages, in which the categories are:

- IF0: Ignore/don’t open (based on source, timing, title, circumstances, etc.)
- IF1: Open and classify irrelevant
- IF2: Open and classify relevant/useful
- IF3: Open and classify critical/essential

Information production operates only on the messages passed through the information filtering stages, namely IF2 and IF3. Measuring IP is a somewhat more complicated in that we want to characterize the messages in terms of the cognitive complexity of processes required to produce them. In terms of increasing complexity, outgoing messages may be classified as follows:

- IP1: pass-through: e.g., forward messages intact
- IP2: form judgment: e.g., fuse pieces of information, assess the situation (what-is type of uncertainty reduction)
- IP3: solve problem: e.g., formulate course of action, recommend decision (what-if type of uncertainty reduction)

The IF and IP phases may be particularly useful for measuring communication behaviors when these behaviors are also classified on the following dimensions:

- Direction of communications: Superiors (SUP), Co-Officers Staff (COS), Subordinates (SUB).
- Subject of incoming messages: Enemy or own troops.
- Type of communications: Information/Status (IS), Action/Plan (AP).
- Message classes: Request (REQ), Initiate (INI), and Respond (RES). This classification has been used before in other team performance work and has been proved quite useful (Entin, Serfaty, & Deckert, 1994).

Communication behaviors--who sends what messages to whom--are indicators of how the team processes data and how it adapts to changes in information load.

In the following sections, we discuss specific measures of communications performance based on the IF and IP phases and on these dimensions of them. Some of the measures are regular or global measures, based on counts of particular types of communications. They provide a sense of patterns of information flow. A second, general category of ratio formulas provides measures that are “normalized” and therefore highly diagnostic of performance.

Information Filtering (IF) Measures

The information filtering measures are based on several variables, defined as follows:

- II: Incoming information messages
- IF0: Ignored messages
- IF1: Read and irrelevant messages
- IF2: Read and useful messages
- IF3: Read and essential messages

Thus:

$$II = IF0 + IF1 + IF2 + IF3$$

All of these variables (IF0…IF3) can be decomposed using the categorization scheme described above. For example:

- IF2 (RES) = number of response messages deemed useful by the recipient
- IF0(SUB & IS) = number of information/status messages received from subordinates that have been ignored (unopened)
- IF3 (SUP & AP & REQ) = number of requests for actions or plans coming from superior commanders that have been read and classified essential

The lower bound of decomposition is determined by the sparsity of the data matrix (crossing the various dimensions) from a particular scenario run.

Information Filtering (IF) Ratios

A second set of measures in the IF process can be constructed as ratios of the previous variables. We have found in the past (Serfaty, Entin & Deckert, 1993; Serfaty, Entin, & Volpe, 1993) that ratios are superior indicators of behavior because they are more sensitive to changes in coordination strategies. Examples of these filtering ratios follow with hypotheses concerning the effects of STIM training under dynamic information loads.

- (IF-R1). Ignore ratio (IF0/II): An indicator of the strength of the first information management filter, which is not based on message content, but on external message markers, such as message type or source. Related hypotheses: As II increases, IF-R1 remains constant until some threshold of II is reached, at which point IF-R1 increases disproportionately to the number of incoming messages. STIM training should hold this increase constant (or at a constant growth rate) as II increases.
- (IF-R2). Informed rejection ratio (IF1/(IF0+IF1)): An index of the ability of staff to filter out messages based on content as opposed to surface features (such as the subject line or origin). Related hypothesis: As II increases, IF-R2 should decrease as officers become more selective about the messages they read. STIM training should stabilize IF-R2 at a high level.
- (IF-R3). Hierarchical information index (IF3(SUP)/IF3). An indicator of the focus on critical classification of messages coming from superiors. Related hypotheses: As II increases, IF-R3 increases. S3’s become more narrowly

focused on messages coming from above and less aware of criticality of messages coming from other sources. STIM training should remedy this. Assessment updates should help officers focus on message content, rather than the rank of the source.

- (IF-R4). Information reduction ratio ((IF2+IF3)/(IF0+IF1)): An indicator of strength of the second information management filter, based on relevance of message content. An alternative measure for this ratio is (IF2+IF3)/II. Related hypotheses: As II increases, IF-R4 increases first, then decreases. STIM training should maintain IF-R4 constant.
- (IF-R5). Content filtering ratio (IF1/(IF1+IF2+IF3)): An indicator of ability to dismiss what is not relevant to the current tactical situation. It may be an indicator of situational awareness. Related hypotheses: IF-R5 decreases as II increases. STIM training should increase IF-R5, because it supports informed dismissing of irrelevant information.

Information Production (IP) Measures

Information production (IP) measures concern the ability of staff to act on messages that pass the filtering stage (IF2 and IF3). Again, in this case, some definitions and raw measures are required:

- IO: Outgoing information messages
- IP1: Forwarded messages
- IP2: Messages that constitute judgment
- IP3: Messages that represent problem solving

Thus:

$$IO = IP1 + IP2 + IP3$$

As in the IF process, these variables can be further decomposed using the categories above. For example:

- IO (IS): Number of information/status messages produced and sent
- IP1 (REQ) = number of request messages passed through and sent
- IP2(IS&SUB&RES) = number of information/status messages sent to subordinate officer as a response to a previous request
- IP3 (AP&INI& SUP) = number of action or plan recommendations messages voluntarily initiated and sent to superior commanders.

Information Production (IP) Ratios

A second set of measures in the IP process can be constructed as ratios of the previous variables.

- (IP-R1). Information compression ratio (IO/II): An indicator of the ability of the S3's to reduce the volume of information they send out as a function of the information load they absorb. Related hypotheses: As II increases, IP-R1 remains constant. However, training should reduce IP-R1 in high information load (high II) cases. In this case, STIM training should act as an information volume stabilization device, aimed at controlling information inflation in the C2 organization as a whole.
- (IP-R2). Information/Action ratio ((IO(IS)/IO(AP)): An initial indicator of implicit coordination under high information load. Accurate, shared mental models can be used to infer the actions other team members should take when they receive an IS message. Staff trained in using assessment updates should be less likely to request actions if those actions will be performed anyway without the request. Related hypothesis: IP-R2 should remain stable or decrease with II. With STIM training, IP-R2 should increase.
- (IP-R3). Anticipation ratio ((IO(INI)/II(REQ)): Secondary indicator of implicit coordination strategies under high load. Such ratios--central to team adaptation theory--are indicators of anticipatory behavior and are very good predictors of performance. The anticipation ratio is a rich measure and can be broken down by destination or content. For example an upward information anticipation ratio, indicator of a staff member's ability to anticipate the information need of the commander can be calculated as ((IO(INI&SUP&IS)/II(REQ&SUP&IS). The basic hypothesis here is that as information load (II) increases, the anticipation ratio (IP-R3) decreases or remains stable. STIM training should foster an increase in IP-R3, especially under high information load. More detailed hypotheses can be developed for other variants of the anticipation ratio.
- (IP-R4). Responsiveness ratios (IO(RES)/IO(REQ)). An indicator of a staff member's ability to respond to the information or action/plan needs of the other team members. Related hypotheses: As II increases, IP-R4 decreases, i.e., officers are too overloaded to answer the needs of the others. STIM training should increase IP-R4 or hold it constant.
- (IP-R5). Pro-action ratio (IO(INI)/IO(RES)). An indicator of the ability of staff to remain pro-active in terms of initiating transfer of information and communication of

action as opposed to being reactive in terms of responding only to specific requests. Related hypotheses: As II increases, IP-R5 decreases. STIM training should encourage a high IP-R5 ratio as a function of an officer's situation awareness level.

- (IP-R6). Information Processing/Forwarding ratio ((IO2+IO3)/IO1): An index of the tendency to process information by forming judgements or solving problems, rather than simply forwarding data. Related hypothesis: As II increases, IP-R6 decreases. STIM training should stabilize this ratio.

Workload

A measure of workload has two potential uses in STIM. During formative or summative evaluation of STIM, the content validity of the decision making and communications measures could be assessed by demonstrating that performance on the measures varies as predicted between trained and untrained officers at a given workload level (as tested in the pilot study) and that performance varies as hypothesized within officers as workload levels change. During fielded use, an instrument for measuring workload could be used to adjust the quantity or quality of messages. This could optimize the difficulty of a practice or test scenario for the individual or the group.

There are three main ways to infer or assess workload in cognitively complex tasks. Physiological measures assess stress, i.e., the human response to workload (or other factors), by monitoring heart rate variability, pupil diameter, galvanic skin response, evoked potentials, etc. Performance-based measures indicate the effect on task work as a function of changes in workload. In this approach, a second task, such as auditory tracking, is superimposed on the main task. Performance decrement on the secondary task is an indicator of the workload generated by the primary task. Subjective measures, the third approach, are used to elicit participants' reports of the intensity of the task, or workload, using rating scales. Note that physiological metrics directly but measure stress (a physiological response), while the performance-based and subjective approaches measure workload, the determinant of stress that is of interest here. In research efforts in which it is important to minimize intrusion into the main task, we have found that subjective measurement methods provide both ease-of-use and reliability.

Two measures of workload have been extensively used in cognitively-demanding task contexts: the Subjective Workload Assessment Technique (SWAT) and the NASA Task Load Index (TLX). The SWAT (Reid & Nygren, 1988) uses three dimensions of workload: mental effort, time demand, and stress. The TLX (Hart &

Staveland, 1988) has six dimensions. The first three (mental demand, physical demand, and temporal demand) are viewed as relating to the demands imposed on the participant and the other three (performance, effort, and frustration level) to the interactions of a participant with the task. Both measures involve a procedure by which the workload dimensions are calibrated to an individual's perception of the most relevant dimensions for a particular type of task.

We recommend the TLX for two reasons. First, it requires less time from the participant than the SWAT to administer the calibration ratings, and it involves very little post-processing. In addition, the six TLX subscales provide more specific diagnostic information about the sources of workload than does the SWAT. Users of STIM can periodically complete a simplified TLX rating form (See Appendix H) to describe workload along the dimensions.

In this research effort, we adapted or developed measures of decision accuracy, decision making, or critical thinking skill, information filtering, information production, and workload. Most or all of these measures are designed with an eye toward future automation; they can be taken using computerized instruments during the run of messages that constitutes a scenario, or at breaks in a scenario, as discussed in the section concerning further development of STIM. In the next section, we present the results of a study that employed a selected set of these measures.

A PILOT TEST OF STIM CONCEPTS

Hypotheses and Research Questions

An experiment was conducted to pilot test the core training concepts of STIM, to establish the content validity of selected performance measures, and to elicit feedback concerning STIM from individuals with staff experience. We made the following predictions concerning the combined effects of STIM training and STIM's graphical, argument construction interface:

- H1: STIM will improve the accuracy of decisions participants make in response to requests for tactical recommendations at scenario breaks.
- H2: STIM will improve decision-making processes. This was operationalized as a test of training on SME ratings of the persuasiveness of arguments.
- H3: STIM will improve the structure of arguments. That is, it will enable participants to generate arguments that

contain more of the fundamental components of a sound argument.

- H4: STIM will improve information filtering behaviors. In particular, STIM will improve performance on several of the previously defined measures: the ignore ratio (IF-R1), the informed rejection ratio (IF-R2), and the hierarchical information index (IF-R3).
- H5: STIM will improve information production behaviors, specifically overall information production (IO), the information compression ratio (IP-R1), the information/action ratio (IP-R2), the information processing/forwarding ratio (IP-R6), and the pro-action ratio (IP-R5).

In addition, we explored several research questions on which the validity of other measures hinged or which were of value in developing STIM further.

- Q1: Can officers reliably parse their responses into argument categories? If officers could do so, then STIM training was clear regarding the argument syntax and representation, and prospects were good for automating structural analysis of arguments in future trials.
- Q2: Does the test scenario impose an appropriately heavy workload on staff? An answer to this question could guide the development of scenarios for future experiments with STIM.
- Q3: Does training influence the perceived workload level?
- Q4: What are users' impressions of STIM?
- Q5: What audience might benefit most from using STIM?

Experimental Design

The experiment manipulated one composite variable between subjects: the provision of STIM training, STIM's graphical argument construction tool, and assessment updates. Participants in a training treatment received these putative benefits, controls did not.

Subjects

The participants in this study were 11 former active-duty Army officers[5] with an average of 19.8 years of combined active and reserve Army duty (Standard Error of the Mean (S.E.M.) = 1.007). Approximately two-thirds of the participants had completed Basic and Advanced officers training, Combined Armed Services Staff School (CAS3), and Command and General Staff College (CGSC). All had some staff experience. The participants were training developers located at Ft. Knox, and all but one (a control) had written, vetted, played, administered or modified the scenario used in testing STIM. In sum, the participants were a relatively homogeneous group of experienced staff officers who were experts concerning the scenario used in testing. Four participants served as controls and seven received the experimental treatment.

Materials

The materials used in the study were a training booklet, a scenario studied by trained participants and controls, a test scenario, and debriefing materials.

Trained participants and controls studied the same scenario prior to testing. Called the Frankfurt scenario, it concerned an American battalion within a brigade-sized contingency force tasked to hold a port under threat of attack from two enemy Motorized Rifle Regiments (MRRs), one to the northwest and one to the southwest. The scenario briefing and numerous messages made it ambiguous which of the enemy forces, if either, might attack the port. The scenario was drawn verbatim from parts of the training book.

The test was a single segment of a defense-in-sector (DIS) scenario, 23 minutes and 30 seconds in length. This DIS scenario had been extensively evaluated and refined during its development for the Staff Group Trainer simulator (previously referred to as Commander Staff Trainer) (BDM Federal, 1996) and previous Army training projects. It was further modified by an SME for this experiment. The DIS scenario was chosen because it was reputed to present a high workload to the S3 (the role that participants played in this experiment), and because the defensive posture of the blue forces offered great potential for uncertainty and surprise. Materials for the scenario were four briefing documents, a message stream, and a situation map. The briefing

[5]A 12th participant argued strongly that he had no prior training or experience relevant to situation assessment and tactical decision making of the sort addressed in this study. Furthermore, this individual could not touch-type, a distinct handicap in this experiment. He was dropped from the analysis.

documents (only the briefest and most important of which are included in appendices to this report) were a short (two-page) Brigade Commander's Guidance (see Appendix A), a Brigade Area Defense Order, a short Battalion Task Force Commander's Guidance (see Appendix B) and a Task Force Order (key pages of which are in Appendices C, D, and E). The situation map was a set of three large maps (scale 1:50,000) of part of the National Training Center prepared by the Defense Mapping Agency Topographic Center, plus overlays of phase lines, critical areas of interest, and red and blue positions, which the test administrator updated on the overlays as the scenario progressed. The message stream consisted of scripted email from virtual task force elements concerning scenario events (sightings of enemy units, engagement reports, calls for fire, etc. (see Appendix F).

The debriefing materials were designed to elicit participants' evaluations of the training strategy (see Appendix G), their subjective ratings of the level of workload imposed by the test scenario (Appendix H), and biographical information (Appendix I).

Procedure

Each experimental session was four hours long and was attended by four participants situated at networked computer workstations. The session schedule began with brief introductory remarks, approximately 100 minutes of training or control activities, a 15-minute break, a scenario-based test lasting about 100 minutes, and a 15-minute debriefing.

Training and Control Activities

Officers in the experimental condition received the STIM training (see Appendix J). Instruction and demonstration sections of each unit were presented by the experimenter as a lecture with overhead transparencies[6].

The training was integrated with an introduction to the STIM interface. During training concerning assessment updates, the experimenter familiarized participants with the email application. The lesson concerning critical thinking skills introduced the graphical argument-construction application. In addition, the experimenter provided tips on managing space in the drawing application by minimizing nodes, overlapping nodes, and

[6]Due to time constraints, we eliminated the parts of the training material formally labeled practice. Participants practiced on parts of the material originally intended for demonstration. Expert responses in the demonstration material were presented as feedback to the participants as they completed each practice session.

moving nodes to adjacent pages. The presentation of training material varied slightly between the two, trained classes as the trainer developed his delivery style.

The control group performed two tasks in the time allotted to training other participants. These tasks were designed to expose controls to the scenarios used in experimental training and to the same instructional concerns (information management in the digital Army), but without presenting explicit instruction or key elements of the STIM interface. In the first task, the experimenter presented the Frankfurt scenario (see Appendix K) and asked officers to prepare a message describing their assessment of the situation and initial plans. Three blocks of messages were then delivered via email, to which the participants were asked to compose an email message describing appropriate S3 responses. In the second task, the group discussed information overload and information management issues regarding the digital environment.

Testing

At the beginning of testing, participants received and reviewed the four scenario briefing documents. The experimenter read the brief Battalion Task Force Commander's guidance and directed participants to review any other material they wished, and to pay special attention to three, one-page appendices to the Task Force DIS Order: the Task Force Execution Matrix (see Appendix C), the Synchronization Matrix (Appendix D), and the Decision Support Template (Appendix E). In addition, the experimenter described the status of forces at the start of the exercise with reference to a situation map. The briefing lasted approximately 30 minutes.

During the 23-minute and 30-second scenario run of the scenario, participants received 32 messages, delivered by email an average of 45 seconds apart. Those in the training condition received an additional two messages, each an assessment update, 2 minutes 30 seconds and 9 minutes 31 seconds into the scenario. The assessment updates provided no new information concerning scenario events. Like the assessment updates participants studied, however, these messages alerted staff to the Commanding Officer’s (CO’s) concern about troubling tactical issues, namely the size of the enemy force at the first break and the potential for fratricide while shelling the enemy at the second break. The two assessment updates in the message stream were provided to trained participants to replicate the effects of working in a team with a CO trained using STIM. Thus, these two messages were an independent variable designed to reflect team-oriented aspects of STIM training that otherwise could not be evaluated given the small available sample of participants.

Participants were asked to handle the incoming messages as if they were the Falcon task force S3. As messages streamed in, participants responded using an email application plus an address book listing other (virtual) officers in the scenario. (The email application is pictured on the right side of Figure 5.) To obviate the need for participants to acknowledge every message, we told them that opening an email automatically acknowledged its receipt.

At three points in the scenario (7:30, 13:50 and 23:30), we stopped the message stream and asked officers to respond to the last message they had received, which was a request from the task force commander (06) for a tactical recommendation. At each break, we repeated instructions to (a) answer the question asked in the message, (b) defend your answer, and (c) indicate any actions you wish to take. Control participants responded to each question in writing using the email application. Trained participants responded by constructing an argument using the graphical interface. (The template of shapes available in the graph-builder appears on the left side of Figure 5. Participants dragged these shapes to a window containing a blank worksheet, linked them with arrows, and filled the contents by typing in free text or dragging in email messages.) Officers were given eight minutes to complete their answers to the questions at each break.

The experimenter posted reports of unit sightings in real-time on a full-scale sitmap in view of all participants. Participants were invited to get up from their seats to look at the map if they wished. Few did, and these did so only once or twice during the scenario.

At the conclusion of each break, participants were given a printed page listing each of the email messages they had just received. The experimenter asked them to rate each message (excepting the final message (the question) at each break) on a scale of 0 to 3, indicating the importance of the message:

0 = ignore (messages not worth reading)
1 = irrelevant (messages worth reading but not of much importance)
2 = relevant or important
3 = critical

Trained participants in the last experimental session were given a one-page summary of the four steps of the IDEA method and the questions asked by the crystal ball at each step, when it became apparent that the previous group of experimental participants would have benefited by this reminder.

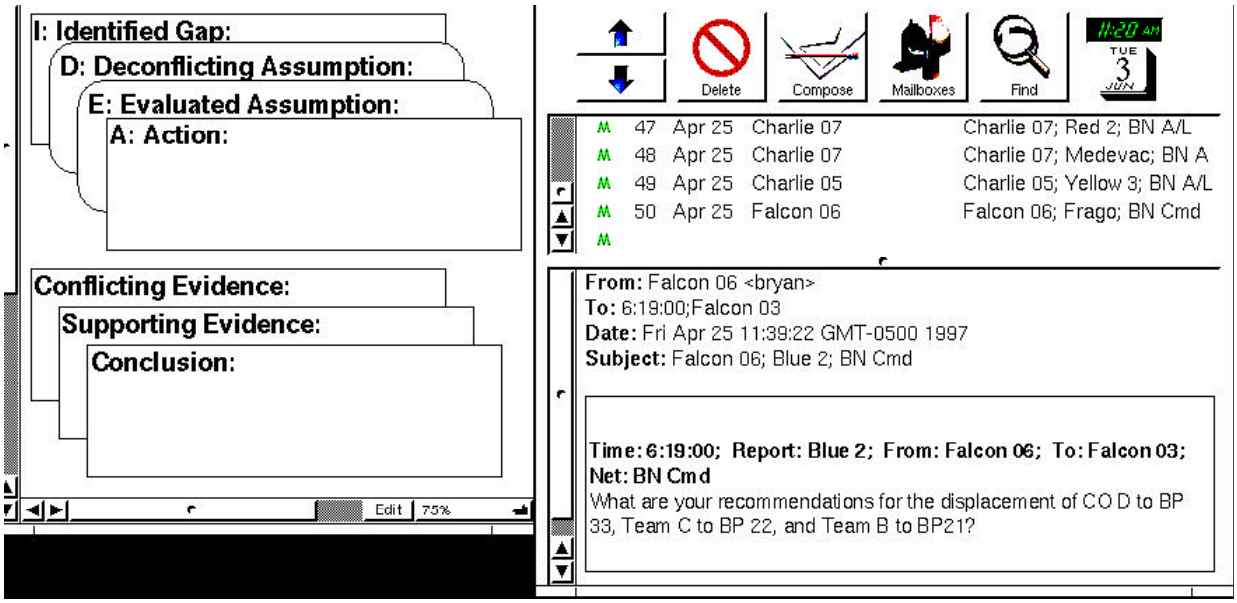

Figure 5. The STIM interface.

Debriefing

At the conclusion of the test, officers were asked to fill out a debriefing form consisting of several parts:

- A modified version of the NASA TLX form for eliciting subjective ratings of the workload in the test scenario;
- A questionnaire concerning the effectiveness of the training and the usefulness of aspects of a training system interface; and
- A biographical information form.

Apparatus

Each participant trained and tested at a Pentium-based personal computer. These workstations plus a server were linked to form a five-station network. The network architecture simplified test administration and data collection. The software at each station was the NeXTSTEP®[7] graphical interface to UNIX®[8], a simple, graphical email utility with an address book, and (for officers who received experimental training only) a drawing application (DIAGRAM!®[9]). Message streams were presented across the network as incoming mail under the control of a Perl script.

---

[7]NeXTSTEP is a registered trademark of NeXT Software, Inc., a division of Apple Computer.

[8]UNIX is a registered trademark of UNIX System Labs, Inc.

[9]DIAGRAM! is a registered trademark of Lighthouse Designs, Ltd.

The experiment was conducted at the Mounted Warfare Test Bed at Ft. Knox, a large facility used by the Army for training and force development.

Results

The small sample of participants available for this study constrained the statistical power of the experiment. In light of these factors, effects above $p = 0.05$ and as weak as $p = 0.20$ are reported as trends. All t-tests are pooled, two-tailed tests unless otherwise described.

In general, these results should be interpreted with caution, given the small size of the sample, the use of a single test scenario, homogeneous characteristics of the participants, the variance in presentation of materials during training between groups, short duration of training, and minimal individual feedback.

Decision Accuracy

The fundamental test of training concerned its effect on decisions. We hypothesized that STIM training would improve participants' tactical recommendations (H1).

Qualitative analysis of the responses was performed by a professional decision analyst and retired LTC with 27 years of military experience. This SME was a graduate of the Command and General Staff College, a former faculty member of the US Military Academy, and a former adjunct faculty member of the National War College. To blind the SME to experimental conditions, the argument graphs created by trained officers were converted to text, the responses of controls were parsed into argument components like those of the trained officers, and responses on each break by each participant were given a unique, random identifying number (to prevent the SME from inferring subject condition from patterns of responses across breaks). In the analysis of decision accuracy, the SME considered only the conclusion of each argument.

Participants were scored on the accuracy of the conclusions they presented in their responses to the three break questions. On the first break question, participants received a request from Falcon 06 (the task force commander) asking whether the task force was in contact with the enemy's forward support element (FSE) or with it's main body, the motorized rifle regiment (MRR). The SME awarded one point for a conclusion stating the enemy force was the FSE and a score of zero for other responses. On the second break, the 06 inquired whether fires should be stopped because of the possibility that they were striking the task force's own unit, Charlie. The SME gave one point for the

conclusion that fires should not be stopped, and a score of zero otherwise. On the third break, the 06 presented a more open-ended request for recommendations concerning the displacement of forces. The SME awarded a score of one to any answer that was clear, complete and tactically reasonable, and zero to inadequate responses.

Over all breaks, trained participants were 36% more likely than controls to reach accurate or reasonable conclusions. Ninety percent of all responses by trained participants contained good conclusions (representing a mean total score over three breaks of 2.714 out of a possible 3, S.E.M. = 0.286), versus 67% for controls (mean (M) = 2.000, S.E.M. = 0.408). This benefit of training represented a trend, in a two- tailed t-test with pooled variance (t_9 = -1.467, $p < 0.20$)[10] (see Figure 6). Trained participants also produced better conclusions than controls at each individual break. On break 1, 86% of conclusions by trained participants were reasonable vs. 75% for controls; on break 2: 86% vs. 50%; and on break 3: 100% vs. 75%. None of these differences was statistically reliable.

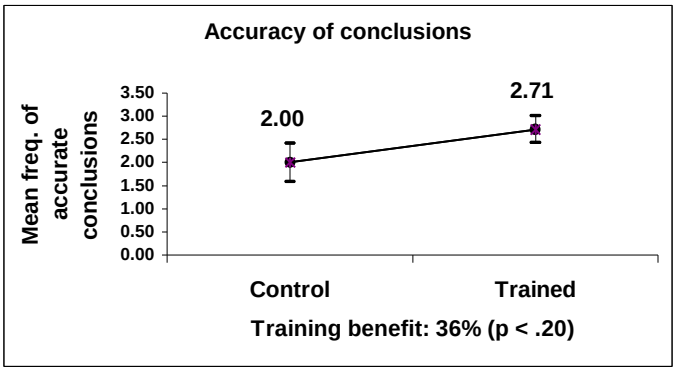

Figure 6. The accuracy of tactical conclusions.

Persuasiveness of Arguments

One potential explanation for the increase in accuracy among STIM-trained participants is that the training and the graph construction software supported sound processes of tactical decision making. An global measure of this effect was the SME's score of the persuasiveness of the arguments participants gave in support of their conclusions. The SME scored responses to each of the three break questions on an 11-point scale, where 0 = very weak argument (unpersuasive) and 10 = very strong argument

[10]Z-scores for skew and kurtosis were not extreme for the data of either group.

(highly persuasive)[11].

Consistent with the prediction (H2), responses by trained participants were 93% stronger than those of controls when persuasiveness scores were totaled over all three breaks. Trained participants scored a total of 19.071 on average (S.E.M. = 3.165), while the mean control score was 9.875(S.E.M. = 4.943) (t_9 = -1.647, p < 0.15) (see Figure 7).

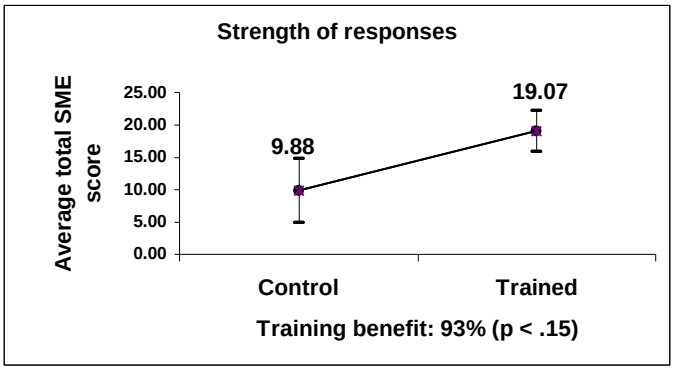

Figure 7. The persuasiveness of arguments.

Responses by trained participants were also more persuasive than those of controls on each individual break, on average. On the last two breaks, when participants were presumably more comfortable with the testing procedure, these differences were statistically reliable (see Table 2).

It might be argued that trained individuals faced greater task demands than controls. During the brief breaks, trained participants had to operate both the email application and the graphing application simultaneously (if they wished to drag prior messages into their graphs, as most did); they had to manage the layout of nodes and links in a relatively small drawing space; and they bore the general burden of operating a newly learned drawing tool. When asked, participants in the training condition stated that they needed more than the eight minutes allotted to record their concepts. We attempted to compensate for the apparent lack of time in the following manner. At the first break, trained participants were given precisely eight minutes to complete their answers. At the second break, they were told they

[11]To blind the SME to experimental conditions, the responses of trained officers were converted to text, the responses of controls were parsed to resemble the phrased responses of trained officers, and responses on each break by each participant were given a unique, random identifying number (to prevent the SME from inferring subject condition from patterns of responses across breaks).

Table 2. Mean Persuasiveness of Participants' Arguments.

Break	Control	Trained	Reliability
1	M = 4.750 S.E.M. = 1.931	M = 6.143 S.E.M. = 1.189	t_9 = -0.653 not significant (n.s.)
2	M = 2.50 S.E.M. = 1.555	M = 6.214 S.E.M. = 0.975	t_9 = -2.141 p < 0.10
3	M = 2.625 S.E.M. = 1.675	M = 6.714 S.E.M. = 1.079	t_9 = -2.153 p < 0.10
Total	M = 9.875 S.E.M. = 4.943	M = 19.071 S.E.M. = 3.165	t_9 = -1.647 p < 0.15

would have exactly eight minutes, and at the end of that time, their diagrams were saved to disk. However, we then granted them an additional three minutes with the proviso that this was a one-time arrangement. At the third break, we again announced they would have only eight minutes, saved their work, and then announced that the test would end with a final three-minute extension to complete their responses to the current question. A comparison of argument persuasiveness by break and over breaks indicated that, with additional time at breaks two and three, trained participants improved their performance further. They scored 105% higher on persuasiveness than controls over all breaks (t_9 = -1.874, p < 0.10). Trained individuals outperformed controls on each of the three breaks, as well, and the differences were significant on the second and third breaks (see Table 3).

Accuracy of Classification of Argument Components

One goal of the present study was to establish whether participants could correctly classify components of their own arguments using the graph-construction tool (Q1). Did they, for example, present evidence supporting their conclusion in a node labeled "Supporting Evidence"? If participants could not correctly classify their own statements, then STIM training was unclear and the prospects for automating measurement of argument quality based on these structural data were dim.

To assess the accuracy with which trained participants classified components of their arguments, the experimenter and SME generated correct classifications of those components. The accuracy of each respondent at each break was the ratio of argument components classified correctly by the participant to all argument components the participant generated. Over all

Table 3. Mean Persuasiveness of Arguments Given Additional, Compensatory Response Time.

Break	Time allotted	Control	Trained	Reliability
1	Controls: 8 minutes Trained: 8 minutes	M = 4.75 S.E.M. = 1.931	M = 6.143 S.E.M. = 1.189	t_9 = -0.653 n.s.
2	Controls: 8 minutes Trained: 11 minutes	M = 2.5 S.E.M. = 1.555	M = 6.857 S.E.M. = 0.918	t_9 = -2.6 p < 0.05
3	Controls: 8 minutes Trained: 11 minutes	M = 2.625 S.E.M. = 1.675	M = 7.214 S.E.M. = 1.09	t_9 = -2.402 p < 0.05
Total		M = 9.875 S.E.M. = 4.943	M = 20.214 S.E.M. = 3.103	t_9 = -1.874 p < 0.10

breaks, trained participants correctly classified 82% of all argument components on average. (The mean score was 2.447 out of a possible 3 points, representing three perfectly classified sets of arguments, S.E.M. = 0.159.)[12] They were least accurate in applying three classifications: conclusions (this label was correctly used in 76% of all instances), conflicting evidence (used correctly 67% of the time), and deconflicting assumptions (40% accuracy). The bulk of what participants called deconflicting assumptions were classified by the experimenter and SME as assumptions unrelated to conflicting evidence (35%), or as supporting evidence (20%) or gaps (5%) (see Table 4). We conclude that participants classified argument components with reasonable accuracy, particularly given the brevity of training.

Effects of Training on Argument Structure

Given that participants were reasonably accurate in their classification of argument components (and assuming that this classification could be improved with training), we asked whether the structure of responses differed between control and trained participants. If it did not do so, then there was little point in assessing specific differences in argument structure (H3).

[12]Accuracy of argument component classification was virtually identical when trained officers were given three additional minutes to complete their responses on the last two breaks.

Table 4. Accuracy of Trained Participants at Classifying the Components of Arguments.

Corrected coding (columns) Participant coding (rows)	C	SE	CE	D	E	I	A	Other	Grand Total
Conclusion (C)	76%	8%			4%	4%	8%		100%
Supporting evidence (SE)	2%	96%						2%	100%
Conflicting evidence (CE)		11%	67%		11%	6%		5%	100%
Deconflicting assumption (D)		20%		40%	35%	5%			100%
Evaluated assumption (E)				7%	85%		8%		100%
Identified gap (I)		12%				82%	3%	2%	100%
Action (A)	11%						89%		100%

Note. Cells indicate the percentage of correct classifications by participants. Blank cells represent zero confusion errors. Rounding errors may result in row totals other than 100%.

To evaluate the effects of training on argument structure, the break responses of control participants were parsed and categorized using the scheme described above. The SME vetted all categorizations. A comparison was then made of the distribution of responses by argument category over all breaks for the control and treatment groups. The distributions showed striking differences.

Only trained participants specified conflicting evidence in their arguments (M = 1.571 points of conflicting evidence per trained respondent over all three 8-minute breaks, S.E.M. = 0.528) or deconflicting assumptions and assertions (M = 1.286, S.E.M. = 3.402). These participants also offered more supporting evidence for their recommendations (M = 8.429, S.E.M. = 1.288) than did controls (M = 3.250, S.E.M. = 1.377). This was a reliable effect (t_9 = -2.578, $p < 0.05$). Trained participants more often identified the gaps or missing information in their arguments (M = 3.714, S.E.M. = 0.993) than did controls (M = 0.500, S.E.M. = 0.500), a reliable effect (t_9 = -2.308, $p < 0.05$). Trained participants specified more assumptions (M = 2.000, S.E.M. = 1.254) than controls (M = 0.500, S.E.M. = 0.500). The actions trained participants listed were fewer in number (M = 3.429, S.E.M. = 1.088) (but better in quality, on average, see the analysis, below) than the actions of controls (M = 4.750, S.E.M. = 3.772). Effects on assumptions and actions were not statistically reliable.

Trained participants generated almost twice as many arguments on average over all three breaks (M = 23.714, S.E.M. = 4.162) than did controls (M = 12.250, S.E.M. = 5.603). However, this difference between groups was much weaker than other trends reported here (t_9 = -1.651, p < 0.30). Participants varied substantially in the length and complexity of their arguments. Overall, however, it appears that training improved the structure of arguments participants generated (H3) (see Figure 8).

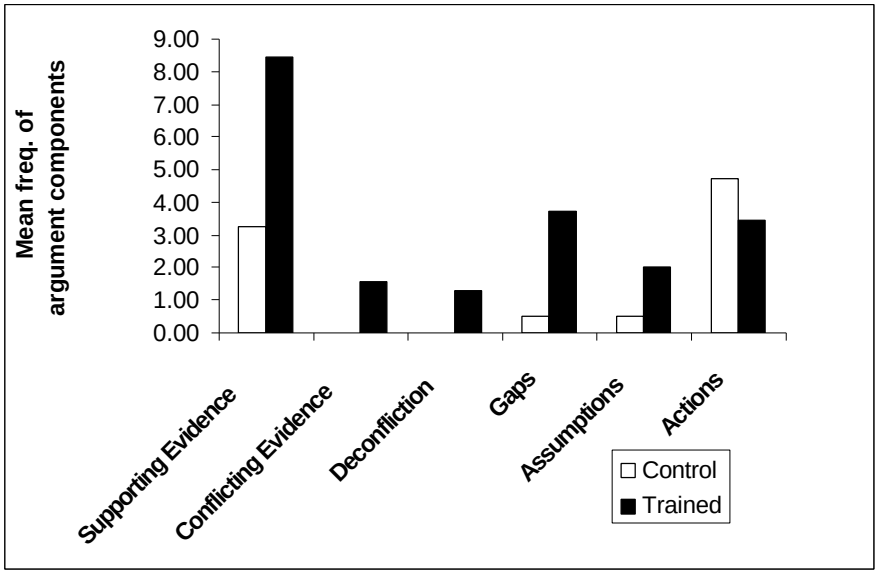

Figure 8. The variety of argument components used.

Other Qualitative Factors

In rating responses, the SME spontaneously considered several factors, including (a) whether participants cited messages as evidence, (b) whether they went beyond the evidence in articulating their reasoning, (c) whether they considered alternative hypotheses or challenged assumptions, and (d) the reasonableness of their actions. The effects of training on some of these factors are analyzed here, though these are not proposed as automated STIM measures.

In order to serve well, staff must competently gather data to inform themselves, fellow staff, line officers and others, and they should recommend appropriate actions. In grading responses to the break questions, the SME awarded each response a score of one for reasonable actions, such as appropriate requests for information or the recommendation of sound tactical maneuvers, or a zero otherwise. Trained participants committed themselves to reasonable actions on over half of all breaks (M = 1.714 out of 3, S.E.M. = 0.421), while controls did so on only one-third of

breaks (M = 1, S.E.M. = 0.707), a 71% difference in performance. However, this effect was not statistically reliable.

The SME awarded one point if a participant cited one or more incoming messages as evidence in their responses to break questions, and zero otherwise. Trained participants cited messages as evidence in 86% of responses, while controls did so on only 42% of breaks. This difference was statistically reliable (t_9 = 2.993, $p < .05$). There is a simple explanation for this pattern: it was easier for trained participants than controls to cite incoming email messages because STIM enabled users to simply drag email into their argument graphs.

Trained participants were more likely than controls to go beyond the evidence, that is, to state assumptions and inferences in their responses. Specifically, trained participants went beyond the evidence in 86% of their responses, while controls did so only 17% of the time. Trained participants were also more likely than controls to consider alternative hypotheses or challenge assumptions. Trained participants exhibited this behavior on 86% of responses, while controls did so on 17% of their responses. However, neither of these findings was statistically reliable.

Communications

Information Filtering Behaviors

Tests of information filtering behaviors concerned the effects of training on the perceived criticality of incoming messages. As defined above, the measures employed participants' subjective ratings of the relevance of incoming scenario messages as ignored, irrelevant, relevant, or critical[13].

Overall, the trained group classified messages and filtered information in ways that were significantly different from controls (χ^2_{33} = 54.4 $p < .015$) in a Friedman test (Siegel, 1956). The detailed results were generally in line with our predictions (H4).

Compared with controls, trained participants were inclined to read, rather than simply ignore a larger proportion of the least relevant messages. The informed rejection ratio (IF-R2) for trained participants was 75% (S.E.M. = 0.033) vs. 57% (S.E.M. = 0.036) for controls (t_9 = -1.92, $p < .05$) (see Figure 9).

[13]This analysis considered only the subjective ratings of message criticality. Future studies should compare participants' judgements of message criticality with those of an SME.

In rating incoming messages, the trained group was also less influenced than the control group by the rank of the message sender. This was evident in two measures. Trained participants were 44% less likely to indicate they ignored messages from the lower echelon than were controls. At the mean, trained participants ignored 5% of these messages (S.E.M. = 0.005); the figure was 9% for controls (S.E.M. = 0.008). This pattern on IF-R1 represented a respectable trend in the predicted direction (t_9 = -1.41, p = .09). In addition, trained participants were 45% less likely to rate messages from superiors as critical than were controls. Among trained participants 18% of superiors' messages were rated as critical, on average (S.E.M. = 0.027), while the mean among controls was 33% (S.E.M. = 0.033). This pattern on IF-R3 approached statistical reliability (t_9 = -1.64, p < .07) (see Figure 10). In summary, it appears that trained participants attended to message content more than to parameters such as origin of the information.

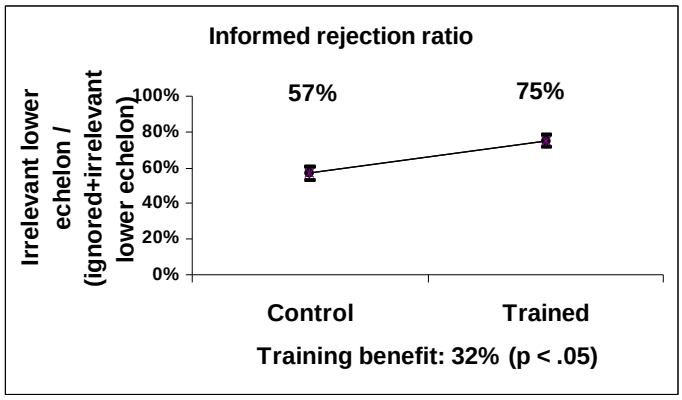

Figure 9. The informed rejection ratio (IF-R2).

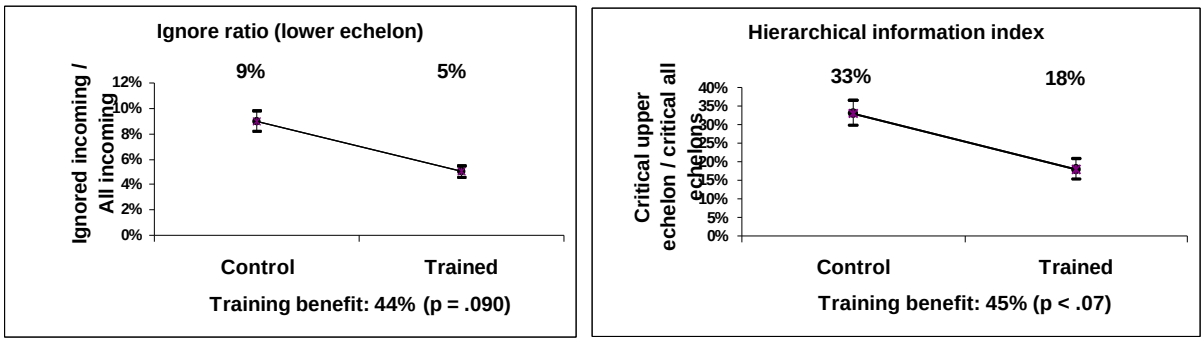

Figure 10. The influence of rank on subjective ratings of message.

Information Production

As predicted, STIM improved message production (H5). A key effect was that the trained group generated 45% fewer messages, on average, than controls. (For trained participants, M = 11.3, S.E.M. = 1.266; for controls, M = 20.700, S.E.M. = 2.141; t_9 = 1.82, p < .05). As a result, the information compression ratio (IP-R1) was reliably 83% higher for trained participants (M = 3.00, S.E.M. = 0.121) than for controls (M = 1.35, S.E.M. = 0.075; t_9 = -1.71, p < .05) (see Figure 11).

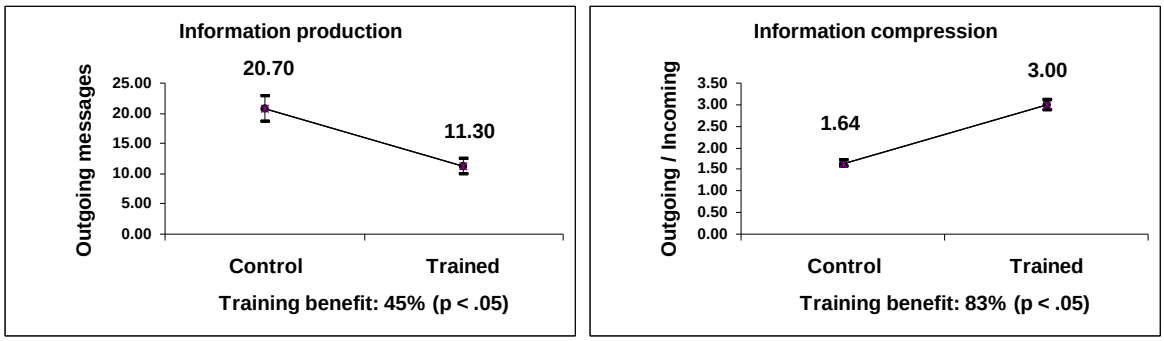

Figure 11. The number of messages generated and the compression ratio (IP-R1).

Trained participants also produced 83% more information or status messages for every action message or plan they produced (M = 2.47, S.E.M. = 0.109 on the information/action measure IP-R2) than did controls (M = 1.35, S.E.M. = 0.124; t_9 = -1.84, p < .05) (see Figure 12). As discussed below, this effect has been interpreted as evidence of implicit coordination within the team. Training may sensitize participants to the understanding that passing information is often sufficient to trigger actions, and they may infer that some requests for action are unnecessary.

When the frequency of messages by class was examined, the trained participants were found to generate more messages on their own initiative (INI) for every message that constituted a response (RES) to a request for information. (IP-R5: M = 3.10, S.E.M. = 0.124), relative to controls (M = 2.30, S.E.M. = 0.115; t_9 = -1.39, p < .10) (see Figure 13). As predicted, then, the training group was more proactive in its communications.

Finally, trained participants generated more messages that involved judgement or problem solving per message forwarded (IP-R6: M = 2.12, S.E.M. = 0.112) relative to controls (M = 1.25, S.E.M. = 0.060; t_9 = -1.95, p < .05) (see Figure 14).

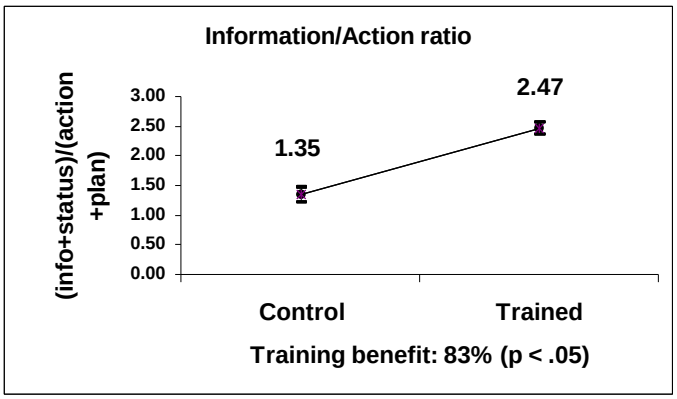

Figure 12. The ratio of information or status messages to action or planning messages (IP-R2).

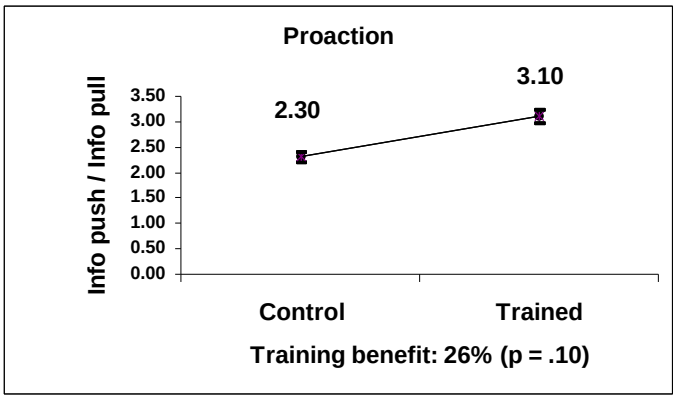

Figure 13. Proactive communications (IP-R5).

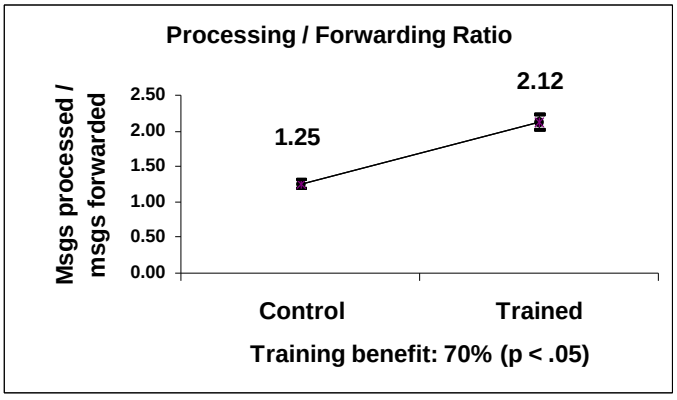

Figure 14. The ratio of processed to forwarded data (IP-R6).

Workload

A variant of the NASA TLX workload questionnaire was administered immediately after participants completed the

scenario. This form asked participants to "rate the scenario you've just completed with respect to your experience concerning:" mental demand, physical demand, time pressure, effort, and frustration. In addition, we asked officers to rate their performance. All ratings were on a scale from 0 (very low) to 10 (very high). Analysis of the results bore on the question of whether the test scenario imposed an appropriately heavy workload on staff (Q2), and on the effects of training on perceptions of workload (Q3).

Trained participants perceived slightly higher task demands than did controls, as indicated by a mean rating of physical demands 71% higher among trained participants (M = 2.143, S.E.M. = 0.738) than controls (M = 1.25, S.E.M. = 0.25), and ratings of time pressure that were 24% higher among trained participants (M = 6.857, S.E.M. = 0.829) than controls (M = 5.5, S.E.M. = 1.19). However, differences between groups on these two measures and the measure of mental demand (approximately 7.00 in both groups) were not statistically reliable (see Figure 15).

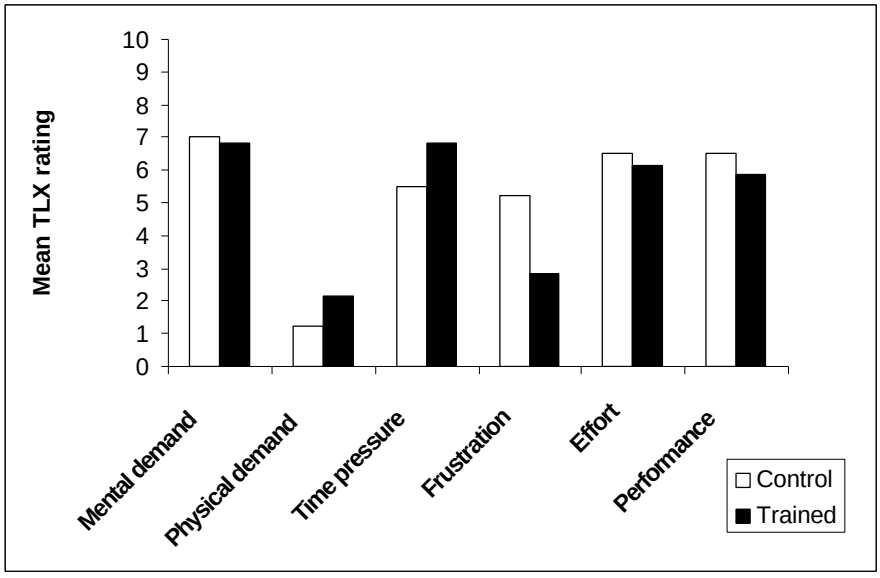

Figure 15. TLX ratings of workload.

Trained participants perceived slightly lower levels of workload on measures of the interaction between the task and the individual. Ratings were 46% lower on frustration among trained participants (M = 2.857, S.E.M. = 0.508) than controls (M = 5.25, S.E.M. = 0.25) (t_9 = 3.362, $p < .01$). Effort was 5% lower (trained M = 6.143, S.E.M. = 0.508, control M = 6.500, S.E.M. = 0.289). Self-assessed performance was 10% lower among trained participants (M = 5.857, S.E.M. = 0.404) than controls (M = 6.5, S.E.M. = 0.866). Neither effort nor performance scores differed reliably between groups.

The absolute values of the means on all measures were low to moderate, suggesting that the test scenario did not impose a heavy workload on participants, contrary to expectations.

Participants' Evaluations of STIM

When asked to "rate the training overall" on a scale of 1 to 10, participants responded with a modestly positive rating of 6.571 (S.E.M. = 0.429). The responses to three other debriefing questions provide more specific insight into their estimates of STIM's potential value (Q4).

We asked participants, "Did the training and/or the interface influence your performance on this test?" Of the six participants in the experimental condition who responded to this question, four (67%) answered yes. Three answers worthy of note were these:

- Yes. I had never used this type of system but was able to send and forward messages as appropriate.
- The boxes for supporting & conflicting evidence were useful. Didn't use crystal ball. I can question others, not myself.
- Like the interface better than Staff Group Trainer (SGT). But requirements for no map edits cuts down on interface requirements.

Participants who claimed the training and/or the interface did not help them on the test simply answered "no" or "not really" in response to this question.

Participants told us that STIM training was likely to influence how battalion staff officers solve problems in the field in Force XXI. Five of the six responses (83%) to a question on this topic were positive. Several typical or interesting responses were:

- Yes, it makes people evaluate how they think.
- Yes. Any mental exercise + staff thinking would help.
- Seems to be a useful technique. Would like to have additional time for the prep.
- The one negative response was, simply, "No."

We asked trained participants to give us their general comments concerning the training. All of the seven responses to this question were positive, though one participant noted the need for longer training and another the need for more focus on "content," possibly indicating a desire for more feedback and demonstration and practice opportunities, or perhaps denoting an

interest in training focused on Force XXI technologies. The responses concerning the general value of training were:

- This is an excellent approach.
- Good for training environment
- Training would be an excellent tool to use in a classroom environment.
- Good beginning.
- I liked it. A bit more formalized version of brainstorming and what-iffing.
- Helpful. Additional streamlining would improve acceptance. Less on mechanics of process and more on contents.

In sum, participants were generally positive in their assessments of STIM. Most believed that it improved their test performance on a scenario with which they were (with the exception of one officer) already highly familiar. Most felt it would improve decision making in the field and all had positive overall comments.

Potential Audiences for STIM Training

The overall effects of STIM training appeared to be positive. However, we wished to learn for whom they might be most valuable (Q5). There was no meaningful correlation between the accuracy of conclusions or argument persuasiveness and the SME's rating of participants' career experience with S3 responsibilities[14]. Thus, we turned to comments from participants to help determine where STIM training might be best applied. One participant made the following suggestions in his debriefing notes:

> This vehicle should look at the Operations Other Than War (OOTW) arena. This is an area which is only effectively taught at CMTC in Europe and Joint Readiness Training Center (JRTC) in Continental United States (CONUS) (thereby missing a significant part of active Army and the Reserve Component (RC) element)...Feel that it can be targeted at the Officer Advanced Course, Advanced Non-Commissioned Officers Course ANCOC, Battle Staff Non-Commissioned Officers (NCO) courses & provide benefits to the Army's Advanced Individual Training.

[14] The qualification rating was a single score (0 to 10) generated by the SME for each participant. It was based data from the biographical questionnaires. Prior S3 experience weighed heavily in the SME's ratings.

In informal discussion after the experimental sessions, another participant indicated that the training was particularly appropriate for Captain's instruction. He noted that STIM training has the potential to make training in rapid decision making more interesting and productive by encouraging officers to exercise judgement rather than relying on memorized, doctrinal responses. However, an additional contribution of this approach is that it requires officers to consider when to invest precious time in argument-based decision making and when to rely on rote, doctrinal responses.

Discussion

The study was designed to provide preliminary data concerning the effects of core aspects of STIM: staff training in critical thinking and coordination, the use of a graphical notation and tool for representing tactical reasoning, the content validity of measures, and the face validity of the training and interface.

Several factors compel us to interpret these findings with reserve. The available sample of participants was quite small, and this limited the power of the tests. Many of the effects represented only trends ($0.05 < \underline{p} < 0.20$). Furthermore, the participants were a relatively homogeneous group, from whom measures were taken on a single scenario, factors that limit the generalizability of the findings. Finally, minor variance in the presentation of materials during training, the short duration of training, and the minimal level of individual feedback during training are potentially sources of error variance that should be controlled in larger, future studies. With these caveats in mind, however, we observe that the data generally supported theoretically grounded hypotheses that training would improve staff decision making and communications behaviors (see Table 5).

Among the key findings was the trend that STIM increased decision accuracy by 34%. The simple, multiple choice measure of the effect was easy to implement, and modifying an existing scenario to challenge performance on this measure was reasonably straightforward.

Decision-making processes also tended to benefit from training. The persuasiveness of arguments was 93% greater with training than without. Furthermore, there were positive structural differences in the arguments generated by control and trained participants. STIM helped participants to apply more of the evidence in arguments defending their conclusions. Particularly noteworthy was that trained participants cited and dealt with more of the evidence that seemed to conflict with their conclusions. This indicates that STIM may be a prophylactic for confirmation bias, the frequently observed effect in which

people underweight data that conflict with their beliefs (Nisbett & Ross, 1980). STIM also helped participants to identify gaps and assumptions and to cite more of the available supporting evidence to reason about tactical issues. While some of these effects were only statistical trends, they were all in the predicted direction.

Table 5. Summary of STIM Training Effects.

Measure	Effect	Reliability
Decision accuracy	Increased by 34%	$p < 0.20$
Decision-making processes (argument persuasiveness)	Increased by 93%	$p < 0.15$
Decision-making processes (argument structural integrity)	Increased recognition of supporting evidence, conflict, gaps, and assumptions	Mixed significant and n.s.
Information filtering	Improved	$0.05 < p < 0.10$
Information production	Improved	$p < 0.05$ (one effect: $p = 0.10$)

Trained participants were moderately accurate in classifying the argument components they generated: they classified 82% of argument components correctly. Given the brevity of training, this is a reasonable accuracy rate. However, we predict that accuracy could be improved considerably with improvements to the training and increased feedback. If more reliable classification of argument components can be achieved, then it may be possible to automate metrics of argument quality that employ data concerning argument structure.

It might be argued that the comparison of performance by controls and trained participants was invalid because trained participants used a tool (the graphical argument builder) on the test that supported them in constructing arguments, while controls did not. This critique is most clearly relevant to the issues of argument persuasiveness and structure. Though all participants were told to defend their conclusions, trained participants also had the support of the STIM interface for formulating arguments. This support was in the form of a template with blank nodes labeled to remind them to consider supporting evidence, conflicting evidence, assumptions, and so forth. However, the email editor used by controls to respond to questions gave them the freedom to employ the same, simple and common elements of argument, and to go beyond the STIM syntax, if necessary. Controls could potentially have composed arguments that were as strong as or stronger than those of trained

participants. However, the arguments of controls were weaker than those of trained participants on average when the SME rated argument persuasiveness “overall,” without reference to the STIM argument syntax and just as a commander might assess an S3’s defense of a tactical recommendation. (Furthermore, the SME was blind to experimental condition and reviewed the responses only after they were transposed into textual form.) In sum, controls in this experiment performed much like the participants in Kuhn's (1991) extensive studies across the life span: given the opportunity to make strong arguments, her informants produced weak ones. The control participants in this study generally ignored much of the supporting evidence, the conflicting evidence, gaps in the given data, and some assumptions even though they were not constrained from recognizing it or reporting it.

The same critique (that the difference in interfaces biased responses) does not directly bear on the difference in the accuracy of conclusions between the experimental groups. Controls and trained officers received effectively the same, minimal support concerning the formation of conclusions: the instruction to both groups to answer the given question. The STIM interface presented trained participants with a blank node labeled “conclusion,” but this cannot reasonably be viewed as a support tool. Nor did trained officers receive any direct instruction concerning formulating reasonable or accurate tactical recommendations. Despite this equality of treatment regarding conclusions, trained participants were much more likely to reach a reasonable conclusion than controls.

It is also inappropriate apply the critique to the results concerning information filtering and production. The trained group outperformed controls with respect to communications measures. However, the STIM interface did not support information filtering and production, and so differences in interfaces used during testing probably did not contribute to differences in communications performance. The effects on communication appear to be side effects of STIM, like the side-effect on the accuracy of conclusions. They were intended but not directly addressed in training.

This said, larger, future studies involving STIM should be designed to neutralize this objection to the validity of data concerning argument persuasiveness and structure. Such a design would employ within-subjects comparisons of the performance of trained participants using the graphical editor on some breaks of a much longer scenario, and a simple text editor on others. We predict that the effects of STIM training on arguments generated with and without the graph editor would be equivalent or proportional. Such an outcome would provide further justification for the use of the graph editor in situations where the Army

desires the benefits of the graph editor, namely support for automated, real-time assessment and feedback (see the section concerning future development of STIM). When those functions are not needed, students need not use the graph editor for testing, and the current SME rating procedures could be applied.

As noted above, STIM training did not define or address accuracy in tactical recommendations. How then, did there emerge a trend for participants in the training condition arrive at better conclusions than controls? It is likely that, in studying how to construct better arguments, the trained group learned to think through tactical problems more thoroughly, and thus they reached better conclusions. This claim is consistent with the model of adaptive decision making defined above. It is also supported by the data. There was a strong correlation between the persuasiveness of arguments and the accuracy of conclusions (Pearson $r = 0.751$, $p < .01$). Though neither the direction nor the source of causality can be firmly established from a correlation, the simplest interpretation is that training targeted at critical thinking skills helped participants to critique the possible conclusions and make better selections from among them. In addition, the structural analysis indicated that trained participants reported more of the evidence in their arguments, and this suggests that they considered more evidence than did controls. Training may have helped them to think about problems more thoroughly, and this may have led to more accurate conclusions. In sum, there is reason to believe that STIM training in critical thinking skills, in particular, may help staff officers to make better tactical decisions.

STIM appears to have improved information filtering behaviors. As predicted, trained participants attended more to the content of messages and less to their source. This training may help students focus on message content, not surface features of messages.

Several effects on information production were detected. Their combined effect suggested that trained participants maintained a quieter network (that is, they generated fewer messages, and thus dampened rather than boosted the overall load of message traffic), that the messages they did send more often reflected thoughtful data interpretation than simple forwarding of data, that they made and acted on inferences concerning the information needs of others, and that they avoided making what may have been unnecessary requests for action. The latter finding can be interpreted as an indicator of implicit coordination under the interpretation that trained participants pass information to other staff, infer that the information will trigger needed actions, and thus do not make unnecessary requests for those actions. This strategy can be highly efficient and effective

under high information load[15]. Most of these effects were statistically reliable.

Analyses of the TLX workload measures indicated that STIM lowered frustration with the task of performing the S3 duties in these scenarios. However, there is a hint in these data that the demands of using the STIM interface may vary considerably between users, suggesting the need for better interface training and improvements to the interface. The most impressive aspect of the workload data is that self-assessed performance did not reliably degrade with training, as might have been expected if the training conflicted with habitual decision-making processes of these expert participants, as is often observed in training studies (Lajoie, 1986).

The strength of the results overall is surprising for several reasons. First, the participants were expert with respect to staff duties and to the scenario used in testing. There might well have been no room for improvement. However, training in generic critical thinking skills may have helped participants leverage their domain-specific knowledge. Second, training lasted less than two hours, yet it produced marked effects on performance in an area in which participants were relatively expert. Third, the interface used in the experiment was domain-independent. It could have been used to test decision making in medicine, financial analysis, or legal reasoning (though the test scenarios, obviously, could not). The interface did not resemble Force XXI staff tools, nor did it provide support for tactical decision making. Yet, participants using the interface made better tactical decisions and communicated tactical information more efficiently. It is intriguing to consider the impact of modifying the interface to resemble familiar communications and decision making tools such as the All Source Analysis System (ASAS) Remote Workstation, Maneuver Control System (MCS), Applique, or their successors. Fourth, participants received very little personal feedback concerning their performance during training. Feedback might have benefited the lowest scorers most, thus reducing variability among trained participants overall and increasing statistical reliability of effects.

In sum, data from this small and preliminary study indicate that STIM training and the STIM interface may improve decision accuracy, decision making, and communications, even with a highly experienced sample of subjects. The measures used here were

[15]Tests of the efficiency and effectiveness of communications in future studies should consider the importance of messages, perhaps as rated by a subject matter expert, relative to participants' ratings of importance and their handling of messages.

responsive to the training manipulation, indicating content validity. Participants were generally enthusiastic about STIM.

CONCEPTS FOR FURTHER DEVELOPMENT OF STIM

Though results of the pilot test were generally positive, there are a number of ways in which STIM can be improved. In general, we are interested in developing a more automated staff training system, one that reflects the technology and needs of Force XXI staff, and one that is available to physically dispersed students across an internet.

Training

The experimental training focused on team coordination (using assessment updates) and decision-making skills. STIM training might be enhanced by addressing other aspects of coordination and critical thinking, or by training the other skills specified in the adaptive team performance model: team restructuring and tool modification. Such training can be somewhat generic in character, or highly specified to staff positions. A few examples follow.

Coordination

In previous research, for example, Serfaty and colleagues (Entin, Serfaty, & Deckert, 1994) have demonstrated explicitly that instructing staff to push information to line officers and others (rather than await requests for information) improves communications skills. This is a promising avenue.

An indirect benefit of the training tested here and training evaluated by Serfaty and colleagues (Entin, Serfaty, & Deckert, 1994) was that officers were less likely to request actions that should be performed automatically. That is, they did not make unnecessary, action-oriented communications. Explicit instruction on this point may be helpful.

Computerized white boards may be an integral component of the Force XXI information technology suite (Schatz, 1996). If they are, then staff may benefit from training in strategies for effective white-board briefs and assessment updates.

Teams may benefit from training in detecting idle periods and using them to plan team responses to anticipated events.

Decision Making

In research with Navy and Army command staff, the authors have found that experienced staff officers consider a variety of interesting, but domain-specific issues during decision making

(Cohen, Freeman, et al., 1995; Cohen, Freeman, & Thompson, in press; Cohen, Freeman, & Wolf, 1996; Freeman & Cohen, 1996; Cohen, Freeman, & Thompson, 1997). For example, an S2 analyzing intelligence data may benefit by considering (a) the accuracy of the initial observation, (b) the honesty and accuracy of the reporting source, (c) the reliability of the communications link(s) from the source, and (c) the validity of the analysis of the data by the source or subsequent processors. Junior staff officers may benefit by explicit instruction concerning frameworks for critiquing intelligence, assessments of enemy intent, friendly plans, and other tactical matters.

Reports of the AWE suggest a number of areas in which decision-making instruction might be customized to the Force XXI environment[16]. One example is that the S2 might benefit from explicit training in balancing battle tracking with intelligence analysis and production (Bruce Sterling, personal communication, April, 1997).

Team Restructuring

Overall team performance might be enhanced with training that emphasizes how to recognize information overload among fellow staff and how to ameliorate the problem by reallocating burdensome tasks to subordinates or fellow staff (e.g., offloading selected tasks from the Battle Captain to the S3).

Tool Selection and Parameterization

AWE reports indicate that the S3 might benefit from training in methods of quickly composing consolidated graphics of the tactical situation using Force XXI data. This data fusion task is apparently not directly supported by current Force XXI technology (Bruce Sterling, personal communication, April, 1997).

The Battle Captain might receive training in strategies for using (or, in select cases, avoiding) the complexly formatted Applique message system (Bruce Sterling, personal communication, April, 1997).

In recent research, Cohen, Parasuraman, Serfaty, & Andes (1997) have proposed that knowledge of the strengths and shortcomings of a decision-support system may enable Army helicopter pilots to better discern when to rely on these systems and how much trust to vest in their output. Force XXI staff might benefit from instruction of this sort (specific to the decision

[16]Instruction concerning issues specific to Force XXI battalion staff would require field studies and cognitive task analyses, a task we have proposed for Phase II research and development.

aids Army staff may use). Similarly, staff might value training concerning the extent to which various data display modes help or hinder reasoning about specific types of problems.

Emulation of Force XXI Technology

It might also be useful to design STIM interfaces that emulate specific Force XXI technology. In the pilot experiment conducted in Phase I, training was delivered by an instructor, and practice and test scenarios were delivered on a simple interface consisting of a generic e-mail application and an application for drawing and annotating node-link graphs. This strategy allowed us to flexibly develop and test training and measurement instruments applicable to a range of staff positions. However, the face validity of training, retention of instruction, and transfer effects might be enhanced by presenting demonstration, practice and test scenarios using Force XXI interfaces. Particularly good candidates for this are the interfaces for the core staff team: the Maneuver Control System (MCS) interface for the S3, the All Source Analysis System (ASAS) Remote Workstation interface for the S2, and Applique (or its successor) for the Battle Captain. Interface emulation would be a modest but important step towards embedded training. It might, in fact, be more valuable than an embedded system because STIM could be delivered on virtually any personal computer or workstation attached to the internet.

Instructional Strategy

The Phase I research concerning instructional strategy addressed several topics: performance assessment, feedback, and system adaptation. Here we describe methods of automating many of the measures used above, describe other measures of interest, and address automated feedback and adaptation concepts.

Automated Assessment

Automated Assessment of Decision making

The measures of decision making evaluated in this Phase I project used data concerning the structure and content of arguments. It would be relatively simple to automate the measure of conclusion accuracy employed in this study by requiring users to choose conclusions from among a menu of options, or assemble them using a constrained, possibly menu-based vocabulary. The measure of argument structure could be automated simply by developing software that tallies the number of graph nodes (or argument components) of each type and computes a score that is a weighted sum of the total number of nodes used (i.e., the total amount of evidence cited) and the number of nodes of each type used (i.e., the variety of classes of evidence used).

The problem of automating the evaluation of argument persuasiveness requires a more complex solution. Argument persuasiveness was graded manually in the pilot study by an SME. This was a laborious process, as is the rule with SME rating. STIM could break this rule. It could automate qualitative SME grading. Our approach capitalizes on the structure or syntax of arguments, described above. Syntax powerfully constrains meaning, so powerfully that it makes it possible to automate the analysis of the textual content of argument. Specifically, STIM could incorporate a hybrid engine capable of matching student arguments (or responses) to SME graded arguments, and returning grades for the persuasiveness of the argument, argument components (such as individual pieces of supporting evidence) and sub-arguments (chains of argument components such as conflicting evidence, deconflicting assumption and action). The engine would wed statistical algorithms for encoding text with an inferential neural net (INN, a class of artificial neural net) capable of recognizing approximate matches between encoded student responses to previously observed, graded responses. While this is a sophisticated approach, it is not conjectural. CTI has previously applied this technology to indexing and retrieving briefing documents in a related Army training system for ARI, Ft. Leavenworth (Cohen, Thompson, et al., 1995). Below, we describe the two parts of a hybrid assessment engine in detail. These parts are a statistical text classification system and an INN pattern-matching and grading system.

Statistical text analysis. STIM could encode (or classify) the text of arguments using factor analysis or principle components analysis (PCA). PCA is typically used by statisticians to reduce a large number of observed variables to a smaller number of abstract factors. The input is a matrix of cases (such as subjects) by variables (such as scores on test questions). The output is a relatively small set of principle components or factors (the term we will use to avoid confusion with argument components) whose presence or influence in each variable is represented by a coefficient. When applied to texts, PCA is often known as Latent Semantic Indexing (LSI) or Latent Semantic Analysis (LSA) (Landauer, Foltz, & Laham, 1997). In this context, LSI is a technique for representing the conceptual content of texts. LSI builds a matrix that crosses documents with the terms they contain. This large, sparse matrix is then reduced using singular value decomposition (SVD) to obtain an optimal, lower-rank approximation of the matrix.

PCA would be used in two phases in STIM. During construction of STIM, PCA would be applied to a large body of argument components elicited in pilot testing. This would produce a set of PCA factors. Because argument components of different types (e.g., supporting evidence, conflicting evidence, actions) would be submitted for analysis separately, the factor lists would

effectively be customized for each component type, making them sensitive both to the structural role of argument components, as well as to content. During its use as a training platform, the second phase of statistical analysis, STIM would compute the weights of the PCA factors (derived in stage one) on the text of each argument component submitted in a response. Thus, it would essentially impose a common coding scheme on arguments that constitute the network set and those elicited from students (see Figure 16).

Inferential neural nets. The problem of grading (PCA-encoded) arguments is essentially one of matching new arguments to known, previously graded ones. Inferential neural networks are an ideal tool for this task. Traditional connectionist models excel at identifying stimuli that only roughly approximate known patterns. Inferential neural networks (INN) add systematicity to this capacity for soft-matching (c.f., Shastri & Ajjanagadde, 1993). By systematicity, we mean that an INN represents structural aspects of data. The notion of inter-linked argument components, presented above, is precisely the type of structure that can be represented in an INN. Systematicity enables an INN to identify matches of structure and content between networks of prior, graded responses and newly input student responses.

An INN, like PCA, would be applied in two stages. To build the INN, a representation of arguments elicited in pilot testing would be constructed consisting of a predicate representing the name of an argument component (e.g., supporting evidence), the PCA factor weightings representing its textual content, pointers to argument components linked to it (e.g., the conclusion and actions), and SME ratings of the value of the argument component, the substructure of which it is a part and the overall argument. All of the components of a given argument would be submitted to a version of the INN in a linked data structure until all arguments in the data set were entered. The resulting compilation of graded argument structures would constitute the argument rating engine.

During staff training with STIM, the INN would receive PCA-encoded student arguments as input. It would attempt to match each argument to all or parts of prior, scored arguments. For each argument component or sub-structure that an officer generated to defend a conclusion, and which was recognized by the network at some threshold of similarity, SME scores would be read directly off the network. For argument components or sub-structures that the student failed to cite, and which were highly rated by the SME for the given conclusion, the INN would generate a code representing the missing component and a score for the omission. Truly novel responses, which the system couldn't judge as sufficiently similar to any prior, known argument, would be archived for later analysis by an SME. We anticipate that most responses could be analyzed by the engine in real time. To

support feedback, the INN might also be used to retrieve (a) the best known response for the given conclusion or (b) the known response with the closest match and the highest rating. This would enable the student to review dramatically or incrementally better solutions to the problem. In sum, the INN would function as an SME with the ability to recognize and retrieve ratings for the concepts and structures of student arguments, as well as examples of better responses (see Figure 16).

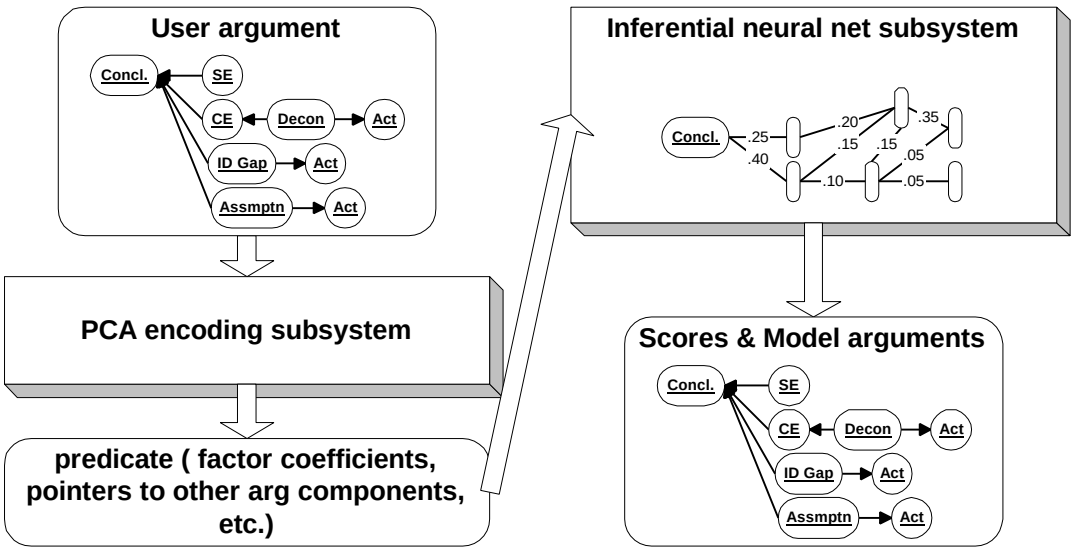

Figure 16. The hybrid argument assessment engine.

Ratings of the overall argument and argument substructures would not be redundant with the measure of structure suggested earlier, which considers the variety of argument components. The latter is a general measure of the mastery of specific critical thinking skills. The INN effectively scores critical thinking skills in the context of a specific conclusion. This specificity of context means that INN scores do not support general inferences about an officer's cognitive skills. However, the INN scores can help researchers identify interactions of the test problems with training and aptitude, and these scores could help the system provide feedback in the form of concrete examples that the officer may be able to interpret easily.

There are, of course, simpler approaches to interpreting the content of responses, but they impose unsatisfying constraints on the ways in which users can express themselves. The simplest approach is to restrict arguments to multiple choice selections. A related approach is to allow users to compose arguments from pre-graded lists of material, such as incoming messages or text from orders, estimates, and other database material. The second of these approaches is potentially quite useful to staff because it minimizes the labor required to weave extant material into an

argument as, for example, supporting or conflicting evidence. However, it is too artificial to appeal to the likely student body, we suspect, and constrains responses so severely that measures of their persuasiveness might be of very low validity.

We find the hybrid approach to be not only potentially powerful, but intellectually intriguing. The INN's soft-matching of new student text to prior responses can be thought of as generalization from learned examples. The INN would generalize in several interesting ways. It would generalize across training and test problems by applying what it learns concerning one problem to interpreting responses to another problem. It would generalize across arguments for a given conclusion to a problem such that similar arguments receive similar scores. It would also generalize across textual expressions of a given concept in an argument component. The INN would perform this complex pattern-matching activity in parallel, which would ensure rapid feedback to students and timely adaptation of training and tests. New methods of leveraging or limiting this capacity generalization could be explored in future research.

In sum, STIM would measure the quality of decision-making processes with ratings derived by matching student responses against prior responses rated by SMEs. This approach relies on advanced statistical and neural processing algorithms that CTI is currently applying in other projects.

Automated Assessment of Communications

Some of the communications measures defined in this paper can be readily automated because (a) the required data are generated naturally during electronic communication (e.g., such data might include the recipient(s) of a message, which the student must indicate when generating a new message, or the time of transmission of a new message, which the email system automatically indicates for each newly transmitted message); (b) the data can be elicited with only minor intrusion into the natural workflow (e.g., ratings of the importance of incoming messages can be gathered by requiring students to rate messages after reading them); or (c) the data are known at the time of scenario design (e.g., the rank of the author of a scripted message or the SME's rating of the importance of the message are specified during scenario design and need not be gathered from students during scenario runs).

However, some aspects of outgoing (IP) messages require content analysis. These include categorization of messages by type of communication (information/status vs. action/plan), class (request, initiate, and respond), and level of processing (pass-through, form judgement, and solve problem). There are several ways to perform this type of categorization in STIM. First, a PCA

engine mated to a simple artificial neural net might categorize messages officers generate. Because the categorization scheme would be rough, the accuracy of the pattern-matching ANN engine could be quite high. Second, students could be trained to classify messages as they transmit them. This approach might have instructional value in that it makes officers aware of distinctions between forwarding, judgement, and problem solving. Officers may need to make such distinctions in order to adapt to changes in workload or to the management style of their commanding officer. Finally, it may be most efficient simply to provide an SME with a rating form, such as the one used to code data in this study, training, and analysis software with which to code messages as they are produced and to analyze them at breaks for use in an After Action Review.

Measures of expertise. The expertise of staff is potentially an important predictor of training effects. As demonstrated in Cohen, Freeman, & Wolf (1996), some staff may benefit more than others from training as a function of their military tenure, prior military training, or battlefield experience. Some may require more extended or more elaborated instruction in some aspects of information management training. Staff in specialized positions may require specialized instruction or access to particular reference materials during training. To discern these needs, STIM would request biographical information concerning students by presenting on-screen, biographical questionnaires. Information concerning the user's training goals might also be of value. Initially, these data could be used to test the effects of instruction at different levels of experience. Potentially, the data could be used to adapt training and testing to individual differences.

Measures of user satisfaction. Users of STIM may have strong opinions and useful comments concerning training concepts, scenarios, and the system interface. These can be gathered on-line for manual, qualitative assessment by trainers and researchers. Comments concerning interface problems might be validated by examining the context in which students use help and "undo" features. Such keystroke level data might be particularly helpful during the evaluation of new STIM modules.

Feedback

The potential strength of STIM's automated performance assessment subsystems present opportunities for implementing sophisticated feedback. However, any strategy for presenting feedback must consider several issues: what feedback will be presented, when will it be issued, and in what form.

In training cognitive skills, strong effects have been found for feedback that flags errors (but does not explain them), and

is presented immediately upon commission of the error (Corbett & Anderson, 1990; Anderson, 1992). However, this approach may be more appropriate for training procedural skills, such as LISP programming or constructing geometric proofs, than for training strategies for critical thinking and communications in complex scenarios. In this context, it may be beneficial to present students with their own work, an example of relevant expert work to which to compare their efforts, and a score, critique or guiding principles with which to improve future performance. Such strategies can be highly effective, as evidenced in Bangert-Drowns's (1991) meta-analysis of feedback in 40 studies, in which the author found that providing answers or explaining answers was more effective than simply flagging errors.

Precisely how would this type of feedback be implemented in STIM? Feedback concerning communications strategies, could be presented to students periodically, perhaps at breaks or in an After Action Review, rather than immediately upon commission of individual errors. This would be done in part to preserve the flow of practice and test scenarios, and in part because many of the communications measures must be computed over multiple messages, rather than in response to a single message. The form of feedback might be an overall performance score on the measure, a target score specified by an SME, examples of messages that raised the score and those that lowered it, and a canned principle or rule to guide the student in the future. This feedback might be presented as text. However, there may be opportunities for graphical feedback. For example, feedback concerning patterns of message traffic within staff, to subordinates, and superiors could be readily represented as a network with density of traffic denoted by the thickness of arcs. Histograms might be used to display comparisons of student scores and target scores.

Feedback concerning decision accuracy might be presented simply as a list of possible conclusions concerning a given break question. The student's choice from among the list would be highlighted and annotated with a brief, canned SME critique.

Feedback concerning decision-making processes might take the following complex but instructive form, or any simplification of it. After evaluating a student's argument, the INN would immediately, or in an After Action Review, present the officer with:

- Their own graphical argument annotated with scores for argument components, argument substructures (consisting of linked components), and the overall argument;
- A graph of the best known response for the given conclusion and its scores;

- A graph of an argument that is highly similar to the student's and highly rated, with its scores; and
- Highlighting or arrows on the graphs indicating evidence missing from the student's argument.

By providing students with scored arguments to which to compare their own work, we give them concrete examples to model in future responses. The best response for a given conclusion may differ radically from the student's, and this may elicit insight at best or confusion at worst. The response that is most similar to the student's and most highly rated may be more accessible to the student but less informative. These are interesting tradeoffs in feedback that might be explored in future research. It may also be possible to have SMEs label the PCA factors that most commonly appear in arguments. Such labels may be useful in retrieving canned critiques of student argument components and substructures. This, too, presents interesting research opportunities.

Adaptation

There are three areas in which STIM might adapt to individual students or to teams: instruction, practice scenarios, and tests. The simplest form of instructional adaptation is for STIM to allow users to simply replay instruction and demonstration. This is a strategy that we recommend. We focus here on concepts for adapting practice and test scenarios and offering additional practice scenarios to deficient teams.

The difficulty of practice and test scenarios might be adapted in several ways. The system could increase difficulty by boosting the number of messages per unit time or the variance in the number of messages per unit time. The former manipulation would help officers to select and practice a performance strategy appropriate to a static workload; the latter would test their ability to shift strategies as workload changes. The system would increase the number of messages it issues by increasing the number of messages it draws from the pool of optional messages. Alternatively, it may be appropriate to break messages (such as large spot reports or status reports) into discrete, independent messages.

Scenario difficulty could also be altered by manipulating qualitative aspects of the message stream. Some core messages might be written in several versions, each designed to introduce more or different forms of uncertainty into the scenario, or the optional message pool could be seeded with messages that invoke uncertainty. The content of such messages might conflict with the current situation assessment or sitmap or bias staff to make unwarranted assumptions. Removing specific messages might

introduce information gaps. The ability of staff to detect and deal with these gaps, unreliable assumptions, and conflict could be directly measured using the analysis of argument structure described above.

Other manipulations of the scenario might also boost the difficulty of tests and practice. "Noise" might be introduced in the form of mis-delivered messages (messages addressed to the wrong staff officer or the incorrect rank level), requests for low-priority, administrative support, or brief equipment or communications failures requiring officers to repeat previously completed tasks. Equally interesting is the prospect of altering the context in which messages are interpreted by manipulating the accuracy of briefing materials, degrading the quality of assessment updates issued by the system in the name of the CO or XO, or changing force ratios in the field.

Many of these adaptations could be made for an individual user, independently of other users. (For example, the number or variance in the number of messages per unit time could be adapted for an individual.) However, most could be administered to the overall team, as well.

The trigger conditions under which STIM would adapt training and test scenarios would be relatively simple. Those who perform well on the measures described above might find scenarios becoming more difficult as they execute them. Those who do not would find scenarios becoming simpler. In addition, it may be desirable to allow students to select the level of difficulty at which they wish to train and test. There are situations under which it is not advisable to adapt test scenarios to the user. For example, if test results are used to compare the performance of teams, then all teams must test on identical scenarios and no adaptation should be allowed. When this is not an issue, however, adaptation of practice and test scenarios may aid learning and retention by providing an appropriate challenge, rather than one that is too formidable or too simple.

These adaptation strategies will increase the work of scenario designers. They would require designers to write scenarios to the maximum level of difficulty, and parameterize individual messages to indicate which are appropriate for lower levels of difficulty (e.g., easy, medium and hard) or each type of challenge (e.g., increased conflict in the message stream or diminished completeness of data). However, scenarios that can be automatically adapted can also be recycled, that is, presented repeatedly in modified forms to the same students. This benefit of developing fewer scenarios may compensate for the cost of more complex scenario design.

System Design Concepts

In the pilot study, we evaluated key components of the STIM training and interface. Some of those components--such as the training and analysis of measures--were implemented manually. Our concept for a full STIM prototype differs considerably from this. We envision an inter-networked training system that presents small staff teams with multi-media training, fully automated practice and test scenarios, and automated assessment and feedback. Though the staff CO or XO might provide additional instruction or feedback (by applying lessons learned from a train-the-trainer package[17]), the emphasis here is on automation[18]. Internet delivery could facilitate distance learning by geographically distributed groups, or opportunistic training by non-distributed groups without sacrificing the benefits of centralized maintenance of databases and system code.

We do not attempt a detailed architectural description of STIM here. However, the basic modules of STIM could be these:

- Scenario databases--Contains scenario message streams, sitmap data, briefing materials, and other data.
- Instructional databases--Contains multimedia training material, such as textual instructions, animated lesson illustrations, audio clips to accompany animated material or video clips.
- Workstation interface manager--Formats scenario material for presentation on emulated Force XXI interfaces (such as MCS for S3, ASAS Remote Workstation for S2, and Applique for Battle Captain). Formats training material and other material (such as performance feedback and scenario break response screens) for display in a generic interface common to all trainee workstations. Captures user actions, such as menu or window selections, manipulation of map or diagram objects, and textual input. Forwards selected user actions to the server interface manager.

---

[17]A train-the-trainer package might describe STIM's instructional objectives, the practice and test scenarios, provide model responses to key questions and indicate how and when to apply remedial training.

[18]For example, a white cell (an SME with scripts for responding in the role of missing players during scenarios) often enhances the realism and seeming dynamism of scenarios. However, provision of a white cell complicates training and reduces the opportunities to make it available on demand. It may or may not not be worthwhile. This, however, must be evaluated in future research.

- Server interface manager--Forwards input concerning user actions to scenario manager, instruction manager, or performance measurement manager. Coordinates the presentation of material from the instruction and scenario managers to the workstation interface manager.
- Instruction manager--Retrieves instructional material from the instructional databases and presents it to the server interface manager.
- Scenario manager--Retrieves scenario messages, sitmaps and other material from the scenario databases and formats it for presentation to the server interface manager.
- Performance measurement manager--Processes user actions relayed by the server interface manager and passes encoded output to the assessment engine.
- Assessment engine--Analyzes data from the performance measurement manager concerning communications strategies, decision accuracy, decision processes, and other skills. Independent assessment sub-engines process data concerning each skill.
- Feedback manager--Formats output from assessment engine and passes the result to the server interface manager. Maintains an archive of assessment results and feedback.
- Test manager--Administers test and debugging scripts that verify the integrity of modules and the interfaces between them.

STIM would be developed using a web-based client/server model. The client side application would contain all functionality relating to the workstation interface management. Other modules, which control the sequencing of training materials, scenario administration, the analysis of trainee responses, provision of feedback, and so forth, would reside on the server.

The client-based workstation interface manager could be developed as a Java application and would interact with the servers via Hyper-Text Markup Language (HTML) and perhaps other Transmission Control Protocol/Internet Protocol (TCP/IP). Most of the server modules could be built using NeXTSTEP® (recently renamed OpenStep®) with the WebObjects®[19] code library, a development environment that provides a very flexible basis for constructing dynamic web-based information servers. This environment is available on many platforms commonly used by the

[19]NeXTSTEP, OpenStep and WebObjects are registered trademarks of Apple Computer.

Army, including Intel Pentium-based systems, and computers by Sun, Hewlett Packard and, soon, Apple. The server-side assessment sub-engine responsible for analyzing the text could be built using available algorithms for the computation of principle components analysis (PCA) for large, sparse matrices and the code base for an Inferential Neural Net called SHRUTI (Shastri & Ajjanagadde, 1993) which CTI is currently applying in other research projects for the Office of Naval Research (Thompson, Cohen & Freeman, 1995).

The training might be presented in multi-media training, consisting of "slides" augmented with audio, and possibly motivational or instructional video clips featuring experienced officers. Additional training content might be developed to address specific problems in Force XXI staff operations at the brigade level and below, based on field research and cognitive task analyses (two significant research needs we have not addressed here). Practice and test scenarios could be adapted, as they were in this phase of research, from the Staff Group Trainer. The medium of message presentation would be improved. STIM could deliver scenario messages on interfaces that emulate Force XXI technology such as ASAS, MCS, or Applique (or their successors)[20]. The graphical argument construction utility would be retained and used at scenario breaks, as it was in the pilot study, to gather data concerning argument structure and content. Selected break questions might also be administered without the graphical tool, but with a simple text editor or the Force XXI emulators, in order to test the transfer of argument construction skills from the highly supportive STIM system to the field environment.

CONCLUSIONS

In Phase I of the STIM project the research team 1) developed scenario-based training in information management for staff officers, 2) conceived instructional strategies and performance measures that lend themselves to automation, 3) conducted a pilot study of key training, interface components and performance measures, and 4) developed concepts for the software and hardware architecture of STIM.

Results of the pilot test suggest that the STIM training system may help improve information management skills. The tactical judgements of trained participants were more accurate than those of controls by 34%, more persuasive by 93%, and trained participants tended to take actions that were more reasonable. Trained participants also were more cognizant of gaps

[20]The use of these highly structured interfaces may provide new opportunities for measuring performance.

in their knowledge, assumptions, and conflicting evidence than were controls. While some of these results were suggestive trends, rather than conventionally significant, they were in the direction predicted from theoretical models and their size was large. The training did not directly address information filtering or production, key issues in resolving information overload, but, as predicted, it improved performance in those areas. Training enabled participants to evaluate incoming messages less on the rank of the sender and more, apparently, on the content of the messages. Trained participants processed incoming data more thoroughly before generating messages, produced fewer messages overall, more often engaged in information pushing, and made fewer unnecessary requests of the virtual staff to whom they sent messages. In short, the pilot data indicate that STIM may improve tactical decisions, tactical reasoning processes, and team communication.

There remain a number of challenges in this line of research and development. The principle challenge is to go beyond generic digital interfaces and the training tested here to explore individual, team, and human-computer interface problems specific to the Force XXI digital environment. This will involve field studies and cognitive task analyses, a program of research that we have not discussed here. It will also require development of interfaces that emulate Force XXI equipment on which to present practice and test scenarios. These tasks are planned for a proposed Phase II effort.

We have also found weaknesses in the training that we tested in Phase I. Trained participants acknowledged conflicting evidence, but did so rarely, and they had great difficulty grasping the notion that arguments can be deconflicted (by making assumptions or assertions that at least temporarily explain the conflict). In addition, the notion of linking argument components in graphs was not well understood. This must be remedied to improve the potential for accurate, automated argument assessment.

The measures of decision-making processes used here were revealing but also intrusive. They require that a scenario be halted in mid-run while trainees respond to tactical queries using a very unusual interface: a graphical argument construction kit. This tradeoff seems worthwhile because it supports a potentially powerful measurement instrument and useful feedback, and because participants in this experiment largely endorsed it. However, it may be possible to devise less intrusive means of eliciting responses in a structured format, perhaps by issuing a stream messages in the form of questions designed to elicit responses (such as lists of supporting, conflicting evidence, assumptions, or actions) that are equivalent to specific argument components. Users would respond in free text. At the least,

students should be given practice generating persuasive arguments both with and without the graph editor. These issues should be explored.

This experiment did not test the notion, central to the adaptive team process model and the decision-making model, that teams adapt to changes in workload. Workload was not manipulated in this scenario. Future research should explore the interactions between training effects and varied workload.

Neither did the experiment attempt to discriminate between the effects of training and the effects of the STIM interface, or the effects of the main components of the training: assessment updates and critical thinking (which, itself has several components). These should be explored, as should the effects of training on communications efficiency and effectiveness relative to the objective judgements of SMEs.

Though the pilot data concerning STIM is only preliminary, it is very encouraging. STIM appears to improve decision accuracy, decision-making processes, information filtering, and information production. The measures used here exhibit construct validity and most can be fully automated to drive feedback and to adapt training and testing to the individual user or the team. In sum, STIM training and training support software are promising tools for training and evaluating the information management skills of Army staff.

REFERENCES

Anderson, J.R. (1992). General principles for an intelligent tutoring architecture. In J.W. Regian & V.J. Shute (Eds.), Cognitive approaches to automated instruction. Hillsdale, NJ: Erlbaum.

Bangert-Drowns, et al. (1991, Summer). The instructional effect of feedback in test-like events. Review of Educational Research, 2, 213-238.

BDM Federal, Inc. (1996). Commander/Staff Trainer (C/ST) project design. Ft. Knox, KY: Author.

Cannon-Bowers, J.A., Salas, E., & Converse, S. (1990). Cognitive psychology and team training: Training shared mental models of complex systems. Human Factors Society Bulletin, 33 (12), 1-4.

CECOM. (1997, September 15). What is Force XXI? [On-line]. http://www.monmouth.army.mil/cecom/lrc/exfor/general.html#general

Cohen, M.S., Freeman, J.T., & Thompson, B.T. (in press). Critical thinking skills in tactical decision making: A model and a training method. In J. Canon-Bowers, & E. Salas (Eds.), Decision-making under stress: Implications for training and Simulation. Washington, DC: American Psychological Association.

Cohen, M.S., Freeman, J.T., Marvin, F.F., Bresnick, T.A., Adelman, L., & Tolcott, M.A. (1995). Training metacognitive skills to enhance situation assessment in the battlefield (Technical Report 95-1). Arlington, VA: Cognitive Technologies, Inc.

Cohen, M.S., Thompson, B.B., Adelman, L., Bresnick, T.A., Tolcott, M.A., & Freeman, J.T. (1995). Rapid capturing of battlefield mental models (Technical Report 95-3). Arlington, VA: Cognitive Technologies, Inc.

Cohen, M.S., Freeman, J.T., & Thompson, B.T. (1997). Integrated critical thinking training and decision support for tactical anti-air warfare. Proceedings of the 1997 Command and Control Research and Technology Symposium, Washington, DC.

Cohen, M.S., Freeman, J.T., & Wolf, S. (1996). Meta-recognition in time-stressed decision making: Recognizing, critiquing, and correcting. Journal of the Human Factors and Ergonomics Society.

Cohen, M.S., Parasuraman, R., Serfaty, D., & Andes, R.C. (1997). Trust in decision aids: A model and a training strategy (Technical Report USAATCOM TR 97-D-4). Fort Eustis, VA: Aviation Applied Technology Directorate: Aviation Research, Development & Engineering Center (ATCOM).

Corbett, A.T., & Anderson, J.R. (1990). The effect of feedback control on learning to program with the Lisp tutor. Proceedings of the 12th Cognitive Science Conference, Cambridge, MA.

Entin, E.E., Serfaty, D., & Deckert, J.C. (1994). Team adaptation and coordination training. Burlington, MA: AlphaTech, Inc.

Freeman, J.T., & Cohen, M.S. (1996). Training for complex decision-making: A test of instruction based on the recognition/metacognition model. Proceedings of the 1996 Command and Control Research and Technology Symposium, Monterey, CA.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), Human mental workload. Amsterdam: North-Holland.

Kuhn, D. (1991). The skills of argument. New York: Cambridge University Press.

Kuhn, D. (1992). Thinking as argument. Harvard Educational Review, 62(2), 155-178.

Lajoie, S.P. (1986). Individual difference in spatial ability: A computerized tutor for orthographic projection tasks. Unpublished doctoral dissertation, Stanford University.

Landauer, T. K., Foltz, P. W., & Laham, D. (1997, September 17). Introduction to latent semantic analysis. [On-line]. Available: http://samiam.colorado.edu/~lsi/Home.html.

Naylor, S.D. (1997, April 28). Mission accomplished. Army Times, 12.

Nisbett, R.E., & Ross, L. (1980). Human inference: Strategies and shortcomings of social judgment. Englewood Cliffs, NJ: Prentice Hall.

Rasmussen, J. (1990). Human error and the problem of causality in analysis of accidents. In D.E. Broadbent, J. Reason, & A. Baddely (Eds.), Human factors in hazardous situations. Oxford: Clarendon Press.

Reid, G.B., & Nygren, T.E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. In P. A. Hancock & N. Meshkati (Eds.), Human mental workload. Amsterdam: North-Holland.

Schatz, J. (1996, April 19). Battlefield Information Transmission System (BITS) and Battlefield Awareness and Data Dissemination (BADD) for Task Force XXI (TF XXI). [On-line]. http://www.monmouth.army.mil/cecom/lrc/forcexxi/lsp/bits.html

Serfaty, D., Entin, E.E., & Deckert, J.C. (1993). Team adaptation to stress in decision making and coordination with implications for CIC team training, Volumes I & II (TR-564). Burlington, MA: AlphaTech, Inc.

Serfaty, D., Entin, E.E., & Volpe C. (1993). Adaptation to stress in team decision-making and coordination. Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting. Santa Monica, CA: Human Factors Society.

Shastri, L., & Ajjanagadde, A. (1993). From simple associations to systematic reasoning. Behavioral and Brain Sciences, 16(3), 417-494.

Siegel, S. (1956). Non parametric statistics for the behavioral sciences. New York: McGraw-Hill.

Terino, J. (1997, May). The shape of the Army's 21st-century soldier. The Retired Officer, 33-36.

Thompson, B.B., Cohen, M.S., & Freeman, J.T. (1995). Metacognitive behavior in adaptive agents. Proceedings of the World Congress on Neural Networks, Washington, DC.

Toulmin, S., Rieke, & Janik. (1984). An introduction to reasoning. New York: MacMillan.

van Creveld, M. (1985). Command in war. Cambridge, MA: Harvard University Press.

Wilson, G.C. (1997, May 28). EXFOR performance comes under scrutiny. Army Times, 3.

Wilson, J.E. (1997, June). The information age Army. Army, 14-22.

APPENDIX A. BRIGADE COMMANDER'S GUIDANCE

As you all know, the Krasnovian forces launched a full scale attack into Mojave with the 19th Combined Arms Army (CAA), followed by the 16th CAA. The 19th CAA's attack was a supporting attack for the Krasnovian main attack by the KERN Front on its right (south) flank. Our Division, the 55th Infantry Division (Mech), defeated their advance guards, and occupied the objectives as outlined in the Division OPORD. Bde 21 engaged elements of the 231st Motorized Rifle Division (MRD) which are estimated to have suffered 30% losses in both men and fighting vehicles. The 231st MRD has established a typical defense in contact with the enemy and its regimental second echelon forces have halted behind their lead elements except for elements of the BTR-equipped 218th Motorized Rifle Regiment (MRR) which continue to move through the difficult terrain of their mountainous zone at a slow rate. All indications are that enemy forces throughout the Corps sector are preparing for a return to offensive operations; an attack by the 231st MRD against our Brigade from current positions can be expected within the next 24 hours. The 231st main effort is expected along Phase Line Davis, but because the Front's main effort is south of our sector, Frontal aviation and artillery assets will probably be committed elsewhere.

The 55th ID (Mech) Mission is as follows: conduct an area defense (NK 2527 to MJ 6547) from Phase Line (PL) QUINCY to PL HANCOCK (Note: PL HANCOCK is off the map to our south) not later than 170500 March 97 to defeat enemy forces in sector; on order, counterattack. The Division Commander, Major General Johnson intends to defeat the attacking enemy forces in sector by drawing enemy forces into the natural kill zone east of Barstow. The Division will defeat the enemy attack by containing the enemy west of PL HANCOCK; then, attacking the concentrated enemy forces along I-15 with a combination of attack helicopters and local counterattack. The end state of this operation is the clearing of the Division sector out to PL PHOENIX and positioning of forces to continue offensive action.

The mission of our brigade, Brigade 21, is to defend from NK 233256 to MJ 996909 (off the map) NLT 170500 March 97 to defeat attacking Kransnovian forces and prevent penetration of the Division right (north) flank. My intent is as follows:

This Brigade will retain control of the dominant terrain along PL PHOENIX to secure the Division northern flank and guard Division counterattack avenue. I intend to accomplish this by conducting an area defense to defeat the Krasnovian attack into our sector. I will use a Brigade security force along PL DAVIS to win the counter-recon battle. In the center and north of the Brigade Main Battle Area, I am prepared to accept risk to be able

to mass the combat power of up to three task forces against the enemy main effort in the south. The end state for this operation is the destruction of all enemy first echelon formations; defeat of second echelon formations between PL QUINCY and PL PHOENIX; and, the retention of defensible terrain along PL PHOENIX to insure that the western flank is secured. This mission will be conducted in three phases. PHASE I is the security force battle; PHASE II is the structuring of the Brigade's MBA defense; PHASE III is the defeat of the enemy attack.

In PHASE I (Security Force Battle), we will establish a strong security force using TF OUTLAW and B-14 Cav to establish a forward screen along PL DAVIS. The security force will destroy all enemy recon elements and regimental forward security elements forcing enemy lead regiments to deploy out of march formation into attack formation. Additionally, TF FALCON deploys an internal security force to screen Nelson Lake (NK 2020) and destroy enemy recon. The brigade will accept moderate risk in the Brigade rear area.

In PHASE II (Structure MBA), Brigade 21 conducts its main defense with TF FALCON defending in sector in the north, B-14 Cav screening across the center sector, TF OUTLAW and TF SEAHAWK defending to mass combat power against the enemy main effort in the south. TF EAGLE acts as the Brigade counterattack force. As the enemy attack echelon enters our area, the Brigade will counterattack to destroy enemy combat forces and artillery groups while limiting risk to friendly forces.

In PHASE III (Defeat Enemy Attack), TF FALCON attacks to defeat enemy forces in sector, while TF OUTLAW and TF SEAHAWK defend in sector. Upon defeat of enemy attack, on order, TF EAGLE attacks to destroy enemy forces. TF OUTLAW and TF SEAHAWK place fires on lead enemy forces to support TF EAGLE's attack.

APPENDIX B. BATTALION TASK FORCE COMMANDER'S GUIDANCE

The 55th ID (Mech) Mission is as follows: conduct an area defense (NK 2527 to MJ 6547) from Phase Line (PL) QUINCY to PL HANCOCK (Note: PL HANCOCK is off the map to our south) not later than 170500 March 97 to defeat enemy forces in sector; on order, counterattack. The mission of our brigade, Brigade 21, is to defend from NK 233256 to MJ 996909 (off the map) NLT 170500 March 97 to defeat attacking Kransnovian forces and prevent penetration of the Division right (north) flank. The Brigade Commander's concept is as follows:

The Brigade will retain control of the dominant terrain along PL PHOENIX to secure the Division northern flank and guard Division counterattack avenue. The end state for this operation is the destruction of all enemy first echelon formations; defeat of second echelon formations between PL QUINCY and PL PHOENIX; and, the retention of defensible terrain along PL PHOENIX to insure that the western flank is secured. This mission will be conducted in three phases. PHASE I is the security force battle; PHASE II is the structuring of the Brigade's MBA defense; PHASE III is the defeat of the enemy attack.

Our task force, TF Falcon, defends the northern flank of Brigade 21. We are facing a weakened but still dangerous 231st Motorized Rifle Division which is at approximately 70% strength.. The 23lst consists of the 218th MRR (BTR), the 269th MRR (BMP), and the 166th MRR (BMP) in the first echelon, with the 33rd Tank Regiment (TR) in the second echelon. Our specific mission is as follows:

TF FALCON defends in sector from NK 200109 to NK 232256 NLT 170500 March 9X to defeat attacking Krasnovian forces and prevent penetration of the Brigade's right (north) flank; on order, reestablish FEBA west of PL PHOENIX.

It is my intent to support the brigade's scheme of maneuver; the task force must defeat all enemy attacks forward of PL PHOENIX. I intend to conduct the defeat of the enemy attack in the vicinity of PL AUSTIN by utilizing the dominant terrain located there. I want to structure the defense to take advantage of natural chokepoints to disrupt and defeat the enemy as it attempts to deploy. The obstacle plan must turn the enemy's main effort into the northern part of the sector and deny it the ability to flank our Main Battle Area (MBA) along PL AUSTIN. The task force security force will withdraw before it is decisively engaged and form the task force reserve. The task force counterattack plan will include the re-occupation of our positions along PL AUSTIN for preparation of our follow-on

defense. End state is the destruction of all enemy forces cast of PL QUINCY with the FEBA re-established along PL AUSTIN.

The mission will be conducted in four phases. PHASE I is the counter-recon battle to destroy all enemy reconnaissance forces vicinity PL QUINCY; PHASE II is the structuring of the MBA by blocking the southern regimental avenue of approach, tuning the enemy's main attack into the north portion of the task force sector, and massing task force fires into EA BAYOU; PHASE III is the defeat of the enemy attack in the vicinity of PL AUSTIN by utilizing integrated defensive fires in EA BAYOU, then displacing to positions to destroy enemy forces in the chokepoints at EAs MILK and GUITAR. If forced back from PL AUSTIN, then we will use a combination of on order integrated defenses at PL PHOENIX and/or PL COCHISE plus a brigade counterattack into the enemy's rear; PHASE IV is the re-establishment of the FEBA along PL AUSTIN by counterattacking the remnants of the enemy MRR, re-occupying initial battle positions, and preparing to defend against follow-on forces.

I want you to be prepared to answer the following priority intelligence requirements:

1. Will enemy in sector be BMP or BTR equipped?
2. Along which avenue of approach will the enemy attack develop?
3. Where will enemy main force deploy?
4. Will 33 TR be committed in sector?
5. Where will lead battalions deploy into attack formations?
6. Where and when will enemy elements begin to withdraw from contact?

I would like you to pay particular attention to the Task Force Execution Matrix. Teams A and B will initially be in the front of our sector and will initially engage the enemy recon forces. The idea is to draw the enemy into Engagement Area (EA) BAYOU where teams C and D can use terrain to channel the mechanized forces into kill zones favorable to us and to defeat the enemy in detail. If necessary, we can trade time for space back to phase line PHOENIX. By then, we need to re-establish the sector with a counter attack. We can’t let the MRR split our forces, so everyone make sure you fully understand the Synchronization Matrix and the Decision Support Template. I’ve laid out the decision criteria for moving from one phase of the battle to another, and I’m particularly concerned about the disengagement criteria. Things will be happening quickly, and it won’t be obvious when the criteria will be met.

Now, take some time to review the Operations Order. I want to make sure that there are no questions. Let’s get back together in 30 minutes.

APPENDIX C. TASK FORCE EXECUTION MATRIX

APPENDIX D. SYNCHRONIZATION MATRIX

APPENDIX E. TASK FORCE DECISION SUPPORT TEMPLATE

APPENDIX F. ANNOTATED DIS TEST SCENARIO MESSAGE STREAM

Notes concerning column headings:

- The function column contains annotations for the reader indicating the role of a message in the scenario: evidence (supporting or conflicting), assessment update (delivered to trained participants only) or break question.
- Message headers, reproduced in this table, consisted of the Time, Report, Originator, Addressee and Net.
- The contents of each email message is in the Message column.

Function	Time	Report	Originator	Addressee	Net	Message
Evidence	T5:55:30	Blue 1	Bravo 05	Falcon 03	BN Cmd	4-5 enemy dism obstacle P-11 CRP we destroy enemy pressure obstacle.
Evidence	T5:56:00	Green 2	Strike 03	All Stations	BDE O&I	Intel indicate moving 20-40km debrief of POW
Evidence	T5:56:01	Green 2	Strike 02	Falcon 02	BDE O&I	DIV collection size Mech forc NK1923. Believ the 218th MRR.
Evidence	T5:56:25	Blue 1	Alpha 06	Falcon 03	BN Cmd	Refit complete
Evidence	T5:56:50	Blue 2	Sapper A	Falcon 03	BN Cmd	BP 20 and 22 c BP23
Evidence	T5:57:00	Green 2	Strike 03	All Stations	BDE O&I	1ST Bde (SLICE main body elem Regiment and E receiving nume
Evidence	T5:57:30	Blue 1	Bravo 05	Falcon 03	BN Cmd	20 vehicles mo AG0026 LEFT2 D
Assessmt Update	T5:58:00	Blue 2	Falcon 06	Falcon 03	BN Cmd	This looks lik getting to be expected.
Evidence	T5:58:01	Green 2	Strike 02	Falcon 02	BDE O&I	JTF MOJAVE rep Co at NK1545,
Evidence	T6:00:00	Green 2	Falcon 03	Falcon 33	BN Cmd	20 vehicles ob the FSE. Bravo appropriate.
Evidence	T6:01:30	Blue 1	Bravo 05	Falcon 33	BN Cmd	Destroyed 2 BT continuing to
Evidence	T6:02:00	Blue 1	Sct 02R	Falcon 02	BN 0&I	Confirm 15 MTI at NK270180
Break question	T6:03:00	Blue 2	Falcon 06	Falcon 03	BN Cmd	Lots of actior concerned that

						contact with t MRR. What do y
Evidence	T6:04:00	Blue 1	Falcon 33	Charlie 05	BN Cmd	We have lost c Last known loc Can you see th
Evidence	T6:04:30	Blue 1	Charlie 05	Falcon 33	BN Cmd	Negative -- ca 01.
Evidence	T6:04:35	Blue 1	Delta 05	Falcon 33	BN Cmd	Sorry -- no co
Evidence	T6:05:00	Blue 1	Charlie 05	Falcon 03	BN Cmd	Observing enem vehicles movir position at we BAYOU.
Assessmt Update	T6:05:01	Blue 1	Falcon 06	Falcon 03	BN Cmd	We may need to has become dec the FSE, but w some positive and we haven’t
Evidence	T6:06:45	Blue 1	Charlie 06	Delta 06	BN Cmd	Can you observ elements I rep engage?
Evidence	T6:07:00	Blue 1	Delta 06	Charlie 06	BN Cmd	We observe the reported; will range.
Evidence	T6:08:30	Blue 1	Bravo 05	Falcon 03	BN Cmd	Spotted large vehicles heade RIGHT TWO, app
Evidence	T6:09:00	CFF	Bravo FIST	DS-FA	BN Cmd	NK 230215; DDP
Break question	T6:09:20	Blue 1	Falcon 33	Falcon 03	BN Cmd	We’re getting close to the l Scout 01, and commo with the fire missions
Evidence	T6:10:00	Green 2	Strike 03	Falcon 03	BDE O&I	EAGLE heavily Have you estab your boundary observing NAI
Evidence	T6:11:00	Blue 1	Charlie 05	Falcon 03	BN Cmd	2 and 3 PLT's

						enemy vehicles
Evidence	T6:12:00	Blue 1	Charlie 05	Falcon 03	BN Cmd	Destroyed 6 BT continuing to
Evidence	T6:13:00	Blue 1	Charlie 05	Falcon 03	BN Cmd	Observing heav throughout pos M2 .
Evidence	T6:13:45	Yellow 1	Charlie 05	Falcon 04	BN A/L	Have 2 M2 dest
Evidence	T6:14:00	Blue 1	Charlie 05	Falcon 03	BN Cmd	Destroyed 3 BT stopped out of infantry.
Evidence	T6:14:25	Red 2	Charlie 07	Falcon 01	BN A/L	2-KIA'S, 4-WIA wounds. Will a of WIA's. If I this, I will j
Evidence	T6:14:28	Medevac	Charlie 07	Falcon 01	BN A/L	Medic en route
Evidence	T6:16:00	Yellow 3	Charlie 05	Falcon 04	BN A/L	Request resupp
Evidence	T6:18:00	Frago	Falcon 06	Falcon 05	BN Cmd	Jump the Main Acknowledge.
Break question	T6:19:00	Blue 2	Falcon 06	Falcon 03	BN Cmd	What are your the displaceme Team C to BP 2 BP21?

APPENDIX G. DEBRIEFING FORM

1. Did the training and/or the interface influence your performance on this test? Please comment.

2. Is this kind of training likely to influence how battalion staff officers problems in the field in Force XXI? Please comment.

3. What areas of decision making under information overload conditions do you believe need more attention for Force XXI staff at the battalion level?

4. What are your general comments about this training?

5. Please rate the training overall (circle a number):

Very bad Very good

1-----2-----3-----4-----5-----6-----7-----8-----9-----10

APPENDIX H. TLX WORKLOAD QUESTIONNAIRE

Please rate the scenario you've just completed with respect to your experience concerning:

6. Mental demand (0 = very low, 10 = very high): _____

7. Physical demand (0 = very low, 10 = very high): _____

8. Time pressure (0 = very low, 10 = very high): _____

9. Effort (0 = very low, 10 = very high): _____

10. Frustration (0 = very low, 10 = very high): _____

Please rate your performance (0 = failure, 10 = perfect): _____

APPENDIX I. BIOGRAPHICAL SURVEY

Name: _____

Brief description of current job:

Rank: _____

Years of service:
Active duty: _____ years
Reserves: _____ years

Military education (check courses you've completed)
_____ Basic course
_____ Advanced course
_____ CAS3
_____ CGSC
_____ Other staff-related training. Please describe:

Staff experience
List staff positions have you held at the battalion level or higher

_____

_____

Number of major exercises: _____

Names of combat assignments

Experience with Defense in Sector scenario.
_____ Wrote or vetted it
_____ Played the scenario
_____ Administered or taught the scenario to other officers
_____ Other (please explain):

APPENDIX J. TRAINING MATERIALS

APPENDIX K. PRACTICE MATERIALS FOR CONTROLS