Effects of Decision Support Technology and Training on Tactical Decision Making[1]

Jared T. Freeman, Marvin S. Cohen, & Bryan Thompson

* Cognitive Technologies, Inc.
4200 Lorcom Lane, Arlington, VA 22207
(703-524-4331; www.cog-tech.com)

Abstract

The effects on tactical decision making of decision support technology (SPAWAR's DSS) and critical thinking training (CTT) were tested. CTT improved the plausibility of officers' assessments of track intent, and the strength of their arguments supporting those assessments. It reliably helped them to find flaws in assessments and conceive alternative assessments. DSS and DSS+CTT helped officers to improve the strength of arguments supporting assessments, and produced (non-significant) improvements on every other measure of decision making process. No effects on planned actions were found. The treatments did not increase workload or lower confidence. Implications for improving training and display design are described.

1. Introduction

Domain experts employ two naturalistic modes of decision-making in complex, tactical scenarios. Recognitional decision making is, as the name implies, based on rapid recognition of a problem and an appropriate response. There is abundant evidence of recognitional decision-making in laboratory tasks, games [Chase and Simon, 1973; Ericsson and Charness, 1994], and real-world task performance in military and civilian settings [Klein, 1993]. Critical thinking, in contrast, denotes deliberate interpretation of evidence and careful evaluation of conclusions drawn from that evidence, such as assessments and plans. While critical thinking may not be used as frequently as recognitional decision making, it is nonetheless evident at key junctures in virtually every critical incident described to us by experienced officers in the Navy [Cohen, et al., 1996] and Army [Cohen, et al., 1993].

The Recognition/Metacognition (R/M) framework [Cohen, et al., 1996] integrates both of these decision making processes into a single model. It acknowledges the role of pattern-matching, or recognitional decision-making, and specifies that it is the preferred mode when the circumstance is highly familiar, stakes are low, or time is short. More formally, the framework specifies a mechanism that (1) tests for the presence of uncertainty that is significant enough that its reduction could potentially change a decision; (2) tests whether a change of decision could decisively affect the outcome (where the swing in the outcomes defines the stakes); and (3) compares the potential benefits of reducing uncertainty to the cost of taking time to do so [Cohen, Parasuraman, & Freeman, 1998] (this volume). When the situation is novel, stakes are high, and time is available, critical thinking is the

[1] This research was performed under contract to U.S. Navy Air Warfare Center / Training Systems Division, contract Number: N61339-96-C-0066. We are indebted to Dr. Joan Hall-Johnston, her colleagues at NAWC/TSD, and staff and contractors at SPAWAR for their contributions to this research.

preferred mode of decision-making. In this mode, decision-makers exercise skills that help them to identify and reduce uncertainty in their understanding of the situation or their response plans. These skills help officers to identify and handle three types of uncertainty: gaps in critical information, unreliable assumptions, and conflicting interpretations of the evidence. Decision-makers often employ critical thinking to refine mental models (or situation models) that represent the events at hand and the causal forces that drive them. Stories are one example of such models. In numerous critical incident interviews, such as those conducted for the Navy's program on Tactical Decision-Making Under Stress, experienced military officers have described instances in which they wove together events concerning a suspect air track, evaluated how well past events predicted present or future events, and then correctly reinterpreted the situation to avoid a fratricide, an inappropriate engagement, a strike on own ship, or another disaster.

The R/M model provides a useful foundation upon which to design decision-aiding interventions, whether in the form of training or decision aids. It suggests that such aids should:

- Support rapid recognition – Displays can integrate related data or use graphics to convey relationships such as rate of change over time or distance [Tufte, 1983; Kosslyn, 1993]. Training can focus foster recognition and rote responses to simple, common problems.
- Identify opportunities for critical thinking — Displays and training can help officers determine when to engage critical thinking and when recognitional decision making or plan execution is more appropriate. For example, less experienced officers may benefit from a simple display of the time available for resolving a key decision, such as whether to divert a commercial airliner to avoid a fuel shortage [Freeman, et al., in press]. More powerful displays might represent the predicted change in stakes over time as a function of the system's certainty concerning the current assessment of a suspect track. Training may heighten awareness of all three factors. The goal, in any case, is to alert officers to opportunities to critique their understanding of important tactical situations, and, conversely, to warn them when there is little or no opportunity to do so.
- Help officers think critically — Critical thinking displays and techniques (training) should help officers to build complete and coherent mental models of the tactical situation. For examples, displays may represent the events and causal relations between them explicitly, whether pictorially, as node-link graphs [Freeman, et al., 1997; Cohen, et al., 1995], or in some other structured format. Training may help officers to understand what causal forces bear on a situation, and what types of information are key. Displays and training can also help officers to rapidly find gaps in their knowledge, conflicting assessments of a situation, or weak assumptions that have been or must be made to support an assessment or plan.

The experiment reported here tested the effectiveness of training in critical thinking skills as well as technology to support recognitional decision-making and critical thinking. Taken together, the training and technology implemented interventions in each of the three categories, above. The question at hand was whether the training, the technology, or both in combination improved tactical decisions or decision-making processes.

The technology was the Decision Support System (DSS), developed by the Space and Naval Warfare Systems Center, San Diego. In brief, the DSS consists of two displays typically presented on dual monitors. The left

display is a standard geoplot (a plan view of local geography and cultural features, tracks and their history) as well as a panel of geoplot controls. The right display is organized into three general areas. The top and middle left portions present critical information about a single track: the Track Summary contains kinematics data and historical alerts; the Track Profile graphically represents changes in altitude and range over time; the Response Manager shows actions officers may take as a function of the distance of the selected track from ownship). The middle right portion of the screen is the Basis for Assessment module, which presents assessments, evidence supporting the assessment, evidence conflicting with it, and assumptions that may neutralize the conflict. The bottom section of the left and right displays is a set of tactical data summaries for each track (the miniCRO).

The interactive training in critical thinking consisted of a simple procedure for generating, testing, and communicating assessments; detailed guidance on evaluating attack assessments; a procedure for generating alternative assessments and resolving conflict within an assessment; and guidelines for determining when to invest time in critical thinking and when to act immediately on the best available plan.

2. Method

2.1 Design

Participants executed one of two test scenarios as a pretest, received critical thinking training or a control treatment, and then performed the second test scenario as a posttest. Thus, the study crossed two treatment conditions (DSS+CTT or critical thinking training and DSS alone) and two orders of test scenarios (Charlie then India, or India then Charlie) between subjects. All participants used a deprived version of the DSS (described below) on the pretest, and the full DSS on practice scenarios and on the posttest. This design gave us (1) the effect of the full DSS (within subjects), (2) the effect of the DSS+CTT (within subjects) and (3) the effect of CTT (DSS+CTT less DSS, between subjects). Dependent measures of effectiveness and process are described specifically in the results section, below.

2.2 Participants

Thirty-four active-duty Navy officers participated in the experiment in teams of two, playing the roles of CO and TAO. (Most analyses used the dyad or one member of the dyad (the CO) as the unit of measure.) Officers in the DSS condition had served longer in the Navy (mean = 12.609, S.E.M. = 1.376) than officers in the DSS+CTT group (mean = 8.847, S.E.M. = 1.085), a reliable difference ($t32 = 2.170$, $p = 0.038$). However, the groups did not differ with respect to three specific measures of expertise in CIC AAW operations: as a group, they averaged 10 weeks of TAO experience, 54.5 weeks of AAW experience, and 44% reported that they had served in the weapons department.

2.3 Materials

Two test scenarios (Charlie & India) and two practice scenarios were administered. Each was designed to require critical thinking regarding key tracks at specific times. Information concerning those tracks was incomplete or conflicting, and the behavior of tracks was likely to elicit unreliable assumptions relevant to assessments of track intent. The available time for decision making also varied between key tracks, such that in one practice scenario, critical thinking was appropriate for some tracks early in the scenario, but inappropriate shortly before they launched missiles at ownship. The scenarios were adapted from well-vetted scenarios developed by SPAWAR.

The test scenarios were presented on two versions of the DSS. A deprived DSS, intended to emulate the existing AEGIS C&D

display, provided only the geoplot, track summary, and screen control modules (described above). All officers used the deprived DSS on the pretest and the full DSS during their treatment and posttest.

Several paper forms were administered during each experimental session. All officers completed a biographical information form (used in part to assign officers to roles as CO or TAO), a test questionnaire, a variant of the NASA TLX workload questionnaire, a questionnaire concerning practice scenarios (for the DSS group only)2, and a debriefing form. The test form (which was presented at the end of each test scenario) consisted of six questions intended to elicit assessments of tracks specified by the experimenter, the evidence supporting or conflicting with those assessments, and actions concerning the tracks.

2.4 Procedure

Each experimental session lasted eight to nine hours. Each session began with a pre-brief, in which an SME informed the officers that they were helping to evaluate prototype CIC technology and training in its use, and presented a brief concerning the overall geopolitical context of the practice and test scenarios. Officers filled out biographical questionnaires. The SME assigned the officer with more tactical decision making experience to play the technically demanding role of TAO, and the other officer to play the role of CO.

Officers were then given individual training on the deprived DSS. A criterion test was administered and deficiencies in skills using the device were remediated. Next, officers were read a tactical brief concerning the pretest scenario, and the experimenter reviewed the test questionnaire with them to orient them to their key task — critically assessing track intent — and to resolve any questions concerning test wording. The test forms were collected, and the scenario was run by the system administrator(s). The system administrator(s) responded to participants' requests for information and reports in the expected roles (AAWC, Gulf Bravo, etc.) and an SME provided technical guidance in the few instances when it was needed. All communications were recorded on audio tape, and interactions with the DSS (deprived and full) were logged to disk. At the end of this scenario, the experimenter handed out the test booklets. Participants were told that the given tracks were selected because they elicited useful responses, not necessarily because they were true innocents or villains, and that the given assessments similarly produced useful responses but were not necessarily ground truth or extreme fictions. Officers were asked to interact as much as they wished but to complete separate test booklets. After completing the test form, the officers filled out a TLX workload assessment.

Following the pretest, all officers received training in the use of the full DSS, performed a criterion test, and received any additional training needed. Next, officers in the DSS+CTT treatment received instruction from a trainer concerning critical thinking, read supporting training slides, and executed two practice scenarios in which they discussed the intent of selected tracks with the experimenter and each other. Officers who received the DSS treatment did not get the structured training, but did execute the same practice scenarios, after which they completed a printed form asking them to identify key events and instructional lessons to be learned from the scenario.

The posttest followed the same procedure as the pretest, consisting of a tactical briefing, a review of test questions, the scenario run including two queries from higher command, a

[2] The control practice survey consisted of two questions intended to focus officers on surface features of the scenarios and the instructional potential of the scenarios, not issues that might involve critical thinking on tactical issues. These questions were "List the key events in this scenario." and "What lessons can be learned from this scenario?"

25-minute period for completing the test form, and a few minutes for completing the TLX form. The session concluded with a debriefing, in which officers provided their assessment of the training treatment (or the DSS treatment practice regime) and evaluated the technology.

2.5 Analytic Strategy

This study was designed principally to test two hypotheses:

- H1: Providing the full DSS gives officers an advantage over more standard equipment, represented by the deprived DSS
- H2: Providing the full DSS plus critical thinking training gives officers an advantage over using standard equipment.
- H3: Providing critical thinking training augments the benefits of the full DSS.

These hypotheses were applied to specific dependent measures concerning decision outcomes, specifically SME ratings of assessment plausibility and the appropriateness of actions, and decision processes, including the number of points of evidence offered in defense of assessments, the number of conflicting points of evidence identified, and the number of alternative assessments generated. These measures are discussed in detail below.

In most of the analyses, we tested these hypotheses using a repeated measures ANOVA. The within-subjects repeated measure was test score (pretest vs. posttest). Between-subjects independent variables were treatment (DSS+CTT vs. DSS), scenario order (Charlie followed by India vs. India then Charlie), and the interaction of treatment and scenario order. Also included in the model were the interactions of the within-subjects variable with all of the between-subjects variables. Three specific tests were performed using this model. The test of H1 determined whether the posttest scores of officers who received only the DSS treatment differed reliably from their scores on the pretest (using the deprived DSS). The test of H2 determined whether the posttest scores of the DSS+CTT group reliably differed from their pretest scores (using the deprived DSS). Regarding H3, we tested for an interaction of test scores (pre vs. post) with treatment (DSS vs. DSS+CTT). The presence of such an interaction indicated that the change in scores between tests varied by treatment, an effect that was attributable only to training, not display technology.

2.6 Decision Products: Effects on Assessments and Actions

Participants were asked to make two fundamental decisions on the written tests. The first was to assess the intent of a given track. The second was to describe what actions they planned to take concerning tracks of interest to them. Most officers planned a response regarding the given track, among others. SMEs reviewed these decisions. They rated the plausibility of assessments that officers formulated, the appropriateness of the actions they planned to take given those assessments, and the strength of reasoning (if any) that they offered in defense of those decisions. All ratings were made on appropriately anchored scales of 0 to 10[3].

Five subject matter experts participated in the analyses of the effects of the treatments on decision outcomes. A single set of SME judgments was generated by a team consisting of two U.S. Navy Captains (ret'd.) with 30 and 34 years experience, respectively, and a retired officer with 5½ years experience (SME team A). The two senior officers in this group were highly familiar with the specific scenarios involved and had helped to conduct this and

[3] For ratings of assessment plausibility the anchors were 0 = implausible, 10 = highly plausible. For ratings of assessment strength, anchors were 0 = very weak, 10 = very strong. For ratings of action appropriateness given the assessment, anchors were 0 = very inappropriate, 10 = very appropriate. For ratings of strength of defense of actions, anchors were 0 = very weak, 10 = very strong.

other experiments. A second set of ratings was produced by a U.S. Navy Captain (ret'd.) with 30 years experience (SME B). This officer was introduced to the scenarios and the experiment while analyzing these data. All SMEs were blind to the treatment condition and test (pre vs. post) on which responses were given.

Inter-rater reliability over all rating tasks was low (r = 0.315, p = .001). Anecdotal evidence suggests that this is the norm in this domain [Jentsch, personal communication], as it is for other ill-structured domains [Shanteau, 1997]. Despite the low inter-rater reliability, all SME ratings were used in this preliminary analysis. However, the potential for large differences in ratings between judges was recognized by adding to the repeated measures model a covariate that represented the effect of the particular rater. That covariate contributed reliably to the model in several instances. Only responses by the CO were evaluated.

2.6.1 Effects on Assessments

Officers gave a wide range of responses when asked to assess the intent of one, given track in each scenario. For scenario Charlie, six COs said that a P-3 in the vicinity of incoming military aircraft was providing targeting, three claimed it was participating in a nearby SAR, one stated it was testing ROE, five said it was on patrol, and two assessed it as unknown. In scenario India, one CO asserted that one of a pair of incoming military aircraft was on an attack run, one labeled it assumed hostile, two claimed the aircraft was performing reconnaissance, and thirteen said it was in transit.

SME ratings of the plausibility of those assessments declined for the DSS-only group by 16% between tests from 5.312 to 4.437 (S.E.M. = 0.457) [4]. However, the plausibility of assessments by officers who received DSS+CTT rose 28% between tests from 4.512 to 5.762 (S.E.M. = 0.433). The decline in scores by officers in the DSS group was not statistically reliable, indicating that providing the full DSS had no reliable impact over the deprived DSS (H1). The rise among CTT+DSS participants represented a weak trend (H2: $F_{1,29}$ = 1.998, p = 0.168). However, the interaction between test scores of the two groups was reliable (H1: $F_{1,29}$ = 5.227, p = 0.03). Training officers in critical thinking helped them to improve their assessments above any effect of the full DSS alone.

SMEs rated the strength with which COs defended their own, favored assessments. The arguments presented by the DSS group declined 24% in strength from 5.125 to 3.875 (S.E.M. = 0.515), while those of officers who received DSS+CTT rose 27% from 4.187 to 5.338 (S.E.M. = 0.489). Neither effect was reliable. Thus, providing the full DSS alone did not reliably affect the strength of arguments (H1), nor did providing DSS+CTT (H2). However, the interaction was significant; training improved the strength of officers arguments in defense of their assessments (H3: $F_{1,29}$ = 5.549, p = 0.025).

2.6.2 Effects on Planned Actions

COs responded to the P-3 at the end of scenario Charlie with plans that ranged from monitoring to illuminating and covering. A fighter aircraft in scenario India elicited a more aggressive set of responses ranging from engagement to covering, illuminating, and monitoring.

The appropriateness of these actions, according to SMEs, declined 6% among officers who received only the full DSS from 7.833 to 7.333 (S.E.M. = 0.976) and rose 4% among officers who received DSS+CTT, from 6.939 to 7.197 (S.E.M. = 0.75). The effect on planned actions of providing technology was not reliable (H1), nor was the effect of technology plus training (H2), or the effect attributable to training alone (H3).

[4] Standard Error of the Mean is given for posttest scores only, unless it varies strongly from the pretest S.E.M.

The reasons officers gave for their actions declined in strength among all COs between tests. The mean decline was 10% for the DSS treatment (from 6.75 to 6.083, S.E.M. = 0.794) and 11% for those who received DSS+CTT from (6.75 to 6, S.E.M. = 0.842). None of the tested effects — of the DSS, DSS+CTT, or CTT alone — was statistically significant.

2.6.3 Discussion

Critical thinking training alone was responsible for improving the plausibility of tactical assessments made by officers in this experiment. Training also improved the strength of arguments officers gave in defense of their assessments. That training should raise performance on both measures is consistent with findings in prior studies of critical thinking training [Freeman, et al., 1997; Cohen and Freeman, 1997].

However, neither the DSS, CTT, nor the combination improved the actions officers planned to take regarding the given track nor the reasoning they offered for those plans.

It is worth noting here that there was a strong correlations between SME ratings of assessment plausibility and argument strength for SME team A ($r = 0.980$, $p < .001$) and a moderate one for SME B ($r = 0.567$, $p < .001$), and a similar set of correlations between action appropriateness and strength of rationale for SME team A ($r = 0.931$, $p < .001$) and SME B ($r = 0.533$, $p < .013$). These correlations provide some backing for the claim, upon which we rely, below, that argument strength is a valid measure of process and a predictor of the quality of decision-making products such as assessment plausibility and appropriateness of actions.

2.7 Decision Processes: Use of Evidence in Critical Thinking

Though the treatments improved decision outcomes, we wished to know whether decision processes improved as well. To measure these processes, we examined the amount of evidence that officers mustered to rebut and defend assessments, and the number of alternative assessments they generated. Their skill at controlling these processes was measured by their ability to discriminate between assessments on the basis of confidence.

2.7.1 Identifying conflicting evidence

Officers were asked to list the evidence that conflicted with the assessment that a given track would attack. Participants' responses were parsed into arguments. The number of arguments (or pieces of evidence) listed by each participant on the pretest and posttest served as the within-subjects dependent variable.

Scores for officers in the DSS-only condition declined 45% between tests, from 3.625 to 2.000 (S.E.M. = 0.442), indicating that providing the DSS-2 alone may have hindered officers with respect to identifying conflicting evidence (H1: $F_{1,11} = 4.718$, $p = 0.053$). Scores for officers in the DSS+CTT group rose 44% from 2.375 to 3.425 (S.E.M. = 0.342), however this effect was not a reliable indicator that DSS+CTT improved performance over the deprived testing condition (H2). The obvious interaction of treatment with test represented a reliable benefit of critical thinking training (H3: $F_{1,11} = 13.846$, $p = 0.003$).

2.7.2 Explaining conflicting evidence

After identifying the evidence conflicting with the given assessment, officers were asked to defend the assessment anyway, as a test of their ability to resolve conflicting evidence.

The dependent variable was the number of arguments issued in defense of the given assessment per conflicting point of evidence raised. This proportion was quite low for the DSS group on the pretest, at 0.175, but rose to 0.938 on the posttest (S.E.M. = 0.262). This was a statistically reliable improvement in

performance, attributable to the use of the full DSS (H1: $F_{1,10} = 6.292$, p = 0.031). Scores for the DSS+CTT group rose 31% from 1.125 on the pretest to 1.469 on the posttest (S.E.M. = 0.214). This effect was nearly reliable (H2: $F_{1,10} = 4.038$, $p = 0.072$), indicating that DSS+CTT was an effective treatment on this measure. However, the interaction of test and treatment was not reliable (H3). Training did not have an effect on the ability to explain conflicting evidence.

2.7.3 Defending assessments

Officers were asked to make the case for their own assessments of a given track. On each test, they defended an assessment they favored and one they did not favor. For each of these two responses we examined the number of arguments.

When defending their favored assessment, officers who received only the DSS treatment produced roughly the same number of arguments on the pretest and posttest: 2.500 at the mean (S.E.M. = 0.614). Thus, there was no impact of the technology on this measure (H1). The performance of the DSS+CTT group rose 13% from 2.825 arguments on the pretest to 3.200 arguments (S.E.M. = 0.582). This effect was not reliable either (H2), nor was the weak interaction that represented any advantage of training above and beyond that attributed to the DSS (H3).

When asked to defend an assessment they did not favor, officers in the DSS group produced 90% more arguments on the posttest (mean = 2.375, S.E.M = 0.627) than the pretest (mean = 1.250). This was a reliable improvement attributable to the DSS (H1: $F_{1,13} = 14.846$, $p = 0.002$). Officers who received DSS+CTT produced 56% more arguments on the posttest (mean = 3.550, S.E.M. = 0.595) than the pretest (mean = 2.275). This effect was a highly reliable indicator of the benefits of DSS+CTT combined (H2: $F_{1,13} = 18.303$, $p = 0.001$). However, there was no treatment x test interaction, suggesting that training did not improve performance (H3).

2.7.4 Generating alternative assessments

Officers were asked to list alternatives to their favored assessment of the given track. The number of assessments in this list declined 20% within the DSS group from 3.75 to 3.000 on the posttest (S.E.M. = 0.468). This effect of the full DSS was not reliable (H1). Trained officers showed a 19% increase in alternative assessments between tests, from 3.7 to 4.4 (S.E.M. = 0.444). This effect of DSS+CTT also was not reliable (H2). The pronounced difference in treatment effects between groups was statistically significant, however, indicating that training improved the ability to conceive alternative assessments (H3: $F_{1,13} = 6.323$, $p = 0.026$).

2.7.5 Controlling critical thinking: Confidence

Participants were asked to generate assessments of the intent of a specified track, and to rate their confidence in those assessments on a scale from 0 to 100 (where 0 denoted no confidence and 10 extreme confidence). These ratings were used to create an index of discriminability between assessments by taking the average of the difference in confidence between all of the ranked confidence ratings for the favored assessment and all alternates. For officers who used all or almost all of the scale on each test (and many did), the average difference between confidence ratings was often inversely proportional to the number of alternatives listed[5]. To control for this coincidental effect of the number of alternatives listed and the

[5] For example, for an officer who produced five alternative assessments with confidence ratings between 100 and 0, the averaged difference in confidence ratings was 20. An officer with ten alternatives rated between 100 and 0 produced an average difference in ratings of 10. Thus, the ability to conceive of more assessments that could account for the behavior of a given track apparently diminished the ability of the officer to discriminate one assessment from another.

differences in ratings, the difference in the number of alternatives between tests was appended as a covariate to the repeated measures model used elsewhere.

Neither treatment affected the ability of officers to discriminate between assessments on the basis of confidence. Within the DSS-only group, discriminability declined very slightly from 10.173 on the pretest to 10.095 (S.E.M. = 0.998). Among trained officers, discriminability increased from 10.551 to 10.759 (S.E.M. = 0.944). The ability to discriminate between assessments on the basis of confidence did not diminish as a result of providing the DSS (H1) or the DSS plus training to help officers to critique their assessments (H2), nor was there an independent effect of training (H3).

2.7.6 Discussion

In general, the DSS+CTT treatment improved performance on every process measure examined here. This was a reliable effect in some instances, and a trend, at best, in others, but it was a consistent pattern across all of the data described above.

Officers in both treatment conditions used more of the available evidence to explain conflicting evidence and to defend a disfavored assessment (but not a favored one) on the posttest than they did on the pretest. These were reliable benefits of both the DSS alone (H1) and the DSS plus critical thinking training (H2)

Critical thinking training was beneficial (above and beyond the DSS) in two interesting respects (H3). It helped officers to identify sources of conflict in given assessments and generate alternative assessments of a given track. This suggests that, whatever the improvements of information representation in the DSS, the displays in and of themselves did not help officers to recognize problems with assessments nor to explore alternatives. Training or additional decision-aiding technology may needed to help them with these, literally critical tasks.

Finally, officers' ability to discriminate between assessments on the basis of confidence did not diminish. The DSS could conceivably have evoked information overload, and an accompanying sense of helplessness. It did not. Critical thinking training encouraged officers to find fault with their own assessments, yet it did not diminish their ability to discriminate between assessments. Officers can think critically about highly ambiguous circumstances and still be decisive.

2.8 Effects of Treatments on Perceived Workload

The scenarios used in the present study were designed to be complex and demanding. An analysis of responses to the TLX rating form, administered after each test, gave some insight into the effects of the treatments on perceived workload.

After the pretest and posttest, officers filled out a TLX form describing their experienced in the scenario using 20-point rating scales anchored at "low" and "high." They provided one rating on each of the following dimensions: mental workload, physical workload, temporal demand, effort, performance (quality of work), and frustration. The same repeated measures ANOVA was used to model these data, and we performed the same three tests of effects of DSS, DSS+CTT, and CTT, respectively. Data from all participants (not just those playing CO) were used in these analyses.

Providing the DSS alone (H1) had no reliable effect on any TLX measure, though there was a trend for it to lower frustration by approximately 16% (H1: $F_{1,30} = 2.64$, $p = 0.115$). The DSS+CTT treatment produced a reliable, 25% decrease in frustration (H2: $F_{1,30} = 4.946$, $p = 0.034$) and a trend towards physical workloads that were 26% higher (H2: $F_{1,30} = 2.481$, $p = 0.126$). Critical thinking

training alone produced no reliable effects on TLX ratings (H3).

2.9 Evaluations of the Treatments and Technology

2.9.1 Ratings of the Training

At the conclusion of the experimental session, all participants were asked to rate the instruction. They used the following scale: 1 (strongly negative), 2 (negative), 3 (neutral), 4 (positive), or 5 (strongly positive).

Officers in the group that received the training group rated their treatment at 4.56 at the mean (S.E.M. = 0.120), a value that was reliably higher even than a positive rating of 4 (t_{16} = 4.642, p < 0.001). However, ratings by DSS+CTT officers did not differ reliably from ratings by DSS members, who gave the DSS training and practice scenarios a mean rating of 4.267 (S.E.M. = 0.206).

Trained officers were asked whether critical thinking training influenced their approach to the posttest. Fourteen, or 77% of the 18 trained officers reported that it did so more than a little. They reported that it helped them to critique their decisions before acting on them, be more alert to the time available for decision-making and thus "step out of crisis mode", and consider alternative interpretations of events. Not coincidentally, these were some of the key goals of the training. One officer felt that only the technology, not the training, had value. Three felt the training had only a small value in executing the test, and two of these noted that time pressures prohibited them from applying the training more thoroughly.

Officers were asked whether the training was likely to influence their decision making in the field. All but one, or 94%, reported that it would do so more than a little, arguing that it helped them to structure their decision making, consider alternative interpretations of events, and critique their assessments more thoroughly. The lone dissenter indicated that the training differed little from current practice. However, this can also be read as an endorsement of the training as a vehicle for conveying best current practices.

2.9.2 Ratings of the DSS

We asked officers "How important was the information in each DSS module for your tasks?" and officers responded by rating each of the modules on a five point scale anchored thus: 1 = Not important, 3 = Somewhat important, 5 = Very important. Ratings served as the dependent variable in ANOVA. Treatment, posttest scenario, the role of each officer (TAO vs. CO) and their interactions were independent variables.

The ratings did not differ reliably by treatment except in the case of the geoplot. The DSS group gave this module a mean rating of 4.500 (S.E.M. = 0.128), while those who received DSS+CTT gave it a mean rating of 4.938 (S.E.M. 0.122) ($F_{1,26}$ = 5.780, p = 0.024). There was a trend for the DSS+CTT group officers to award lower ratings to the response manager on average (mean = 3.788, S.E.M. = 0.194) than did the DSS group (mean = 4.250, S.E.M. = 0.204) ($F_{1,30}$ = 2.936, p = 0.099).

One main effect of role (CO vs. TAO) emerged. Officers in the CO role rated the response manager higher (mean = 4.313, S.E.M. = 0.191) than did those in the TAO role (mean = 3.725, S.E.M. = 0.191) ($F_{1,26}$ = 4.738, p = 0.039). An examination of elapsed time per module (from DSS data logs) indicated that COs also spent more time per scenario using the response manager (mean = 3.214 minutes, S.E.M. = 0.417) than did TAOs (mean = 0.258, S.E.M = 0.417) ($F_{1,20}$ = 25.084, p < .001).

3. Conclusion

In this experiment, we tested the hypotheses that tactical decision-making could be improved by providing (H1) advanced decision support displays (SPAWAR's DSS), (H2) the

DSS plus critical thinking training (CTT), or (H3) critical thinking training alone.

Critical thinking training alone was responsible for improving the plausibility of officers' assessments of track intent, and the strength of their arguments supporting those assessments. It reliably helped officers to think against themselves by improving their ability to identify conflicting evidence and alternatives to their preferred assessments, yet it (like the DSS and DSS+CTT treatments) did not lower their ability to discriminate between assessments on the basis of confidence. Future implementations of the critical thinking training should focus on these benefits, because this is where the training appears to pay off, and where the literature on decision biases suggests that improvement is needed. In contrast, training had little effect on the ability to generate arguments supporting assessments. There may be little room for improvement in that skill.

DSS+CTT produced a striking overall pattern: it improved performance on almost every measure of decision process examined here (though only a few effects were statistically reliable). This suggests that there is a valuable synergy between the technology and training.

Both DSS+CTT and DSS alone helped officers to defend assessments other than the ones they preferred. Thus, treatments involving the DSS helped officers to explore specific assessments when explicitly asked to do so. However, it is clear that the DSS does not by itself help officers to find weaknesses in their assessments. At a minimum, good training is needed to help officers think against themselves and the DSS in order to improve their assessments. A better solution may be to enhance the DSS so that it notifies users of opportunities for and targets of critical thinking. Opportunities for critical thinking might be signaled by annotating or color-coding the mini-CROs at the bottom of the display. This would help users identify tracks where the degree of uncertainty, the available time, and the stakes warrant critical thinking. It may also be beneficial to represent time (e.g., to technical or most likely engagement range) in some displays (such as the track profile) in order to emphasize awareness of available decision time. Targets (or topics) for critical thinking might be flagged with tags, in a manner analogous to "post-it" notes. Such tags would appear automatically in various display modules (which would otherwise remain unchanged) to signal gaps, conflicting evidence, or unreliable assumptions pertaining to the selected track. The current Basis of Assessment window might be used to select an assessment for evaluation, and to display detailed information regarding any tag that the user selects.

In general, the treatments had no effect on officers' plans for acting against given tracks. Given that there are relatively few actions officers can take regarding a suspect track, this may not be surprising.

We were pleased to find that the new technology and the training did not increase workloads. The sole reliable effect was a decline in frustration levels among officers in the DSS+CTT condition. This suggests that officers may learn to apply DSS+CTT relatively quickly, and that there is relatively low overhead in doing so or that the DSS and CTT improve efficiency dramatically. Larger and longer term studies may help to verify whether this finding is reliable.

Participants rated the training highly and cited benefits identical to those the training was designed to produce. The DSS modules received high ratings, with the exception of the Basis for Assessment module and, arguably, the Mini-CRO. Data used in these modules presented minor problems noted by a few officers. The value of these modules might be enhanced with the modifications mentioned above. Participants playing the role of CO

made heavier use of the Response Manager and considered it more valuable than did officers in the TAO role. Future modifications to the Response Manager, if any, should consider that use of this module varies by role. That officers' roles did not affect their ratings of other modules suggests that the system generally serves TAOs and COs equally well.

Two caveats are in order concerning the findings reported here. First, the analyses are preliminary. Second, it is important to recognize that the design used here was quasi-experimental. Effects cannot be clearly attributed to the use of the DSS or DSS+CTT alone because the administration of these treatments was conflated with practice between the pretest and posttest. There were too few subjects available to allow for a control group that received the deprived DSS on the pretest and posttest and no training. Thus, declines in some scores may be due to exhaustion over the course of the day-long experiment. Improvements may be a function of hours of team practice at tactical decision-making tasks. However, there is no predominant pattern of exhaustion or practice effects over the treatments, nor did any subjects mention these issues. We are inclined to interpret the effects primarily as functions of the treatments.

It is clear that training in selected critical thinking skills is crucial if officers are to use the DSS well, though design modifications might also be productive. Critical thinking training, too, can be improved by focusing on identifying sources of conflict in assessments and generating alternative assessments. However, the DSS and critical thinking training offer benefits to officers now. Use of the DSS with or without training improves tactical decision making processes, and training alone improves processes and the tactical assessments that officers maker in complex scenarios.

References

[Chase and Simon, 1973] Chase, W.G. & Simon, H.A. Perception in Chess. *Cognitive Psychology*, 4(1), 55-81, 1973.

[Cohen and Freeman, 1997] Cohen, M.S. & Freeman, J.T. Improving Critical Thinking. In Flin, R., et al. (eds.), *Decision Making Under Stress: Emerging Themes and Applications.* Brookfield, VT: Ashgate Publishing Co., 1997.

[Cohen, et al., 1993] Cohen, M.S., Adelman, L., Tolcott, M.A., Bresnick, T.A., & Marvin, F.F. (1993). *A cognitive framework for battlefield commanders' situation assessment* (Technical Report 93-1). Arlington, VA: Cognitive Technologies, Inc., 1993.

[Cohen, et al., 1995] Cohen, M.S., Thompson, B.T., Adelman, L., Bresnick, T.A., Tolcott, M.A., and Freeman, J.T. *Rapid Capturing of Battlefield Mental Models.* U.S. Army Research Institute, Ft. Leavenworth, KS. Contract #DASW01-95-C-0062. Arlington, VA: Cognitive Technologies, Inc., 1995

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

[Cohen, et al., 1998] Cohen, M.S., Parasuraman, R., & Freeman, J.T. Trust in Decision Aids: What is it and how can it be improved? *Proceedings of the 1998 Command and Control Research & Technology Symposium*, Monterey, CA., 1998

[Ericsson and Charness, 1994] Ericsson, K. Anders & Charness, N. Expert performance: Its structure and acquisition. *American Psychologist*, 49(8), 725-47, 1994

[Freeman, et al., 1997] Freeman, J.T., Cohen, M.S. and Serfaty, D. Information Overload in the Digital Army: Simulator-based Training for Prevention, Detection & Cure.

Proceedings of the 1997 Command and Control Research and Technology Symposium, Washington, D.C., 1997.

[Freeman, et al., in press] Freeman, J., Cohen, M.S., Smith, C., and Thompson, B.T. Time-Stressed Decision-Making in the Cockpit. *Proceedings of the Association for Information Systems 1998 Americas Conference*, Baltimore, MD., 1998

[Klein, 1993] Klein, G. A Recognition-Primed Decision (RPD) Model of Rapid Decision Making. In Gary Klein, J. Orasanu, R. Calderwood, and C.E. Zsambok (eds.), *Decision Making in Action: Models and Methods*. Norwood, NJ: Ablex., 1993.

[Kosslyn, 1993] Kosslyn, Stephen M. *Elements of Graph Design*. New York: W H Freeman & Co., 1993.

[Shanteau, 1997] Shanteau, James. Why do experts disagree? Presented at the SPUDM 1997, Leeds, England. 1997.

[Tufte, 1983] Tufte, Edward R. *The Visual Display of Quantitative Information*. Cheshire, CT: Graphics Press. 1983.