Training for Complex Decision-Making: A Test of Instruction Based on the Recognition / Metacognition Model[1]

Jared T. Freeman, Ph.D.
Marvin S. Cohen, Ph.D.
Cognitive Technologies, Inc.
4200 Lorcom Lane
Arlington, VA 22207
cti@access.digex.net

Abstract

A model of decision making is proposed that integrates recognitional and metacognitive processes. The metacognitive functions supplement recognitional skills in situations in which events are novel or the meaning of events is unclear for other reasons. Such situations are common in military settings, such as Naval anti-air warfare, where it is often the case that stakes are high, time is short, and events do not fit standard patterns. Training based on the model was designed to help Navy officers assess such situations. Two experimental tests of the training were conducted using highly realistic, computer-based simulations. Results indicate that the training improved decision processes and decision accuracy. Implications for team training and the design of decision aids are discussed.

1. Introduction

Decision-making by Naval officers engaged in anti-air warfare (AAW) challenges overly simple theories of how decisions are made, our understanding of the skills officers exercise when making tough decisions, and our designs for teaching those skills.

The empirical data concerning Naval decision-making have strong implications for decision theory. In interviews from the Navy's investigation of Tactical Decision-Making Under Stress (TADMUS), we find that experienced officers rarely use the techniques prescribed by decision-analytic theorists (e.g., [Raiffa, 1968, Keeney and Raiffa, 1976]), which involve exhaustively specifying and quantifying evidence and outcomes. There are several possible explanations for this, all having to do with the poor fit between analytic methods and the demands of real-world decision-making. Analytic methods presume that the decision maker's knowledge of probabilities and values is thorough and precise, though imperfect and approximate knowledge is the norm; the methods assume a fixed, serial process, though decision makers frequently generate new options and criteria opportunistically; and, finally, that decision makers rarely use analytic approaches suggests that the cost (in time or effort) is rarely worth the benefit in actual practice. Thus, training officers in decision-analytic techniques needs to targeted to very specific types of problems, at best, and may be generally ill-advised, at worst.

There is evidence in the TADMUS protocols to support a different theory of decision making: pattern recognition (sometimes called rule-based

[1] This work was supported by Contracts No. N61339-92-C-0092 and N61339-95-C-0107 with the Naval Air Warfare Center / Training Systems Division.

behavior [Rasmussen, 1981] or procedural knowledge [Anderson, 1982]. As simple recognition models predict, senior officers are familiar with characteristic track patterns (such as the constant bearing and decreasing range of a potentially hostile aircraft), and they map these events to standard responses, such as those specified in the rules of engagement (ROE), or to responses taken in similar encounters. However, simple theories of pattern recognition do not account for some of the most interesting decision-making behaviors of senior officers, such as the ability to develop reasonable interpretations of events they have never before encountered, to generate and resolve conflicting interpretations of events, to change their minds, and to use all of the available time (no less and no more) for thinking about ambiguous situations.

We have developed a model that accounts for these behaviors. It does so by acknowledging the role of recognition in decision-making, and by specifying critical thinking processes that refine the assessments and plans that are the products of recognition. Called the Recognition / Metacognition (R/M) model (to capture the complementary roles of recognition and metacognition in decision-making), it is a useful resource in the design of training for high-stress, high-stakes, high-novelty decision environments, such as AAW operations in a shipboard Combat Information Center (CIC). We have conducted two experimental tests of this training with experienced Navy officers. In this paper, we describe the R/M model, the training, and the results of our experiments.

2. The Recognition / Metacognition Model

The basic level of cognition in the R/M model is recognitional. Events are observed and they cue recall of related knowledge, goals, and plans [Neisser, 1976, Connolly and Wagner, 1988]. For example, an officer observes a pair of slow, low-flying aircraft. This activates a memory of intel warnings concerning the enemy's preparation of light aircraft for suicide missions against American warships. The pattern of events cues retrieval of a *situation template*, a cognitive model of goals, plans, and information relevant to the class of situations to which this one may belong: hostile intent scenarios. Many of the officers interviewed for the TADMUS project told stories of confrontations with potentially hostile aircraft. Thus, much of the current research focuses on such situations.

Situation templates categorize information and facilitate inferential reasoning. The issues that officers addressed in deciding whether a track was hostile and how to respond to it are captured in the "hostile-intent template" (see Figure 1).

When recognitional processes retrieve a situation template, the information compiled concerning the current situation instantiates it, transforming it into a *situation model* that represents the specific situation at hand. However, this initial situation model may be incomplete, laden with conflicting evidence, and/or contain unreliable assumptions. To handle these sources of uncertainty, the decision maker must do more than simply fill in the slots in the template. He or she must formulate a model that makes sense of the hostile intent assessment with respect to each piece of information. To do this, the decision-maker generates *arguments* that link slots containing evidence to the slot containing the assessment of intent. It is with arguments that officers build and defend their assessments and their plans. For example, the most plausible arguments concerning hostile intent show:

- how high-level goals of the enemy motivate an attack,
- how the enemy asset is a logical choice as an attack platform given the overall capabilities of the attacking country,
- how own ship is a logical target for attack given the enemy's high-level goals and any other potential targets,
- how the contact might detect own ship's location,
- how the enemy track's actions make sense as ways of getting to an attack position quickly and safely, and
- whether the outcome of an attack will fulfill the enemy's goals.

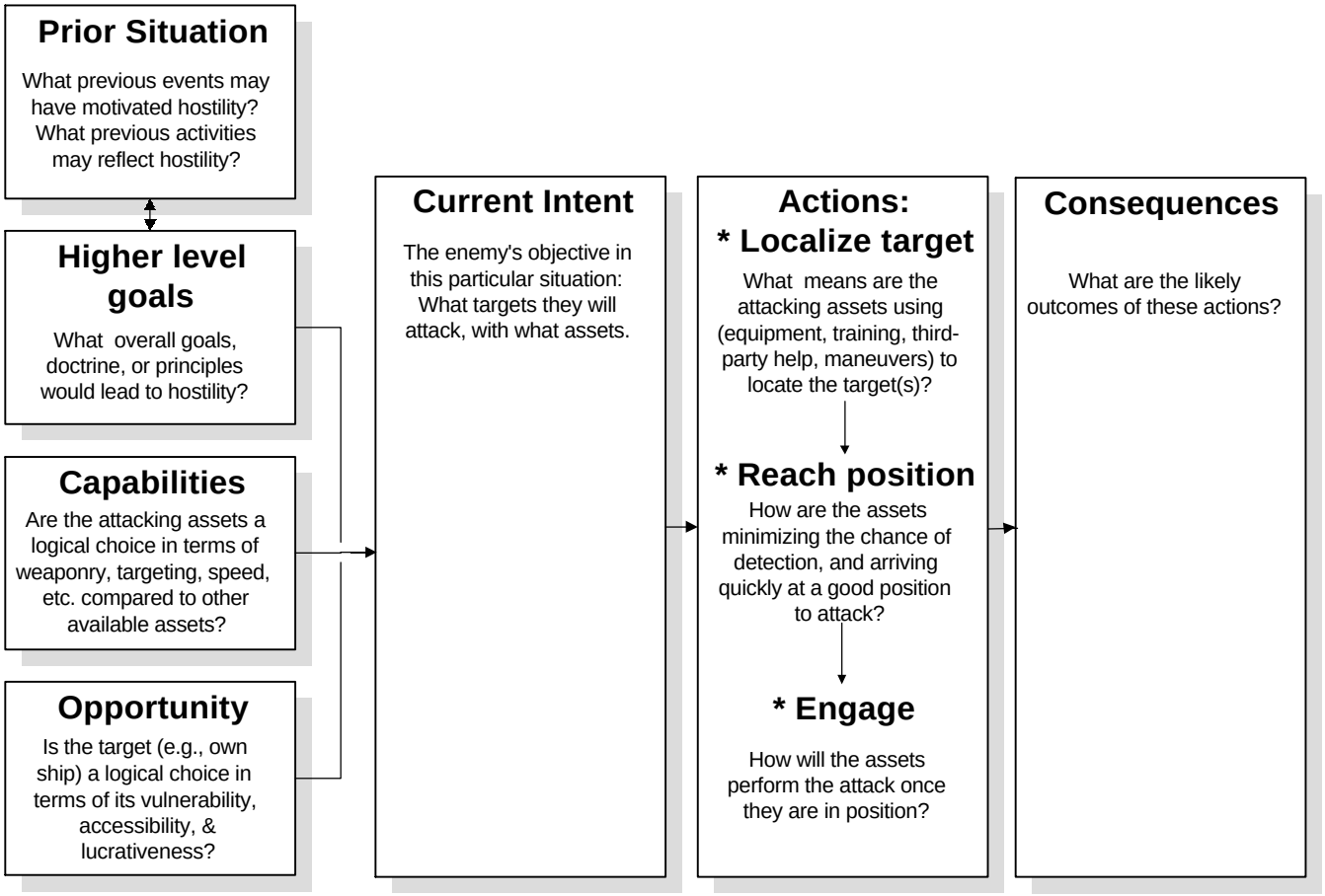

Figure 1. Template for constructing a hostile intent story. To assess whether a contact is hostile, the decision maker tries to fill in these boxes convincingly and generate plausible arguments to the intent slot from evidence in other slots.

Several metacognitive processes help the decision-maker to make sense of a situation, that is, to refine recognitionally instantiated situation models. *Critiquing* processes identify sources of uncertainty in the situation model, and *correcting* process attempt to reduce uncertainty. Three types of uncertainty, alluded to above, can weaken a situation model:

- When arguments are missing or cannot be completed for lack of evidence, the situation model is *incomplete*.
- When arguments are founded on implausible assumptions, the model is *unreliable*.
- When arguments from the same body of evidence support contradictory conclusions, the model exhibits *conflict*.

The specific skills with which decision-makers search for uncertainty (critique) and attempt to reduce it (correct) are common enough between domains that they have familiar names. However, their usefulness within any one domain hinges on the decision-maker's recognitional knowledge of that domain [Kuhn, et al., 1988]. Thus, these skills are termed *meta-recognitional* to denote that they manage or leverage recognitional knowledge.

In the AAW domain, critiquing is implemented by a repertoire of skills that includes using checklists such as the Rules of Engagement (ROE) to ferret out incompleteness, mental simulation to uncover conflict, and variants of the devil's advocate technique to identify unreliable assumptions.

Corrective techniques are of two general types: gathering information (from memory or from the world) and changing assumptions. However, both require specific knowledge of the domain. Officers must know from which systems, individuals, and texts they can collect needed information, and they

must understand the costs of these options. For example, scanning a radar display for tracks that potentially comprise a coordinated attack is a low-cost and rapid action, while illuminating an unresponsive, incoming track with fire control radar is somewhat time-consuming and may trigger an unplanned attack from a frightened, inexperienced pilot. In the absence of hard data, officers may also take the corrective action of dropping or making assumptions. Gathering information and changing assumptions can fill in gaps in argument structures (and evidence), produce more reliable assumptions, and resolve conflicting arguments.

Critiquing and correcting are not appropriate at all times in all situations. In a given situation, there may barely be time to implement a plan, and little or none to critique it. Alternatively, the cost of implementing the wrong plan may be so low that some aspect of the situation does not warrant time-consuming study. Finally, the situation and correct response may be thoroughly familiar, and thus critiquing the assessment or the plan is not worth the while (at least when more pressing decisions must be made). The *Quick Test* is the process responsible for testing for these conditions. When any one is true (time is short, stakes are low, *or* the situation is highly familiar), critiquing and correcting are suspended in favor of implementing the decision-maker's current plan. When all are false (time is plentiful, stakes are high, *and* the situation is sufficiently novel), then the Quick Test activates critiquing and correcting processes until the status of time, stakes, or familiarity changes.

In sum, the R / M model explains how experienced decision makers successfully apply recognitional pattern matching to situations that require flexible, adaptive thinking. Recognition primes the assessment of situations and construction of plans, meta-recognitional critiquing and correcting functions strategically apply recognitional knowledge to refine a situation model, and the Quick Test halts this thinking process when there is no opportunity or need to continue with it.

3. Training Based on the Recognition/Metacognition Model

We have developed and tested training that is based on the R / M model for the Navy [Cohen, et al., 1995, Cohen, et al., in preparation] and the Army [Cohen and Freeman, 1995]. The most recent version of the Navy training has four units. In the first unit, officers study a simple procedure for building situation models (which we simply call stories, in training). The *STEP* procedure consists of building a *S*tory that addresses most of the issues found in, for example, the hostile intent template; *T*esting that story to identify conflicts that arise from the evidence and resolve them, if possible; *E*valuating the assumptions on which the story is based; and formulating contingency *P*lans to protect against unreliable assumptions identified during the preceding steps. In the second unit of training, officers study and apply the hostile intent template, described above. The third unit presents a variant of the devil's advocate technique that is particularly useful for explaining conflict within a story and generating alternative interpretations of evidence that may improve, or at least inform, assessment and planning. The fourth and final unit describes how experienced officers apply criteria concerning time, stakes, and familiarity (i.e., the Quick Test) to decide when to halt critiquing and correcting and to implement their plans.

The training uses the rich stories gathered in TADMUS interviews as textual examples and exercises. In other exercises, and in tests, it capitalizes on realistic scenarios implemented on the DEFTT computer-based simulator (Decision-making Evaluation Facility for Tactical Teams). The scenarios focus on one of the most difficult AAW decisions: assessing the intent of an unknown air track. The scenarios are complex, the information provided is incomplete and often ambiguous, and the action is fast-paced. For example, officers are presented with many tracks to consider, the identity of almost all of the tracks is unknown, and some high speed air tracks change course unexpectedly and in a threatening manner. Each of the textual and simulator scenarios is designed to exercise a specific skill

taught in a single unit (such as Quick Test or evaluating assumptions), but there is ample opportunity in each scenario for students to integrate the skills they have learned in other units. In sum, students study cases based on real-life incidents in which air tracks exhibit threatening or suspicious behavior, they attempt to construct accounts of that behavior, they identify weaknesses in these stories, they attempt to generate other interpretations of the tracks' behaviors, and they decide when to stop this reflective behavior and act on the plans they have developed.

We would expect two types of effects from training of this sort. First, training should improve decision making processes. Trained officers should be better able to identify sources of uncertainty in their assessments, they should be better able to reduce uncertainty, and they should be able to generate more interpretations of a given body of information. Second, and more important, training should improve the decisions officers make. In particular, this training should enable junior officers to make assessments that are more like those of senior officers. Such outcomes would support the argument that the R/M model is psychologically valid, and that it is a useful foundation upon which to develop training in decision-making for officers who face high-stakes, high-novelty situations under severe time-stress.

The R / M training has been tested twice with large groups of experienced Naval officers, once at the Surface Warfare Officer's School (SWOS) in Newport, Rhode Island, in 194-95 and once at the Naval Postgraduate School (NPS) in Monterey, California, in 1996. Here, we will describe the two studies together in order to make the presentation concise. We note important differences between the experimental designs, subjects, materials and procedures where they occurred. We present findings from the SWOS experiment and provisional results of the NPS study.

4. Methodology

4.1 Design

The study at SWOS crossed two treatment conditions (training vs. control) and two test sequences (to counterbalance two test scenarios between pretest and posttest positions). All factors were between subjects. The study at NPS was also a pretest/posttest design with two counterbalanced scenarios, but did not employ a control group.

4.2 Subjects

Sixty officers at SWOS participated in the training (40 in the training condition, 20 as controls). These officers were being trained to serve as department heads in engineering, operations, or weapons. At NPS, 35 officers took part in a second training experiment. These officers were studying for advanced degrees in operations research and other fields. While officers at SWOS and NPS both had performed approximately 9.5 years of military service, they differed in several other respects.

- All of the officers at SWOS were in the Navy. Among officers at NPS, 51% were Navy, 14% were Marines, 29% were Army, and 6% were Air Force.
- Of the officers at SWOS, 92% had performed shipboard duty in the CIC, while only 46% of officers at NPS had worked in the CIC or in similar tactical positions in the Marines, Air Force, or Army.
- Approximately 32% of the officers at SWOS had served as Tactical Action Officer (TAO). The TAO is the second-in-command in the CIC, responsible for coordinating information sources and helping the commanding officer assess the air, surface and subsurface picture. Only 14% of officers at NPS had served as TAO.
- Among the officers at SWOS, about 33% had worked in AAW posts in the CIC, while only 14% of officers at NPS had done so.

4.3 Materials

The R/M training materials used in these experiments were a training text, brief explanatory lectures and exercises, as described above. All exercises were presented on paper to the officers,

at SWOS and executed by each class as a group. At NPS, four exercises were presented on the DEFTT simulator. Subjects executed these scenarios individually, while the remaining paper exercises were addressed in class groups.

The pretest and posttest in each experiment were DEFTT simulations punctuated by breaks in which participants responded to written questions concerning their assessments and plans. The DEFTT simulations began with oral and written briefings concerning the geopolitical context of the scenario and the military forces involved. Each participant then turned to a personal computer that simulated a command and display (C&D) station. The C&D presents symbology concerning the identity, speed, bearing and range of air and surface tracks, as well as textual details concerning these characteristics and the track's response to electronic interrogation (IFF). Virtually all tracks in these scenarios except own ship and accompanying surface craft were symbolically marked as unknown (rather than as friend or foe) and were unresponsive to electronic interrogation. Thus, subjects had to make assessments of track intent based on the behavior of tracks (e.g., their flight patterns), as well as audio communications that simulated internal and external comms. Most communications concerned the location of tracks, their presumed identity (e.g., F-4, Mirage), and electronic warfare (EW) data received from the tracks (such as search radar signals or fire control radar). For NPS officers, these communications were edited to so that they could be understood by non-Navy officers. For officers at SWOS, EW data was also provided on a large display in the center of the classroom.

Participants performed scenarios in groups of five or six, but each subject worked independently. Subjects could not consult with their classmates during tests nor take any actions that would alter events in the scenario for themselves and their classmates. Specifically, they could not maneuver own ship, fire weapons, interrogate tracks, or initiate communications during scenarios, though they could indicate their intent to perform such actions in response to questions during testing. In sum, the role of subjects during scenarios was to monitor events.

4.4 Procedure

For subjects at NPS, the experiment began with a lecture concerning the function of the CIC, the role of the TAO, and the operation of the DEFTT simulator. This lecture was not needed by the SWOS officers, whose knowledge of the domain was, in general, considerably deeper, and who had all executed DEFTT simulations before.

The pretest scenario was paused at three points. During each break, participants received a test booklet consisting of five questions. The questions asked participants to 1) assess the intent of a single track and defend that assessment; 2) defend an assessment of their own choosing that they did not favor; 3) critique and defend a weak assessment provided by the experimenter; 4) generate alternative assessments of that track; and 5) describe actions they would have taken immediately prior to the break. Subjects provided confidence ratings for most assessments. Participants did not know which track would be the focus of attention until after the relevant segment of the scenario was completed and test booklets were handed out.

Following the pretest, participants received training (unless they were SWOS officers in the control condition) and executed a posttest. In place of training, SWOS control subjects completed a psychological battery and discussed challenging problems in their jobs as weapons officers, engineers, or operations officers.

In addition, all subjects completed a biographical survey form concerning military experience, and all trained officers evaluated the instruction.

Testing procedures were quite similar for SWOS and NPS officers. However, training differed in several ways. Training at NPS utilized DEFTT exercises, as indicated above, and thus resembled the conditions of testing, while instruction at SWOS did not incorporate DEFTT. In addition, the time-span of the experiments differed. At SWOS, training and testing occurred in a single day: training was conducted in 90 minutes, preceded by a two-hour pretest and

followed by a two-hour posttest. At NPS, events were broken into five, two hour sessions on separate days over two weeks. The introduction to the CIC and DEFTT occupied the first session, the pretest occurred in the next session, two training sessions followed, and the experiment ended with the posttest in the final session.

5. Results

Two general classes of results were of interest in these studies. These were the effects of training on decision-making processes, and training's impact on the accuracy of decisions. Process issues concerned the effects of training on the identification of conflicting evidence, the handling of conflict, the generation of arguments more generally, and the generation of alternative assessments. These skills are all employed in critiquing or correcting, as posited in the R/M model. In addition, measures of the Quick Test process were examined.

Analysis of results from the NPS study began as this paper went to press. However, preliminary findings from that study are reported below. Where no results from NPS are reported, it indicates that the analysis had not been performed as of press time.

5.1 Effects of training on processes of decision-making

5.1.1 Identification of conflicting evidence

The training was in part designed to improve an officer's ability to identify evidence that conflicted with an assessment of a situation. One of the test questions specifically asked subjects to list evidence that conflicted with a weak assessment given by the experimenter.

For officers at NPS, training boosted the number of points of conflicting evidence identified by 58%, from 1.620 items on the pretest to 2.554 on the posttest ($t_{32} = 5.481$, $p < .001$). Among these officers, the greatest benefits of training accrued to those who needed it most. That is, officers who identified the fewest pieces of conflicting evidence on the pretest made the greatest gains on the posttest (Pearson's $r = -.444$, $p = .008$).[an05d, model2]

Among officers at SWOS, those who received training identified 52% more conflicting evidence than controls. Trained subjects identified 1.36 items per break, on average after adjusting for pretest score, while controls identified 0.897 items[2] . This effect was statistically reliable ($F_{1,55} = 6.236$, $p = .015$). Training tended to increase the number of points of conflicting evidence identified whether or not these participants happened to agree with the given assessment[3] . To the extent that there was any observable difference in the effect of training due to agreement, it had a slightly larger impact when participants agreed with the assessment. This is where the benefit was most needed to counter potential instances of confirmation bias [Kahneman, et al., 1982], in which officers do not fully consider evidence that might discredit assessments they favor.

5.1.2 Explanation of conflicting evidence

In situations characterized by ambiguity and complexity, it is often possible to find some information that conflicts with any assessment. Yet, one assessment is correct. Therefore, it is inappropriate to discard an assessment in the face of conflicting evidence, because this strategy would force one to reject all assessments. Instead, conflict can be considered a cue to think more deeply about aspects of a situation. Seemingly conflicting evidence may simply point to an exceptional circumstance. For example, an officer is considering an air track that is incoming at descending altitude and that does not respond to warnings. However, the track is moving too slowly to be a jet fighter. One exception condition that accounts for the slow speed of a hostile track is that the plane may be a light aircraft on a suicide mission. The same test question in which officers identified evidence conflicting with a weak assessment required them to consider exception

[2] Raw means were 1.342 for trained participants, and 0.933 for controls

[3] Agreement was determined by comparing the given assessment with assessments generated by the subject. If the given assessment was the same as the assessment that the subject generated and in which he or she was most confident, then the subject was scored as agreeing with the given assessment. Otherwise, the subject was scored as disagreeing with the given assessment.

conditions and use these to defend the assessment, despite its flaws.

Among officers at NPS, training boosted the number of explanations generated by 27%, from 2.566 on the pretest to 3.250 on the posttest (t_{32} = 4.920 , p < .001). Training had the greatest effect on officers who made the fewest explanations on the pretest (Pearson's r = -.535, p = .001)[an05I, model2]

SWOS officers generated 68% more explanations with training (.679 explanations per posttest break) than without (.400). However, variation within groups was quite large, and so this positive pattern was not statistically reliable.(t_{52} = 0.46, p = .643)[test.out]

5.1.3 Generating arguments

Arguments about evidence are the stuff of which assessments are made. Therefore, good training should improve officers' ability to generate arguments about assessments. We asked officers to generate their own assessments of a specified track at each break in the test scenarios and to present arguments in defense of those assessments. Specifically, subjects defended an assessment of their own creation that they preferred, and one of their own that they did not prefer.

For officers at NPS, training increased the number of arguments subjects presented in defense of their preferred assessments by 23%, from a mean of 3.389 per break on the pretest to 4.177 on the posttest (t_{33} = 3.807 , p = .001). Officers who generated the fewest arguments on the pretest benefited most from the training (Pearson's r = -.420, p = .012).[an05a, model4]

Among officers at SWOS, training increased the number of arguments generated by 6.5% for favored hypotheses and 8.6% for disfavored hypotheses. This positive trends, but not statistically significant ($F_{1,47}$ = 2.953, p = .092).

5.1.4 Generating alternative assessments

The training was designed to elicit not only deeper reasoning about any one assessment, an effect tested above, but also broader consideration of alternative assessments. Considering alternative assessments serves two functions. First, an alternative assessment may be better than the current hypothesis. Second, contrasting different assessments can reveal assumptions in the preferred assessment that would otherwise remain hidden.

The effect of training on officers at NPS was beneficial and statistically reliable. The number of alternative assessments generated on a given break rose 41%, from 2.689 on the pretest to 3.792 after training (t_{33} = 4.178 , p < .001). There was a nearly reliable trend for training to benefit most the officers who generated the fewest alternative assessments on the pretest (Pearson's r = -.319, p = .062).[an05c, model4]

There was a non-significant trend for subjects at SWOS to benefit from training. Officers who received instruction generated 9% more assessments per break (3.6, on average), than controls (3.3)(t_{59} = 1.498, p = .140).

5.1.5 Confidence in assessments

Given that training increased the number of assessments officers generated and their ability to find evidence conflicting with assessments, it is natural to wonder whether training might not weaken officers' confidence in their assessments, and whether this in turn might make them less decisive. However, the effects reported above can also be interpreted to mean that, with training, officers explored scenarios more deeply. Deeper understanding should not diminish confidence, and might even enhance it. This was one expectation of training.

As a metric of confidence, we took the difference between confidence ratings for the two assessments the subject generated in which he was most confident. This reflected the subject's ability to discriminate between the preferred assessment and another.

Among officers at SWOS, confidence ratings rose 12.5% with training. While this was not a statistically reliable increase, it indicates at the least that training did not reliably *lower* confidence.

Another indicator of confidence is the willingness of participants to engage a track. If training reduced confidence, trained subjects might be less likely to take such irreversible action.

In analyses conducted using data from officers at SWOS, engagements were rare among both trained and control participants. However, trained participants were twice as likely as controls to identify explicit contingencies (or tripwires) for engagement. Over both scenarios, 6% of the control participants developed contingency plans for engagement on each break, on average, but 13% of the trained participants did so. This difference was highly reliable ($F_{1,57} = 8.362$; $p = .005$). Thus, training may have increased decisiveness with respect to contingency planning for engagements.

5.2 Effects of training on decision outcomes

The findings above demonstrate that training based on the R/M model alters the ability of officers to generate, defend and rebut assessments. However, it does not speak to the accuracy of those assessments. Since the accuracy of assessments is assumed to influence the quality of plans and actions, the effect of training on assessment quality is a critical issue. As a preliminary step in this analysis, we examine whether training changed the types of assessments officers generated. It did.

The assessment in which each subject was most confident was categorized as either hostile, not hostile, or unknown. A loglinear model was then applied to test for non-independence of training with assessment category. For officers at NPS, training reliably affected the type of assessment subjects preferred ($\chi^2_6 = 24.05$, $p = .001$). This was the case for officers at SWOS, as well ($\chi^2_8 = 24.17$; $p = .002$).

We then evaluated the accuracy of assessments by 1) comparing them with the assessments of a subject matter expert (SME) and 2) by measuring consensus among subjects. We would expect consensus to grow if training helped officers to converge on a correct assessment. In addition, we examined whether the actions officers proposed to take changed with their assessments.[an03]

5.2.1 Accuracy of assessments

The standard for judging the accuracy of assessments by subjects in these experiments was the assessment of tracks at each break by the man who customized scenarios for us, a retired senior Navy officer. In both experiments, training produced large improvements in accuracy on one scenario, but no change in the other scenario. We report effects for the one scenario in each experiment that elicited effects.

Among officers at NPS, training boosted agreement with the SME from 60% on the pretest to 81% on the posttest ($\chi^2_2 = 6.791$, $p = .034$).[an03] Of the officers at SWOS, 77% of those who received training were in agreement with the assessments of the SME, compared with 43% of controls ($\chi^2_2 = 6.337$, $p = .013$).

5.2.2 Consensus

An alternative index of the accuracy is the level of consensus among subjects regarding their assessments. Training that improves accuracy should raise consensus among subjects as they converge on a common interpretation of events. We used as a measure of consensus a metric from information theory, called "average uncertainty," (e.g., Garner, 1962) which is defined as:

$$U(x) = - \Sigma\, p(x) \log (p(x))$$

Here, $p(x)$ is the relative frequency with which members of the group picked hypothesis x. U is zero when members of a group all agree, and grows larger with disagreement.

Training appeared to increase consensus among trained officers at both NPS and SWOS. Training lowered average uncertainty 41% among officers at NPS (i.e., it raised consensus), from $U(x) = 0.31$ to 0.22. Among SWOS officers, average uncertainty was 14% lower overall with training ($U(x) = 0.911$) than without (1.042).

5.2.3 Actions

Different actions should flow from different assessments of the intent of a track. Thus, to the degree that training affected assessments, we might also expect effects on whether and how subjects prosecuted tracks. This is precisely what we found among officers at SWOS in the scenario that elicited large training effects on assessment quality.

In that scenario, an Iranian helicopter calls May Day as its engines fail, and confirmation of the helicopter's crash is received from a nearby merchant vessel. Subjects are asked to evaluate

approaching tracks. Among the possible intents of those tracks are that they intend to participate in the search and rescue (SAR) operation, or that they plan to use the SAR operation as cover to close on own ship and attack. Controls were more likely than trained officers to assess tracks as hostile, and they took actions that reflected this, such as vectoring CAP and illuminating the tracks ($F_{1,28} = 2.635$, $p = .081$). Trained participants (like the SME) were far less likely than controls to regard the designated tracks as hostile, and were more likely to offer assistance in the search and rescue ($F_{1,28} = 3.382$, $p = .077$). In sum, training improved the accuracy of situation assessments, and officers' actions changed accordingly.

5.3 Subjective evaluations of training

The participants in both experiments provided us with quantitative and qualitative measures of the training.

During debriefing, subjects rated the training on a scale from 1 (strongly negative) to 5 (strongly positive). A rating of 3 indicated neutral. The average rating among participants at SWOS was 3.7. Twenty-six of the 40 officers in the trained group gave the instruction a positive rating (4). There were no strongly negative ratings (1), only four participants were negative (2), seven were neutral (3), and three gave it a strongly positive rating (5).

The average rating of training by officers at NPS was also 3.7. Twenty two of the 35 subjects rated the training as positive (4). One subject gave a rating of strongly negative, two marked it negative (2), six were neutral (3), two rated the course between neutral and positive (3.5), and 3 were strongly positive (5). NPS officers with some tactical experience in the Navy or other military services were more likely to rate the training positively (73% did so) than were officers with no such experience (58% did so).[an05h]

Qualitative evaluations of the training were also similar between the two experimental groups. Most participants found the training useful in solving the test problems and anticipated that it would be useful in the field. Participants said the training would help "organize what I have been doing previously and take it to another level," "stop me from making assumptions," "reinforce the concept that the obvious answer may not be the correct answer," and "keep tunnel vision to a minimum." Participants mentioned favorably the processes of organizing information in stories and using the devil's advocate to generate alternative interpretations of evidence.

6. Discussion

Training based on the Recognition / Metacognition model appears to enhance the ability of Naval officers to assess and act in situations characterized by complexity, ambiguity, high stakes, and severe time pressure. The training tested here enabled officers to critique assessments more effectively. Trained officers identified more evidence that conflicted with a given assessment. It did so whether the officers agreed or disagreed with the assessments in question. In addition, trained officers generated more arguments in defense of an assessment, and more alternatives to their preferred assessment. These effects did not diminish officers' confidence in the assessments they made, and it appeared to increase their tendency to take decisive action, such as setting tripwires for engagement. In addition, the assessments that trainees made agreed better with the views of the scenario designer, a retired senior Navy officer, and to reflect a greater degree of consensus within the group. Finally, the assessments appeared to influence actions in a consistent and appropriate manner. Trainees themselves approved of the training on the whole and perceived many of the same advantages detected in our analysis.

The benefits of this training were not restricted to the least experienced officers (such as those at NPS) or the most experienced (such as those at SWOS). The performance of both groups improved with training. However, larger benefits accrued to the officers at NPS, and among these subjects, those with the poorest pretest performance experienced the largest gains on some measures. These results suggest that training had a larger impact on less-experienced participants. However, two other factors may also account for these findings. First, the limited time available to

complete tests may have imposed a ceiling effect on more-experienced officers; they may have worked quite quickly on the pretest to record their relatively abundant thoughts and found that the posttest allowed them no more time to record the additional arguments and assessments they generated using newly acquired skills. Second, training at NPS may have been more effective owing to any or all of several unique characteristics of that experiment: the use of DEFTT simulations as instructional exercises, the increased time devoted to training, and spreading training over two weeks (rather than compressing it into a single day).

Further evaluation of this training may help to target the best methods of administering the training and evaluating officers' gains. However, the bottom line is that this training improved both decision processes and decision quality among officers with a wide range of tactical experience. This, in turn, supports the argument that the R/M model is a useful foundation upon which to construct decision training.

It is important to bear in mind that the decisions made in the CIC, and in many other military command environments, are often made by small teams consisting of an executive, his advisors, and technical specialists, or by interdependent teams such as the elements in an air wing. We have begun to explore the implications of the R/M model for the performance of teams such as these, and we are currently designing team training and experiments to test that training.

The R/M model may also be useful in the design of decision aids. For example, the situation template described above specifies information that is used by experienced officers and that might enhance a decision aid, such as evidence concerning the appropriateness of the attacking platform and its ability to localize targets. In addition, a decision aid based on R/M principles should draw the user's attention to conflicting cues, remind the user of exception conditions under which seemingly discrepant evidence is consistent with the current hypothesis, and highlight the assumptions implicit in a given assessment so that they can be evaluated. We are currently designing tests of these and other hypotheses.

In sum, the R/M model provides leverage for understanding and improving decision making. In future research, we will explore the implications of the model for team training and decision aiding.

Acknowledgments

We thank our sponsors, the Naval Air Warfare Center, Training Systems Division, of Orlando, Florida, for supporting this work. Dr. Janis A. Cannon-Bowers and Dr. Joan Johnston-Hall, at NAWC/NTSC, have made numerous insightful comments and have encouraged this research.

Steve Wolf of Klein Associates provided invaluable help in designing the DEFTT scenarios. His colleagues, Laura Militello and Becky Pliske also aided us in this project.

We are indebted to the staff of Sonalyst for their assistance. Al Koster modified DEFTT scenario scripts to meet our needs and served as a subject matter expert on matters large and small. John Poirier developed and presented training to NPS students concerning the CIC, the TAO's duties and DEFTT. Both of these men, and Robin Watters, also of Sonalyst, served as system administrators, operating the DEFTT system for us during training and testing.

References

[Anderson, 1982] Anderson, J.R. Acquisition of cognitive skill. *Psychological Review*, 89(4), 369-406, 1982.

[Cohen and Freeman, 1995] Cohen, Marvin S. and Freeman, Jared T. *Methods for Training Cognitive Skills in Battlefield Situation Assessment.* U.S. Army Research Institute, Ft. Leavenworth, KS. Arlington, VA: Cognitive Technologies, Inc., 1995.

[Cohen, et al., 1995] Cohen, Marvin S., Freeman, Jared T., Wolf, Steve, and Militello, L. *Training Metacognitive Skills in Naval Combat Decision Making.* Naval Air Warfare Center, Training Systems Division, Orlando, FL. Arlington, VA: Cognitive Technologies, Inc., 1995.

[Cohen, et al., in preparation] Cohen, Marvin S., Freeman, Jared T., et al. *Validation of a Method of Training Metacognitive Skills in Naval Combat Decision Making.* Naval Air Warfare Center, Training Systems Division, Orlando, FL. Arlington, VA: Cognitive Technologies, Inc., in preparation.

[Connolly and Wagner, 1988] Connolly, T. and Wagner, W.G. Decision Cycles. In R.L. Cardy, et al. (Eds.), *Advances in Information Processing in Organizations (Vol. 3)*. Greenwich, CT: JAI Press, 1988.

[Kahneman, et al., 1982] Kahneman, D., Slovic P., and Tversky, A. (Eds.). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge, UK: Cambridge University Press, 1982.

[Keeney and Raiffa, 1976] Keeney, R. and Raiffa, H. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs.* NY: John Wiley and Sons, 1976.

[Kuhn, et al.] Kuhn, D., Amsel, E., and O'Loughlin, M. *The Development of Scientific Thinking Skills.* San Diego: Academic Press, 1988.

[Neisser, 1976] Neisser, U. *Cognition and Reality. Principles and Implications of Cognitive Psychology*. San Francisco: W.H. Freeman and Company, 1976.

[Raiffa, 1968] Raiffa, H. *Decision Analysis: Introductory Lectures on Choices Under Uncertainty.* Reading, MA: Addison-Wesley, 1968.

I. [Rasmussen, 1981] Rasmussen, J. Models of mental strategies in process plant diagnosis. In Rasmussen & Rouse (Eds.). *Human Detection and Diagnosis of System Failures*. New York: Plenum Press, 1981.