The Impact of a Response Management Tool on Air Warfare Tactical Decision Making

Mark St. John (stjohn@spawar.navy.mil)
Pacific Science and Engineering Group, Inc., 6310 Greenwich Dr., San Diego, CA 92122 USA

Abstract

Responding appropriately to multiple, fast evolving, high risk situations is difficult. We investigated the design of a decision support tool called a Response Manager within the context of naval air warfare. Our question was whether support, in the form of presenting response options for consideration, should be generic to all aircraft or specifically tailored to different types of aircraft. Specifically tailored options limit clutter on a display, but perhaps at the price of constraining options and decision making to an inappropriately small set. In our experiment, air warfare-trained naval officers saw snapshots of air warfare situations consisting of a map of an airspace, detailed data about one aircraft, plus a set of response options. We varied the contents of the response sets. The results indicate that participants tended to give orders to execute responses that were congruent with the presented response set. Highly experienced officers showed as large an influence of response set presentation as did less experienced officers. Separate threat assessment ratings, however, indicated that the specifically tailored response sets were not influencing threat assessments; they were influencing response selection only. We concluded that the response manager works more like a memory aid, by orienting attention toward the presented response options, rather than by biasing situation interpretations.

Introduction

Automation is becoming increasingly common everywhere from manufacturing to military command and control. As a result, fewer people are doing more by becoming supervisors of automated systems. For these changes to be successful and safe, we must understand the cognition involved in supervision and naturalistic decision making, and we must design effective tools to support this cognition.

Supervision and naturalistic decision making are in many ways similar to problem solving in general (Klein, 1993). The situation is first encoded, then assessed, and then one or more responses are selected for execution. Decision making cycles through these steps continually as the situation evolves. Here, we have focused on supporting *response management*: how to choose the most appropriate response at each point.

Response management is made difficult by having to 1) monitor and respond to multiple simultaneous situations as they evolve over time, 2) assess and respond to highly ambiguous situations, 3) consider and choose among large numbers of response choices, and 4) cope with time pressure and high risk.

In our research, we have approached these issues from a behavioral perspective: how do different versions of a response management support tool influence actual decision making behavior? Naval air warfare, defending against threatening aircraft while maintaining safe passage for friendly and civilian aircraft, provides an excellent example of response management, and it serves as the focus for this research. First, many aircraft fly through an area simultaneously, and they range from helicopters visiting oil platforms, news helicopters, private planes, commercial airliners, maritime patrols, to military aircraft of many varieties. Managing multiple situations simultaneously is difficult because data must be quickly and accurately assessed and remembered for each situation. Failing to remember previous responses, current states, and previous assessments of each evolving situation can lead to disaster.

Second, because an aircraft's intent is invisible, a Navy ship's officers must rely on frequently ambiguous electronic and voice information, always ambiguous geopolitics, and good judgment for deciding how to handle each aircraft. Assuming the worst case and responding accordingly may actually cause a situation to deteriorate. Yet, waiting for clarifying information that may never come can allow a problem to develop beyond easy recovery. This ambiguity leaves the decision maker vulnerable to a host of decision biases.

Third, a Navy ship's Commanding Officer and its Tactical Action Officer, have at their disposal an array of response options, both defensive and offensive, that vary from benign to provocative to overtly hostile. The first response that leaps to mind may not be the best, yet culling through a large set of inappropriate responses takes time and effort better spent elsewhere.

Fourth, air warfare is clearly high risk, as the incidents involving the *USS Stark* and *USS Vincennes* have shown us. The unsuspecting *Stark* was struck by two Exocet missiles fired from an Iraqi jet, and the *Vincennes* shot down an Iranian airliner mistakenly believed to be an attacking Iranian F-14. Air warfare is also very time-pressured: only seven minutes elapsed between Flight 655's take off and its subsequent destruction (Rogers, Rogers, & Gregston, 1992).

A first generation response management tool was developed by the TADMUS project (Morrison, Kelly, & Hutchins, 1996) as part of a suite of air warfare decision support tools. The TADMUS response manager presented 1) a history of responses already taken toward each aircraft, 2) a list of responses for the decision maker to consider taking, and 3) the recommended (broad) ranges, between one's own ship and the aircraft in question, during which to take each response.

The TADMUS response manager was well received by naval officers who used it during realistic tactical simulations and by senior officers and defense analysts who observed demonstrations of the TADMUS system. However, a number of questions about both the content and form of the response manager were identified.

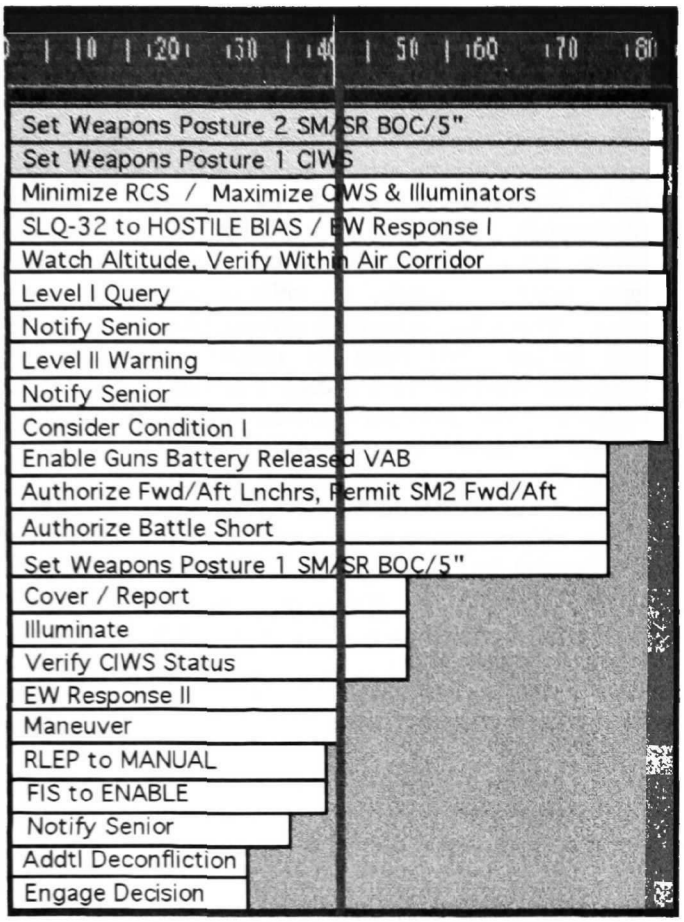

Figure 1. A response manager for air warfare. Numbers across the top indicate range from own ship. Bars indicate when each response is appropriate for execution. Greyed-out bars indicate responses that have already been taken.

Next generation: How specific? We imagined a next generation response manager that would provide more specific response recommendations as more became known about an aircraft. The response manager would start out by providing only a generic set of response options; then, as information about the aircraft's identity and intent accrued, the response manager could provide increasingly specific response sets and recommendations. We will refer to this response manager as aircraft-specific or Track-Specific.

One might well argue that experienced air warfare officers are expert decision makers who decide which responses to take on the basis of available information without regard to any response set, generic or specific. The basic ideas of naturalistic decision making theories (e.g. Klein, 1993) predict this result. The expert decision maker assesses the situation, matches the situation to a similar situation in memory, and retrieves an associated response. There is no place in this model for the response sets to operate. In this simple model the decision maker does not consider alternative options. The response sets, in this view, are superfluous, so any differences between sets would be moot.

On the other hand, one might just as well argue that since recognition is better than recall, laying out a set of appropriate response options is better than requiring decision makers to dredge them up from memory.

Further, it may be the case that more and less experienced decision makers utilize aids such as the response manager differently. Specifically, while more experienced participants might feel confident in disregarding a response set and following their own minds, less experienced participants might adhere strongly to the set of presented responses.

To answer these questions, we used a simple method in which naval officers with various levels of air warfare experience viewed and responded to "snapshots" portraying air warfare situations. Each snapshot contained a variety of information including a response manager that presented either a Track-Specific or Generic response set. If the presented response set influences decisions, then an aircraft-specific response set should produce different responses than a generic response set.

When does the effect occur? If the decision maker's responses vary, the next question is why? Does presenting aircraft-specific responses bias the decision maker to interpret an ambiguous situation in terms of the presented response set? Or does presenting aircraft-specific responses orient the decision maker's attention toward evaluating those responses, but not others, without actually altering the decision maker's interpretation of the situation? We address this question by asking participants to rate each aircraft's level of threat. If the threat ratings vary in the same way that the responses vary, then the presented response sets would be altering interpretations.

Which manager is best? Since the identification and intent of an aircraft can never be certain, response management requires reasoning under uncertainty. This uncertainty raises issues of strength and detail of recommendations, trust in automation, and the potential for errors due to decision biases. An aircraft-specific aid may lead a decision maker to over-interpret an ambiguous situation; then again, a generic aid might lead a decision maker to treat subtly different cases too similarly. An aircraft-specific aid might focus a decision maker's attention to too few or the wrong potential responses, but a generic aid might focus attention on either too many or the wrong responses.

To address these issues, we polled the participants for their preferences, and we marshalled a number of arguments based on their decision data. Based on information from a prior experiment, (St. John & Seymour, 1997), we also created a third version of the response manager in an attempt to combine the generality and prudence of the Generic manager with the relevance and specificity of the Track-Specific manager. The result was the Combination response manager.

The idea of the Combination response manager is to present the entire Generic response set, but to color-code

which specific responses are most appropriate for a given aircraft. Participants could use the color codes or examine the entire set.

Method

Participants The participants were 16 US Navy officers. We classified 8 of these participants as "more experienced" with air warfare and the Aegis air defense system. These officers had all had tours of duty as Tactical Action Officers or Commanding Officers. We classified the remaining 8 participants as "less experienced", though all had stood watch in the Combat Information Center aboard an Aegis equipped ship and were familiar with the Aegis system.

Materials The stimuli were a set of "snapshots" portraying air warfare situations during an imagined scenario in the Persian Gulf during a time of increasing tensions. The snapshots took the form of printed color screen images of experimental air warfare displays. Each snapshot contained a window displaying a map of the airspace surrounding own ship overlayed with aircraft symbols, rather like an air traffic control display. The snapshot also contained several other windows containing more detailed information about one aircraft, such as altitude, bearing, electronic warfare emissions, and a course history. One of these windows contained a response manager that displayed a set of responses, both defensive and offensive, that could be taken against the aircraft. Each snapshot indicated which responses had already been taken toward that aircraft by the point in time illustrated in the snapshot.

Five snapshots portrayed "assumed commercial" aircraft (two airliners and three helicopters), and three snapshots portrayed "assumed enemy" aircraft (one P-3, one F-4, and one unknown jet), all behaving *somewhat* provocatively. These identifications and categorizations to commercial or enemy are assumptions based on available (simulated) information such as altitude, speed, and radar emissions, and they are always uncertain to some degree. Instructions to participants emphasized that these aircraft identifications and classifications were based only on available data that they themselves could verify.

The experimental manipulation involved inserting different response managers into the snapshots. The same eight snapshots were used in all three conditions. In the Generic condition, all eight snapshots contained the same generic response set. This response set was developed by subject matter experts from battle orders taken from Navy ships. The Generic response set is shown in Figure 1.

In the Track-Specific condition, the five assumed commercial aircraft snapshots contained the Assumed Commercial response set: just the top five responses from the Generic set, all of which were defensive. The three assumed enemy aircraft snapshots contained the Assumed Enemy response set: all except the top five responses from the Generic set (See figure 1).

In the Combination condition, the Combination manager presented both the Assumed Commercial and Assumed Enemy response sets, but designated which set was more appropriate for a given aircraft. The assumed commercial responses were shown with a green background, and the assumed enemy responses were shown with a yellow background. A darker green and a darker yellow were used to indicate responses that had already been taken. The designation of which set was more appropriate for a given aircraft was indicated by text stating "Unknown Assumed Commercial" or "Unknown Assumed Enemy" and by coloring the text background either green or yellow. The designation allows the user to concentrate on the more appropriate actions while the presentation of the alternative set allows the user to consider alternative courses of action.

Procedure Participants were first oriented toward the information in the snapshots, including the response manager. They were also provided with a "mission statement" and Rules of Engagement for the operational context. They were then instructed about their task of evaluating snapshots and making response decisions. They were told that the responses presented in the response manager were merely suggestions and that they were free to give orders to execute any response from the set, off the set, or make no response at all—whatever they felt was most appropriate.

All eight snapshots of each condition were shown sequentially, in the same order for each condition. Participants saw the conditions in different orders. Each snapshot was shown for 15 seconds, or until a response was given. Participants also rated each aircraft's level of threat on a seven point scale, where 1 meant the aircraft was "neutral" or no threat and 7 meant the aircraft was a definite threat.

Participants' responses were scored on a scale of 0 to 5, according to their location on the Generic response set (See Figure 1). Responses residing near the top of the set are basic defensive and monitoring actions, so they received a score of 0. Responses further down the list are more aggressive, and they received progressively higher scores.

Participants occasionally gave no order to execute any response in other words, the participants decided to "wait and see". In these cases, the responses on the snapshot shown to have been taken already were scored. When a participant gave orders to execute more than one response, the highest scoring response was used.

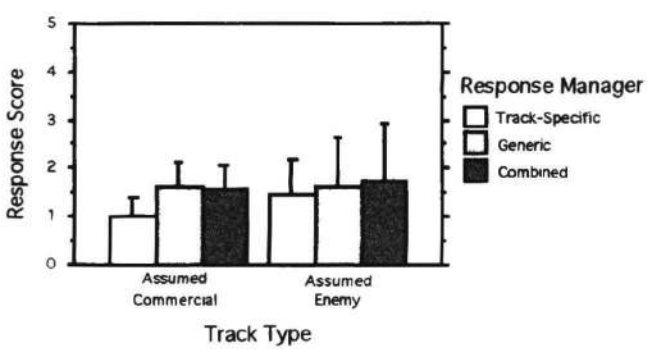

Figure 2. Averaged response scores for each version of the response manager. Error bars indicate one standard deviation.

Results

Response scores Figure 2 shows the averaged response scores. The response scores were submitted to a 3x2x2 repeated measures ANOVA. Response set (Generic, Track-Specific, Combination) was a within participant variable; aircraft-type (commercial, enemy) was a within participant variable; experience (more, less) was a between participant variable.

Overall, the different response sets elicited different response scores: $F(2, 28) = 10.3$, $p < .001$. This main effect of response sets is more clearly understood in terms of the significant response set by aircraft type interaction: $F(2,28)=4.1$, $p=.03$. For assumed commercial aircraft, responses varied depending on which response manager was presented: $F(2, 28) = 17.2$, $p = .0001$. For assumed enemy aircraft, however, responses did not vary: $F(2, 28) = 1.8$, $p = .71$. In a prior experiment (St. John & Seymour, 1997), the response presentation effect obtained for both commercial and enemy aircraft.

The large response presentation effect found for assumed commercial aircraft is most likely due to the large difference between the Track-Specific commercial response set and the Generic response set. The commercial set contains only five responses, and all of the responses were assigned a response score of 0 because they come from the very top part of the generic set.

We found no difference between more and less experienced participants: $F(1, 14) = .003$, $p = 1.0$. This result means that the more experienced participants were just as influenced by the presentation of different response sets as were the less experienced participants. This finding is somewhat surprising, since one might well think that more experienced participants would weigh their own ample experience heavily and pay little credence to the response manager. Nonetheless, we found the same null result in a prior study (St. John & Seymour, 1997).

One explanation is that captains and commanders might be "rusty" on the fine details of air warfare and appreciate the support that the response manager provides. Working against this idea, however, is the fact that even a current Tactical Action Officer on an Arleigh Burke class destroyer (top of the line) showed strong response presentation effects. We conclude that experience really does not reduce the influence of response presentation.

Finally, how did the new Combination response manager compare with the Track-Specific and Generic response managers? Had participants used the color coding to focus selectively on the suggested commercial or enemy subset, then the Combination response manager would have elicited responses similar to the Track-Specific response manager. The results, however, indicate that the Combination response manager elicited responses more like the Generic response manager. Though we believe it is unlikely that participants entirely ignored the color coding and the suggested subset, participants were clearly less impacted by this color-coding presentation scheme than by the present/absent presentation scheme of the Track-Specific response manager.

Preference ratings All participants reported that they remembered the individual snapshots and their responses to them from one condition to the next. Nonetheless, all participants reported re-analyzing the situations, and of course the results indicate that different orders were ultimately given.

Most reported interpreting the response sets as being based on a ship's Battle Orders, and treating the response sets as "serious suggestions" or "recommendations", though some reported treating the sets as simple "memory joggers".

Ten of the participants were asked which of the three response managers they would prefer to use while standing watch. Seven of the ten participants preferred the Combination response manager and three preferred the Generic response manager. No participants preferred the Track-Specific response manager. Apparently, the benefits of specificity provided by the Track-Specific manager, that some participants preferred in a prior study, were co-opted by the Combination manager which simultaneously benefits from showing all responses. Meanwhile the Generic manager was again preferred by those participants who distrusted any suggestions from the response manager as potentially biasing their decision making.

Threat ratings Participants' threat ratings were averaged and submitted to a 3x2x2 ANOVA. The mean ratings are shown in Figure 3. The most important question is whether the threat ratings show the same response presentation effect as did the response scores. For the threat ratings, aircraft type and response manager did not interact: $F(2, 28) = .7$, $p = .48$. Especially for assumed commercial aircraft, where the response presentation effect was pronounced for response scores, no threat rating differences were found. Based on this null effect, we argue that the threat assessment process is not affected by the content or form of the response manager.

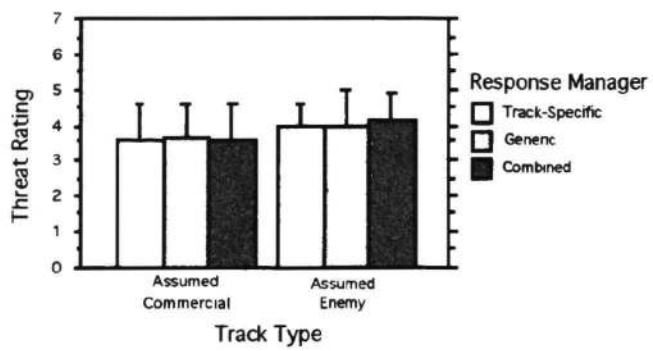

Figure 3. Threat ratings.

Discussion

There were two major questions posed in this research: 1) how specific to different aircraft should an air warfare response manager be, and 2) how does response management work and how might a support tool enhance it?

Design of the Response Manager

Determining which response manager is best is a difficult but important question. First, the preference data were strong but not overwhelming: two out of three participants preferred the Combination response manager. The Combination response manager benefited from the generality of a complete set of responses, as well as the relevance of having an assigned specific set that was more appropriate.

The unbiased nature of the Generic manager, however, remained an important benefit for some participants. Though no participants disagreed with the assignments of the snapshots, these participants described themselves as more skeptical of the response manager and more cautious against falling prey to either their own or the system's biases. Their opinion was that the Track-Specific and Combination managers drew conclusions and classified aircraft into assumed commercial or assumed enemy too early; they preferred to "stay on the fence" longer.

Of course this issue of when the manager classifies aircraft will ultimately be under the decision maker's control. The classification could be performed manually by either the decision maker or a subordinate, or the classification could be performed automatically by the system, but the decision maker could tune the system to wait for higher certainty before making any assignments.

But preference data can only take us so far. What does the participants' *behavior* tell us? First, it tells us that choosing a response manager truly matters because it affects the responses that even expert decision makers order for execution. The Track-Specific manager elicited less assertive—more hands off—responses toward the assumed commercial aircraft than did the other managers.

One might have argued that this influence is simply due to the category assignments influencing how the participants perceived the level of threat of each aircraft. This argument, however, is not supported by the data. First, both the Combination response manager and the Track-Specific response manager made the same category assignments, but the response presentation effect did not obtain with the Combination response manager. Therefore, the assignments alone were not sufficient to influence the participants' decisions. Second, and most compelling, the actual threat assessments did not vary among managers. Therefore, the response manager is not manipulating threat assessments; it is manipulating the later decision making stage of response selection (See figure 4). The response selection nature of the presentation effect suggests that the effect was more a result of the availability for evaluation of certain potential responses rather than a result of interpretation biases.

This interpretation also fits with our observation that participants frequently ran through the visible response set – sometimes using a forefinger – deciding as they went, which responses to order for execution. Under this interpretation, the response manager acts more as a memory aid than as a recommendation or decision aid. In Distributed Cognition and Ecological Psychology terms (e.g. Gibson, 1979, Hutchins, 1995), the presented response set *affords* orienting to each response in turn and evaluating it for execution.

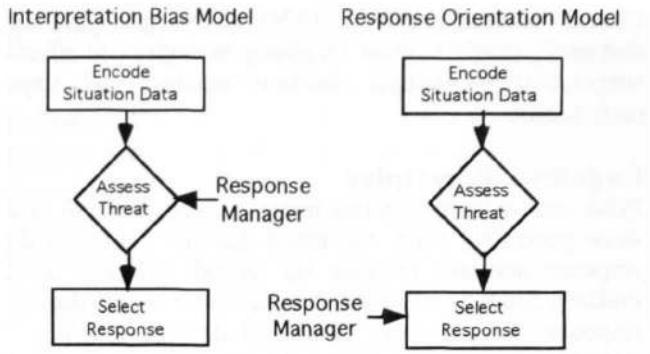

Figure 4. Two possible models of when the response presentation effect occurs: early, during the threat assessment stage, or late, during the response selection stage.

But what does this influence tell us about which version of the manager is better? First, since the Combination manager shows the same classifications and response sets as the Track-Specific manager, then users could order for execution the same responses. The fact that they did not—they responded like users of the Generic manager—implies they *chose* not to. This implication that users choose not to when they don't have to suggests that it is the Track-Specific manager that limits or biases responses. This limitation occurs because the Track-Specific manager presents fewer response options for consideration. When more responses were shown, either in the Combination manager or in the Generic manager, users availed themselves of the opportunity and ended up choosing differently.

If we are correct that the Track-Specific manager limits choice by orienting the user to fewer response options, then we should rule out the Track-Specific manager. That leaves us with the Generic and Combination managers. The Combination manager was preferred by most participants. More importantly, it provides more information. Especially with the ability to tailor the classification function, the Combination manager seems the superior choice.

What would be especially impressive is if we could show that even though participants using the Combination or Generic response managers chose similar responses, the participants using the Combination response manager were more thoughtful or faster or less error prone. Any such demonstration awaits further investigation.

Beside choosing a manager, an equally important issue derives from the finding that participants tend to follow the response sets. In cognitive terms, the response sets orient the user to those responses and make the presented responses more mentally available—and in turn more frequently chosen. We can use this finding to improve decision making skill, regardless of the particular response manager we choose, by including in the response sets additional options for users to notice and perform, such as a) additional deconfliction responses, and b) actual Critical Thinking strategy actions, such as "consider alternative explanations for aircraft behavior" and "consider outcomes". While these examples may appear clumsy or hokey, they may prove to be valuable reminders for critical thinking.

Cohen, Freeman, and Wolf (1996), for example, have shown that easily taught Critical Thinking strategies can effectively supplement naturalistic decision making and improve performance.

Cognitive Principles

What can we learn from this research about decision making more generally? First, we found that the responses that a response manager presents do indeed influence decision making. Second, we found that this effect occurs during the response selection stage of tactical decision making. The response presentations affected the responses made, but they did not influence the threat ratings.

Third, since no effect was found when participants used the Combination manager, which presented all the responses but labeled some as more appropriate, we concluded that the presentation effect was due to which responses were made available to the participant for consideration rather than which responses were labeled as more appropriate by the response manager. It's the presentation that mattered; not the manager's recommendations *per se*.

The response presentation effect should transfer to other response management domains. What this means is that making available to decision makers a list of response options will affect their decisions. The decision makers' actual assessments of the situation will not change, but presenting responses will orient users' attention toward those responses. they will receive first or most consideration, and they will be preferentially selected.

Acknowledgements

This research was sponsored by the Office of Naval Research through the Space and Electronic Warfare System Center, San Diego. George Edw. Seymour, the project manager, Richard T. Kelly, Ronald A. Moore, and Robert Farrington provided valuable comments and technical information.

References

Cohen, M. S. (1993). The naturalistic basis of decision biases. In G. A. Klein, J. Orasanu, R. Calderwood, & C. E. Zsambok (Eds.), *Decision making in action: Models and methods*. Norwood, NJ: Ablex.

Cohen, M. S., Freeman, J. T., & Wolf, S. (1996). Metarecognition in time-stressed decision making: Recognizing, critiquing, and correcting. *Human Factors, 38*, 206–219.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton-Mifflin.

Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: MIT Press.

Klein, G. A. (1993). A recognition-primed decision (RPD) model of rapid decision making. In G. A. Klein, J. Orasanu, R. Calderwood, & C. E. Zsambok (Eds.), *Decision making in action: Models and methods*. Norwood, NJ: Ablex.

Morrison, J. G., Kelly, R. T., & Hutchins, S. G. (1996). Impact of naturalistic decision support on tactical situation awareness. *Proceedings of the Human Factors and Ergonomics Society 40th Annual Meeting*. Santa Monica, CA: HFES. 199–203.

Rogers, W., Rogers, S., & Gregston, G. (1992). *Storm center: The USS Vincennes and Iran Air flight 655*. Annapolis, MD: Naval Institute Press.

St. John, M. & Seymour, G. E. (1997). Do displays of tactical response options influence expert decision making? In *Proceedings of the 41st Annual Meeting of the Human Factors and Ergonomics Society*. Santa Monica, CA: HFES. p. 1406.