DOT/FAA/AM-97/22

Office of Aviation Medicine
Washington, D.C. 20591

The Role of Memory in Air Traffic Control

Scott D. Gronlund
Michael R. P. Dougherty
Daryl D. Ohrt
Gary L. Thomson
M. Kathryn Bleckley

University of Oklahoma
Norman, Oklahoma 73019

Dana L. Bain
Faith Arnell

Federal Aviation Administration Academy
Oklahoma City, Oklahoma 73125

Carol A. Manning

Civil Aeromedical Institute
Federal Aviation Administration
Oklahoma City, Oklahoma 73125

November 1997

Final Report

This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161.

U.S. Department of Transportation

Federal Aviation Administration

19980406 041

DTIC QUALITY INSPECTED 3

NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents thereof.

Technical Report Documentation Page

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/AM-97/22		

4. Title and Subtitle	5. Report Date
The Role of Memory in Air Traffic Control	November 1997
	6. Performing Organization Code
7. Author(s)	8. Performing Organization Report No.
Gronlund, S.D., Dougherty, M.R.P., Ohrt, D.D., Thomson, G.L., and Bleckley, M.K.[1]; Bain, D.L. and Arnell, F.[2]; and Manning, C.A.[3]	
9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
[1]University of Oklahoma, Department of Psychology, Norman, OK 73019 [2]FAA Academy, P.O. Box 25082, Oklahoma City, OK 73125 [3]FAA Civil Aeromedical Institute, P.O. Box 25082, Oklahoma City, OK 73125	11. Contract or Grant No.
12. Sponsoring Agency name and Address	13. Type of Report and Period Covered
Office of Aviation Medicine Federal Aviation Administration 800 Independence Ave., S.W. Washington, DC 20591	14. Sponsoring Agency Code

15. Supplemental Notes

This research was supported by Contract #DTFA-02-93-D-93088.

16. Abstract

We tested air traffic controllers currently serving as instructors and tried to manipulate their memory for various aircraft flight data. In Experiment 1, the *amount* of control exercised (the number of control actions or communications) had little effect on memory for flight data, although we did find excellent memory for the position of aircraft on the radar display. We argued that this was the basis for the mental representation of the aircraft in the sector and may serve as the foundation for situation awareness. In Experiment 2, neither the *type* of control exercised nor the *importance* of the aircraft in the scenario consistently affected memory. We considered several reasons why we were unable to manipulate memory for flight data, including how important memory is to successful task performance and whether we tapped the relevant characteristics of the situation. Resolution of these issues will contribute to improved techniques that assess situation awareness from memory performance.

17. Key Words	18. Distribution Statement
Air Traffic Control, Situation Awareness, Memory	Document is available to the public through the National Technical Information Service, Springfield, Virginia 22161

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	25	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

ACKNOWLEDGMENTS

We thank Jenny Perry for her help with Experiment 2. We are also grateful to Dick Pollock for his cooperation. We appreciate the improvements to this manuscript that resulted from the comments of Bob Blanchard, William Collins, Shelia Kennison, Scott Mills, Mark Rodgers, Dave Schroeder, and Mike Wayda. This research was supported by Contract #DTFA-02-93-D-93088 to the first author.

The Role of Memory in Air Traffic Control

With the rapid advance of technology, complex dynamic systems have evolved that tax the cognitive abilities of their human operators. In the en route air traffic control (ATC) environment (involving the high-speed and high-altitude cruise between takeoff and landing), the complex dynamic system that confronts the air traffic controller is comprised of a large number of aircraft coming from a variety of directions, at diverse speeds and altitudes, heading to various destinations. Like most complex, dynamic systems, this one cannot be periodically halted while the controller takes a brief respite. The ability to remain in control of such a complex, dynamic system requires that the controller maintain situation awareness (SA).

According to Dominquez (1994), SA involves the continuous extraction of environmental information, the integration of this information with prior knowledge to form a coherent understanding of the present situation, and use of that coherent understanding to direct perception and anticipate future events. The three levels of Endsley's (1995a) model of SA parallel this definition. Level 1 involves the perception of elements in the current situation. Level 2 involves the comprehension of that current situation; controllers refer to this as *getting the picture*. Level 3 involves the projection of the current situation into the future.

There is currently no agreed-upon methodology for measuring SA. Endsley (1995b) critically reviewed various methods, including physiological techniques, performance measures, and subjective techniques. The most commonly used method, according to Adams, Tenney, and Pew (1995), is the query technique (e.g., Endsley, 1987; Marshak, Kuperman, Ramsay, & Wilson, 1987). In this technique, the task simulation is suspended, the system displays are blanked, and the participant answers a series of questions about the situation.

Query techniques tap what the participant can recall from memory. According to Endsley (1995b), "SA, composed of highly relevant, attended to, and processed information, should be most receptive to recall." Endsley believes that the vast majority of a participant's SA can be assessed in this manner. Irrespective of the exact correspondence between SA and memory, it is requisite to understand more about the role of memory in air traffic control. Only then can we clarify the correspondence between memory and SA.

The relationship between memory and air traffic control is currently unknown (Mogford, 1994; Rantanen, 1994). Data and opinions about the importance of memory to controlling air traffic run the gamut. Bisseret (1971) found that highly skilled controllers had better recall for aircraft data than average controllers. On the other hand, Stein and Garland (1991) observed that controllers need not process information as thoroughly as it might appear: Because of their extensive knowledge base, the information typically matches their expectations (Rantanen, 1994). This might mean that memory is necessary only to the extent that the information derived from knowledge structures contradicts the current situation. Sperandio (1978) observed that controllers dealt with an increasing workload by changing their operating strategies. They became increasingly selective of the information they processed, which allowed them to deal with only the most relevant information about an aircraft. Hopkin (1980) argued that *forgetting* information may be just as vital a skill as remembering it. He observed that, in a dynamic memory situation like air traffic control, the information to be remembered changes so frequently that it may in fact be to the controller's advantage to be able to forget the previous altitude for an aircraft, or it might interfere with memory for the *n*th (current) altitude.

Means et al. (1988) conducted one of the few studies to empirically examine the role of memory in air traffic control. Means et al. studied three expert air traffic controllers. After controlling traffic for a period of time, the controllers completed a traffic drawing task in which they indicated the location of each aircraft on a paper copy of the sector map (see also Vortac, Edwards, Fuller, & Manning, 1993). Controllers performed exceedingly well on this task, correctly recalling upwards of 90% of the aircraft and

correctly placing about 95% within 10 nautical miles of their actual positions. The ability to position the aircraft on the sector map stood in marked contrast to the recollection of many details regarding the aircraft. Means et al. found that controllers, when cued with the call sign, recalled only 28% of the aircraft types and only 6% of the ground speeds. Controllers obviously have excellent memory for some information (position on the Planned View Display or PVD) and poor memory for other information. What variables affect memory for various pieces of information?

Means et al. (1988) proposed two hypotheses regarding what information controllers remember. One hypothesis was that the probability of recalling information about an aircraft was related to the *amount* of control exercised on the aircraft. This was operationalized as the number of control actions directed to a particular aircraft. There is ample support in the memory literature for the positive effect of frequency and repetition on memory (see Anderson, 1995). Means et al. (1988) found that twice as much flight data was recalled about "hot" aircraft (defined as aircraft for which controllers "exercised a great deal of control") than "cold" aircraft. We operationalized *amount* of control in two ways: 1) by the number of interactions with an aircraft, and 2) by the number of control actions taken. An interaction was defined as any communication with an aircraft that did not result in a change to the aircraft's flight data; control actions were defined as any interaction that resulted in a change to the aircraft's altitude, speed, or heading. The second hypothesis was that the *type* of control exercised was related to the information recalled. For example, vectoring an aircraft was found to lead to better retention of its routing information.

It is important to reveal which variables lead to good recall of flight data because that would lead to refined use of the query technique to measure SA. For example, it may be unreasonable for controllers to remember the same information about all aircraft. Furthermore, to not remember the altitude of a "hot" aircraft might be of greater concern, and indicate poorer SA, than not remembering the altitude of a "cold" aircraft.

Experiment 1

Is *amount* of control the causal factor affecting the recallability of flight information, as Means et al. (1988) suggest? To answer this question, we manipulated the number of interactions and the number of control actions to produce four experimental conditions, denoted: Control3, Control1, Interaction3, and Interaction1. Control3 aircraft received three control actions, Control1 aircraft received one control action, Interaction3 aircraft received three communications, and Interaction1 aircraft received one communication.

In the Control3 condition, the pilot might request an altitude change to 10,000 feet, then to 12,000 feet, and finally to 12,500 to get above a layer of clouds. In the Interaction3 condition, the pilot might report light chop (turbulent air), later asks if there have been other reports, and finally report that it has smoothed out. Although the controller need not attend to any flight data, we thought that this communication would at least highlight the altitude information for the controller. This was informational for the controller because no control actions were warranted. In the Control1 condition, the pilot might request one altitude change. In the Interaction1 condition, the pilot might establish communication with the controller by reporting on at flight level 220 (22,000 feet).

We predicted that controllers would recall more about the Control3 and Interaction3 ("hot") aircraft than about the Control1 and Interaction1 aircraft ("cold"). In addition, performance in the Interaction3 condition might be better than Control3 because the same altitude was interacted with three times for the Interaction3 aircraft, but three different altitudes had been assigned to the Control3 aircraft. On the other hand, performance in the Control3 condition might be better than in the Interaction3 condition because the controller would have to expend more cognitive effort to make sure the requested control action did not conflict with other aircraft.

In Experiment 1, we focused on altitude information because we knew it was important (Leplat & Bisseret, 1966) and we knew it was not remembered so

well that we might have a problem with a ceiling effect (e.g., PVD position). We added one more condition to begin to test Means and associates' (1988) second hypothesis—that *type* of control affected what was remembered. Aircraft in the Traffic condition were put into conflict (*a priori*) with other aircraft. For half of the Traffic aircraft, altitude was the relevant factor that put the aircraft in conflict. For the remaining Traffic aircraft, the aircraft were in conflict for other reasons (e.g., one aircraft overtaking another and both landing at the same airport—controller will probably use speed adjustment or vectoring to resolve the conflict). The former was the Traffic-Relevant condition and the latter was the Traffic-Irrelevant condition. We expected that the altitude of an aircraft would be better remembered in the Traffic-Relevant condition because the altitude control action was relevant to the resolution of the conflict.

Method

Participants. Eighteen full-performance level (FPL) en route air traffic controllers participated. They had been FPL controllers for an average of 12.4 years. They last worked in the field an average of 3.5 years before, with a range of 1.6 to 6 years. All participants were air traffic control instructors at the FAA Academy and were familiar with the AeroCenter airspace used in the experiment.

Materials. The experiment was conducted at the Radar Training Facility (RTF) at the Mike Monroney Aeronautical Center in Oklahoma City, Oklahoma. The RTF provides high-fidelity training simulations using the fictitious AeroCenter airspace. Communications between the controllers and the aircraft take place in the same manner as in the field, although the aircraft were "piloted" by ghost pilots who controlled the simulated aircraft based on the controller's instructions.

The equipment consisted of the radar display (the Planned View Display or PVD), a keyboard and trackball, and a computer readout display (CRD). The PVD shows the 2-D location of the aircraft with an attached data block containing information including the aircraft's call sign, altitude, and ground speed. In addition, a flight progress strip (FPS) for each aircraft was stacked vertically in a strip bay adjacent to the radar display. Flight strips are 20 x 3 cm rectangular paper strips. Participants had one for each aircraft on the radar display. The FRSs have 31 fields of information with the call sign, aircraft type, requested altitude, requested speed, route of flight, etc. The controllers mark on these strips to update this information. In addition, flight data can be referenced on the CRD.

Participants worked the R-side, or radar position. Our SME (Subject Matter Expert) worked the Radar Associate's position and performed all its normal functions (strip marking, communicating with other centers, serving as a second pair of eyes to aid the radar controller). The experiment did not require any deception on the part of the SME; in fact, the integrity of the experiment required that the participant rely on the Radar Associate for reliable information. In addition, providing a Radar Associate allowed us to measure what the participants *could* remember, as opposed to overloading them and measuring what they could *not* remember.

Three high-complexity, 30-minute scenarios were developed with the help of the SME. They were designed around the constraints necessary to test the hypotheses of interest, yet were required to be as realistic as possible. We relied on the judgment of our SME regarding the appropriate level of complexity; there is no agreed-upon, objective method for measuring complexity. The scenarios included a mean of 28.7 aircraft, 9 of which were overflights (not taking off or landing in the sector), 8.7 were arrivals, and 11 were departures. On average, there were 13 aircraft displayed simultaneously.

Procedure. The participants completed a set of sample questions prior to beginning the experiment. They were told that the scenarios would be stopped periodically and that they would be asked questions about various aircraft. However, we did ask them to control traffic as they normally would because that would be most beneficial to us.

The experiment began with the SME working the first minute of the scenario and then giving a position-relief briefing to the participant. During the position-relief

briefing, responsibility for the sector was transferred from one controller (the SME) to another (the participant).

Three times during the 30-minute scenario, at approximately 10-minute intervals, the scenario was paused and the participant was turned away from the radar display and strip bay to complete two tasks. The first task was Map Recall, for which we provided a paper copy of the sector map (no aircraft present). Participants placed an "X" at the location of each aircraft at the time the scenario was paused, and wrote down the call sign or any other identifying information. After they recalled all that they could, they had to "circle the planes that you would consider a group and tell us why they went together." Map Recall was videotaped.

After completing Map Recall, participants moved to the computer to answer a battery of questions about various aircraft. A paper copy of the sector map was provided, which contained all the aircraft in the sector at the time the scenario was paused. The call signs were included because controllers do not generally remember the call signs very well.

Three types of questions were asked about a given aircraft, in the following order: 1) informational—what was American 123's (AAL123's) altitude (or ground speed, route, destination, departure point, or aircraft type); 2) metamemorial—rate your confidence in your answer (a range from 0—absolutely no idea, to 100—absolutely certain); 3) source—do you *remember* this information (memory was the source) or do you *know* it (answer was based on past experience). An example was provided: they might *remember* (type 'r') the aircraft type of AAL123, but they might *know* (type 'k') that Southwest 456 was a Boeing 737 because all Southwest aircraft are 737's.

Questions regarding altitude were of primary interest. They made up one-third of all informational questions. Questions on other flight data were included to discourage the participants from unduly focusing on altitude. The questions regarding altitude were phrased so that it was unambiguous what information was requested (assigned altitude, requested altitude, current altitude). We always asked about the altitude information that was considered most relevant at the time the scenario was paused. For example, if an aircraft was climbing, it was more important to know its assigned altitude than its current altitude. Inadvertently, two altitude questions did not specify which type of altitude was being requested. For these, we counted either the assigned or the current altitude as correct. After completing the battery of questions, participants were allowed as much time as they wanted before resuming the scenario.

Five aircraft were selected in advance. The participants did not know which aircraft (out of an average of 13 on the radar display) would be queried. Of these five aircraft, three were from one of the five conditions of experimental interest: Traffic, Control3, Interaction3, Control1, and Interaction1. Two were filler aircraft included to disguise the experiment. The Traffic, Control1, and two filler aircraft were present in each 10-minute interval. The other three conditions occurred once per scenario, each in a different 10-minute interval.

For the Control3 aircraft, the pilot made three requests that would result in control actions, and those requests were separated by approximately three minutes. This was also true for the three interactions in the Interaction3 condition. The control action required of the Control1 aircraft was scheduled to occur near the end of each 10-minute interval and its completion was the signal to pause the scenario. We could not stop at fixed 10-minute intervals because we could not control when the requested control action would be issued. The Control1 aircraft was the first or second aircraft asked about half the time and the last or next to the last aircraft asked about the remainder of the time (for reasons no longer important). The remaining conditions were ordered randomly.

Three secondary dependent measures were administered. Thirty seconds after the participant took over responsibility for the sector during the second scenario, a surprise Map Recall was administered. The participants returned to the scenario upon completion of this Map Recall. After the completion of each scenario, the SME completed a performance measure called a post-scenario analysis (developed by Vortac et al., 1993). The SME examined the current status of each aircraft still in the sector and determined the number of route, speed, and altitude changes required to get the aircraft safely out of the sector. The re-

searchers reasoned that the more efficient the controller, the fewer control actions remaining. After completion of the experiment, a short questionnaire was administered. We collected biographical data and asked the participants to rate the importance of various pieces of flight data.

Participants were rotated through the six possible orderings of these two scenarios. They completed two of the three 30-minute scenarios, receiving a 30-minute break between scenarios.

Results

On the background questionnaire, participants reported how important it was to remember various pieces of information. The most important pieces of information were altitude and position on the PVD: 83% (altitude) and 67% (PVD position) of the participants responded *Very Important* to these questions. Most participants responded *It Depends* to questions about destination, route, call sign, type of aircraft, and speed (on average, 74% of the responses). *Not Important* was the typical response (80% of the responses) for remembering an aircraft's computer identification (CID) and the time over a fix. These results were expected, which was why we focused on altitude and PVD position in Experiment 1.

Battery of questions. The primary dependent measure from the battery of questions was the recall accuracy for altitude information. Altitude was correctly recalled 71% of the time averaged across the five conditions, which was much better than for the questions about other flight data[1] (average 42%, $t(17) = 8.2$). The mean percent correct for altitude across all five conditions is given in Table 1. A one-way repeated-measures ANOVA found no significant difference among conditions.

These data do not support the notion of better memory for "hot" aircraft (Control3 and Interaction3) when "hot" was operationalized by the frequency of interaction or the frequency of control action. There was a hint that performance was worse when a control action was taken, with recall accuracy slightly better for the conditions involving interaction only. Perhaps this was because changing the altitude resulted in confusion between the current altitude and the prior altitudes (a source monitoring problem, see Johnson & Raye, 1981). This confusion would be especially profound in the Control3 condition. However, we found no support for this hypothesis; only once was the incorrectly recalled altitude one of the prior altitudes.

We examined the Traffic condition in more detail. Overall, there was no difference in recall accuracy between the Traffic-Relevant (83%) and the Traffic-Irrelevant condition (76%, $t(17) = 1.1$). This was contrary to the predictions of Means and associates' (1988) second hypothesis. However, we do not view this as a strong test of this hypothesis because more altitude control actions were actually initiated on the Traffic-Irrelevant aircraft (2.5 vs. 2.0). Perhaps the altitude control actions were initiated for different reasons in the two Traffic conditions. Nevertheless, apparently in these scenarios, even when altitude was not the reason that two aircraft were in conflict, it was still important to resolving the conflict.

Table 1

Experiment 1: Percent Correct for Altitude by Condition

	Traffic	Control1	Control3	Interaction2	Interaction1	Overall
Altitude	80	72	66	83	83	71

[1] Route was dropped from the analysis because of the variety of ways the question was answered (abbreviations, idiosyncratic shorthand) and our inability to accurately classify them as correct or incorrect.

Figure 1 gives the percent correct as a function of the average number of altitude control actions an aircraft received. The number of control actions had opposing effects for the Traffic-Relevant and the Traffic-Irrelevant conditions. In the Traffic-Relevant condition (the altitude was relevant to the resolution of the conflict), the more altitude changes that were made, the better the altitude was remembered. Moreover, three altitude changes to the Traffic-Relevant aircraft resulted in significantly better performance than three altitude changes to a Control3 aircraft (100% vs. 66%, $t(7) = 2.37$). In the Traffic-Irrelevant condition, the opposite was true. Recall performance fell off sharply after more than two altitude control actions (and did not differ from Control3 performance). Clearly, the number of control actions did not determine memorability. However, the pattern suggested that the reason for initiating the control action might determine memorability. We explored this issue in Experiment 2 by focusing on sequencing conflicts that involve separation by speed changes or vectoring.

Confidence. After each recall response, participants estimated their confidence that the answer was correct. We analyzed the confidence data by folding the 100-point scale in half, which made 75% *sure your answer was correct* equivalent to 25% *sure your answer was wrong*. We constructed an individual calibration index (CI_j, Yates, 1990) for each condition j (Equation 1), as well as an overall calibration index for each participant (Equation 2).

$$CI_j = n_j(f_j - \bar{d}_j) \qquad (1)$$

$$CI = \sum \frac{CI_j}{N} \qquad (2)$$

The individual calibration index (CI_j) was a function of the difference between the expressed confidence (f_j) and the percent correct ( ), weighted by the number of judgments (n_j). The overall calibration index (CI) was simply the average of the individual calibration indices for each condition for each of the N participants. These indices are bounded by 1 and 0, with 0 indicating perfect calibration. Using Equation 1, we found no differences in $\bar{d}$ calibration across conditions ($F(4, 14) = 1.69$, $p > .05$), but according to Equation 2, the participants were generally overconfident ($t(17) = 7.29$).

Know-Remember. We asked the participants to specify whether their answers resulted from memory or knowledge. They spontaneously adopted a third response alternative—"guess." We suspected that guesses were based on knowledge, although the knowledge may not have been explicit or may have been knowledge for which they were not very confident. Table 2 shows the proportion of Guess, Know, and Remember responses as a function of question type. It was apparent that, in the scenarios we utilized, participants felt that they had to *remember* the altitude; they seldom based their responses on their knowledge, as they did for the speed where 56% of the responses were based on knowledge or were guesses. Overall, participants reported relying on their memory much more often than their knowledge to answer these questions (of all responses, 72% remember responses vs. 8% know responses).

It was possible that the percentage of *remember* responses was an overestimate, compared to what is true of controllers in the field. It was clear that this experiment was focused on memory, which might have affected the absolute level of *remember* responses. However, it probably would not affect the relative differences across question types.

Participants were most accurate when they reported that they remembered the answer (66% correct), and less accurate when they reported knowing the answer (27%) or making a guess (18%). This was a significant difference, $F(2,10) = 85.8$, and all pairwise differences were significant. (Post hoc tests always divided a by the number of comparisons.) There was also a significant difference in perceived confidence among the three responses ($F(2,10) = 75.1$). (Not all participants used all three response categories, hence the reduced degrees of freedom.) They were more confident in *remember* than in *know* responses, which did not differ from guesses.

Map Recall and PVD Position. Participants were extraordinarily accurate at their placement of aircraft on the paper sector map. Eighty-four percent of the aircraft recalled were placed within 2.5 cm of their actual location (within about 8 nautical miles). Overall,

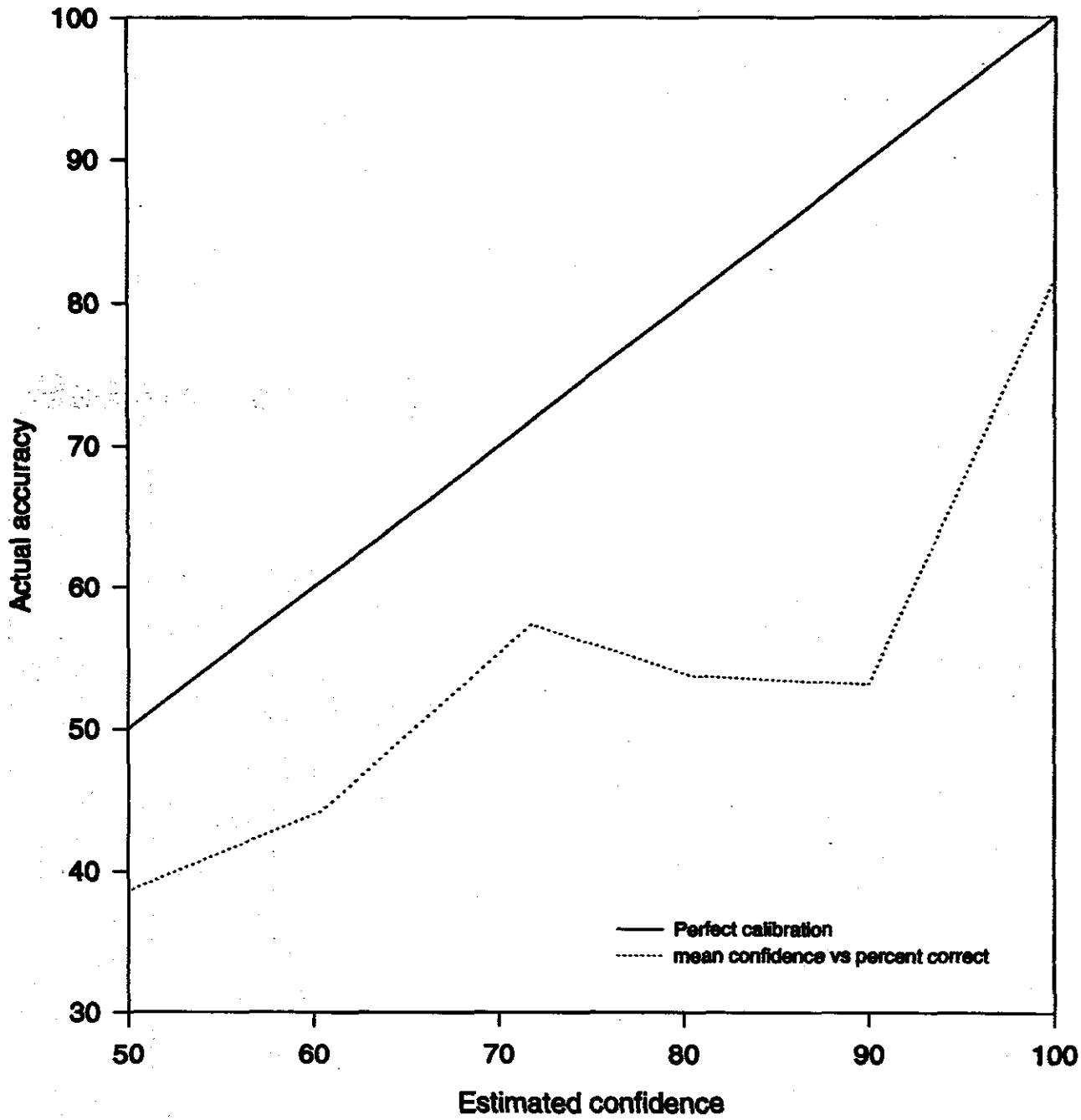

Figure 1. Percent correct for altitude as a function number of altitude control actions for the Traffic-Relevant and Traffic-Irrelevant conditions.

Table 2
Experiment 1: Percent of Guess, Know, and Remember Responses as a Function of Question Type

	Guess	Know	Remember
Altitude	4	2	94
Destination	16	7	77
Departure Point	35	9	56
Ground Speed	32	24	44
Aircraft type	32	9	60
Total	20	8	72

the average missed distance was 1.5 cm, or 5 nautical miles. Ninety percent of all aircraft were recalled. Projection of aircraft position into the future may also be an important part of memory for PVD position, but we tapped only memory for current position.

The results were very similar for the 30-s Map Recall. Participants recalled 95% of the aircraft (4.8 possible) with an average missed distance of 2.4 cm, which did not differ from the missed distance in the regular Map Recall. This suggested that the participants already had a very accurate representation of the position of the aircraft when they took control of the sector.

We examined two variables to determine if either affected the missed distance or recall likelihood: 1) was the aircraft on- or off-frequency (were they talking to the aircraft or was it about to enter or leave the sector), and 2) the class of aircraft (commercial, general aviation, or military). Whether the aircraft was on- or off-frequency affected percent correct (93% vs. 79%, $t(17) = 3.42$), but not missed distance. (All statistical tests are significant at $p < .05$ unless otherwise indicated.) It was not surprising that on-frequency aircraft were recalled better; responsibility for off-frequency aircraft had already been transferred to the next sector or involved aircraft that had not yet entered the sector. Contrary to Vortac et al. (1993), we found no differences due to class of aircraft.[2]

After the completion of Map recall, we asked participants to report which of the recalled aircraft "went together as a group." They recalled an average of 2.1 groups containing 2.4 aircraft, which corresponded closely to what Means et al. (1988) found in a similar task (1.8 and 2.7). The size of the groups was as expected; conflicts between aircraft typically involve only two aircraft (Bisseret, 1971). However, the small number of groups made us question the extent to which groupings of related aircraft were the primary means by which aircraft were mentally represented.

To assess the extent to which these groups reflected the mental representation of the aircraft, as opposed to reflecting a post-hoc grouping done to satisfy an experimenter's request, we determined how often the aircraft within a group were: 1) recalled consecutively, and 2) in close temporal proximity (the time between successive recalls was determined from the videotape). Sixty-nine percent of the groups resulted in the consecutive recall of its members. This was less than what Means et al. found (98%), but still quite high. However, the average time between successive recalls was 7.1 s, which was relatively slow if one aircraft was triggering the recall of another.

We believe that these groupings did not reflect the primary means by which aircraft were mentally represented. If it was, we would have expected to find either: 1) more groups, or 2) a shorter duration between successive recalls of aircraft within a group. The majority of recalled aircraft (over 60%) were not part of any group.

We tried a second method to find evidence of groupings: We examined the timing of aircraft recall. Quick bursts of successive retrievals should mark the existence of underlying organizational units (chunks). This more on-line measure might be more sensitive to relationships among aircraft than requiring participants to circle related aircraft at the conclusion of recall.

We defined a chunk as a set of aircraft recalled sequentially with less than t seconds between successively recalled aircraft.[3] We varied t over a wide range and examined the mean number of chunks and the mean chunk size. It was not until t equaled 4 s that we found an average of one chunk (of size 4) per participant. When t equaled 7 s we found an average of two chunks, but they were of size six. A chunk of this size was probably too large to correspond to a meaningful unit. Furthermore, chunks of this size did not correspond to the participants' groupings (two chunks of

[2] Vortac et al. (1993) found large differences among class of aircraft in recall of FPS information (commercial better than military better than general aviation). Because class of aircraft was not randomly assigned to condition in the present experiment, it was possible that this factor could contribute to any recall differences found across conditions. However, we found no difference in recall accuracy as a function of class of aircraft (commercial 50.3% vs. general aviation 49.5%, we had very few military aircraft).

[3] The timing data were not as uncontaminated as one might like. Rather than have controllers simply make a mark at the location of a remembered aircraft, they were instructed to simultaneously identify the mark by writing the call sign or other identifying information. This obviously inflated the time between successive recalls and may have hindered finding chunks in the output.

size two). Finally, 7 s was a relatively long time between successive recalls to assume that one aircraft triggered the recall of the next (that meant that perhaps 35 s elapsed during the recall of these six aircraft).

An examination of the timing of aircraft recall uncovered little evidence for groupings of related aircraft. What does this mean regarding how controllers mentally represent aircraft in their sector? To answer that question, we summarized the timing data as a cumulative output function—the number recalled over time.

A cumulative output function takes one of two general shapes (e.g., Gronlund & Shiffrin, 1986). A curvilinear shape (well described by a negative exponential, see Bousfield & Sedgwick, 1944) results when the growth of recall is initially very rapid but gradually slows. This occurs when there are a limited number of cues, each connected to a large number of items. For example, if asked to generate as many "fruits" as possible, assume that the only cues you can think of are fruits you like, fruits at the grocery store, and types of pies. The growth of recall is initially very rapid because these cues provide access to a large number of items, but the output rate eventually slows because no new cues are generated. Instead, the same cues are reused, resulting in the resampling of already recalled items.

The other general shape of a cumulative output function is linear. This shape results when retrieval is guided by a large number of cues that each subsume only one or two items. The initial growth of recall is slower because relatively more time is spent switching cues than retrieving items from cues. However, recall continues to grow throughout the recall period because new cues are generated that grant access to additional items, thereby limiting resampling of already recalled items.

A curvilinear shape would result if the mental representation of the aircraft was mediated by aircraft-to-aircraft links, as argued by Means et al. (1988). Each of the groupings of related aircraft would be accessed by a cue, and the retrieval of one aircraft in the group should quickly trigger the retrieval of the next. However, unless there was some strategy that continued to provide access to new cues and new groups throughout the recall period, thereby preventing the resampling of the already exhausted cues, the output rate would gradually slow.

We examined the cumulative number of aircraft recalled as a function of time (see Figure 2). We truncated the data at 13 aircraft because beyond that point we lost a significant number of participants. The most striking result was the linearity in the growth of recall (overall $r^2 = .99$). Each participant's cumulative output function was consistent with this overall function (the individual r^2's ranged from .88 to .99). We computed the average time between successive recalls (i.e., time between 1st and 2nd recall, 2nd and 3rd, etc.) and found that this function was linear ($r^2 = .87$) and remarkably flat. Although the regression equation indicated a significant positive slope, it showed only a 900-ms increase for each successive recall. The recall of aircraft was not governed by extensive groupings of related aircraft, so what could account for the linear rate of output?

We think the participants capitalized on their excellent memory for PVD position and let their knowledge of the sector guide retrieval. This evidently provided a large number of cues to help recall aircraft. The adoption of this strategy might have been the result of the participants being required to recall the aircraft on the paper map, as opposed to verbalizing them or writing them down on a sheet of paper. However, we think that the resulting output function would still remain linear if verbal or written recall was required if the linkages in memory that govern recall are not from aircraft-to-aircraft but are instead from a mental representation of the airspace to the aircraft.

Discussion

The participants in this study believed that the two most important pieces of information to remember were an aircraft's position on the PVD and the altitude. We found memory for aircraft position was excellent; 84% of the aircraft recalled were placed within 2.5 cm of their actual location. Altitude was also well recalled (71% accurate). The two together would provide the controller with a 3-D representation of the airspace.

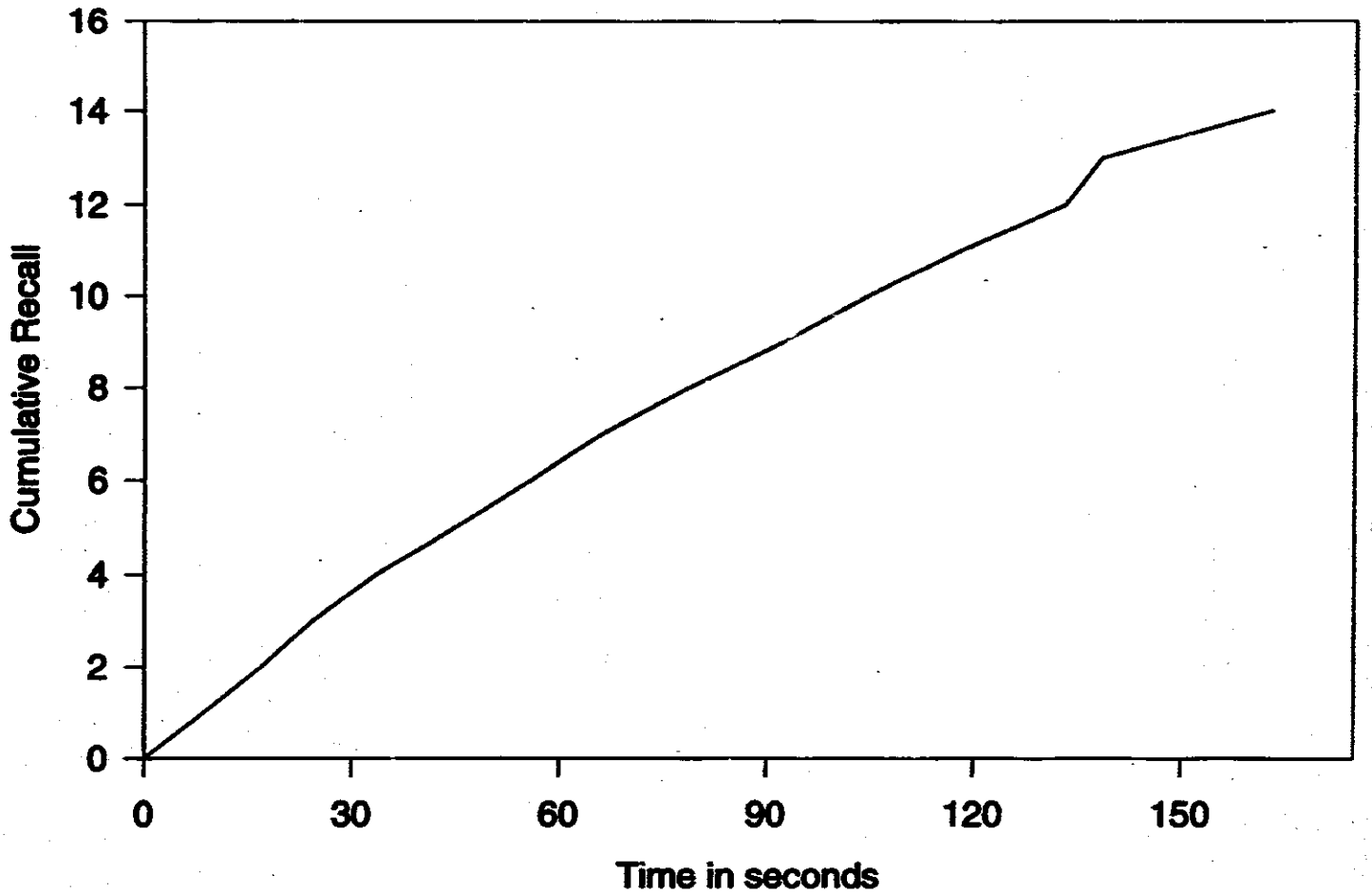

Figure 2. Cumulative number of aircraft recalled as a function of time in seconds.

We found no support for the Means et al. (1988) hypothesis that the number of control actions affected the likelihood of the recall of altitude information. One possible explanation for the null effect was that altitude was so important that participants always tried to encode it. Consequently, we might have to look at other flight data to determine which variables affect memory in air traffic control. We do so in Experiment 2. Perhaps the Means et al. (1988) "hot" aircraft hypothesis holds for other types of "less critical" flight data.

The participants were overconfident in the accuracy of their memory for altitude. This was not surprising; overconfidence characterizes the memory of many experts (Ayton, 1992) and the judgments of most laypersons (Lichtenstein, Fischhoff, & Phillips, 1982). Shanteau (1992) analyzed various domains where overconfident expert performance was documented and argued that the calibration of the expert depended on certain task characteristics. The job of the controller shares many task characteristics with other poorly-calibrated experts, including having to deal with dynamic stimuli, less predictable problems, few errors allowed, and unique tasks (a similar conflict may be resolved in different ways by the same controller at different times). Ayton (1992) found that receiving prompt and unambiguous feedback differentiated well-calibrated from overconfident experts. The feedback in air traffic control is neither prompt nor unambiguous.

There was little evidence that the mental representation of the aircraft under control involved aircraft-to-aircraft links in memory. The linear output rate was consistent with the use of a strategy that provided new cues throughout the recall period, perhaps a strategy that relied on the sector itself to guide retrieval. This reliance on spatial information to remember large quantities of information is in keeping with other cognitive experts studied by Ericsson and Kintsch (1995). For example, an expert waiter remembered orders by location around the table; chess experts remembered board configurations after being told what piece occupied what square on the board, despite never actually viewing the whole configuration. This retrieval structure may serve as the foundation for SA. Flach (1996) defined SA as the congruence between the subjective interpretation of an event and the objective measures of the actual event.

Experiment 2

According to Experiment 1, whatever was strengthened by repeated interactions involving the altitude or repeated control actions changing the altitude, it was not memory for that altitude. However, frequent contact might result in increased familiarity of an aircraft's call sign. Consequently, in Experiment 2, we checked to see if the call signs of aircraft that received more control actions were remembered better. If so, this would rule out the possibility that the range of altitude changes we manipulated in Experiment 1 (from 1 to 3) was insufficient to affect memory.

Because traditional memory variables, such as the number of repetitions (operationalized as number of interactions or number of control actions) and study time (length of time in the airspace)[4], did not affect the likelihood of recalling an aircraft's altitude, perhaps we need to examine the system at a deeper level to ascertain which variables affect memory, the function of an aircraft in a scenario.

The Traffic condition was carried over from Experiment 1, to which we added a Not-traffic and a Pre-traffic condition. The Traffic condition involved the resolution of a sequencing problem. The Traffic aircraft were the aircraft the participants were actively separating and monitoring to ensure that separation was maintained. The Not-traffic condition involved two aircraft that were physically close to one another (like the Traffic aircraft) but were not traffic for one another. There was no compelling motivation to remember much flight data about these aircraft. The Pre-traffic condition involved two aircraft that might become traffic for one another in the near future. Little might be known about these aircraft because they would have just entered the airspace.

[4] The Interaction3 aircraft averaged 14 minutes in the airspace and the Interaction1 aircraft averaged 6 minutes, but their recall accuracy was equal.

An informal polling of controllers (none of who participated in the study) indicated that they would remember more about the Traffic aircraft because these were the aircraft that they were actively separating; they were the important aircraft. The text comprehension literature contains related findings. For example, the likelihood of recalling a fact from a text is little affected by the repetition that fact receives, compared to the position that fact occupies (the role the fact plays) in the propositional structure of the text (e.g., McKoon, 1977).

Means and associates' (1988) second hypothesis—that the type of control exercised influenced what was recalled—makes a similar prediction. The effect of two aircraft being in conflict in the Traffic condition should be to highlight some piece of flight data, increasing its likelihood of being recalled. Although a variety of types of control might be exercised on the various Traffic aircraft, and various types of control would highlight different types of flight data, the effect should be to raise the overall recall level for these flight data, and as a result, recall of flight data for the Traffic aircraft as a whole.

We included questions that tapped both static and dynamic flight data. Questions regarding dynamic flight data included Altitude, Ground speed, and Altitude status (was the aircraft currently climbing, level, or descending). We asked questions about three pieces of static flight data. We dropped departure point used in Experiment 1 and replaced it with Relationship to sector (arrival, departure, or overflight regarding your sector); it was considered more important to know whether an aircraft was a departure than to know from where it departed. We also asked about Direction of flight and Destination.

Experiment 1 showed that what was done with an aircraft did not affect memory for its flight data. In Experiment 2, we try to determine if the role the aircraft played affected memory for its flight data.

Method

Participants. Fourteen full-performance level (FPL) en route air traffic controllers participated. They had been FPL controllers for an average of 11.5 years. They last worked in the field 2.8 years ago, with a range of .2 to 7.3 years. All participants were instructors at the FAA Academy and all but one were familiar with the AeroCenter airspace. Six had participated in Experiment 1.

Materials. The experiment was conducted at the Radar Training Facility (RTF) at the Mike Monroney Aeronautical Center. Participants worked the R-side position and the SME worked the Radar Associate's position. The experiment required no deception on the part of the SME.

Ten high-complexity scenarios were created with the help of the SME. Each was constructed around a sequencing problem and was designed to require more extensive use of speed control to achieve separation than in Experiment 1. The scenarios included a mean of 10.6 aircraft, 5.9 of which were overflights, 2.8 were arrivals, and 1.9 were departures. The scenarios in Experiment 2 were probably of higher-fidelity than in Experiment 1 because no scripted control actions or interactions were necessary.

Procedure. The SME specified a starting point for each scenario that was just prior to the point that control actions were necessary to begin to solve the sequencing problem. The participants sat down at this point, received a position-relief briefing, and assumed control of the sector. They were instructed to control traffic as they normally would. At the conclusion of the experiment, three participants indicated that they sometimes tried to commit more to memory than normal. However, their data did not appear to differ from the remaining participants and was retained.

A scenario was stopped at a predetermined stopping time; an average of 6.8 minutes elapsed between the starting and stopping point for each scenario. The participant was turned away from the PVD and strip bay and completed two tasks.

The call sign recognition task required judgments regarding whether an aircraft was on the PVD at the time the scenario was stopped. Twelve aircraft were tested, six that were not on the PVD (called distractors) and six that were (targets). All six of the target aircraft were under the control of the controller. The set of distractors was created by taking all the target call signs, changing the number (e.g., AAL23 became AAL96), and randomly assigning them to one of the ten scenarios. The target and distractor call signs for a given scenario did not vary across participants.

There were two targets from each of three conditions (Pre-traffic, Traffic, and Not-traffic). The targets from the same condition were tested sequentially. Each pair of targets was preceded by and followed by a distractor, otherwise the ordering of tests was random.

The Pre-traffic condition consisted of two aircraft that were on routes that would cross at some point soon. Typically, they had entered the airspace near the end of the scenario and were quite far apart from one another (in two-dimensional space, about 55 miles or 17.4 cm on the PVD). The Traffic condition consisted of the two aircraft that would probably (as judged by the SME) be the first two aircraft in the sequence (the primary conflict the participant had to solve). The Not-traffic condition consisted of two aircraft that were close together (like the Traffic condition), but were not traffic for one another. As it turned out, the Not-traffic aircraft were physically closer to one another at the time that the scenario was stopped (5.7 cm) than the Traffic aircraft (7.9 cm). This difference was significant ($F(2,12) = 4278.9$); post-hoc tests showed that all pairwise differences were significant.

The aircraft in the different conditions indeed served different roles in the scenarios, as measured by the number of control actions they received during the experiment (altitude: $F(2, 10) = 20.2$; speed: $F(2, 10) = 10.7$). There was an average of .72 altitude changes and .4 speed changes in the Traffic condition, which was significantly greater than in the Not-traffic condition (altitude: 0.40; speed: 0.08), which was significantly greater than in the Pre-traffic condition (altitude: 0.14; speed: 0).

The second task to be completed, the recall task, immediately followed the call sign recognition task. We provided a paper copy of the sector map that showed the location of each aircraft and its call sign. The target planes from the three conditions were used again. We asked six questions about each plane: a) altitude; b) ground speed; c) current altitude status (level, climbing, or descending); d) relationship to the sector (arrival, departure, and overflight); e) direction of flight; and f) destination. The first three tapped dynamic flight data; the last three tapped static flight data. All six questions about a given aircraft were asked consecutively, although in a random order. The order of the six aircraft was randomized.

We collected confidence judgments after question a), b), or f) (randomly selected), and after one of the other three questions (randomly selected), for each of the six aircraft. Participants indicated their confidence in the accuracy of their previous answer by sliding a tick mark along a bar whose endpoints were labeled 0% and 100%. We thought that this method of judging confidence would overcome the problem observed in Experiment 1 where participants failed to distinguish among mid-range confidence judgments (i.e., anything between about 51% and 99% confidence was treated as equivalent, turning our continuous scale into a three-alternative forced-choice among guess, probably correct, and absolutely correct).

Each participant completed ten scenarios. The order of scenarios was counterbalanced across participants. There were 15-minute breaks after the third and seventh scenarios.

Table 3

Experiment 2: Percent Correct for the Six Questions Types for Each of the Three Conditions

	Not-traffic	Pre-traffic	Traffic
Altitude	66	67	69
Ground speed	19_b	25	29_a
Altitude status	89_b	94_a	82_c
Relationship to sector	[illegible]	96_a	97_a
Direction of flight	[illegible]	82	75
Destination	51_b	47_b	93_a

Note: Means with different subscripts were significantly different across conditions.

Results

Table 3 shows accuracy (percent correct) for each question type for each condition. A MANOVA[5] showed a main effect of condition ($F(2, 12) = 16.61$)[6], question type ($F(5, 9) = 220.66$), and an interaction ($F(10, 4) = 23.36$). Means in Table 3 with different subscripts were significantly different across conditions.

There were no differences among conditions for the altitude question. As in Experiment 1, the greater number of altitude control actions for the Traffic aircraft did not result in better recall for altitude. This was not caused by a lack of statistical power (a potential criticism of Experiment 1) because there were significant effects for other questions.

For altitude status, we found that performance was best in the Pre-traffic condition, next best in the Not-traffic, and worst in the Traffic condition; for relationship to sector, Not-traffic was worse than the other two conditions; for ground speed, Not-traffic was worse than Traffic. The only question type for which the Traffic condition was significantly better than both the Not-traffic and the Pre-traffic conditions was Destination. Unfortunately, this result was probably an artifact. Performance for the Traffic aircraft was inflated because both Traffic aircraft always had the same destination; that was why these aircraft had to be sequenced. Also, it was usually true that several other aircraft in the scenario, also part of the sequencing problem, were going to that destination.

To facilitate comparisons across question types, we subtracted an estimate of chance performance from the percent correct given in Table 3. We assumed that chance was 1/3 for altitude status and relationship to sector, 1/8 for direction of flight, and 1/10 for ground speed and altitude (according to the SME, there were about 10 possible altitudes or speeds that were reasonable for a given aircraft). Destination was dropped because of the problem with the Traffic condition. There was a significant main effect of condition ($F(2, 12) = 4.65$) and question type ($F(3, 11) = 7.66$), and an interaction ($F(6, 8) = 5.79$). Post hoc comparisons showed that ground speed was remembered significantly worse than everything else (but significantly better than chance), and that direction of flight was remembered significantly better than altitude or altitude status (minimum $t(13) = 3.24$).

In addition to remembering the exact speed or altitude, there were two additional ways the participants could demonstrate some degree of memory for

[5] A MANOVA was used because repeated-measures ANOVA's assume sphericity. The MANOVA does not require this assumption and is generally a more conservative test of significance.

[6] Because we cannot regulate the control actions a participant would take, and because there were methodological controls that had to be sacrificed to maintain scenario fidelity, we were unable to achieve an equal distribution of correct answers across the various response options. As a result, the Altitude status of every one of the Pre-traffic aircraft was level, but only 73% of the Not-traffic and 56% of the Traffic aircraft were level. That meant that performance differences across conditions could have been the result of guessing "level" and being correct almost all the time in the Pre-traffic condition, correct next most often in the Not-traffic condition, and correct least often in the Traffic condition (exactly the pattern observed). However, that was not the case. The controllers were equally accurate when they responded "level" across the three conditions (93%, 95%, and 93% correct for Not-traffic, Pre-traffic, and Traffic, respectively).

these flight data. Their response could approximate the correct answer, or they could remember the speed or altitude relationally. In other words, participants might not remember the exact speed (or altitude) of AAL123, but their response might be close to the correct answer or they might know that it was faster, slower, or the same speed (higher, lower, or at the same altitude) as another aircraft.

We first examined whether the estimate of speed and altitude approximated the correct answer. We scored as correct any response within 20 knots (2000 feet)[7] of the correct answer. We also re-scored the data to extract relational information. For example, denote the two aircraft in a condition as plane A and plane B. If plane A was faster than plane B, it was coded a 1, if plane A was slower than plane B, it was coded a 2, and if the two planes had the same speed it was coded a 3. The same procedure was used to score the participants' responses. Any time the answer code matched the response code, it was counted correct.

Figure 3 gives percent correct for ground speed (top panel) and altitude (bottom panel) for the approximation and relational scoring methods, as well as the exact responses (taken from Table 3). Accuracy for approximation responses must be greater than exact responses because they included exact responses as a subset.

The participants seldom remembered the exact ground speed of an aircraft. However, their responses usually approximated the correct answer (within 20 knots). There were significant differences across conditions ($F(2, 12) = 20.75$), with the Pre-traffic and Traffic conditions significantly more accurate than the Not-traffic condition. They were also very often correct relationally. There were differences across conditions ($F(2, 12) = 24.06$), with the Not-traffic and Traffic conditions significantly greater than the Pre-traffic. For altitude, there were no significant differences across conditions for approximation scoring (a possible ceiling effect). The pattern for relational altitude was similar to relational speed ($F(2, 12) = 251.19$), although in this case, all conditions differed significantly. Overall, there seemed to be less emphasis on representing altitude in a relational way, compared to speed.

As in Experiment 1, participants were overconfident in their memory ($t(13) = 7.23$). On those occasions when they were fairly well calibrated, it was probably because, as accuracy approached 100%, their confidence could not exceed 100%. Table 4 gives the calibration scores (Yates, 1990). A MANOVA showed a main effect of question type ($F(5,9) = 12.73$), but post-hoc tests found no significant difference across conditions.

Call-sign recognition. For the call sign recognition task, recognition accuracy was measured by d' (McNicol, 1972). The three conditions differed ($F(2,12 = 4.35$). Post-hoc tests showed that performance in the Traffic condition ($d' = 1.59$) was better than in either of the other conditions (Not-traffic = 1.14 and Pre-traffic = 1.19). Changing an aircraft's altitude or speed made the participant more familiar with the call sign of these aircraft, but no more familiar with the flight data being modified (see also Experiment 1). Responses to Traffic aircraft were also the fastest (although not significantly so), ruling out the possibility of a speed-accuracy trade-off (Pachella, 1974). Furthermore, if the Traffic aircraft were linked in memory as a group, presentation of one member of the group should facilitate the processing of the immediately proceeding member (see, for example, Ratcliff & McKoon, 1978). There was no evidence of any facilitation (1st Traffic aircraft tested = 1648 ms, 2nd Traffic aircraft tested = 1652 ms), which was consistent with the results of the Map Recall in Experiment 1; the mental representation does not consist of aircraft-to-aircraft links.

Multiple regression. We completed an exploratory multiple regression to determine to what extent a given piece of flight data was predictable from other flight data. It was possible that static flight data would be more predictable than dynamic flight data because the former did not change over the course of the scenario. We also thought that flight data based on knowledge might be more predictable than flight data derived from memory. In Experiment 1, the vast majority of altitude responses was judged to be "remember" responses, while many more speed responses were judged to be "know" responses. Was exact ground speed more predictable than altitude?

7 We chose 2000 feet because if the controller remembered the direction of flight, they would capitalize on the fact that East and Northbound aircraft utilize odd altitudes (e.g., FL230, FL250) and West and Southbound aircraft utilize even altitudes.

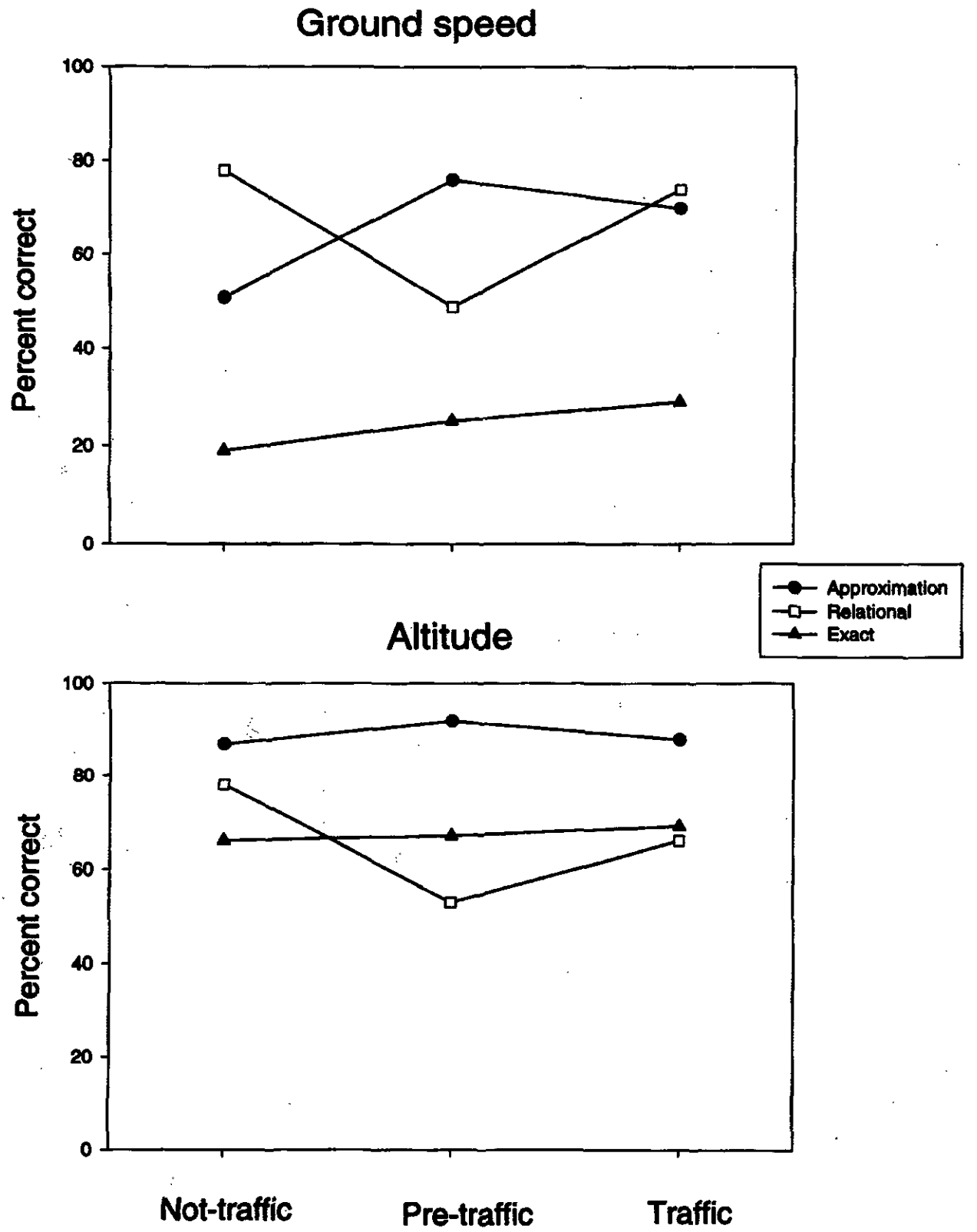

Figure 3. Percent correct for the Not-traffic, Pre-traffic, and Traffic conditions for exact, relational, and approximation scoring. The top panel is for Ground speed and the bottom panel for Altitude.

Table 5

Multiple Regression Analyses for Each Condition Separately

Question Type	Equation	R^2
	Traffic	
Altitude (A)	.29 (RS)	.036
Ground speed (S)	.14 (AS) - .27 (Dest) + .10 (Dir)	.036
Altitude status (AS)	.55 (RS) +.08 (S)	.057
Relationship to sector (RS)	.20 (AS) +.18 (A) + .19 (Dest) + .11 (Dir)	**.130**
Direction (Dir)	.11 (RS) + .11 (S)	.019
Destination (Dest)	.36 (RS) - .08 (AS)	.065
	Pre-traffic	
Altitude (A)	.17 (RS) + .11 (Dest) + .10 (Dir)	.055
Altitude status (AS)	.59 (RS) + .08 (Dir)	**.359**
Relationship to sector (RS)	.57 (AS) + .08 (Dest) + .12 (A)	**.376**
Direction (Dir)	.10 (AS) + .09 (Dest) + .10 (A)	.028
Destination (Dest)	.11 (A) + .09 (Dir) + .12 (RS)	.036
	Not-traffic	
Altitude (A)	.18 (Dest) + .10 (Dir)	.042
Ground speed (S)	.13 (AS)	.007
Altitude status (AS)	.11 (RS) +.11 (S)	.017
Relationship to sector (RS)	.10 (AS) + .35 (Dest) + .17 (Dir)	**.175**
Direction (Dir)	.22 (RS) + .11(A)	.062
Destination (Dest)	.17 (A) + .49 (RS)	**.162**

Note: Adjusted R^2 and standardized beta weights are shown.

We completed one multiple regression for each of the three conditions. Each of the question types was used as a dependent variable and the remaining factors were used as predictors. Table 5 gives the equations with the standardized beta weights. The degree of prediction was given by the adjusted R^2. Except for ground speed in the Pre-traffic condition, each dependent variable was predictable to a significant degree. However, there were only five dependent variables for which 10% or more of the variance could be predicted. These are highlighted in boldface in Table 5.

Three of these dependent variables were relationship to sector, once in each condition. Relationship to sector was also the most frequent predictor overall. If a dependent variable loaded on relationship to sector, it was the strongest (or tied for the strongest) predictor. Not surprisingly, when relationship to sector was eliminated as a predictor, no dependent variables had an R^2 better than .06. The least predictable dependent variable was ground speed (average R^2 = .01), followed by direction (average R^2 = .04) and altitude (average R^2 = .04).

The Pre-traffic aircraft were the most predictable overall. The average R^2 for Pre-traffic aircraft was .14 (including R^2 = 0 for ground speed); it was .08 and .06 for Not-traffic and Traffic, respectively. This was primarily due to the relatively high predictability of altitude status and relationship to sector. For altitude status, this was due entirely to *level* flights being well predicted; for relationship to sector, it was due entirely to *overflights* being well-predicted. Apparently there was a prototypical Pre-traffic aircraft in these scenarios (the level overflight), which was by definition, fairly predictable. Whether this is true in the field as well is unknown. There was no prototypical Traffic or Not-traffic aircraft, and consequently, these were poorly predicted.

We also completed a descriptive discriminant analysis using the flight data as response variables. The discriminant analysis yielded a function that discriminated among Pre-traffic, No-traffic, and Traffic aircraft as a function of these response variables. When we excluded relationship to sector and altitude status from the discriminant analysis, there was still sufficient structure in the data to correctly classify a sizable proportion of the Pre-traffic aircraft as Pre-traffic aircraft, based on their direction of flight and altitude. This provided additional support for the prototypical nature of these Pre-traffic aircraft. Although it may be the result of the particular scenarios we used, it is nevertheless an illustration of the type of subtle information on which the controller might capitalize.

Discussion

The increased number of control actions initiated on Traffic aircraft did affect memory. It improved recognition of the call sign of the aircraft. It did not, however, improve memory for the flight data from that aircraft. The fact that recognition was used for the call sign task while recall was used for the other task may have been a contributing factor, except that recall in these experiments was really forced-choice recognition. Take altitude status, for instance: The participant knew the three possible answers, and only had to "recognize" the correct answer from among those possibilities.

Flight data from the Traffic aircraft were not the best remembered. This was contrary to expectations and contrary to a generalization of Means and associates' (1988) second hypothesis. Assuming that the Traffic aircraft were more important to the controller, that importance did not manifest itself in improved memory for the flight data. We do not know if that was because these aircraft were really not important (unlikely), were all equally important, or differed in importance but our measures failed to tap that importance. We take up the latter two suggestions in the General Discussion.

The overall low level of performance for ground speed was surprising given that these scenarios were designed to require the use of speed control. However, the poor memory for the exact speed might be caused by the phraseology controllers use. Although controllers instruct pilots to climb or descend to a *specific* flight level, they often tell them to increase or decrease their speed by (for example) 10 knots. Consequently, the controllers remember exact altitude fairly well because that was how they interacted with altitude information, but because they did not deal with exact speed, they do not remember it.

It would be wrong to conclude that the participants remembered nothing about the ground speed of the aircraft under their control. Their exact responses usually preserved the ordinal relationship between the Traffic aircraft and between the Not-traffic aircraft. Moreover, when the participants failed to remember the exact ground speed of both aircraft, we observed that some of them always got the correct ordinal relationship, although others never did. We wonder if this might not be diagnostic of good SA. In other words, none of the participants remembered the exact speeds very well, but some reliably preserved the correct ordinal relationship.

How could the relatively poor memory for the ground speed of Pre-traffic aircraft (according to exact and relational scoring) result in accurate approximation responses? Perhaps it was because these were not responses from memory but guesses that took advantage of the fact that these were "prototypical" Pre-traffic aircraft. The multiple regression showed that these were the best predicted aircraft, primarily due to the predictability of *level overflights*, which would require minimal control actions.

General Discussion

Situation awareness is assumed to be central to successful air traffic control performance (e.g., Endsley, 1995a). The products of memory are viewed as central to achieving SA (Endsley, 1995b). What have we learned about the role of memory in air traffic control?

We had little success manipulating the memorability of flight data about aircraft. We examined two hypotheses. One hypothesis proposed that flight data about "hot" aircraft (which we operationalized by the number of communications and/or the number of control actions) would be recalled better. This was not

supported. The other hypothesis was that the type of control exercised would affect what was recalled. In Experiment 1, there was no difference in recall of altitude as a function of whether altitude was more or less relevant to the resolution of a conflict. In Experiment 2, ground speed was made central to the resolution of conflicts for the Traffic aircraft, however, ground speed was no better recalled in the Traffic than in the Not-traffic condition. Furthermore, despite the greater importance of speed control in Experiment 2, altitude was still recalled about as well as in Experiment 1 (71% accuracy in Experiment 1, 68% accuracy in Experiment 2). Finally, flight data about aircraft that were being actively separated (i.e., Traffic aircraft) were no better remembered than flight data about aircraft that were not traffic.

Why were we unsuccessful in finding variables that affected the recallability of flight data? We consider four possibilities.

One possibility is that we have yet to discover the correct variables that affect recall. We view this as unlikely given that we tested variables that past research indicated were important. These included peripheral (hot versus cold—frequency and repetition, length of time in airspace) as well as more central, meaning-based variables (type of control, role aircraft played, importance). There is ample evidence in the literature for the positive impact of variables like frequency, repetition, study duration, and importance, on memory (e.g., Crowder, 1976).

A second possibility for why these variables did not affect memory was because memory for the flight data was so vital to task performance that the flight data were not highlighted further by these manipulations. However, except for memory for PVD position, no flight data was recalled at a level that suggested that it was vital to task performance.

A third possibility is that memory is irrelevant to the performance of the controller and consequently, irrelevant to SA. There are reasons to question the importance of memory to air traffic control. The air traffic control situation is so dynamic that it is probably not good to remember flight data for long because it will interfere with the current flight data. In addition, the controller does not need to commit a lot of information to memory because of the extensive external aids that are available (the FPSs, the CRD, and the data blocks on the radar display). The information from external displays is always at least as reliable as memory and, if it can be located quickly, may be preferred to reliance on memory. Durso (personal communication, April 15, 1996) proposed that the latency to find requested flight data using external aids might be a good measure of SA. The controllers' excellent memory for the locations of aircraft in their sector would allow this rapid access to information.

If either of the previous two possibilities were true, query techniques for measuring SA that assumed that all aircraft were equivalent would be appropriate (Endsley, 1987). Consequently, flight data about different aircraft would be expected to be equally well (or poorly) recalled. This would be contrary to our hypothesis that controllers should remember a lot about some aircraft, but could remember very little about others.

The final possibility we consider is that memory is important to air traffic control and SA, but the wrong measures were used in these studies. Do the controllers need to remember the *exact* altitude and ground speed of an aircraft (i.e., the verbatim details) to be able to perform their job and to be considered to have good SA? Research on cognitive development suggests that gist information (i.e., memory for meaning), and not verbatim information, is crucial for reasoning (Brainerd & Reyna, 1993).

Cognitive developmentalists discovered that verbatim memory for critical background information in a reasoning problem is independent of the quality of reasoning that results. For example, memory for the exact premises of a transitive inference problem (A > B, B > C) is unrelated to the likelihood of making the correct inference (A > C) (Brainerd & Kingma, 1984). Furthermore, this memory-independence effect continues into adulthood and appears to hold across a wide range of situations (e.g., attitude change, Hastie & Park, 1986; numerical reasoning, Klapp, Marshburn & Lester, 1983).

There are several memorial advantages to the encoding of gist over verbatim details (Brainerd & Reyna, 1990; Reyna & Brainerd, 1992). These include stability, ease of retrieval, and ease of manipulation. There are also several processing advantages,

including simplified processing, increased accuracy, and reduced effort.

What are the implications of the independence between reasoning ability and memory (the so-called memory-independence effect) for the role of memory in air traffic control? First, the number of verbatim details that controllers remember about an aircraft should be independent of their ability to separate aircraft. Moreover, good memory for specific flight data (the kinds of questions we asked) might actually lead to poorer performance. This was what Brainerd and Reyna (1993) found for children solving reasoning problems. Adelson (1984) found that novice programmers sometimes had better memories for the specific (irrelevant) details of a task than did experts.

A second implication of the memory-independence effect is that understanding what controllers need to remember to perform their jobs will require alternate methods for tapping memory. Consequently, we need measures to tap the gist traces that support reasoning and decision-making, not measures that tap only exact altitude and speed.

De Groot (1946/1978) found that world-class chess players accessed the best chess moves during their initial perception of the situation, suggesting that pattern-based retrieval from memory was fundamental to expertise. We think that controllers continually scan the PVD looking for patterns that signal a conflict. Like the chess expert, they have learned countless patterns (e.g., two aircraft converging at the same altitude, one aircraft climbing through an other's airspace) that signal a potential problem. However, exact flight data are not part of these patterns. Two aircraft crossing at the same altitude is a problem, regardless of the exact altitude. In other words, rather than encoding that AAL123 is at FL230 and SWA456 is at FL270, controllers encode only the "gist" (i.e., SWA is higher than AAL, or no one else is at the same altitude as AAL123).

If Brainerd and Reyna (1993) are correct, and if we are right about the applicability of their theory to air traffic control, gist and not verbatim traces support SA. This means that future methodologies that measure SA in air traffic control, and perhaps in other domains as well, should tap memory for the information that actually supports task performance.

References

Adams, M.J., Tenney, Y.J., & Pew, R.W. (1995). Situation awareness and the cognitive management of complex systems. *Human Factors, 37*, 85-104.

Adelson, B. (1984). When novices surpass experts: The difficulty of the task may increase with expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 484-95.

Anderson, J.R. (1995). *Cognitive psychology and its implications* (4th ed.). New York: W.H. Freeman and Company.

Ayton, P. (1992). On the competence and incompetence of experts. In G. Wright & F. Bolger (Eds.), *Expertise and decision support*. New York: Plenum Press.

Bisseret, A. (1971). Analysis of mental processes involved in air traffic control. *Ergonomics, 14*, 565-70.

Bousfield, W.A., & Sedgwick, H.W. (1944). An analysis of sequences of restricted associative responses. *Journal of General Psychology, 30*, 149-65.

Brainerd, C.J., & Kingma, J. (1984). Do children have to remember to reason? A fuzzy-trace theory of transitivity development. *Developmental Review, 4*, 311-77.

Brainerd, C.J., & Reyna, V.F. (1990). Gist is the grist: Fuzzy-trace theory and the new intuitionism. *Developmental Review, 10*, 3-47.

Brainerd, C.J., & Reyna, V.F. (1993). Memory independence and memory interference in cognitive development. *Psychological Review, 100*, 42-67.

Crowder, R.B. (1976). *Principles of learning and memory*. Hillsdale, NJ: Erlbaum.

de Groot, A.D. (1946/1978). *Thought and choice and chess*. The Hague, Netherlands: Mouton.

Dominquez, C. (1994). Can SA be defined? In M. Vidulich, C. Dominquez, E. Vogl, & G. McMillan (Eds.), *Situation awareness: Papers and annotated bibliography*. (pp. 5-15). AL/CF-TR-1994-0085, Armstrong Laboratory.

Endsley, M.R. (1987). SAGAT: A methodology for the measurement of situation awareness (NOR DOC 87-83). Hawthorne, CA: Northrop Corporation.

Endsley, M.R. (1995a). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37*, 32-64.

Endsley, M.R. (1995b). Measurement of situation awareness in dynamic systems. *Human Factors, 37,* 65-84.

Ericsson, K.A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102,* 211-45.

Flach, J.M. (1996). Situation awareness: In search of meaning, *CSERIAC Gateway, 6,* 1-4.

Gronlund, S.D. & Shiffrin, R.M. (1986). Retrieval strategies in recall of natural categories and categorized lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 643-8.

Hastie, R., & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. *Psychological Review, 93,* 258-68.

Hopkin, V.D. (1980). The measurement of the air traffic controller. *Human Factors, 22,* 547-60.

Johnson, M.K., & Raye, C.L. (1981). Reality monitoring. *Psychological Review, 88,* 67-85.

Klapp, S.T., Marshburn, E.A., & Lester, P.T. (1983). Short-term memory does not involve the "working memory" of information processing: The demise of a common assumption. *Journal of Experimental Psychology: General, 112,* 240-63.

Leplat, J., & Bisseret, A. (1966). Analysis of the processes involved in the treatment of information by the air traffic controller. *The Controller, 5,* 13-22.

Lichtenstein, S., Fischhoff, B., & Phillips, L.D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases.* Cambridge, UK: Cambridge University Press.

Marshak, W.P., Kuperman, G., Ramsay, E.G., & Wilson, D. (1987). Situational awareness in map displays. In *Proceedings of the Human Factors Society 31st Annual Meeting.* Santa Monica, CA: Human Factors and Ergonomics Society.

Means, B., Mumaw, R.J., Roth, C., Schlager, M.S., McWilliams, E., Gagne', E., Rice, V., Rosenthal, D., & Heon, S. (1988). *ATC training analysis study: Design of the next-generation ATC training system.* (Report No. FAA/OPM 342-036) Washington, DC: Department of Transportation/Federal Aviation Administration.

McKoon, G. (1977). Organization of information in text memory. *Journal of Verbal Learning and Verbal Behavior, 16,* 247-60.

McNicol, D. (1972). *A primer of signal detection theory.* London, George Allen & Unwin, Ltd.

Mogford, R.H. (1994). Mental models and situation awareness in air traffic control. In D. Gilson, D.J. Garland, & J.M. Koonce (Eds.), *Situation Awareness in Air Traffic Control. Situational Awareness in complex systems,* (pp. 199-207). Daytona Beach, FL: Embry-Riddle Aeronautical University Press.

Pachella, R.G. (1974). The interpretation of reaction time in information-processing research. In B. Kantowitz (Ed.), *Human information processing,* (pp. 41-82). Potomac, MD: Erlbaum Press.

Rantanen E. (1994). The role of dynamic memory in air traffic controllers situation awareness. In R. D. Gilson, D.J. Garland, & J.M. Koonce (Eds.), *Situation Awareness in Air Traffic Control. Situational Awareness in Complex Systems.* (pp. 209-215). Daytona Beach, FL: Embry-Riddle Aeronautical University Press.

Ratcliff, R., & McKoon, G. (1978). Priming in item recognition: Evidence for the propositional structure of sentences. *Journal of Verbal Learning and Verbal Behavior, 17,* 403-17.

Reyna, V.F., & Brainerd, C.J. (1992). A fuzzy-trace theory of reasoning and remembering: Paradoxes, patterns, and parallelism. In A.F. Healy, S. Kosslyn, & R.M. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (pp. 235-60). Hillsdale, NJ: Erlbaum.

Shanteau, J. (1992). Competence in experts: The role of task characteristics. *Organizational Behavior and Human Decision Processes, 53,* 252-66.

Sperandio, J.C. (1978). The regulation of working methods as a function of workload among air traffic controllers. *Ergonomics, 21,* 195-200.

Stein, E.S. & Garland, D.J. (1991). Air traffic controller working memory: considerations in air traffic control tactical operations. DOT/FAA CT-TN93/37, Atlantic City, NJ: Federal Aviation Administration Technical Center.

Vortac, O.U., Edwards, M.B., Fuller, D.K., & Manning, C.A. (1993). Automation and cognition in air traffic control: An empirical investigation. *Applied Cognitive Psychology, 7,* 631-51.

Yates, J.F. (1990). *Judgment and decision-making.* Englewood Cliff, NJ: Prentice Hall.

☆U.S. GOVERNMENT PRINTING OFFICE: 1997 -660-008/60046