Role of Memory in Air Traffic Control

Scott D. Gronlund, Daryl D. Ohrt, Michael R. P. Dougherty, and Jennifer L. Perry
University of Oklahoma

Carol A. Manning
Civil Aeromedical Institute

En route air traffic controllers serving as instructors at the Federal Aviation Administration Academy were tested to determine what they remembered about the aircraft in their sector. The study focused on memory for flight data (aircraft altitude and ground speed) and aircraft position on the radar. Aircraft importance, but not frequency of interaction, affected memory for flight data; neither variable affected recall of the aircraft's radar position. It was hypothesized that controllers use their memory for aircraft position to classify aircraft as important (potential traffic) or not. Information that is not represented spatially (e.g., altitude) is filtered by importance. In addition, a task-relevant filter facilitates encoding other information (ground speed) if it is pertinent. Results have implications for improving the interpretation of techniques that assess situation awareness by assessing the amount remembered.

With the rapid advance of technology, complex and 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 system that confronts the air traffic controller comprises a large number of aircraft coming from a variety of directions, at diverse speeds and altitudes, heading to different destinations. Like most complex, dynamic systems, this one cannot be periodically halted while the controller takes a brief respite.

The ability to manage such a system requires the controller to maintain situation awareness, or SA (for a review see Durso & Gronlund, in press). 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 the use of that coherent understanding to direct perception and anticipate future events (see also Endsley, 1995a). Although a number of factors can lead to failure to perform a task successfully, several case studies (e.g., Cheung, Money, & Sarkar, 1995) and analyses of existing databases (e.g., Durso et al., 1996) point to the loss of SA as an important precursor to performance failure. Failure in such complex cognitive tasks due to loss of SA can have devastating results. For example, controlled flight into terrain killed nearly...
5,000 people from 1978 to 1992 (Woodhouse & Woodhouse, 1995), and 74% of these accidents were hypothesized to be due in part to insufficient SA.

The benefits to public safety are obvious if techniques for assessing SA can be developed. However, there is currently no agreed-upon methodology for its assessment (Endsley, 1995b). 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 the participant's ability to recall information about the situation from memory. According to Endsley (1995b), "SA, composed of highly relevant, attended to, and processed information, should be most receptive to recall" (p. 72). Endsley believed the vast majority of a participant's SA can be assessed in this manner.

The goal of our experiments was to determine the variables that affect two important classes of information used by air traffic controllers to maintain their SA: (a) the recall of aircraft position on the radar and (b) memory for various flight data about an aircraft. If we can identify some of the variables that affect memory for this information, it would modify how the query technique is interpreted. For example, currently, all aircraft are considered equally important and therefore information from all aircraft should be equally well remembered. However, an alternative possibility is that there should be some aircraft about which the controller should remember more, and other aircraft for which it would be acceptable (i.e., safe) that little was remembered. In fact, remembering more information about these latter aircraft may actually signal poorer SA. We begin by reviewing a study that examined one variable that might affect memory for the two classes of information.

Means et al. (1988) conducted a study with 3 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, 19931). Controllers correctly recalled upwards of 95% within 10 nautical miles (nmi) 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 the controllers, when cued with the call sign, recalled only 28% of the aircraft types and only 6% of the ground speeds.

Means et al. (1988) proposed 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. 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. In Experiment 1, we operationalized hot aircraft in two different ways: (a) by the number of interactions with an aircraft and (b) by the number of control actions taken on an aircraft (as in Means et al., 1988). An interaction was defined as any communication with an aircraft that did not result in a change to the aircraft's flight data; a control action was defined as any interaction that resulted in a change to the aircraft's altitude, speed, or heading.

**Experiment 1**

In Experiment 1, we sought to replicate Means et al. (1988) using a larger sample size ($N = 18$ vs. 3). More importantly, we felt that in Means et al., the reason for the control action might have been confounded with the frequency of the control action. We attempted to disentangle the two in Experiment 1 by manipulating the aircraft that were hot or cold (and holding constant the reason for the control action) rather than having those roles determined by the actions of the participants, as was the case in Means et al.

Before describing the experiments, it is important to note that the conclusions we draw about the role of memory in air traffic control are descriptive rather than prescriptive. In other words, we can indicate what controllers remem-
bered about the aircraft they were controlling, but we are unable to indicate what they should have remembered without an adequate measure of SA or performance. Current measures of air traffic control performance were too gross in their assessment to make this determination, summarizing performance over too wide a time frame (e.g., Buckley, DeBaryshe, Hitchner, & Kohn, 1983, aircraft fuel consumption, delays; Vortac et al., 1993, number of remaining control actions needed to exit aircraft from the sector). For example, in Experiment 1, the correlation between the Vortac et al. (1993) global performance measure and recall accuracy was nearly 0. Although this could reflect the inadequacy of the performance measure, we believe the problem involved the time scale over which performance was measured. We believe that prescriptive conclusions can be drawn only when moment-to-moment performance fluctuations are correlated with memory. Nevertheless, to the extent that our participants maintained SA during our experiments, what they did remember might signal in part what they should have remembered.

Method

Participants. Eighteen en route air traffic controllers participated. All were instructors at the Federal Aviation Administration (FAA) Academy and were familiar with the AeroCenter airspace used in the experiment. All were full-performance level (FPL), or expert controllers (meaning that they were certified to work a sector independently, in contrast to a trainee who must work with an FPL). 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 the experiment, with a range of 1.6 to 6 years.

Materials. The experiment was conducted at the Radar Training Facility (RTF) at the Mike Monroney Aeronautical Center in Oklahoma City, Oklahoma. The RTF has high-fidelity air traffic training simulators used to provide radar training. Communications between the controllers and the aircraft took place in the same manner as in the field, although the aircraft were "piloted" by ghost pilots who controlled the simulated aircraft on the basis of the controller's instructions.

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

Three 30-min scenarios were developed with the help of a subject matter expert (SME). The scenarios were judged by the SME to approximate the workload level typically experienced in the field. The scenarios included a mean of 28.7 aircraft, of which 9 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. The scenarios were designed around the constraints necessary to test the hypothesis of interest, yet were required to be as realistic as possible. The conflicting aircraft that were included in the scenario were designed to not interfere with the aircraft of interest (i.e., the hot and cold aircraft).

Procedure. The participants completed a set of sample questions prior to beginning the experiment to familiarize themselves with the procedure. They were informed that the scenarios would be stopped periodically and that they would be asked questions about various aircraft. They were also told that it would be most beneficial if they controlled traffic as they normally did rather than modifying their performance to try to answer all our questions correctly.

Participants worked the R-side, or radar position. Our SME 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 the participant to rely on the Radar Associate for reliable information.

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 a scenario the problem was paused. The
signal to pause the scenario was the execution of
a control action requested by the pilot of a critical
aircraft; this type of request was scheduled to
occur at approximate 10-min intervals. When the
scenario was paused, the participant was turned
away from the PVD and strip bay and completed
two tasks.

The first task was a map recall, for which we
provided a full-size paper replica of the sector
map (no aircraft present). Participants placed an
"X" at the location of each aircraft and wrote
down the call sign or any other identifying
information. After completing the map recall,
participants moved to the computer to complete
the second task, which was a battery of questions
about aircraft flight data. For the battery of
questions, a second paper copy of the sector map
was provided. The locations of aircraft on this
map were copied from the PVD while the partici-
pant was completing the map recall and included
the correct position of all the aircraft with their
call signs. Call signs were included because
controllers do not generally remember them, and
we referenced aircraft by call sign in the questions.

Three types of questions were asked about
each of five aircraft, in the following order: (a)
informational—what was American123's
(AAL123's) altitude (or ground speed, route,
destination, departure point, or aircraft type)?: (b)
metamemorial—rate your confidence in your
answer (a range from 0 = absolutely no idea to
100 = absolutely certain); and (c) 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 explicitly remember the aircraft type
of AAL123, but they might implicitly know that
Southwest456 was a Boeing 737 because all
Southwest Airlines aircraft are 737s.

Questions regarding altitude were of primary
interest. They made up one-third of all informa-
tional questions. Questions on other flight data
were included to discourage the participants from
unduly focusing on altitude. Altitude questions
were phrased so that it was unambiguous what
was requested (i.e., 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 ques-
tions (6.7% of all the altitude questions) did not
specify which type of altitude was requested. For
these, we counted as correct either the assigned or
the current altitude. After completing the battery
of questions, the participants returned to the
scenario. They were allowed as much time as
they wished to prepare before resuming the
scenario.

We manipulated the number of altitude control
actions and the number of altitude-relevant inter-
actions that an aircraft had received in the prior
(approximate) 10-min block to produce four
experimental conditions: Control 3 and Interac-
tion 3 (hot), Control 1 and Interaction 1 (cold). Control 3 aircraft received three control actions,
Interaction 3 aircraft received three interactions (communica-
tions), and Interaction 1 aircraft received one.

For the Control 3 aircraft, the pilot made three
requests that resulted in control actions; the
requests were separated by approximately 3 min.
For example, the pilot might request an altitude
change to 10,000 ft, then to 12,000 ft, and finally
to 13,000 ft, to get above a layer of clouds. The
scenario was designed such that there was no
reason why the controller should not grant the
requested control action, and they customarily
were granted. The control actions were not de-
signed to separate traffic, nor did they introduce
conflicts. The requested control actions were
judged to be reasonable by our SME and no
participant ever reported noticing anything out of
the ordinary about these requests or the scenarios
more generally.

In the Interaction 3 condition, the pilot might
report light chop (turbulent air), later ask if there
have been other reports, and finally report that it
had smoothed out. No control actions were
warranted, although altitude information was
emphasized as a result of the pilot communica-
tions. These communications were separated by
about 3 min. In the Control 1 condition, the pilot
might request one altitude change. In the Interaction 1 condition, the pilot might ask about the ride at a certain flight level.

We asked questions about five aircraft per 10-min block. Three were filler aircraft, one was always a Control 1 aircraft, and the other was either a Control 3, Interaction 3, or Interaction 1 aircraft, with a different one in each block. The aircraft were tested in a different random order for each participant, the only exception being that the Control 1 aircraft was either in the first two or last two positions (for reasons unimportant to this article). A particular aircraft was tested only once in the experiment.

The SME was able to design the scenarios so that the aircraft of experimental interest (Control 3, Control 1, Interaction 1, and Interaction 3) would not conflict with other aircraft (come within 5 mi horizontally or 1,000 ft vertically); there were other aircraft in the scenario that were traffic for one another. Consequently, the aircraft of interest were not likely to require many additional control actions beyond those we preplanned (hence our claim that the reason for the control actions/communications was disentangled from the frequency). For example, the Control 3 aircraft received only 0.85 additional altitude control actions on average during the prior 10-min block while the Control 1 aircraft received only 0.31 additional altitude control actions. In addition, there were minimal heading changes and almost no speed control actions to confound the independent variable. (This emphasis on altitude control actions to separate traffic may not necessarily be true in the field).

Participants completed two of the three 30-min scenarios, receiving a 30-min break between scenarios. We rotated participants through the six possible orderings of two scenarios chosen from the three total.

Two secondary dependent measures were administered. Thirty seconds after the participant took over responsibility for the sector in the second scenario, the scenario was paused and a surprise map recall was administered. The participants returned to the scenario upon completion of this map recall. After completion of the experiment, a short questionnaire was administered.

**Results and Discussion**

On the questionnaire, participants reported how important it was to remember various pieces of information. They indicated that 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. These results were expected, which was why Experiment 1 focused on PVD position (i.e., the location of the aircraft on the 2-D radar) and altitude. Most participants responded it depends to questions about destination, route, call sign, type of aircraft, and speed (on average, 74% of the responses). Participants gave not important (80% of the responses) as the typical response for remembering an aircraft’s computer identification or the time over a fixed point in space.

Participants were fairly accurate in their placement of aircraft on the paper sector map (replicating Means et al., 1988; Vortac et al., 1993). Eighty-four percent of the aircraft recalled were placed within 2.5 cm of their actual location (within about 8 nmi). Overall, the average missed distance was only 1.5 cm, or 5 nmi. This degree of accuracy did not result from the selective recall of just those aircraft for which the controllers were highly confident; 90% of all aircraft were recalled. 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, or 8 nmi, which did not differ from the missed distance in the regular map recall. This suggested that the participants already had an accurate representation of the 2-D position of the aircraft as they took control of the sector.

The primary dependent measure from the battery of questions was the recall accuracy for altitude information. Exact altitude\(^2\) was correctly recalled 76% of the time, which was much better than for the questions about other flight

\(^2\)Assigned altitude is measured in increments of 1,000 feet. For example, an aircraft is assigned 22,000 or 23,000 feet, but never 22,567 feet. Likewise, ground speed is measured in increments of 10 knots.
The mean percentage correct for altitude across all four conditions is given in Figure 1. A one-way repeated-measures analysis of variance (ANOVA) found no significant difference among conditions, \( F(3, 45) < 1, p > .05, MSE = 0.122. \)

Recall in the Control 3 condition was not depressed due to confusion resulting from changing the altitude three times in a block (a source monitoring problem; see Johnson & Raye, 1981). Only once was the incorrectly recalled altitude one of the prior altitudes. In sum, these data do not support the hypothesis of better memory for hot aircraft operationalized either as those aircraft receiving more control actions or those receiving more communications.

Confidence. We analyzed the confidence data by folding the 100-point scale in half; 75% response of sure your answer was correct was equivalent to 25% response of sure your answer was wrong. Although the participants were generally overconfident, \( t(15) = 4.88, s_{\text{diff}} = 0.055, \) mean accuracy = 64.4%, mean confidence = 91.1%, there were no differences across conditions, \( F(3, 45) = 1.56, p > .05, MSE = 0.056. \)

![Figure 1. Percentage correct for altitude recall in Experiment 1. Black bars signify cold aircraft and white bars signify hot aircraft. Error bars signify 1 SEM.](image)

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 experts depended on certain task characteristics. The job of the controller shared many task characteristics with other poorly calibrated experts, including dealing 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).

Know–Remember. We asked the participants to specify whether their answers resulted from memory or knowledge. Many times, participants spontaneously adopted a third response alternative—guess. We suspected that participants' guesses were based on knowledge, although it may have been knowledge in which they were not very confident. The majority of their responses indicated that they perceived memory to be the source of the information (mean 72%, ranging from 44% for speed to 94% for altitude). Participants were most accurate when they reported that they remembered the answer (64% correct) and were less accurate when they reported knowing the answer (21%) or guessing (16%). The difference was significant, \( F(2, 22) = 24.4, MSE = 0.045, \) as were the three pairwise differences. (All post hoc tests divided \( \alpha \) by the number of comparisons.)

3 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 responses as correct or incorrect.

4 Vortac et al. (1993) found large differences among classes 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%, with very few military aircraft in these scenarios).
Although introspections can be misleading (see discussion by Ericsson & Simon, 1993), it appeared that the controllers perceived that memory was important to their job. However, it was possible that the percentage of remember responses was an overestimate compared to what would be true in the field. For instance, it was clear that this experiment focused on memory. Second, in the field, a controller frequently works the same sector of airspace at the same time of day and consequently knows, and does not need to remember, that AAL123 is going to Phoenix when it enters the controller’s airspace at 10 a.m.

In sum, we found no support for the Means et al. (1988) hot/cold hypothesis that the number of control actions (or the number of communications) affected the likelihood of recalling altitude information. We also replicated the Means et al. and Vortac et al. (1993) finding of good memory for the 2-D position of the aircraft. Interestingly, neither the recall probability, \( F(3, 45) = 1.81, p > .1, \text{MSE} = 0.082 \), nor the missed distance, \( F(3, 30) < 1, p > .1, \text{MSE} = 1.072 \), were affected by whether the aircraft was hot or cold.

**Experiment 2**

One potential problem with the Means et al. (1988) hypothesis was that the assignment of aircraft to condition was determined by the actions of the controller rather than being manipulated by the experimenter. Consequently, some other factor might be responsible for making an aircraft hot or cold. In Experiment 2, we tested whether that factor was the importance of the aircraft in the scenario. Thus, Means et al. might not have found that more was remembered about hot aircraft; instead, more might have been remembered about important aircraft (those in conflict with other aircraft), which typically receive or require more control actions.

We operationalized importance by the temporal proximity to losing separation with another aircraft. We reasoned (and our SME concurred) that aircraft that were immediate traffic for one another were more important than aircraft that might become a problem for one another in the relatively distant (10 min) future, which were more important than aircraft that were not likely to be in conflict at all. In Experiment 2, we tested the hypothesis that the importance of an aircraft in the scenario determined how much was remembered about that aircraft.

**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 0.2 to 7.3 years. All participants were instructors at the FAA Academy, and all but one were familiar with the AeroCenter airspace. Four had participated in Experiment 1. At the conclusion of the experiment, 3 participants indicated that they sometimes tried to commit more to memory than normal. However, their data did not differ from the remaining participants’ and were retained.

**Materials**. The experiment was conducted in the en route laboratory at the FAA Academy’s Radar Training Facility in Oklahoma City. 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 scenarios were created with the help of the SME. In Experiment 2, all control actions were initiated by the controller; none were preplanned for the ghost pilots. The scenarios were shorter than the Experiment 1 scenarios (lasting on average 6.7 minutes), which left little opportunity for the independent variable of aircraft importance to be contaminated by other factors. Each scenario was constructed around a sequencing problem: A number of aircraft that had to be put in line for approach to an airport were merging to the same route, and greater than 5-mi horizontal separation is required for aircraft on the same route. A sequencing problem required more extensive use of speed control to achieve separation than was the case in Experiment 1. The scenarios included a mean of 10.6 aircraft for which the controller was responsible: 5.9 were overflights, 2.8 were arrivals, and 1.9 were departures. Many of the overflights required control actions; they were called overflights because they were not landing in the controller’s sector. Six of the 10 scenarios required the sequencing of arrivals; the remaining 4 involved the sequencing of overflights (e.g., setting up approaches to Dallas/Forth Worth). The scenario workload was again
judged by our SME to be comparable to what was typical in the field.

**Procedure.** We defined three levels of aircraft importance: (a) traffic condition, (b) pretraffic condition, and (c) not-traffic condition. The traffic condition involved the resolution of an impending conflict (if nothing was done within the next 2 min, separation would be lost). In some cases, the controllers had already taken action to separate these aircraft at the time the scenario was stopped. The traffic condition consisted of the two aircraft that would be (as judged by the SME) the first two aircraft among those being sequenced. The pretraffic condition consisted of two aircraft that were on routes that would cross 10 or more min in the future. These two aircraft were quite far apart from one another at the time the scenario was stopped (in 2-D space, about 55 mi, or 17.4 cm on the PVD). The not-traffic condition involved two aircraft that were physically close to one another in horizontal space (like the traffic-condition aircraft) but were not traffic for one another (no control action was necessary to separate these aircraft; e.g., they were at different altitudes or flying in opposite directions). This was done to unconfound separation distance on the PVD from aircraft importance. As it turned out, the not-traffic aircraft were physically closer to one another in horizontal space at the time that the scenario was stopped (5.7 cm) than were the two traffic aircraft (7.9 cm), and the aircraft in each of these conditions were closer than the two pretraffic aircraft, $F(2, 26) = 2194.45, \text{MSE} = 0.253$; post hoc tests showed that all pairwise differences were significant.

The number of altitude and speed control actions the aircraft received in the different conditions supported the claim that these aircraft were treated differently by the controllers: altitude, $F(2, 22) = 31.20, \text{MSE} = 0.044$; speed, $F(2, 22) = 17.35, \text{MSE} = 0.046$. An average of 0.8 altitude changes and 0.48 speed changes were given 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 pretraffic condition (for altitude, 0.14; but speed, 0, was only significantly less than traffic). The not-traffic aircraft received more control actions than the pretraffic aircraft because the former had been in the airspace several minutes longer than the latter.

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. At this juncture, the participants sat down, received a position-relief briefing, and assumed control of the sector. The scenario was stopped at a predetermined point that ranged from 6 to 8 min after the starting time. At the stopping point, 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 (called *targets*). The six target aircraft were the two traffic, two pretraffic, and two not-traffic aircraft. 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 10 scenarios. The target and distractor call signs for a given scenario did not vary across participants.

The second task to be completed, the recall task, immediately followed the call sign recognition task. As in Experiment 1, we provided a paper copy of the sector map that showed the location of each aircraft and its call sign. For each pair of aircraft in each of the three conditions, we asked six questions: (a) altitude; (b) ground speed; (c) current altitude status (level, climbing, or descending); (d) relationship to sector (arrival, departure, or overflight with respect to the sector); (e) direction of flight (north, northeast, east, etc.); and (f) destination. All six questions about a given aircraft were asked consecutively, although in a random order. The order of the six aircraft was randomized differently for each participant.

Each participant completed 10 scenarios. The order of scenarios was counterbalanced across participants. There were 15-min breaks after the 3rd and 7th scenarios.

**Results and Discussion**

The small number of control actions occurring in these short scenarios was sufficient to produce measurable differences across conditions. Recog-
ition accuracy measured by $d'$ (McNicol, 1972) showed differences among the three levels of aircraft importance, $F(2, 26) = 6.21, MSE = 0.144$; 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 pretraffic = 1.19). Responses to traffic aircraft were also the fastest (although not significantly so), ruling out the possibility of trading speed for accuracy (Pachella, 1974).

Table 1 shows accuracy (percentage correct) for the various question types for the three levels of aircraft importance. The destination question was dropped because performance for the traffic condition 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 same destination, making it a reasonable guess when the controllers did not remember it. A repeated-measures ANOVA showed a main effect of aircraft importance, $F(2, 26) = 4.63, MSE = 98.905$, question type, $F(4, 52) = 198.94, MSE = 159.238$, and an interaction, $F(8, 104) = 4.77, MSE = 84.562$. Within a particular question type, means with different subscripts were significantly different across aircraft importance.

There were no differences among conditions for the altitude question or for direction of flight. For altitude status, performance was best in the pretraffic condition, next best in the not-traffic, and worst in the traffic condition. However, this was due in part to the fact that the altitude status of every one of the pretraffic aircraft was level, but only 73% of the not-traffic and 56% of the traffic aircraft were level. If the controllers simply guessed level each time, their data would resemble the observed data. There was no question type for which performance in the traffic condition was better than both the pretraffic and not-traffic conditions. However, performance in the traffic condition was better than the not-traffic condition for ground speed and relationship to sector.

It appears that two distinct types of information may underlie controller performance. Altitude and direction of flight were not affected by aircraft importance. Together with position on the PVD (in Experiment 1), altitude and direction of flight all relate to an aircraft's location in space (in this case, dynamic 3-D space). Aircraft importance might not affect these data because the controller needs to know altitude and direction of flight to make a classification of aircraft importance. Those aircraft that are classified as important should recruit more cognitive resources from controllers, resulting in improved memory for their flight data. This was observed for relationship to sector and ground speed, where performance in the traffic condition was better than in the not-traffic condition. The importance of pretraffic aircraft was unclear. It fell between traffic and not-traffic for ground speed, providing some evidence for gradations of aircraft importance. The pretraffic aircraft was equivalent to the traffic aircraft for relationship to sector, which suggests that aircraft importance may be a binary variable (aircraft are either important or not). Experiment 3 followed up on the idea that some information is unaffected by importance because that information is necessary to determine aircraft importance, and also sought to determine if aircraft importance is graded or is a binary-valued variable.

To facilitate comparisons across question types, we subtracted an estimate of chance performance from the percentage correct given in Table 1. Chance would reflect the probability of simply guessing the correct answer. We assumed that chance was 1/3 for altitude status and relationship to sector (there were 3 possible responses), 1/8 for direction of flight (8 possible responses), and 1/10 for ground speed and altitude (according to

<table>
<thead>
<tr>
<th>Question type</th>
<th>Not-traffic</th>
<th>Pretraffic</th>
<th>Traffic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Altitude status</td>
<td>89.9</td>
<td>94.9</td>
<td>82.6</td>
</tr>
<tr>
<td>Ground speed</td>
<td>19.8</td>
<td>25.2</td>
<td>29.6</td>
</tr>
<tr>
<td>Relationship to sector</td>
<td>83.2</td>
<td>96.1</td>
<td>97.5</td>
</tr>
<tr>
<td>Altitude</td>
<td>67.3</td>
<td>67.3</td>
<td>70.1</td>
</tr>
<tr>
<td>Direction of flight</td>
<td>82.2</td>
<td>83.4</td>
<td>75.3</td>
</tr>
</tbody>
</table>

Note. Means with different subscripts were significantly different.
the SME, there were about 10 possible altitudes or speeds that were reasonable for a given aircraft). There was a significant main effect of condition, $F(2, 26) = 4.63, MSE = 98.944$, and question type, $F(4, 52) = 114.65, MSE = 159.219$, and an interaction, $F(6, 104) = 4.77, MSE = 84.505$. Post hoc comparisons showed that ground speed was remembered significantly worse than everything else (but better than chance); direction of flight was remembered significantly better than altitude or altitude status (all $p s < .001$).

The poor recall for exact ground speed might have been caused by the phraseology controllers use. Although controllers instruct pilots to climb or descend to a specific flight level, controllers often tell pilots to increase or decrease their speed by (for example) 10 knots. That would mean that we were asking the wrong question about ground speed. Perhaps controllers do not remember the exact speed, but instead remember the speed approximately or relationally.

We examined whether the estimate of speed (and altitude) approximated the correct answer. We scored as correct any response within $\pm 20$ knots ($\pm 2,000$ feet) of the correct answer. We also rescored the data to extract relational information. For example, we denoted the two aircraft in a condition as Plane A and Plane B. If Plane A was faster than Plane B, the pair was coded a 1. If Plane A was slower than Plane B, it was coded a 2. 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 as correct. To compare performance across the different methods of scoring, we subtracted out the differing levels of chance (1/3 for relational scoring, 1/2 for approximation scoring, and 1/10 for exact scoring). For ground speed, relational scoring was best (relational = 29.6% vs. exact = 14.5%, and approximate = 0.4%, not different from 0). For altitude, all three scoring methods differed from each other and from 0, but exact was clearly superior (exact = 58.0% vs. approximate = 21.9% and relational = 4.0%). Altitude information was best remembered exactly, but ground speed was best remembered relationally. More direct support for this finding was sought in Experiment 3.

There is much evidence to suggest that controllers should remember in ways other than exactly. Bisseret (1971) analyzed controller errors and found that they often reflected a lack of precision rather than an error. For example, rather than encoding that AAL123 is at 23,000 feet (FL230) and SWA456 is at FL270, perhaps controllers encode only the "gist" (i.e., SWA is higher than AAL). Research on cognitive development has shown that the verbatim memory retained for critical background information in a reasoning problem is independent of the quality of reasoning that results (Brainerd & Reyna, 1993). Similar findings have been found with adults and appear to hold across a range of situations (e.g., attitude change, Hastie & Park, 1986; numerical reasoning, Klapp, Marshburn, & Lester, 1983). There are several advantages to the encoding of gist over the encoding of verbatim details, including ease of retrieval, increased accuracy, and simplified processing (Brainerd & Reyna, 1990; Reyna & Brainerd, 1992). A similar argument for the simplification of the controllers' mental load through "gistification" can be found in Moray (1990).

**Experiment 3**

We sought to refine the importance hypothesis by testing whether aircraft importance was binary valued rather than graded. If it was binary valued, we would expect memory resources to be focused on the important aircraft, resulting in superior memory for their flight data. We focused on PVD position and altitude as representative of the type of information controllers need to determine aircraft importance; we focused on ground speed as representative of the type of information that should be affected by aircraft importance.

We utilized a better control condition. In Experiment 2, the not-traffic aircraft sometimes had been traffic for other aircraft earlier in the scenario. We suspect that many of the controllers would have classified these aircraft as important for that reason. This could explain why the flight data from the not-traffic condition was not consistently more poorly recalled than the other two conditions in Experiment 2. It might also be the case that recall of PVD position and altitude are affected by aircraft importance, given the proper control condition. In Experiment 3, we chose aircraft that were not, and never would be,
traffic for any other aircraft; we called these never-traffic.

We also examined the impact on memory performance of one implementation of the concept of “free flight” (Federal Aviation Administration, 1995). Currently, most aircraft fly along structured routes through the sky (so-called “highways in the sky”). This was the case for the scenarios in our first two experiments. As a result, there are particular points in a controller’s airspace where routes intersect and merging aircraft will conflict. However, when aircraft fly on direct or straight-line (unstructured) routes through the sector, aircraft may intersect at any point in the airspace. Because the FAA is planning a move to a direct-route, free-flight environment (Federal Aviation Administration, 1995) and has already implemented some characteristics of free flight at high altitudes, we thought it would be timely to include this manipulation in our study. The use of unstructured routing could increase the workload on the controller, which might lead to more simplification through gist processing.

Method

Participants. Eleven full-performance level (FPL) controllers participated. They had been FPL controllers for an average of 14.0 years. The controllers last worked in the field an average of 4.0 years before, with a range of 0.5 to 6.8 years. All participants were instructors at the FAA Academy and were familiar with the AeroCenter airspace used in the experiment. Eight had also completed Experiment 2 (2 of those participants had also completed Experiment 1). Participants worked alone and were responsible for all aspects of the sector, including communication with pilots and other centers and performing their own strip marking.

Materials. The experiment was again conducted at the RTF. Two 30-min scenarios were created and modified with the help of the SME. One scenario used structured routing, in which most aircraft flew on standard routes. In the other scenario, most aircraft used unstructured routing, which we instantiated as straight-line routes from departure point to destination. We were unable to simulate true free flight (Radio Technical Commission for Aeronautics, 1995; RTCA) because the simulation facilities at our disposal could not support it. The scenarios included a mean of 36.0 aircraft; 33.0 were overflights, 0 were arrivals, and 3.0 were departures. On average, there were 15.0 aircraft displayed at the time of testing, about the same number as in Experiment 1, but each participant worked alone in Experiment 3.

Procedure. There were two traffic conditions. One consisted of two aircraft converging at the same altitude (traffic-level); the other consisted of one aircraft climbing through the altitude of a level aircraft (traffic-climb). We included a pretraffic condition as in Experiment 2. All three of these conditions were considered important. Finally, a never-traffic condition was included as a more effective control than the not-traffic condition used in Experiment 2. Aircraft importance was operationalized as before: temporal proximity to loss of separation. For the two traffic conditions, the time window was approximately 2 min, the pretraffic aircraft were about 10 min apart, and the pair of never-traffic aircraft would never conflict with each other (or with any other aircraft within the time frame of our scenario). A given aircraft served in only one condition per scenario and was not repeated.

After studying and ordering the flight strips to their liking, the participants were told which scenario they were controlling (structured or unstructured routes). They were also told that all the aircraft on the PVD were under their control and none were currently in conflict.

At three points during each 30-min scenario, the scenario was paused. We again used a triggering event as a signal. The signal occurred at approximate 10-min intervals. The participant was turned away from the PVD and strip bay and completed two tasks.

The first task was map recall. The computer presented a replica of the AeroCenter airspace (66% of its normal size) with no planes displayed. The participant used a trackball to move a cursor to the 2-D location of each aircraft on the radar. The participant returned to this task after completing the memory questions to indicate which call sign went with which marked location.

After completing map recall, the participant answered a series of questions about various aircraft. During this portion of the study, a full-size paper replica of the sector was provided with magnets designating all of the aircraft in position and their call signs. Three types of
questions regarding the ground speed and altitude were asked about each pair of traffic-level, traffic-climb, pretraffic, and never-traffic aircraft. The first question concerned relational information. For example, “Is AAL123 higher, lower, or level with DAL246?” The second question concerned range information for each aircraft in the pair (ordered randomly). For example, “Give an upper and lower bound for which you are 95% certain that the actual speed of AAL123 lies between them.” The final question concerned the verbatim, or exact, flight data for each aircraft in the pair (ordered randomly). For example, “What is the exact altitude of AAL123?” We varied whether speed or altitude was asked about first. Because of an oversight, the order of the conditions was fixed: pretraffic, never-traffic, traffic-climb, and traffic-level. We do not think that this was a serious limitation. As we discovered, the two best-performing conditions occurred first and last. Furthermore, these two conditions also occurred in Experiment 2 (where their ordering was randomized), and performance on exact altitude was equivalent between these two conditions in Experiment 3, just as it was in Experiment 2. In addition, the worst-performing condition was tested second (not so late in testing as to be unduly affected by forgetting).

A practice phase was completed prior to beginning the experiment that included a sample map recall and an example of each question type. It began with the experimenter demonstrating how to respond to the various questions using the computer, and concluded with the participant completing the same practice trial. Participants completed two scenarios (randomly ordered) and received a 20-min break between scenarios. After completion of both scenarios, a demographic questionnaire was administered.

Results and Discussion

In the map recall, locations corresponding to 79.6% of the aircraft were recalled. Recall varied slightly with condition and type of scenario producing a significant interaction, $F(3, 8) = 4.37$, $MSE = 0.094$, due to worse recall for the traffic-climb aircraft in the structured scenario (95% vs. 74% respectively, $t(10) = 3.57$, $SE = 0.052$). However, this difference was due to very poor recall (27%) of a single aircraft in that condition. We are not sure why this aircraft was so poorly recalled. Once this aircraft was removed, the interaction was eliminated, $F(3, 8) = 2.09$, $p = .18$, $MSE = 0.086$. In general, the locations of all the aircraft, even the never-traffic aircraft, were equally well remembered.

Next, we computed the distance between the actual 2-D position of the aircraft and the recalled position of the aircraft (see Figure 2). All aircraft were included (excluding the poorly recalled traffic-climb aircraft just mentioned did not change the results). As in Experiment 1, controllers were fairly accurate in recalling the position of the aircraft, only missing the actual position on the radar by an average of 2.85 cm (about 9.6 nmi). The missed distance was less in Experiment 1 (an average of 1.5 cm), which could be the result of the increased workload in Experiment 3 (i.e., the lack of support from a Radar Associate), or because Experiment 1 recall was performed directly on a full-size replica of the radar, whereas Experiment 3 was performed using a trackball on a reduced map. The mean missed distance did vary significantly with condition, $F(3, 7) = 19.80$, $MSE = 1.415$ (one participant’s data were lost due to computer malfunction), contrary to expectations. It was due to better placement of traffic-level versus pretraffic aircraft, $t(10) = 3.67$, $sdiff = 0.213$, which could have been because the pretraffic aircraft usually had just entered the airspace prior to the scenario being stopped. Nevertheless, we are hesitant to claim that aircraft importance affects memory for PVD.

![Figure 2](image-url)
position because none of the important aircraft (pretraffic, traffic-level, traffic-climb) differed from the control (never-traffic).

We conducted three $2 \times 2 \times 4$—scenario type (Structured vs. Unstructured) $\times$ Question type (Speed vs. Altitude) $\times$ Aircraft importance—repeated-measures ANOVAs, one each for the exact, approximation, and relational questions. Mean performance for each question type for the four levels of aircraft importance are shown in Figure 3 (collapsed over scenario type). The percentage of responses to the exact question is shown in the top panel, the approximation question results are in the middle panel, and the relational question results are in the bottom panel. We begin with the exact questions.

There was an interaction between scenario type and type of question, $F(1, 9) = 11.29$, $MSE = 0.004$; the interaction was the result of worse altitude recall in the unstructured scenario: structured = 63%, unstructured = 50%, $t(9) = 4.95$, $s_{diff} = 0.024$. At present, this finding could be due to any number of potential differences between the scenarios (e.g., complexity, type of conflicts to be resolved). For example, it is possible that the unstructured scenario produced more complex flying in the vertical dimension. If this was the case, we would interpret the poorer altitude recall as reflecting this increased complexity. On the other hand, if the unstructured scenario produced more complex flying in the lateral direction, resources normally committed to remembering the altitude might be required to monitor aircraft vectoring. Irrespective of the reason, with the impending implementation of free flight, worse altitude recall in the unstructured scenario raises a cautionary flag and warrants further study.

There was a main effect of type of question, with recall of altitude superior to recall of speed, $F(1, 9) = 11.61$, $MSE = 0.022$, mean altitude = 56.5%, mean speed = 20.6%. There were no differences as a function of aircraft importance for the speed question (see top panel of Figure 3). Moreover, the only condition above chance for speed was the pretraffic condition (and only 3% above chance at that). For altitude recall, post-hoc tests (collapsed over scenario type, which did not interact with aircraft importance) showed that the traffic-level and pretraffic aircraft were superior to the traffic-climb and never-traffic aircraft, minimum $t(9) = 3.63$, $s_{diff} = 0.057$. The pattern of results was very similar for the approximation question (a response was scored as correct if the actual altitude or speed fell within the participant-selected range), except that type of question interacted with aircraft importance rather than scenario type, $F(1, 9) = 3.34$, $MSE = 0.035$, never-traffic speed relatively better than never-traffic altitude. This is shown in the middle panel of Figure 3.

It turned out that the poor performance for the traffic-climb condition, purportedly involving important aircraft, was somewhat misleading. In our

![Figure 3](image-url)
instantiation of this conflict, one aircraft in the pair was level and the other was not. When we scored the accuracy of these two aircraft separately we found that exact altitude recall for the level traffic-climb aircraft was 72.2%, comparable to that of the traffic-level and pretraffic aircraft. However, performance on the climbing traffic-climb aircraft was much worse, at only 18.3%. (The same was true for the approximation question.)

Were the participants aware of this? If so, we might detect some compensation on their part in the approximation question through the choice of a wider confidence range for this aircraft. The width of the participant-selected range did not differ with aircraft importance (altitude mean width = 2,280 feet, speed mean width = 37.6 knots). In fact, the mean range of the climbing traffic-climb aircraft was actually narrower than the level traffic-climb aircraft, \( t(9) = 3.06, S_{\text{diff}} = 8.056; \) mean climbing traffic-climb = 1,303 feet; mean level traffic-climb = 3,280).

The poor memory of controllers for the altitude of climbing traffic-climb aircraft might have been due to the phrasing of the questions: Although we asked for the assigned altitude of the climbing traffic-climb aircraft, some participants might have tried to estimate its current altitude, mid-climb. However, performance did not improve to the level achieved by the level traffic-climb aircraft when we rescored the data, counting as correct either the assigned altitude or \( \pm 2,000 \) feet around the current altitude (it improved by only 12%). Another possible reason for the poor performance of these climbing traffic-climb aircraft might be due to the design of the scenarios; these aircraft typically had just entered the airspace prior to the time that we stopped the scenario. However, the late entry into the scenario was also the case for the pretraffic aircraft, and altitude recall for those aircraft was very good.

Poor memory for the altitude (current or assigned) of aircraft that were climbing might signal poor SA. After all, 83% of the operational errors (loss of 5 mi horizontal or 1,000 feet vertical separation, or both) at en route facilities in 1993 involved aircraft changing altitudes (i.e., descending through an altitude occupied by a level or climbing flight; Durso et al., 1996). On the other hand, it might be sufficient that the controller remembered that the climbing aircraft was below, at, or above the aircraft with which it was in conflict. If so, we should see evidence that the altitude of the traffic-climb aircraft was remembered relationally.

The altitudes of traffic-climb aircraft were not remembered relationally either (see bottom panel of Figure 3). A \( 2 \times 2 \times 4 \) repeated-measures ANOVA was conducted on the relational questions. The three-way interaction was significant, \( F(3, 27) = 9.29, MSE = 0.075, \) as were the two two-way interactions: scenario type with question type, \( F(1, 9) = 7.76, MSE = 0.064, \) of the same form as for the exact question (controllers had poorer altitude recall, but not for ground speed, for the unstructured scenario), and question type with aircraft importance, \( F(1, 9) = 22.35, MSE = 0.033. \) There was also a main effect of type of question, \( F(1, 9) = 53.26, MSE = 0.068, \) (mean altitude = 57.9%; mean speed = 27.9%), and a main effect of aircraft importance, \( F(1, 9) = 5.05, MSE = 0.053. \) Post hoc tests on aircraft importance showed that there were no significant differences for speed. Most importantly, for altitude, traffic-level and pretraffic performance were superior to the never-traffic and traffic-climb conditions, minimum \( t(9) = 5.16, S_{\text{diff}} = 0.075. \) We could not separately analyze the climbing traffic-climb and the level traffic-climb aircraft because the relational question necessarily considered both aircraft.

Performance on the exact, relational, and approximate speed and altitude questions were compared by subtracting out an estimate of chance. Chance was 1/3 for relational responses (three possible responses), and our SME estimated chance to be 1/10 for exact responses. Chance was difficult to compute for the approximation questions as it depended on the width of the participant-selected range. However, one estimate would be 50%—the answer either fell within the range (which on average covered about half of the possible range) or it did not. We subtracted out these estimates of chance and compared performance across exact, relational, and approximate questions. In accord with Experiment 2, exact was best for altitude (exact = 46.5% vs. approximation = 19.2% and relational = 24.9%, all different from 0, exact greater than the other two, which did not differ). Controllers
remembered the exact altitude of the aircraft, especially the important aircraft.

For speed, only exact was above chance (exact = 10.6% vs. approximation = -6.8% and relational = -5.0%; negative percentages indicated that performance was below chance, although not significantly). In contrast to Experiment 2, speed was poorly remembered relationally. This might have been the result of the increased emphasis on speed in the Experiment 2 scenarios (i.e., the sequencing problems). Improved memory for flight data highlighted by a scenario was consistent with a second hypothesis proposed by Means et al. (1988): Type of control determined what flight data were remembered. For example, Means et al. found that vectoring an aircraft led to better retention of its routing information. If this hypothesis is true, speed was better remembered in Experiment 2 because it was more relevant to resolving conflicts. But interestingly, when speed was remembered, it was better remembered relationally. However, note that despite the increased emphasis on speed in Experiment 2, memory for exact speed was remembered equivalently in the two experiments, perhaps reflecting an incidental level of encoding for this information.

In sum, we have support for a hypothesis that assumes that controllers classify aircraft into two categories (important vs. unimportant) on the basis of their knowledge of the 2-D position of the aircraft. The utilization of a better control condition showed that importance affected the likelihood of recalling the exact altitude, but it did not affect memory for aircraft position (i.e., none of the important aircraft differed from the control). A deemphasis on the relevance of speed resulted in negligible recall of that information.

General Discussion

Situation awareness is assumed to be central to successful air traffic control performance (see Durso & Gronlund, in press; Endsley, 1995a), and 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?

In Experiment 1, we found no support for the hypothesis that the amount of control (above a minimum of one control action) affected the recall of aircraft flight data. This was true whether we operationalized amount of control as the number of control actions or simply as the number of interactions. The level of recall and the accuracy of the 2-D placement of the aircraft was good (see also Means et al., 1988; Vortac et al., 1993). The variables we examined did not differentially affect memory for this information. We argue that this was because it was necessary for the controllers to know the 2-D location of every aircraft on the PVD in order to make a determination of whether or not an aircraft was important. Direction of flight may also contribute to a determination of aircraft importance; Experiment 2 showed that this information was well remembered and not affected by aircraft importance. However, because direction of flight was not tested in Experiment 3, we have not established that it would be equally well remembered for never-traffic aircraft.

Because controllers classify aircraft as important or not, they were better able to distribute their mental resources. In Experiments 2 and 3, we found that the importance of the aircraft affected the likelihood of recalling some of its flight data. This result would be consistent with Means et al.'s (1988) original hypothesis if we assumed that the important aircraft received more control actions. Nevertheless, we feel that the operative factor was aircraft importance, not amount of control. Experiment 3 refined the importance hypothesis by showing that aircraft importance was a binary-valued construct; an aircraft either was not traffic for any other aircraft (unimportant), or it was or might be (important). Similar arguments were made by Bisseret (1971), who found that controllers grouped aircraft into two categories, conflicting and other. There are also parallels between our results and the results of a cognitive task analysis of controllers by Seamster, Redding, Cannon, Ryder, & Purcell (1993), which showed that controllers categorize aircraft into sector events.

The one exception to the importance hypothesis was the poor altitude memory for climbing
traffic-climb aircraft. This finding warrants further study, especially given the large percentage of operational errors involving aircraft that are changing altitude (Durso et al., 1996). At present, it is unclear to what extent this finding is a general problem versus something unique to our scenarios and methodology. If our results are indicative of a general problem, monitoring the altitude of climbing (and, presumably, descending) aircraft is a situation in which the controllers would benefit from extra assistance. Perhaps some kind of gist representation of the relevant information might help to alleviate the problem. For example, climbing aircraft could be color coded to signal their altitude relative to another aircraft or relative to various fixed altitudes. Alternatively, an interface could signal when the climbing aircraft was nearing 1,000 feet of another aircraft.

Experiments 2 and 3 showed that the controllers remembered altitude exactly rather than approximately or relationally. This was surprising, given the information reduction advantages of remembering relationally (Brainerd & Reyna, 1990; Moray, 1990). However, it may be a reflection of the way altitude information is presented to the controller—in the form of digital data tags on the radar and on the FPS. It could also be the case that we did not tap the appropriate gist representation for altitude (e.g., are there any other aircraft at the same altitude as AAL1237)? No one has tried to determine if memory for exact altitude is the best way for controllers to represent this information; we have shown that it is the way they presently remember it. There is a great deal of research on cockpit displays that attempts to find better ways to present information to the pilot (e.g., Haskell & Wickens, 1993). No doubt the same endeavor would be beneficial to the controller's work environment as well.

In contrast to altitude, exact ground speed was poorly remembered (see also Means et al., 1988). However, the increased relevance of speed in the sequencing problems of Experiment 2 was sufficient to produce a fairly high level of performance for relational speed. This provided support for the Means et al. hypothesis that the type of control can affect what is remembered. It is possible that the controllers better remembered relational speed because they could rely more on their knowledge than in the Experiment 3 scenarios. In seven of the Experiment 2 scenarios, the problem to be solved involved sequencing aircraft for approach to a nearby airport. Because the controllers remembered the positions of the aircraft on the radar, they knew which aircraft was closer to the airport and which aircraft would likely be slower. However, if ground speed, when it is remembered, is remembered relationally, it would mean that SA measures should tap it accordingly.

In sum, these experiments help delineate the processes whereby air traffic controllers gain and maintain SA. First, the controller considers the spatial position of all the aircraft on the PVD. Then, information that is not represented spatially (in our experiments, the altitude) is filtered by importance (potential conflict or not). Finally, a task-relevant filter can sometimes facilitate encoding of other information (e.g., ground speed) if it is of particular relevance to the task.

There are, of course, countervailing factors that should be considered. For example, in Experiments 1 and 3, neither the stopping points in the scenario nor the information requested was randomly determined (as has been recommended, see Endsley, 1995b), which would have made it impossible for controllers to do any special preparations in advance of the question period. But, neither were the stopping points entirely predictable (occurring at approximate, not fixed, 10-min intervals). Furthermore, the controllers were stopped only six times, and given how busy they were with their normal duties, we suspect that few if any ever attempted to anticipate the stopping times. A second countervailing factor was that any cross-experiment comparisons that one may wish to draw were confounded by workload differences (the SME served as the radar associate in Experiments 1 and 2) and scenario differences (traffic volume, complexity, type of problems to be solved). Because the effects of these factors on memory are unknown, it is possible that other factors besides (in addition to) the relevance of speed could be responsible for the differential memory for relational speed found between Experiments 2 and 3.

Our findings do not address the issue of whether the query technique is the best way to assess SA. Rather, our findings can aid in the interpretation of the results from the use of this technique. There are some aircraft about which the controller should remember more, and other
aircraft about which the controller could remember less. In other words, not all aircraft are equally important to the controller and assessments of SA should not assume that they are. For example, better memory for the altitude of important aircraft would signal better SA. When solving sequencing conflicts, poorer SA might be signaled by a controller who did not know the relative speeds of the aircraft involved; for other types of conflicts, it may not be important (or even be a sign of wasted cognitive resources) to remember the speed. Durso et al. (1998) also argued that information differed in its relevance to the controller; information about the present situation may be less relevant than information about the future situation. Durso et al. found that those controllers with better memory for the present situation performed less efficiently and had poorer appreciation of the future. Wickens (1996) made similar arguments regarding the importance of understanding the different aspects of a mission that were critical to maintaining SA at different times and of using that information to determine the appropriate ways to display information and support SA.

Of course, whether the asymmetric distribution of mental resources to important versus unimportant aircraft is an optimal strategy awaits improved performance measures that reflect moment-to-moment fluctuations. At this point, all we can be certain of is that the controllers do asymmetrically distribute their mental resources. If it did turn out to be optimal, it would be important to determine if the level of recall obtained for the flight data from unimportant aircraft reflected incidental encoding or a reduced level of normal encoding. If it reflected the latter, an improved interface or changes to training procedures could attempt to further redirect “wasted” resources away from the unimportant aircraft.

One finding that warrants additional study is that exact altitude was remembered relatively poorly for unstructured scenarios. This finding could be the result of many possible factors, including drawing attention away from altitude information because of the increased difficulty of detecting conflicts at all points in a sector rather than at designated intersection points. A cognitive aid that facilitates the finding of conflicts might ease the transition to a free-flight environment by freeing up mental resources to focus on altitude.

The cognitive abilities of air traffic controllers will be further stretched by the advent of free flight and the ever-increasing volume of air traffic. The continuing investigation of how controllers perform their jobs, and memory’s role in support of that job, can continue to be profitably applied to the development of cognitive aids and the redesign of interfaces. These modifications will allow controllers to better manage the complex dynamic air traffic control system of tomorrow.

References


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, 258–268.


Received September 17, 1996

Revision received March 9, 1998

Accepted March 12, 1998