HUMAN FACTORS, 1995, 37(1), 65–84

Measurement of Situation Awareness in Dynamic Systems

MICA R. ENDSLEY,[1] *Texas Tech University, Lubbock, Texas*

Methodologies for the empirical measurement of situation awareness are reviewed, including a discussion of the advantages and disadvantages of each method and the potential limitations of the measures from a theoretical and practical viewpoint. Two studies are presented that investigate questions of validity and intrusiveness regarding a query-based technique. This technique requires that a simulation of the operational tasks be momentarily interrupted in order to query operators on their situation awareness. The results of the two studies indicate that the query technique is not intrusive on normal subject behavior during the trial and does not suffer from limitations of human memory, which provides an indication of empirical validity. The results of other validity studies regarding the technique are discussed along with recommendations for its use in measuring situation awareness in varied settings.

INTRODUCTION

In the preceding paper in this issue (Endsley, 1995), I presented a definition and theory of situation awareness (SA), outlining its role in dynamic decision making and establishing a model that depicts the relevant factors and underlying mechanisms affecting the process of achieving SA. In this paper I explore the measurement of operator SA within particular system contexts.

A measure of SA is valuable in the engineering design cycle because it assures that prospective designs provide operators with this critical commodity. For this reason I review various methods that have been proposed or used for measuring SA and address the veracity of each technique in comparison with standard criteria. By presenting experiments evaluating the Situation Awareness Global Assessment Technique (SAGAT), I will establish its validity for measuring SA. A brief example of the use of the technique will be presented, which demonstrates the type of data the technique generates and its contribution in a display design evaluation effort. In addition, considerations for the use of the technique and its applicability within various system domains will be discussed.

SA is an understanding of the state of the environment (including relevant parameters of the system). It provides the primary basis for subsequent decision making and performance in the operation of complex, dynamic systems. SA is formally defined as a person's "perception of the elements of the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future" (Endsley, 1988, p. 97). At the lowest level of SA, a person needs to perceive relevant information (Level 1 SA). Integrating various pieces of data in conjunction with operator goals provides an understanding of the meaning of that information, forming Level 2 SA. Based on

[1] Requests for reprints should be sent to Mica R. Endsley, Department of Industrial Engineering, Texas Tech University, Lubbock, TX 79409.

this understanding, future events and system states can be predicted (Level 3), allowing for timely and effective decision making. Several processing mechanisms have been hypothesized to be related to SA, including attention and working memory limitations, attention distribution, current goals, mental models, schemata, and automaticity (Endsley, 1995, this issue). In addition to characteristics of individuals, the design of a system and the operator interface can affect SA.

The enhancement of SA has become a major design goal for developers of operator interfaces, automation concepts, and training programs in a variety of fields. To evaluate the degree to which new technologies or design concepts actually improve (or degrade) operator SA, it is necessary to systematically evaluate them based on a measure of SA, which can determine those ideas that have merit and those that have unforeseen negative consequences. Only in this way can a design concept's utility be established in accordance with the design goal.

In addition to evaluating design concepts, a measure of SA is also useful for evaluating the effect of training techniques on SA; conducting studies to empirically examine factors that may affect SA, such as individual abilities and skills, or the effectiveness of different processes and strategies for acquiring SA; and investigating the nature of the SA contruct itself. (This type of quantitative measurement—that is, evaluating an operator's level of SA in light of concepts under experimental evaluation—differs significantly from research efforts that seek to directly assess the processes operators use in acquiring SA, as discussed by Adams, Tenney, and Pew [1995, this issue].)

Evaluation of potential design concepts routinely takes place within the context of rapid prototyping or part-task simulation. Several different methods for measuring SA can be considered during this type of testing. In addition, many of these techniques could be used to evaluate the SA of operators working with actual systems. For the most part, efforts at measuring SA thus far have been concentrated in the aircraft environment; however, most techniques are also applicable to other domains.

To establish the validity and reliability of an SA measurement technique, it is necessary to establish that the metric (a) measures the construct it claims to measure and is not a reflection of other processes, (b) provides the required insight in the form of sensitivity and diagnosticity, and (c) does not substantially alter the construct in the process, which would provide biased data and altered behavior. In addition, it would be useful to establish the existence of a relationship between the measure and other constructs, as would be predicted by theory. In this case, the measure of SA should be predictive of performance and sensitive to manipulations in workload and attention. Against these criteria, several potential measurement techniques will be explored, along with an assessment of the advantages and disadvantages of each.

Physiological Techniques

P300 and other electroencephalographic measurements have shown promise for determining whether information is registered cognitively. Although these techniques allow researchers to determine whether elements in the environment are perceived and processed by subjects, they cannot determine how much information remains in memory, whether the information is registered correctly, or what comprehension the subject has of those elements. For the same reasons, eye-tracking devices appear to fall short in their ability to measure SA as a state of knowledge. Furthermore, these devices do not tell which elements in the periphery of the subject's vision are observed, or if the subject has even processed what was seen. (The devices may be useful for exploring the processes operators use in achieving SA, however, as was recently investigated by Smolensky [1993].) Known physiological techniques, though providing useful data for other purposes, are not very promising for the measurement of SA as a state of knowledge.

Performance Measures

In general, performance measures have the advantage of being objective and are usually nonintrusive. Compute simulations can be pro ified performance data the required data relati ever, several limitatio mance data to infer SA

Global measures. Gl mance suffer from diag problems. If overall mance is the only crite design concepts, impo can be masked, despit tom-line measure. Per only the end result of processes, providing why poor performance given situation (assum liably during testing a could occur from a la sampling strategies, i projection, heavy work ing, or action errors, ar of which are not SA rel the distinction among performance, see the p this issue.)

As stated earlier, ove is often masked by ot aircraft environment, f performance is, by nat subject to the influenc besides SA. A new syst with better SA, but i fact can be easily ma loads, the intentional poor decision making mance is used as the o would be desirable, t more directly.

External task measu mance measure that volves artificially chan or removing certain pi operator displays ar amount of time requir act to this event (Sarte from the fact that such

nonintrusive. Computers for conducting system simulations can be programmed to record specified performance data automatically, making the required data relatively easy to collect. However, several limitations exist in using performance data to infer SA.

Global measures. Global measures of performance suffer from diagnosticity and sensitivity problems. If overall operator/system performance is the only criterion used for evaluating design concepts, important system differences can be masked, despite its being a useful bottom-line measure. Performance measures give only the end result of a long string of cognitive processes, providing little information about why poor performance may have occurred in a given situation (assuming it can be detected reliably during testing at all). Poor performance could occur from a lack of information, poor sampling strategies, improper integration or projection, heavy workload, poor decision making, or action errors, among other factors, many of which are not SA related. (For a discussion of the distinction among SA, decision making, and performance, see the previous Endsley paper in this issue.)

As stated earlier, overall system performance is often masked by other factors. In a tactical aircraft environment, for instance, much of pilot performance is, by nature, highly variable and subject to the influence of many other factors besides SA. A new system may provide the pilot with better SA, but in evaluation testing this fact can be easily masked by excessive workloads, the intentional use of varied tactics, or poor decision making if overall mission performance is used as the only dependent measure. It would be desirable, therefore, to measure SA more directly.

External task measures. One type of performance measure that has been suggested involves artificially changing certain information or removing certain pieces of information from operator displays and then measuring the amount of time required for the operator to react to this event (Sarter and Woods, 1991). Aside from the fact that such a manipulation is heavily intrusive, requiring the subject to undertake tasks involved with discovering what happened to the changed or missing data while attempting to maintain satisfactory performance on other tasks, this technique may provide highly misleading results. It assumes that an operator will act in a certain way when, in fact, operators often employ compensating schemes to function under such circumstances. For instance, if a displayed aircraft suddenly disappears, the operator may assume that equipment has malfunctioned, the aircraft was destroyed or landed, or the aircraft's emitting equipment was turned off, rendering it more difficult to detect (a not infrequent occurrence). In any case, the operator may choose to ignore the disappearance, assuming it will come back on the next several sweeps of the radar, worry about it but not say anything, or put off dealing with the disappearance until other tasks are complete. Any of these actions will yield highly misleading results for the experimenter who expects SA to be reflected by the operator's overt behavior.

This measurement technique not only has invalid assumptions but it alters the subject's ongoing tasks, possibly affecting attention and thus SA itself. Anytime one artificially alters the realism of the simulation, it can fundamentally change the way the operator conceptualizes the underlying information (see Manktelow and Jones, 1987), thus altering both SA and decision making. In addition, such a manipulation would certainly interfere with any concurrent workload or performance measurement undertaken during testing.

Imbedded task measures. Some information about SA can be determined from examining performance on specific operator subtasks that are of interest. For example, when evaluating an altitude display, deviations from prescribed altitude levels or time to reach a certain altitude can be measured. This type of detailed performance measure can provide some inferences regarding the amount of SA that a display reveals about a specific parameter. Such measures will be more meaningful than global performance measures and will not suffer from the

same problems of intrusiveness as external task measures. Although selecting finite task measures for evaluating certain kinds of systems may be easy, determining appropriate measures for other systems may be more difficult. An expert system, for instance, may influence many factors in a global, not readily predicted manner.

The major limitation of the imbedded task measure approach stems from the interactive nature of SA subcomponents. A system that provides SA on one element may simultaneously reduce SA on another, unmeasured element. In addition, it is easy for subjects to have biased attention on the issue under evaluation in a particular study (e.g., altitude) if they figure out the purpose of the study. Therefore, because improved SA on some elements may result in decreased SA on others, relying solely on performance measures of specific parameters can yield misleading results.

Researchers need to know how much SA operators have when taxed with multiple, competing demands on their attention during system operations. To this end, a global measure of SA that simultaneously depicts SA across many elements of interest is desirable. To improve SA, designers need to be able to evaluate the entire impact of design concepts on operator SA.

Subjective Techniques

Self-rating. One simple technique is to ask operators to subjectively rate their own SA (e.g., on a 1-to-10 scale). Researchers in the Advanced Medium-Range Air-to-Air Missile Operation Utility Evaluation Test (AMRAAM OUE) study (McDonnell Douglas Aircraft Corporation, 1982) used this method. Pilot (and overall flight) SA was subjectively rated by the participants and a trained observer. The main advantages of subjective estimation techniques are low cost and ease of use. In general, however, the subjective self-rating of SA has several limitations.

If the ratings are collected during a simulation trial, the operators' ability to estimate their own SA will be limited because they do not know what is really happening in the environment (they have only their perceptions of that reality). Operators may know when they do not have a clue as to what is going on but will probably not know if their knowledge is incomplete or inaccurate.

If operators are asked to subjectively evaluate SA in a posttrial debriefing session, the rating may also be highly tainted by the outcome of the trial. When performance is favorable, whether through good SA or good luck, an operator will most likely report good SA, and vice-versa. In a reevaluation of the AMRAAM OUE study, Venturino, Hamilton, and Dvorchak (1990) found that posttrial subjective SA ratings were highly correlated with performance, supporting this premise. In addition, when ratings are gathered after the mission, operators will probably be inclined to rationalize and overgeneralize about their SA, as has been shown when information about mental processes is collected after the fact (Nisbett and Wilson, 1977). Thus detailed information will be lost or misconstrued.

What do such subjective estimates actually measure? I would speculate that self-ratings of SA most likely convey a measure of subjects' confidence levels regarding that SA—that is, how comfortable they feel about their SA. Subjects who know they have incomplete knowledge or understanding would subjectively rate their SA as low. Subjects whose knowledge may not be any greater but who do not subjectively have the same concerns about how much is not known would rate their SA higher. In other words, ignorance may be bliss.

Several efforts have been made to develop more rigorous subjective measures of SA. Taylor (1990) developed the Situation Awareness Rating Technique (SART), which allows operators to rate a system design based on the amount of demand on attentional resources, supply of attentional resources, and understanding of the situation provided. As such, it considers operators' perceived workload (supply and demand on attentional resources) in addition to their perceived understanding of the situation. Although SART has been shown to be correlated with performance measures (Selcon and Taylor, 1990), it is unclear w
table to the workloa
components.

In a new application,
Dominance (SWORD)
has been applied as a s
SA (Hughes, Hassoun, a
allows subjects to mak
ratings of competing c
continuum that expres
one concept entails less
The resultant preferenc
ing the analytic hierarc
provide a linear orderin
In the Hughes et al. stu
fied to allow pilots to r
presented display conce
of workload. Not surpr
was subjectively rated
using SWORD was the
expressed a strong pre
ascertain whether subj
higher SWORD ratings
ther research is needed
such ratings.

Observer-rating. A se
rating involves using i
able observers to rate t
SA. A trained observer
mation than the opera
happening in a simul
knowledge gleaned fro
puter). The observer wo
edge, however, about v
cept of the situation is
about the operator's pe
would come from oper
ded or elicited verbali
(e.g., from voice transm
of the task or from ve
requested by the exper
knowledge can be usef
example, detecting ov
misperceptions or lack
provide a complete
knowledge. The operat
lihood, will—store infor

1990), it is unclear whether this is attributable to the workload or the understanding components.

In a new application, the Subjective Workload Dominance (SWORD) metric (Vidulich, 1989) has been applied as a subjective rating tool for SA (Hughes, Hassoun, and Ward, 1990). SWORD allows subjects to make pairwise comparative ratings of competing design concepts along a continuum that expresses the degree to which one concept entails less workload than the other. The resultant preferences are then combined using the analytic hierarchy process technique to provide a linear ordering of the design concepts. In the Hughes et al. study, SWORD was modified to allow pilots to rate the degree to which presented display concepts provided SA instead of workload. Not surprisingly, the display that was subjectively rated as providing the best SA using SWORD was the display for which pilots expressed a strong preference. It is difficult to ascertain whether subjective preference led to higher SWORD ratings of SA, or vice-versa. Further research is needed to determine the locus of such ratings.

Observer-rating. A second type of subjective rating involves using independent, knowledgeable observers to rate the quality of a subject's SA. A trained observer might have more information than the operator about what is really happening in a simulation (through perfect knowledge gleaned from the simulation computer). The observer would have limited knowledge, however, about what the operator's concept of the situation is. The only information about the operator's perception of the situation would come from operator actions and imbedded or elicited verbalizations by the operator (e.g., from voice transmissions during the course of the task or from verbal protocols explicitly requested by the experimenter). Although this knowledge can be useful diagnostically by, for example, detecting overt errors in SA (stated misperceptions or lack of knowledge), it does not provide a complete representation of that knowledge. The operator may—and, in all likelihood, will—store information internally that is not verbalized. For instance, a pilot may discuss efforts to ascertain the identity of a certain aircraft, but knowledge about ownship system status, heading, other aircraft, and so on might not be mentioned at all. An outside observer has no way of knowing whether the pilot is aware of these variables but is not discussing them because they are not of immediate concern, or whether the pilot has succumbed to attentional narrowing and is unaware of their status. As such, the rating of SA by outside observers is also limited.

A variation on this theme is to use a confederate who acts as an associate to the operator (e.g., another crew member or air traffic controller in the case of a commercial aircraft) and who requests certain information from the operator to encourage further verbalization, as has been suggested by Sarter and Woods (1991). In addition to this technique succumbing to the same limitations encumbering observer ratings, it may also serve to alter SA by artificially directing the subject's attention to certain parameters. Because the distribution of the subject's attention across the elements in the environment largely determines SA, this method probably does not provide an unbiased assessment of operator SA.

Questionnaires

Questionnaires allow for detailed information about subject SA to be collected on an element-by-element basis that can be evaluated against reality, thus providing an objective assessment of operator SA. This type of assessment is a more direct measure of SA (i.e., it taps into rather than infers the operator's perceptions) and does not require subjects or observers to make judgments about situational knowledge on the basis of incomplete information, as subjective assessments do. In a review of such questionnaires. Herrmann (1984) concluded that when perceptions can be evaluated on the basis of objective knowledge, this method has been found to have good validity. Several methods of administration are possible.

Posttest. A detailed questionnaire can be

administered after the completion of each simulated trial. This allows ample time for subjects to respond to a lengthy and detailed list of questions about their SA during the trial, providing needed information about subject perceptions. Unfortunately, people are not good at reporting detailed information about past mental events, even recent ones; there is a tendency to overgeneralize and overrationalize. Recall is stilted by the amount of time and intervening events that occur between the activities of interest and the administration of the questionnaire (Nisbett and Wilson, 1977). Earlier misperceptions can be quickly forgotten as the real picture unfolds during the course of events. Therefore, a posttest questionnaire will reliably capture the subject's SA only at the very end of the trial.

On-line. One way of overcoming this deficiency is to ask operators about their SA while they are carrying out their simulated tasks. Unfortunately, this, too, has several drawbacks. First of all, in many situations the subject will be under very heavy workload, precluding the answering of additional questions. Such questions would constitute a form of ongoing secondary task loading that may alter performance on the main task of operating the system. Furthermore, the questions asked could cue the subject to attend to the requested information on the displays, thus altering the operator's true SA. An assessment of time to answer as an indicator of SA is also faulty, as subjects may employ various time sharing strategies between the dual tasks of operating the system and answering the questions. Overall, this method will be highly intrusive on the primary task of system operation.

Freeze technique. To overcome the limitations of reporting on SA after the fact, several researchers have used a technique in which the simulation is frozen at randomly selected times and subjects are queried as to their perceptions of the situation at the time (Endsley, 1987, 1988, 1989a; Fracker, 1990; Marshak, Kuperman, Ramsey, and Wilson, 1987). With this technique the system displays are blanked and the simulation is suspended while subjects quickly answer questions about their current perceptions of the situation. Thus SA data can be collected immediately, which reduces the problems incurred when collecting data after the fact but does not incur the problems of on-line questioning. Subjects' perceptions are then compared with the real situation according to the simulation computer database to provide an objective measure of SA.

In a study evaluating competing aircraft display concepts, Marshak et al. (1987) queried subjects about navigation, threats, and topography using this technique. The resulting answers were converted to an absolute percentage error score for each question, allowing scores across displays to be compared along these dimensions. Fracker (1990), in another aircraft study, requested information on aircraft location and identity for specifically indicated aircraft in the simulation. Mogford and Tansley (1991) collected data relevant to an air traffic control task. In each study, a measure of SA on only selected parameters was obtained.

The Situation Awareness Global Assessment Technique is a global tool developed to assess SA across all of its elements based on a comprehensive assessment of operator SA requirements (Endsley, 1987, 1988, 1990a). As a global measure, SAGAT includes queries about all SA requirements, including Level 1, 2, and 3 components, and considers system functioning and status and relevant features of the external environment.

Computerized versions of SAGAT have been developed for air-to-air tactical aircraft (Endsley, 1990c) and advanced bomber aircraft (Endsley, 1989a). These versions allow queries to be administered rapidly and SA data to be collected for analysis. The SAGAT tool features an easy-to-use query format that was designed to be as compatible as possible with subject knowledge representations. An example of a SAGAT query is shown in Figure 1. SAGAT's basic methodology is generic and applicable to other types of systems once a delineation of SA requirements has been made in order to construct the queries.

This global approach l
probes that cover only a
items, in that subjects c
queries in advance beca
dress nearly every aspe
which subjects would nc
fore, the chance of subje
ased toward specific item
dition, the global appr
problems incurred when
the fact and minimizes su
from secondary task load
the subject's attention. It
measure of SA that can b
and objectively evaluate
random sampling (both c
presentation) provides u
SA, thus allowing SA sc
pared statistically acros
systems. The primary dis
nique involves the tempo
lation.

I conducted two studie
er or not this imposes u
to determine the best me
this technique. In the firs
cifically determine how l
simulation SA informatio

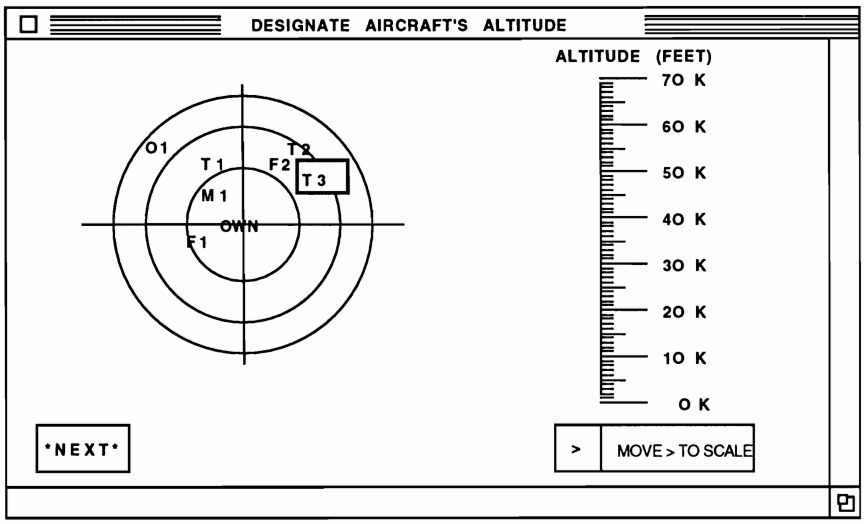


Figure 1. *Sample SAGAT query.*

This global approach has an advantage over probes that cover only a limited number of SA items, in that subjects cannot prepare for the queries in advance because queries could address nearly every aspect of the situation to which subjects would normally attend. Therefore, the chance of subjects' attention being biased toward specific items is minimized. In addition, the global approach overcomes the problems incurred when collecting data after the fact and minimizes subject SA bias resulting from secondary task loading or artificially cuing the subject's attention. It also provides a direct measure of SA that can be objectively collected and objectively evaluated. Finally, the use of random sampling (both of stop times and query presentation) provides unbiased estimates of SA, thus allowing SA scores to be easily compared statistically across trials, subjects, and systems. The primary disadvantage of this technique involves the temporary halt in the simulation.

I conducted two studies to investigate whether or not this imposes undue intrusiveness and to determine the best method for implementing this technique. In the first study I wanted to specifically determine how long after a freeze in the simulation SA information could be obtained. In the second study I investigated whether temporarily freezing the simulation would result in any change in subject behavior, thus specifically evaluating potential intrusiveness.

EXPERIMENT 1

In determining whether SA information is reportable via the SAGAT methodology, several possibilities must be considered. First, data may be processed by subjects in short-term memory (STM), never reaching long-term memory (LTM). If a sequential information-processing model is used, then it is possible that information might enter into STM and never be stored in LTM, where it would be available for retrieval during questioning. In this case, information would not be available during any SAGAT query sessions that exceeded the STM storage limitations (approximately 30 s with no rehearsal).

There is a good deal of evidence, however, that STM does not precede LTM but constitutes an activated subset of LTM (Cowan, 1988; Morton, 1969; Norman, 1968). According to this type of model, information proceeds directly from sensory memory to LTM, which is necessary for pattern recognition and coding. Only those portions of the environment that are salient are then highlighted in STM (either through focalized

attention or automatic activation). This type of model would predict that SA information that has been perceived and/or further processed by the operator would exist in LTM stores and thus be available for recall during SAGAT querying that exceeds 30 s.

Second, data may be processed in a highly automated fashion and thus not be in the subject's awareness. Expert behavior can function in an automated processing/action sequence in some cases. Several authors have found that even when effortful processing is not used, however, the information is retained in LTM and is capable of affecting subject responses (Jacoby and Dallas, 1981; Kellog, 1980; Tulving, 1985). The type of questions used in SAGAT—providing cued-recall and categorical or scalar responses—should be receptive to retrieval of this type of information. In addition, a review of the literature indicates that even under these circumstances, the product of the automatic processes (as a state of knowledge) is available, even though the processes themselves may be difficult to verbalize (Endsley, 1995, this issue).

Third, the information may be in LTM but not easily recalled by the subjects. Evidence suggests that when effortful processing and awareness are used during the storage process, recall is enhanced (Cowan, 1988). SA, composed of highly relevant, attended to, and processed information, should be most receptive to recall. In addition, the SAGAT battery, which requires categorical or scalar responses, is a cued, as opposed to total, recall task, thus aiding retrieval. Under conditions of SAGAT testing, subjects are aware that they may be asked to report their SA at any time. This, too, may aid in the storage and retrieval process. Because the SAGAT battery is administered immediately after the freeze in the simulation, no time for memory decay or competing event interference is allowed. Thus conditions should be optimized for the retrieval of SA information. Although it cannot be said conclusively that all of a subject's SA can be reflected in this manner, the vast majority should be reportable via SAGAT.

To further investigate this matter, I conducted a study to specifically determine how long after a freeze in the simulation SA information could be obtained. I expected that if collection of SA data via this technique was memory limited, this would be evidenced by an increase in errors on SAGAT queries that occurred approximately 30 to 60 s after the freeze time because of short-term memory restrictions. (Although this would not preclude use of the technique, it would limit its use by restricting the number of questions that could be asked at each stop.)

Procedure

A set of air-to-air engagements was conducted in a high-fidelity aircraft simulation facility. A fighter sweep mission with a force ratio of two (blue team) versus four (red team) was used for the trials. The objective of the blue team was to penetrate red territory, maximizing the kills of red fighters while maintaining a high degree of survivability. The red team was directed to fly around their assigned combat air patrol (CAP) points until a blue target was detected in red airspace. The red team was then allowed to leave its CAP point to defend against the blue team. In all cases, specific tactics were at the discretion of the individual pilot teams.

A total of 15 trials were completed by two teams of six subjects. At a random point in each trial, the simulator was frozen and SAGAT data were immediately collected from all six participants. At a given stop, all of the queries were presented once in random order. As this order was different at each stop, each query was presented a variety of times after the stop across subjects and trials. Therefore, performance accuracy on each question could be evaluated as a function of the amount of time elapsed prior to its presentation. After all subjects had completed the SAGAT battery at a given stop, a new trial was begun.

Prior to conducting the study, all subjects were trained on the use of the simulator, the displays, aircraft handling qualities, and SAGAT. In addition to three instructional training sessions on using SAGAT, each subject participated in 18 practice trials in which SAGAT

was administered. (Most
ceived a substantial amou
simulator in the past.) Th
trained prior to testing.

Facilities. A high-fidelity
facility was used for testing
tems, avionics systems, we
stations, and air vehicle pe
eled to provide realistic ai
teristics. A Gould mainfr
trolled the simulations
Graphics–generated hig
graphics displays. This te
tions, each including a sir
play, a tactical situation di
tem controls operated by a
and throttle control switc
and throttle provided prim

Subjects. Twelve experie
fighter pilots participated i
subject age was 48.16 year
Subjects had an average of
6500) and an average of 15
of military flight experienc
7 had combat experience.

Results

Each of the subject's an
queries was compared with
lected by the simulator cor
each stop. Included in the S
time of the test were 26 c
could not be evaluated bec
simulator data were not av
be evaluated only by subje
remaining query answers
heading, ownship location,
craft detections, aircraft
weapon selection, aircraft C
aircraft weapon quantity, a

An error score (SAGAT
actual value) was compute
Absolute error scores for ea
subjects and trials) were
amount of time elapsed bet
simulation and the query, a
calculated. *F* tests for the r

was administered. (Most subjects also had received a substantial amount of training in the simulator in the past.) Thus subjects were well trained prior to testing.

Facilities. A high-fidelity, real-time simulation facility was used for testing. Aircraft control systems, avionics systems, weapons systems, crew stations, and air vehicle performance were modeled to provide realistic aircraft control characteristics. A Gould mainframe computer controlled the simulations and drove Silicon Graphics–generated high-resolution color graphics displays. This test used six pilot stations, each including a simulated head-up display, a tactical situation display, radar, and system controls operated by a touch screen or stick and throttle control switches. A realistic stick and throttle provided primary flight control.

Subjects. Twelve experienced former military fighter pilots participated in the test. The mean subject age was 48.16 years (range of 32 to 68). Subjects had an average of 3310 h (range 1500–6500) and an average of 15.5 years (range 7–26) of military flight experience. Of the 12 subjects, 7 had combat experience.

Results

Each of the subject's answers to the SAGAT queries was compared with actual values, as collected by the simulator computer at the time of each stop. Included in the SAGAT battery at the time of the test were 26 queries. Of those, 11 could not be evaluated because the appropriate simulator data were not available, and 5 could be evaluated only by subjective means. The 10 remaining query answers concerned ownship heading, ownship location, aircraft heading, aircraft detections, aircraft airspeed, aircraft weapon selection, aircraft Gs, aircraft fuel level, aircraft weapon quantity, and aircraft altitude.

An error score (SAGAT query answer minus actual value) was computed for each response. Absolute error scores for each query (across all subjects and trials) were plotted against the amount of time elapsed between the stop in the simulation and the query, and a regression was calculated. F tests for the regressions and Pearson's r^2 for eight of the queries are presented in Table 1. Two of the queries provided a categorical response; therefore, chi-square tests were performed and are presented in Table 2.

None of the regressions or chi-square tests computed for each of the 10 queries was significant (or even close to significant), indicating that subjects were neither more nor less prone to response error as the amount of time between the simulator freeze and the presentation of the query increased. A plot of the regression for altitude error, shown in Figure 2, reveals little or no increase in error over time. (Plots of the regressions for the other variables appeared quite similar.)

Discussion

Based on these data, it would appear that subjects under these conditions were able to provide information on their SA about a particular situation for up to 5 or 6 min. The fact that all 10 of the queries produced flat regressions lends extra weight to this conclusion.

Two explanations can be offered for these findings. First, this study investigated subjects who were expert in particular tasks during a realistic simulation of those tasks. The information subjects were asked to report was important to performance of those tasks. Most laboratory studies that predict fairly rapid decay times (approximately 30 s for short-term memory) typically employ the use of stimuli that have little or no inherent meaning to the subject (nonsense words or pictures). The

TABLE 1

Tests on Regressions of Time of Query Presentation on Query Accuracy

Query	*df*	*F*	*p*	r^2
Own heading	1,19	0.021	0.887	0.001
Own location	1,19	0.940	0.334	0.047
Aircraft heading	1,95	0.040	0.834	0.000
G level	1,90	0.500	0.480	0.006
Fuel level	1,95	1.160	0.284	0.012
Weapon quantity	1,92	0.001	0.981	0.000
Altitude	1,96	0.002	0.965	0.000
Detection	1,524	0.041	0.840	0.000

TABLE 2

Chi-Square Tests of Time Category of Query Presentation on Query Accuracy

Query	*df*	χ^2	*p*
Weapon selection	5 (N = 97)	0.16	>0.995
Airspeed	5 (N = 98)	0.13	>0.995

storage and utilization of relevant information may be quite different from that of irrelevant information (Chase and Simon, 1973).

Second, the results indicate that the SA information was obtainable from long-term memory stores. If schemata, or other mechanisms, are used to organize SA information (as opposed to working memory processes only), then that information will be resident in long-term memory. Many of the 10 items analyzed can be considered Level 1 SA components. The fact that this lower-level information was resident in LTM indicates that either the inputs to higher-level processing were retained as well as the outputs or the Level 1 components were retained as important pieces of pilot SA and are significant components of LTM schemata (e.g., target altitude itself is important to know and not just its implications). Both of these explanations may be correct. These findings generally support the predictions of a processing model in which information passes into LTM storage before being highlighted in STM.

As a caveat, note that subjects were actively working with their SA knowledge by answering the SAGAT queries for the entire period that the simulation was stopped. No intervening period of waiting or any competing activity was introduced prior to administering any SAGAT query. Subject knowledge of SA information may be interfered with if time delays or other activities (particularly continued operational tasks) are imposed before SAGAT is administered. The major implication of these results is that, under these conditions, SA data are readily obtainable through SAGAT for a considerable period after a stop in the simulation (up to 5 or 6 min).

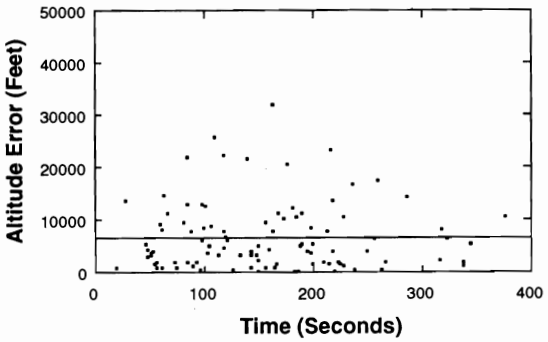


Figure 2. *Altitude error by time until query presentation.*

EXPERIMENT 2

A second study was initiated to address the issue of possible intrusiveness. It is necessary to determine whether temporarily freezing the simulation results in any change in subject behavior. If subject behavior were altered by such a freeze, certain limitations would be indicated. (One might wish to not resume the trial after the freeze, for instance, but start a new trial instead.) The possibility of intrusiveness was tested by evaluating the effect of stopping the simulator on subsequent subject performance.

Procedure

A set of air-to-air engagements was conducted of a fighter sweep mission with a force ratio of two (blue team) versus four (red team). The training, instructions, and pilot mission objectives were identical to those used in Experiment 1. In this study, however, the trial was resumed after a freeze following a specified period for collecting SAGAT data and was continued until specified criteria for completion of the mission were met. Subjects completed as many queries as they could during each stop. The queries were presented in random order.

Five teams of six subjects completed a full test matrix. The independent variables were duration of the stops (½, 1, or 2 min) and the frequency of stops (1, 2, or 3 times during the trial). Each team participated twice in each of these nine conditions. (In any given trial, multiple stops were of the same duration.) Each team also completed six trials in which no stops occurred as a control condition. Therefore, a total

of 30 trials were co
stop condition, and
ducted for each fi
These conditions w
in which no stops c
ministered in rand
Pilot performance
dent measure.

Facilities. This s
multiple-engageme
ment 1.

Subjects. Twenty
itary fighter pilot
(Four of the subjec
one team.) The mea
(range of 32 to 68)
3582 h (range of 97
16.9 years (range
experience. Of the
experience.

Results

Pilot performanc
tions was analyzed
included the num
team losses) and th
(red team kills).
formed on the num
N = 120) = 5.97
losses, χ^2 (2, N = 1
ing between trials
collect SAGAT dat
were no stops, as
There was no signi
formance on either

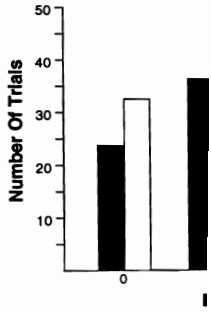


Figure 3. *Blue kills: 1*
out stops.

of 30 trials were conducted for each duration of stop condition, and a total of 30 trials were conducted for each frequency of stop condition. These conditions were compared with 30 trials in which no stops occurred. Conditions were administered in random order, blocked by team. Pilot performance was collected as the dependent measure.

Facilities. This study used the same piloted, multiple-engagement simulator as in Experiment 1.

Subjects. Twenty-five experienced former military fighter pilots participated in this test. (Four of the subjects participated on more than one team.) The mean subject age was 45.16 years (range of 32 to 68). Subjects had an average of 3582 h (range of 975 to 7045) and an average of 16.9 years (range of 6 to 27) of military flight experience. Of the 25 subjects 14 had combat experience.

Results

Pilot performance under each of the conditions was analyzed. Pilot performance measures included the number of blue team kills (red team losses) and the number of blue team losses (red team kills). Chi-square tests were performed on the number of blue team kills, χ^2 (4, N = 120) = 5.973, $p > 0.05$, and blue team losses, χ^2 (2, N = 120) = 0.05, $p > 0.05$, comparing between trials in which there were stops to collect SAGAT data and those in which there were no stops, as depicted in Figures 3 and 4. There was no significant difference in pilot performance on either measure.

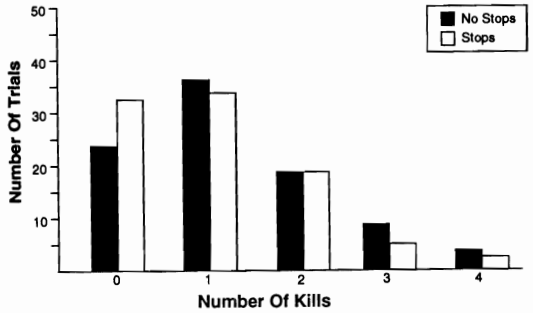


Figure 3. *Blue kills: Trials with stops versus trials without stops.*

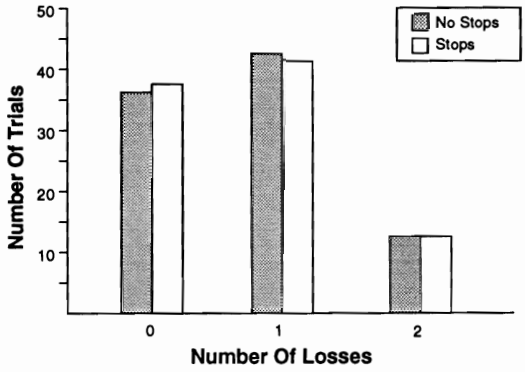


Figure 4. *Blue losses: Trials with stops versus trials without stops.*

Analysis of variance was used to evaluate the effect of number of stops and duration of stops on each of the two performance measures: blue team kills and blue team losses. The number of stops during the trial had no significant effect on either pilot performance measure, $F(3,116)$ = 1.73, $p > 0.05$, β = 0.55, and $F(3,116)$ = 0.20, $p > 0.05$, $\beta > 0.60$, respectively, as shown in Figures 5 and 6. The duration of the stop also did not significantly affect either performance measure, $F(3,116)$ = 2.16, $p > 0.05$, β = 0.40, and $F(3,116)$ = 0.77, $p > 0.05$, $\beta > 0.60$, respectively, as depicted in Figures 7 and 8. In viewing the data, no linear trend appears to be present in

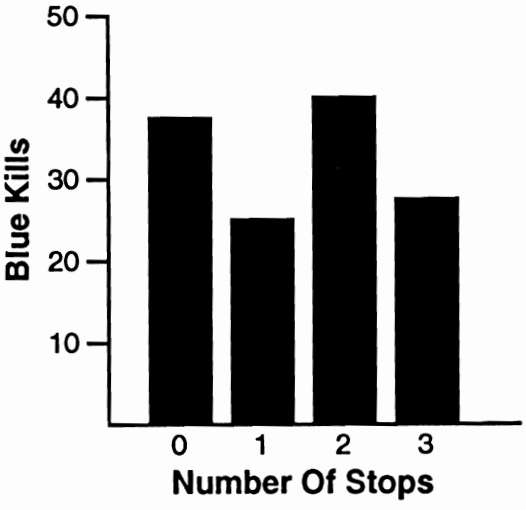


Figure 5. *Blue kills by number of stops in trial.*

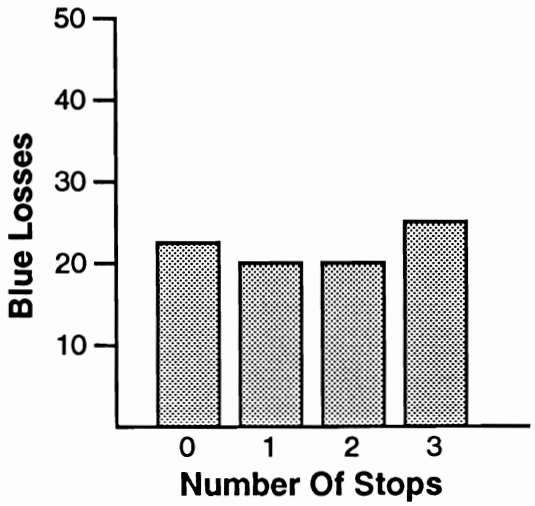


Figure 6. *Blue losses by number of stops in trial.*

either case that would indicate a progressively worse (albeit nonsignificant) effect of increasing number or duration of stops. This indicates that stops to collect SAGAT data (as many as 3 for up to 2 min in duration) did not have a significant effect on subject performance.

Discussion

The lack of a significant influence of this procedure on performance probably rests on the fact that relevant schemata are actively used by subjects during the entire freeze period. Under these conditions, the subjects' SA does not have a chance to decay before the simulation is resumed (as was indicated by Experiment 1). Thus SA is fairly intact when the simulation continues, allowing subjects to proceed with their tasks where they left off. Subjectively, subjects did fairly well with this procedure and were able to readily pick up the simulation at the point at which they left off at the time of the freeze, sometimes with the same sentence they had started before the stop. On many occasions, subjects could not even remember if they had been stopped to collect SA data during the trial, also indicating a certain lack of intrusiveness.

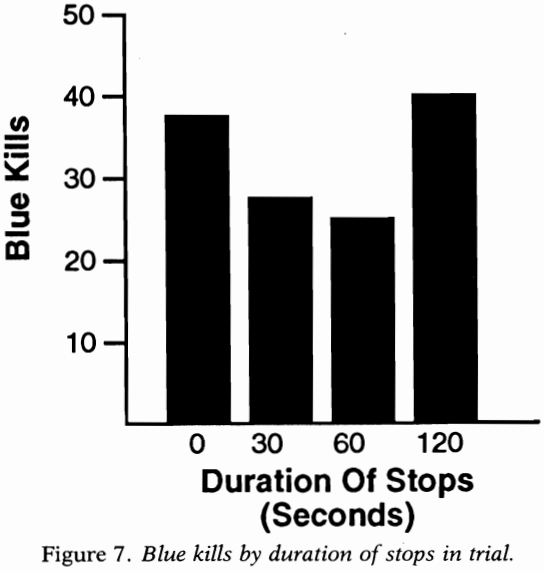


Figure 7. *Blue kills by duration of stops in trial.*

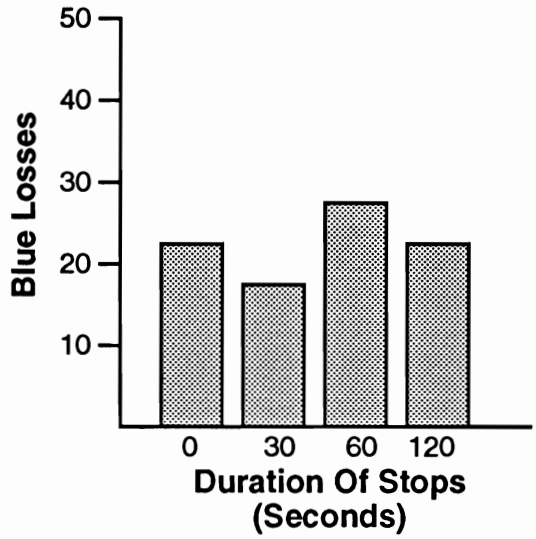


Figure 8. *Blue losses by duration of stops in trial.*

Summary

The use of a temporary freeze in the simulation to collect SA data is supported by Experiments 1 and 2. Subjects were able to report their SA using this technique for as long as 5 or 6 min without apparent memory decay, and the freeze did not appear to be intrusive on subject performance in the simulation, allaying several concerns about the technique. Although it is always difficult to establish no effect of one variable on another (i.e., prove the null hypothesis), it is

reassuring that the find
performance has been
other studies in which
evaluate human-mach
(Bolstad and Endsley,
Northrop, 1988), if the
unpredictable to the su
This finding was also r
which pilots flew in a s
puter-controlled ener
1990d), helping to rule
the finding of no effect
present study could hav
lots on both sides of
equally influenced.

SAGAT has thus far p
criteria for measureme
present studies establish
validity, SAGAT has bee
tive validity (Endsley, 1
ity (Endsley, 1990d). Th
some costs, however, as
requirements is required
battery of queries to b
positive side, this analys
useful for guiding desig

EXAM

To proffer an example
mation provided by S
a study that evaluate
dimensional display for
mation to fighter pilot
ducted in the same h
simulator as in the prio
ing the study, each of si
jects flew a fighter swee
of four digitally controll
subject completed five t
play conditions. The in
and order of presentati
tions were randomized.
collect SAGAT data at a
between 2.5 and 5 min
ately following the colle
Subjective Workload

reassuring that the finding of no intrusion on performance has been repeated in numerous other studies in which SAGAT was used to evaluate human-machine interface concepts (Bolstad and Endsley, 1990; Endsley, 1989b; Northrop, 1988), if the timing of the stop was unpredictable to the subjects (Endsley, 1988). This finding was also repeated in one study in which pilots flew in a simulation against computer-controlled enemy aircraft (Endsley, 1990d), helping to rule out the possibility that the finding of no effect on performance in the present study could have been the result of pilots on both sides of the simulation being equally influenced.

SAGAT has thus far proven to meet the stated criteria for measurement. In addition to the present studies establishing a level of empirical validity, SAGAT has been shown to have predictive validity (Endsley, 1990b) and content validity (Endsley, 1990d). The method is not without some costs, however, as a detailed analysis of SA requirements is required in order to develop the battery of queries to be administered. On the positive side, this analysis can also be extremely useful for guiding design efforts.

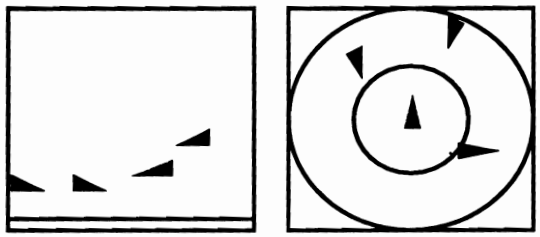
EXAMPLE

To proffer an example of the type of SA information provided by SAGAT, I will present a study that evaluates the use of a three-dimensional display for providing radar information to fighter pilots. The study was conducted in the same high-fidelity, real-time simulator as in the prior two experiments. During the study, each of six experienced pilot subjects flew a fighter sweep mission against a field of four digitally controlled enemy aircraft. Each subject completed five trials in each of five display conditions. The initial starting conditions and order of presentation of the display conditions were randomized. Each trial was halted to collect SAGAT data at a randomly selected time, between 2.5 and 5 min into each trial. Immediately following the collection of SAGAT data, the Subjective Workload Assessment Technique (SWAT) was also administered. In this study, the trial was not resumed after the collection of SAGAT and SWAT data.

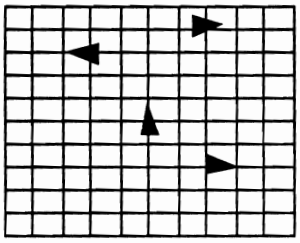
The five display conditions were constructed to evaluate the effect of a three-dimensional (3D) display concept. The 3D display provided a grid representing the ground, above which aircraft targets detected by the radar were presented by a small triangle, pointing in the direction of the aircraft's heading and connected to the grid by a line corresponding to the aircraft's altitude. The plane the subject was flying was presented in the center of this display so that each target's location and altitude were spatially represented relative to the subject's ownship. Four rotations of the grid (0, 30, 45, and 60 deg) were investigated (each providing a different perspective on the scene), as shown in Figure 9. At 0-deg rotation, the grid portrayed a God's-eye view of the scene, thus depicting a traditional two-dimensional display. In addition, these four views were compared with a two-display configuration in which a God's-eye view display was supplemented with a second profile display that pictorially presented altitude. In all conditions, altitude of the aircraft was also presented numerically next to the aircraft symbols.

Following the trials, the collected SAGAT and SWAT data were analyzed. Although the final SAGAT battery for fighter aircraft has 40 queries (Endsley, 1990c), only 23 were administered at the time of the test (as the battery was still in development). The results of 4 pertinent queries are presented here. The subject's reported knowledge of the range, azimuth, altitude, and heading of the enemy targets was compared with the actual values of these elements at the time of the freeze in the simulation. The subject's perceptions were evaluated as correct or incorrect (using tolerance values that had been determined to be operationally relevant).

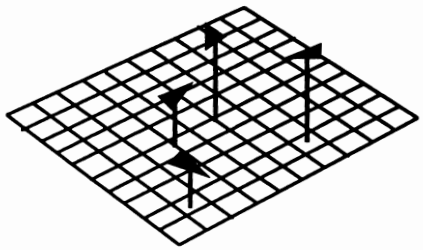
The percentage correct across subjects for each condition on each of these four queries is shown in Figure 10. As the scored data are binomial, an arcsine transformation was applied and an ANOVA performed. The results showed that the display type had a significant effect on pilot



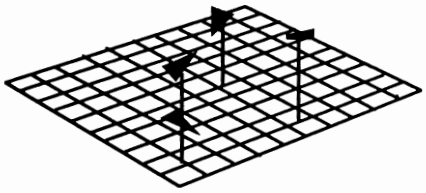
(a) 2D Display



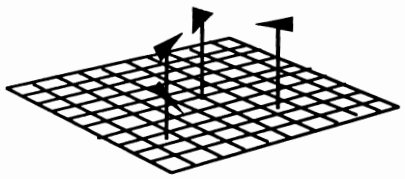
(b)
3D Display
0° Rotation



(c)
3D Display
30° Rotation



(d)
3D Display
45° Rotation

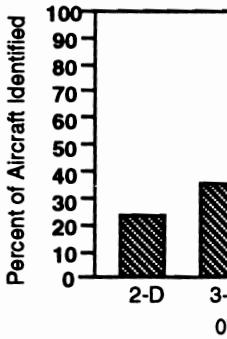


(e)
3D Display
60° Rotation

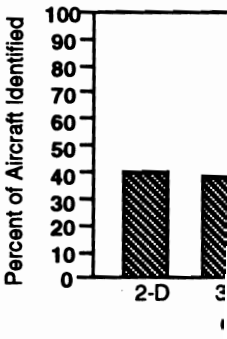
Figure 9. *Two-dimensional display and three-dimensional display at four levels of rotation.*

SA of the aircraft's range, $F(4,288) = 2.575$, $p < 0.05$; azimuth, $F(4,278) = 4.125$, $p < 0.05$; and heading, $F(4,139) = 3.040$, $p < 0.05$; but not altitude, $F(4,94) = 1.83$, $p > 0.05$. The 3D display, when rotated at 45 or 60 deg, provided significantly lower SA on target range than when at 0 or 30 deg, and significantly lower SA of target heading than of any of the other three conditions. A rotation of 60 deg also provided significantly lower SA of target azimuth than did the other four conditions. Although the ANOVA for target altitude was not significant, there was an opposite trend for the highly rotated 3D display to provide more SA of target altitude than did rotations of 0 and 30 deg. The SWAT data supported these findings, with significantly higher ratings for workload in the 45- and 60-deg rotations than in the 0- or 30-deg rotations or the two-dimensional display.

A major point to be gleaned from this example is the importance of examining the impact of a given design across an operator's SA requirements. The proposed 3D display was designed to provide pilots with better information about air-



(a) Knowled



(c) Knowledg

Figure 10. *Situatic*

craft altitude, and it
the trend was not
came at the expense
and heading. If a stu
merely examined su
altitude, this impor
been lost. The SAGA
tant diagnostic inf
understanding the
cept had on pilot
iterations.

IMPLEMENTATI

Several recomm
istration can be m
ence in using the p

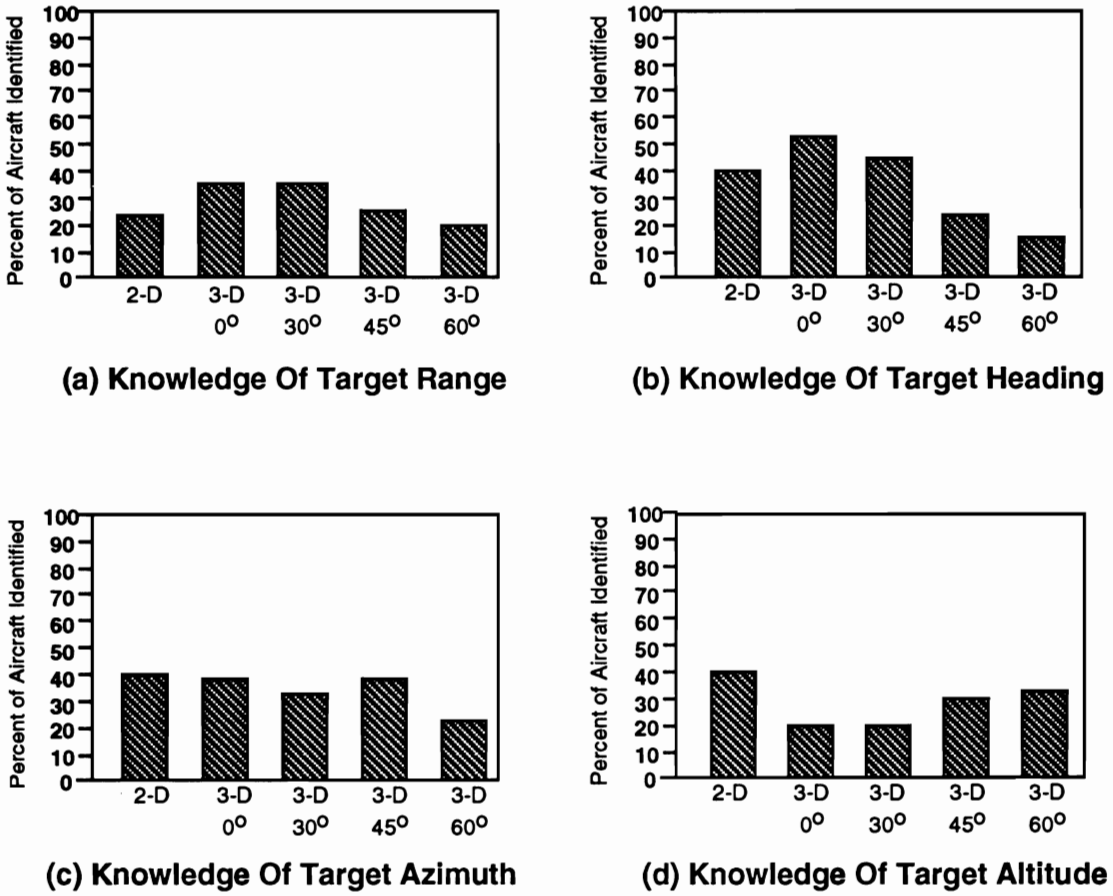


Figure 10. *Situation awareness of target range, heading, azimuth, and altitude in each display condition.*

craft altitude, and it may have done so, although the trend was not significant. However, this came at the expense of SA on location, azimuth, and heading. If a study had been performed that merely examined subjects' knowledge of aircraft altitude, this important information would have been lost. The SAGAT data also provided important diagnostic information that was useful in understanding the exact effect the design concept had on pilot SA in order to make future iterations.

IMPLEMENTATION RECOMMENDATIONS

Several recommendations for SAGAT administration can be made based on previous experience in using the procedure.

Training

An explanation of the SAGAT procedures and detailed instructions for answering each query should be provided to subjects before testing. Several training trials should be conducted in which the simulator is halted frequently to allow subjects ample opportunity to practice responding to the SAGAT queries. Usually three to five samplings are adequate for a subject to become comfortable with the procedure and to clear up any uncertainties in how to answer the queries.

Test Design

SAGAT requires no special test considerations. The same principles of experimental design and administration apply to SAGAT as to

any other dependent measure. Measures of subject performance and workload may be collected concurrently with SAGAT, as no ill effect from the insertion of breaks has been shown. To be cautious, however, if performance measures are to be collected simultaneously with SAGAT data, half of the trials should be conducted without any breaks for SAGAT in order to provide a check for this contingency.

Procedures

Subjects should be instructed to attend to their tasks as they normally would, with the SAGAT queries considered as secondary. No displays or other visual aids should be visible while subjects are answering the queries. If subjects do not know or are uncertain about the answer to a given query, they should be encouraged to make their best guess. There is no penalty for guessing, allowing for consideration of the default values, or other wisdom gained from experience that subjects normally use in decision making. If subjects do not feel comfortable enough to make a guess, they may go on to the next question. Talking or sharing of information among subjects should not be permitted. If multiple subjects are involved in the same simulation, all subjects should be queried simultaneously and the simulation resumed for all subjects at the same time.

Which Queries to Use

Random selection. As it may be impossible to query subjects about all of their SA requirements in a given stop because of time constraints, a portion of the SA queries may be randomly selected and asked each time. A random sampling provides consistency and statistical validity, thus allowing SA scores to be easily compared across trials, subjects, systems, and scenarios.

Because of attentional narrowing or lack of information, certain questions may seem unimportant to a subject at the time of a given stop. It is important to stress that subjects should attempt to answer all queries anyway. This is because (a) even though they think it unimportant, the information may have at least secondary importance; (b) they may not be aware of information that makes a question very important (e.g., the location of a pop-up aircraft); and (c) if only questions of the highest priority were asked, subjects might be inadvertently provided with artificial cues about the situation that will direct their attention when the simulation is resumed. Therefore, a random selection from a constant set of queries is recommended at each stop.

Experimenter controlled. In certain tests it may be desirable to have some queries omitted because of limitations of the simulation or characteristics of the scenarios. For instance, if the simulation does not incorporate aircraft malfunctions, the query related to this issue may be omitted. In addition, with particular test designs it may be desirable to ensure that certain queries are presented every time. When this occurs, subjects should also be queried on a random sampling from all SA requirements and not only on those related to a specific area of interest to the evaluation being conducted. This is because of the ability of subjects to shift attention to the information on which they know they will be tested. What may appear to be an improvement in SA in one area may be merely a shift of attention from another area. When the SAGAT queries cover all of the SA requirements, no such artificial cuing can occur.

When to Collect SAGAT Data

It is recommended that the timing of each freeze for SAGAT administration be randomly determined and unpredictable so that subjects cannot prepare for queries in advance. If the freeze occurrence is covertly associated with the occurrence of specific events, or at specific times across trials, prior studies have shown that the subjects will be able to figure this out (Endsley, 1988), allowing them to prepare for queries or actually improve SA through the artificiality of the freeze cues. An informal rule has been to ensure that no freezes occur earlier than 3 min into a trial to allow subjects to build up a picture of

the situation and th
within 1 min of each

The result of this a
ties occurring at the
randomly selected. S
ing very important ac
to the experimenter,
tivities are occurring
pling of the subjects'
tions. During analys
want to stratify the da
into account.

How Much SAGAT Da

The number of trial
the variability prese
ables being collected
samples taken during
different subjects and
and 60 samplings per
option have previou
within-subjects test d

Multiple SAGAT st
each trial. There is n
ber of times the simu
a given trial. In Exper
found from as many
min trial. In general,
stop last until a cer
elapsed, and then the
less of how many ques
Stops as long as 2 m
with no undue difficu
performance. In Expe
min were shown to a
information without

Data Collection

The simulator co
grammed to collect
ing to the queries at t
sidering that some
higher-level SA requ
available in the com
of the correct answer
rienced observer who
reflecting the SA of a

the situation and that no two freezes occur within 1 min of each other.

The result of this approach is that the activities occurring at the time of the stops will be randomly selected. Some stops may occur during very important activities that are of interest to the experimenter, others when no critical activities are occurring. This gives a good sampling of the subjects' SA in a variety of situations. During analysis the experimenter may want to stratify the data to take these variations into account.

How Much SAGAT Data to Collect

The number of trials necessary will depend on the variability present in the dependent variables being collected and the number of data samples taken during a trial. This will vary with different subjects and designs, but between 30 and 60 samplings per SA query with each design option have previously been adequate in a within-subjects test design.

Multiple SAGAT stops may be taken within each trial. There is no known limit to the number of times the simulator can be frozen during a given trial. In Experiment 2 no ill effects were found from as many as three stops during a 15-min trial. In general, it is recommended that a stop last until a certain amount of time has elapsed, and then the trial is resumed, regardless of how many questions have been answered. Stops as long as 2 min in duration were used with no undue difficulty or effect on subsequent performance. In Experiment 1, stops as long as 5 min were shown to allow subjects access to SA information without memory decay.

Data Collection

The simulator computer should be programmed to collect objective data corresponding to the queries at the time of each freeze. Considering that some queries will pertain to higher-level SA requirements that may be unavailable in the computer, an expert judgment of the correct answer may be made by an experienced observer who is privy to all information, reflecting the SA of a person with perfect knowledge. A comparison of subjects' perceptions of the situation (as input into SAGAT) with the actual status of each variable (as collected per the simulator computer and expert judgment) results in an objective measure of subject SA. Questions asked of the subject but not answered should be considered incorrect. No evaluation should be made of questions not asked during a given stop.

It is recommended that answers to each query be scored as correct or incorrect based on whether or not it falls into an acceptable tolerance band around the actual value. For example, it may be acceptable for a subject to be 10 miles per hour off of actual ground speed. This method of scoring poses less difficulty than dealing with absolute error (see Marshak et al., 1987). A tabulation of the frequency of correctness can then be made within each test condition for each SA element. Because data scored as correct or incorrect are binomial, the conditions for analysis of variance are violated. A correction factor, $Y' = \text{arcsine}(Y)$, can be applied to binomial data, which allows analysis of variance to be used. In addition, a chi-square, Cochran's Q, or binomial t test (depending on the test design) can be used to evaluate the statistical significance of differences in SA between test conditions.

Limitations and Applicability for Use

SAGAT has primarily been used within the confines of high-fidelity and medium-fidelity part-task simulations. This provides experimenter control over freezes and data collection without any danger to the subject or processes involved in the domain. It may be possible to use the technique during actual task performance if multiple operators are present to ensure safety. For example, it might be possible to verbally query one pilot in flight while another assumes flight control. Such an endeavor should be undertaken with extreme caution, however, and may not be appropriate for certain domains.

A recent effort (Sheehy, Davey, Fiegel, and Guo, 1993) employed an adaptation of this technique by making videotapes of an ongoing situation in a nuclear power plant control room.

These tapes were then replayed to naïve subjects with freezes for SAGAT queries employed. It is not known how different the SA of subjects passively viewing a situation may be from subjects actually engaged in task performance; however, this approach may yield some useful data.

Most known uses of SAGAT have involved fighter aircraft simulations. In general, it can be employed in any domain in which a reasonable simulation of task performance exists and an analysis of SA requirements has been made in order to develop the queries. SAGAT queries have been developed for advanced bomber aircraft (Endsley, 1990a) and recently for en route air traffic control (Endsley and Rodgers, 1994). Several researchers have begun to use the technique in evaluating operator SA in nuclear control room studies (Hogg, Torralba, and Volden, 1993; Sheehy et al., 1993). The technique has also been employed in a simulated control task to study adaptive automation (Carmody and Gluckman, 1993). Several new efforts are currently under way to employ the technique in medical decision making (E. Swirsky, personal communication, March 1994) and helicopters (C. Prince, personal communication, November 1993). Potentially it could also be used in studies involving automobile driving, supervisory control of manufacturing systems, teleoperations, and operation of other types of dynamic systems.

CONCLUSION

The use of SA metrics in evaluating the efficiency of design concepts, training programs, and other potential interventions is critical. Without careful evaluation, it will be difficult to make much progress in understanding the factors that influence SA, hindering human factors and engineering efforts. Several techniques have been reviewed here, each with several advantages and disadvantages. The use of one technique, SAGAT, was explored extensively, and two studies were presented that investigated the validity of the measure. It has to date proven a useful and viable technique for measuring SA in numerous contexts. Rigorous validity testing is needed for other techniques proposed for SA measurement. Unless the veracity of these techniques is established, interpreting their results will be hazardous.

In addition to establishing validity of measures of SA, more research is needed regarding the task of SA measurement itself. For instance, as pointed out by Pew (1991), no criteria exist at this time that establish the level of SA required for successful performance. Does an operator need to have SA that is 100% perfect (in both completeness and accuracy), or is some lesser amount sufficient for good performance? This is a complex issue. I would assert that SA can be seen as a factor that increases the probability of good performance but that it does not necessarily guarantee it, as other factors also come into effect (e.g., decision making, workload, performance execution, system capabilities, and SA of others in some cases). How much SA one needs, therefore, becomes a matter of how much probability of error one is willing to accept. Perhaps such criteria should more properly, and usefully, be established as the level of SA needed on each subcomponent at different periods in time. Some guidelines will eventually need to be specified if there is a desire to certify new system designs on the basis of SA.

Overall, the beginnings of the field of SA measurement have been established, allowing researchers in the area of SA to proceed from mainly speculation and anecdotal information to solid empirical findings. The tools for conducting further basic research on the SA construct and developing better system designs are available, allowing needed research on SA in a variety of arenas.

ACKNOWLEDGMENTS

This research was sponsored by the Northrop Corporation. I thank the many people at Northrop, both past and present, who contributed to this effort. In particular, I thank Vic Vizcarra for his assistance in conducting the studies, Paul Cayley and Gerry Armstrong for their programming efforts, and Mike Majors and Del Jacobs for their support in developing an SA research program. In addition, I am indebted to Mark Chignell, who provided helpful support throughout the development of the methodology. I also thank Mark Rodgers, Tom English, Dick Gilson, and several anonymous reviewers for their comments on an earlier version of this paper.

RE

Bolstad, C. A., and Ends
scale range display i
thorne, CA: Northrop

Carmody, M. A., and Glu
fects of automation
mance, workload a
Jensen and D. Neun
enth International Sy
167–171). Columbus
ment of Aviation.

Chase, W. G., and Simo
Cognitive Psychology

Cowan, N. (1988). Evol
selective attention,
the human informa
Bulletin, 104, 163–19

Endsley, M. R. (1987). S
ment of situation awa
CA: Northrop Corp.

Endsley, M. R. (1988). S
technique (SAGAT).
space and Electronic
New York: IEEE.

Endsley, M. R. (1989a).
advanced strategic m
CA: Northrop Corp.

Endsley, M. R. (1989b).
dendum 1 situation
Hawthorne, CA: No

Endsley, M. R. (1990a).
surement of situatio
in aerospace opera
Neuilly-Sur-Seine,
Aerospace Research

Endsley, M. R. (1990b).
sure of situation aw
Factors Society 34t
Monica, CA: Human

Endsley, M. R. (1990c).
technique (SAGAT):
(NOR DOC 89-58, r

Endsley, M. R. (1990d).
decision making: T
doctoral dissertatio
Los Angeles, CA.

Endsley, M. R., and Roc
global assessment te
control version user
University.

Fracker, M. L. (1990). A
ness. In *Situatione*
AGARD-CP-478; pp
NATO—Advisory G
velopment.

Herrmann, D. J. (1984)
Harris and P. E. M
and absent-mindedr

Hogg, D. N., Torralba,
awareness methodo
systems: Studies o
(1993-03-05). Store
Project.

REFERENCES

Bolstad, C. A., and Endsley, M. R. (1990). *Single versus dual scale range display investigation* (NOR DOC 90-90). Hawthorne, CA: Northrop Corporation.

Carmody, M. A., and Gluckman, J. P. (1993). Task-specific effects of automation and automation failure on performance, workload and situational awareness. In R. S. Jensen and D. Neumeister (Eds.), *Proceedings of the Seventh International Symposium on Aviation Psychology* (pp. 167–171). Columbus, OH: Ohio State University, Department of Aviation.

Chase, W. G., and Simon, H. A. (1973). Perceptions in chess. *Cognitive Psychology, 4,* 55–81.

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin, 104,* 163–191.

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

Endsley, M. R. (1988). Situation awareness global assessment technique (SAGAT). In *Proceedings of the National Aerospace and Electronics Conference (NAECON)* (pp. 789–795). New York: IEEE.

Endsley, M. R. (1989a). *Final report: Situation awareness in an advanced strategic mission* (NOR DOC 89-32). Hawthorne, CA: Northrop Corp.

Endsley, M. R. (1989b). *Tactical simulation 3 test report: Addendum 1 situation awareness evaluations* (81203033R). Hawthorne, CA: Northrop Corp.

Endsley, M. R. (1990a). A methodology for the objective measurement of situation awareness. In *Situational awareness in aerospace operations* (AGARD-CP-478; pp. 1/1–1/9). Neuilly-Sur-Seine, France: NATO—Advisory Group for Aerospace Research and Development.

Endsley, M. R. (1990b). Predictive utility of an objective measure of situation awareness. In *Proceedings of the Human Factors Society 34th Annual Meeting* (pp. 41–45). Santa Monica, CA: Human Factors and Ergonomics Society.

Endsley, M. R. (1990c). *Situation awareness global assessment technique (SAGAT): Air-to-air tactical version user guide* (NOR DOC 89-58, rev A). Hawthorne, CA: Northrop Corp.

Endsley, M. R. (1990d). *Situation awareness in dynamic human decision making: Theory and measurement.* Unpublished doctoral dissertation, University of Southern California, Los Angeles, CA.

Endsley, M. R., and Rodgers, M. D. (1994). *Situation awareness global assessment technique (SAGAT): En route air traffic control version user's guide* (Draft). Lubbock: Texas Tech University.

Fracker, M. L. (1990). Attention gradients in situation awareness. In *Situational awareness in aerospace operations* AGARD-CP-478; pp. 6/1–6/10). Neuilly-Sur-Seine, France: NATO—Advisory Group for Aerospace Research and Development.

Herrmann, D. J. (1984). Questionnaires about memory. In J. E. Harris and P. E. Morris (Eds.), *Everyday memory, action and absent-mindedness* (pp. 133–151). London: Academic.

Hogg, D. N., Torralba, B., and Volden, F. S. (1993). *A situation awareness methodology for the evaluation of process control systems: Studies of feasibility and the implication of use* (1993-03-05). Storefjell, Norway: OECD Halden Reactor Project.

Hughes, E. R., Hassoun, J. A., and Ward, F. (1990). *Advanced tactical fighter (ATF) simulation support: Final report* (CSEF-TR-ATF-90-91). Wright-Patterson Air Force Base, OH: Air Force Systems Command, Aeronautical Systems Division.

Jacoby, L. L., and Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General, 110,* 306–340.

Kellog, R. T. (1980). Is conscious attention necessary for long-term storage? *Journal of Experimental Psychology: Human Learning and Memory, 6,* 379–390.

Manktelow, K., and Jones, J. (1987). Principles from the psychology of thinking and mental models. In M. M. Gardiner and B. Christie (Eds.), *Applying cognitive psychology to user-interface design* (pp. 83–117). Chichester, England: Wiley.

Marshak, W. P., Kuperman, G., Ramsey, E. G., and Wilson, D. (1987). Situational awareness in map displays. In *Proceedings of the Human Factors Society 31st Annual Meeting* (pp. 533–535). Santa Monica, CA: Human Factors and Ergonomics Society.

McDonnell Douglas Aircraft Corporation. (1982). *Advanced medium-range air-to-air missile operational utility evaluation (AMRAAM OUE) test final report.* St. Louis, MO: Author.

Mogford, R. H., and Tansley, B. W. (1991). *The importance of the air traffic controller's mental model.* Presented at the Human Factors Society of Canada Annual Meeting, Canada.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review, 76,* 165–178.

Nisbett, R. E., and Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231–259.

Norman, D. A. (1968). Towards a theory of memory and attention. *Psychological Review, 75,* 522–536.

Northrop Corporation. (1988). *Tactical simulation 2 test report: Addendum 1 situation awareness test results.* Hawthorne, CA: Author.

Pew, R. W. (1991). Defining and measuring situation awareness in the commercial aircraft cockpit. In *Proceedings of the Conference on Challenges in Aviation Human Factors: The National Plan* (pp. 30–31). Washington, DC: American Institute of Aeronautics and Astronautics.

Sarter, N. B., and Woods, D. D. (1991). Situation awareness: A critical but ill-defined phenomenon. *International Journal of Aviation Psychology, 1,* 45–57.

Selcon, S. J., and Taylor, R. M. (1990). Evaluation of the situational awareness rating technique (SART) as a tool for aircrew systems design. In *Situational awareness in aerospace operations* (AGARD-CP-478; pp. 5/1–5/8). Neuilly-Sur-Seine, France: NATO—Advisory Group for Aerospace Research and Development.

Sheehy, E. J., Davey, E. C., Fiegel, T. T., and Guo, K. Q. (1993, April). *Usability benchmark for CANDU annunciation—lessons learned.* Presented at the ANS Topical Meeting on Nuclear Plant Instrumentation, Control and Man-Machine Interface Technology, Oak Ridge, TN.

Smolensky, M. W. (1993). Toward the physiological measurement of situation awareness: The case for eye movement measurements. In *Proceedings of the Human Factors and Ergonomics Society 37th Annual Meeting* (p. 41). Santa Monica, CA: Human Factors and Ergonomics Society.

Taylor, R. M. (1990). Situational awareness rating technique (SART): The development of a tool for aircrew systems design. In *Situational awareness in aerospace operations* (AGARD-CP-478; pp. 3/1–3/17). Neuilly-Sur-Seine, France:

NATO—Advisory Group for Aerospace Research and Development.
Tulving, E. (1985). How many memory systems are there? *American Psychologist, 40,* 385–398.
Venturino, M., Hamilton, W. L., and Dvorchak, S. R. (1990). Performance-based measures of merit for tactical situation awareness. In *Situation awareness in aerospace operations* (AGARD-CP-478; pp. 4/1–4/5). Neuilly-Sur-Seine, France: NATO—Advisory Group for Aerospace Research and Development.
Vidulich, M. A. (1989). The use of judgment matrices in subjective workload assessment: The subjective workload dominance (SWORD) technique. In *Proceedings of the Human Factors Society 33rd Annual Meeting* (pp. 1406–1410). Santa Monica, CA: Human Factors and Ergonomics Society.

Date received: January 26, 1993
Date accepted: October 4, 1994

Situation Managem

MARILYN JAGER
Newman Inc., Cam

The issue
systems o
automatic
that comp
dynamics
mercial a
definition
volved in
nitive the
a more a
strengths
toward th

IN

In everyday pa
fers to the up-to-th
to operate or ma
widely used in the
ation communitie
its way into the li
system design. Wi
ation safety, it is c
hold word. But w

We know that i
processes and sta
yet many speak
awareness as th
evident. The chal
the cognitive unc
has important pr

[1] Requests for repri
Bolt Beranek and New
MA 02138.