AL-TR-1991-0128

ARMSTRONG LABORATORY

MEASURES OF SITUATION AWARENESS: REVIEW AND FUTURE DIRECTIONS (U)

Martin L. Fracker, Major, USAF

CREW SYSTEMS DIRECTORATE
HUMAN ENGINEERING DIVISION

OCTOBER 1991

FINAL REPORT FOR PERIOD JANUARY 1990 TO JANUARY 1991

Approved for public release; distribution is unlimited.

AIR FORCE SYSTEMS COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OHIO 45433-6573

NOTICES

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

Please do not request copies of this report from the Armstrong Aerospace Medical Research Laboratory. Additional copies may be purchased from:

National Technical Information Service
5285 Royal Road
Springfield, Virginia 22161

Federal Government agencies and their contractors registered with the Defense Technical Information Center should direct requests for copies of this report to:

Defense Technical Information Center
Cameron Station
Alexandria, Virginia 22314

TECHNICAL REVIEW AND APPROVAL

AL-TR-1991-0127

This report has been reviewed by the Office of Public Affairs (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nations.

The voluntary informed consent of the subjects used in this research was obtained as required by Air Force Regulation 169-3.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

Kenneth R Boff

KENNETH R. BOFF, Chief
Human Engineering Division
Armstrong Laboratory

REPORT DOCUMENTATION PAGE

Form Approved
OMB No 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
October 1991

3. REPORT TYPE AND DATES COVERED
Final Report January 1990-January 1991

4. TITLE AND SUBTITLE
Measures of Situation Awareness: Review and
Future Directions

5. FUNDING NUMBERS
~~C - F33615-89-C-0532~~
PE- 62202F
PR- 7184
TA- 14
WU- 25

6. AUTHOR(S)
Martin L. Fracker, Major, USAF

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Human Engineering Division
Armstrong Laboratory
AL/CFHW
Wright-Patterson AFB OH 45433-6573

8. PERFORMING ORGANIZATION REPORT NUMBER
AL-TR-1991-0128

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Human Engineering Division
AL/CFHW
Wright-Patterson AFB OH 45433-6573

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE
A

13. ABSTRACT *(Maximum 200 words)*

Measures of situation awareness (SA), or what operators know about their immediate situation, are reviewed. Three major approaches to SA assessment are considered: explicit, implicit, and subjective rating. Explicit measures require operators to self-report material in conscious memory. Implicit measures assess the influence of relevant events on subsequent task performance. Subjective ratings require operators to assign numerical values to the self-assessed quality of their SA. These three measurement approaches are evaluated in terms of their reliability and three kinds of validity: construct, content, and criterion. Several problems requiring further research are identified and discussed. In particular, reliability and content validity continue to present serious difficulties, some of which suggest that new approaches to SA measurement may still be needed.

14. SUBJECT TERMS

Attention	reliability	situation awareness
memory probe	signal detection theory	subjective measures
mental workload		validity

15. NUMBER OF PAGES
34

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. Instructions for filling in each block of the form follow It is important to ***stay within the lines*** to meet ***optical scanning requirements***.

Block 1. Agency Use Only *(Leave blank)*.

Block 2. Report Date. Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least the year.

Block 3. Type of Report and Dates Covered State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle A title is taken from the part of the report that provides the most meaningful and complete information When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C	-	Contract	PR	-	Project
G	-	Grant	TA	-	Task
PE	-	Program Element	WU	-	Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es) Self-explanatory.

Block 10. Sponsoring/Monitoring Agency Report Number. *(If known)*

Block 11. Supplementary Notes Enter information not included elsewhere such as: Prepared in cooperation with...; Trans. of...; To be published in... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NOFORN, REL, ITAR).

DOD - See DoDD 5230.24, "Distribution Statements on Technical Documents."
DOE - See authorities.
NASA - See Handbook NHB 2200.2.
NTIS - Leave blank.

Block 12b. Distribution Code.

DOD - Leave blank.
DOE - Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports.
NASA - Leave blank.
NTIS - Leave blank.

Block 13. Abstract. Include a brief *(Maximum 200 words)* factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code *(NTIS only)*

Blocks 17. - 19. Security Classifications. Self-explanatory Enter U.S. Security Classification in accordance with U.S Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page

Block 20. Limitation of Abstract This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.

SUMMARY

Measures of situation awareness (SA), or what operators know about their immediate situation, are reviewed. Three major approaches to SA assessment are considered: explicit, implicit, and subjective rating. Explicit measures require operators to self-report material in conscious memory. Implicit measures assess the influence of relevant events on subsequent task performance. Subjective ratings require operators to assign numerical values to the self-assessed quality of their SA. These three measurement approaches are evaluated in terms of their reliability and three kinds of validity: construct, content, and criterion. Several problems requiring further research are identified and discussed. In particular, reliability and content validity continue to present serious difficulties, some of which suggest that new approaches to SA measurement may still be needed.

PREFACE

The author thanks Michael Vidulich, Gary Reid, Maris Vikmanis, and Mica Endsley for the many helpful discussions which aided in the development of the ideas contained in this technical report. Mark Crabtree's proofreading assistance is gratefully acknowledged. Any errors, whether substantive or technical, are attributable solely to the author.

TABLE OF CONTENTS

Page

INTRODUCTION

Situation awareness (SA) refers to military operators' knowledge of the immediate tactical situation (cf., Sarter and Woods, 1991). SA may be among the most important subjects to be addressed by military psychologists in recent years. Clausewitz (1832/1984) seems to have been referring to SA--what others have called "the fog of war"--when he wrote that the "difficulty of *accurate recognition* constitutes one of the most serious sources of friction in war, by making things appear entirely different from what one had expected" (p. 117). Not knowing the true tactical situation, according to Clausewitz, is one of the principal reasons why even the simplest thing in war is difficult, even though everything in war is very simple (p. 119). The centrality of SA in warfighting is further evident in the importance of surprise in war. Surprise is possible only when the enemy's SA is poor, that is, when the enemy is unaware of the true tactical situation. As Clausewitz observed, preventing the enemy from achieving accurate SA is the means by which one side or the other gains superiority and is so able to prevail (p. 198). The logical corollary is that maintaining good SA is a necessary condition for victory in war.

Given the importance of SA, it is hardly surprising that the Air Force has invested considerable effort in trying to enhance combat pilot's SA, either through pilot training (Eubanks and Killeen, 1983; Thomas, Houcke, and Bell, 1990) or through improved hardware systems (e.g., Arbak, Schwartz, and Kuperman, 1987; Hughes, Hassoun, Ward, and Rueb, 1990; Venturino and Kunze, 1989; Wells, Venturino, and Osgood, 1988). Evaluating the success of attempts to improve SA has been a crucial but difficult step. Assessing the quality of pilots' SA has turned out to be a much larger measurement problem than it first appeared. This article first establishes criteria against which SA metrics may be evaluated and then critically reviews the major approaches to SA measurement that have been developed. Following this review, directions for future research are discussed.

SA MEASUREMENT CRITERIA

The two principle criteria by which SA metrics should be evaluated are their reliability and validity. Additional criteria such as ease of use and operator acceptance should be considered only when choosing between two or more metrics that are approximately equally reliable and valid. *Reliability* concerns whether a metric will remain consistent if the same quantity is measured at different times under the same conditions. *Validity* mainly concerns whether the metric actually measures what it is supposed to measure. Both are important. On one hand, the validity of a measure cannot exceed its reliability. On the other, there is nothing to prevent a highly reliable metric from being invalid. For example, measuring the length of pilots' noses is likely to provide highly reliable but completely invalid assessments of their skill in combat.

Reliability

Reliability theory revolves around the concept of a *true score*, defined as the outcome of all the factors that influence the attribute being measured. Concerning SA, these factors might include characteristics of human operators such as their natural intelligence, training, and experience, as well as characteristics of the environment such as the availability and formatting of relevant information. Any given measure, X, of the attribute is then said to be the sum of the true score, T, and some random error in the measurement, e. Thus,

$$X = T + e.$$

The variability of X, then, is the variability of the sum ($T + e$). Assuming that T and e are uncorrelated, this variability can be re-expressed as the sum of Var(T) and Var(e), denoting the variabilities of T and e, respectively. The reliability, or consistency, of a measure may be defined as the following ratio:

$$\text{Reliability} = \text{Var}(T) / [\text{Var}(T) + \text{Var}(e)].$$

Reliability improves as variability due to measurement error declines. Conversely, any factor that increases measurement error reduces reliability (for extended discussions, see Allen and Yen, 1982; Gulliksen, 1950; Lord and Novick, 1968; Murphy and Davidshofer, 1991).

Methods for Evaluating Reliability

Reliability can be estimated using test-retest, alternate forms, split-half, and internal consistency methods. Test-retest methods require collecting the same measure from the same people under the same conditions at different times. Assuming that the measured attribute does not change over time and that the first measurement does not influence the second, the correlation between the two measurements is a direct estimate of the measure's reliability. In alternate forms methods, two alternate versions of the same measurement technique are used on the same people and compared. Reliability is then estimated by the correlation between the two versions. Split-half methods are appropriate when a measure is aggregated from several response samples, referred to as items. Essentially, the set of items are divided in half and the correlation between the two halves is determined. Internal consistency methods estimate reliability from the intercorrelations among all of the items contributing to a measure.

Improving Reliability

The easiest way to improve the reliability of a measure is to increase the number of observations that contribute to the measure. If the observations are added together to form a composite score, then the sum will be at least as reliable as the least reliable observation. Further, if the observations are correlated, then the reliability of their sum will increase (1) as the number of observations increases and (2) as the correlations among observations are strengthened. Thus, a good way to improve the reliability of a measure is to obtain a larger number of correlated observations and use their sum (or average) as the measure.

In contrast to composite scores (sums or averages), profile scores decrease in reliability as the correlations among observations increase. Profile scores are measures of how one variable differs from another. For example, one might measure pilots' awareness of the locations of enemy aircraft, enemy surface-to-air missiles, and enemy tanks. Some pilots might have good awareness for aircraft locations but poor awareness for missiles and tanks. Other pilots might have poor awareness for aircraft but good awareness for missiles and tanks. Thus, looking at SA profiles might reveal specific weaknesses in SA for specific pilots. Comparing profiles is essentially equivalent to comparing differences between variables (e.g., SA for aircraft versus SA for tanks). If two variables are correlated, then they tend to reflect the same true score. Thus, subtracting one from the other will tend to leave only the random error. As a result, differences between correlated variables will tend to be highly unreliable. In general, then, profile scores should be avoided. When possible, composite scores should be used instead.

Validity

Validity is not a simple concept. At least three types of validity may be identified: construct, content, and criterion.

Construct Validity

A construct is some unobservable psychological attribute such as situation awareness that is hypothesized to account for some aspect of human behavior. Construct validity refers to the degree that a measure can quantify this unobservable psychological attribute. Assessing construct validity involves identifying (1) human behaviors that are logically related to the construct in question, (2) other constructs that are either related or unrelated to the target construct, and (3) behaviors that are logically related to these new constructs (Murphy & Davidshofer, 1991). One then demonstrates that behaviors related to the construct (a) behave as they are supposed to, (b) associate with other related behaviors, and (c) dissociate from behaviors unrelated to the construct. Because statements of the relationship between specific behaviors and a given construct are theoretical in nature, tests of construct validity may also be viewed as tests of the underlying theory. Consequently, failures to establish construct validity are invariably ambiguous. Such failures may mean that the measure is invalid, or that the underlying theory is incorrect. If tests of several alternative measures within the same theoretical framework all fail to establish their construct validity, then one may conclude that the underlying theory is at least not very useful.

Three criteria are proposed in order to establish the construct validity of an SA measure. First, the measure should avoid confusing momentary with reflective knowledge of the situation. Second, the measure should show that SA declines when attention is spread across a larger or more complex situation. Third, the measure should be related to measures of mental effort such that, if situation assessment becomes more difficult, then SA declines, mental effort increases, or both. Each criterion is discussed in turn.

Momentary versus reflective SA. The distinction between momentary and reflective SA is similar to the distinction between battlefield and armchair generals. Battlefield generals must assess what is actually happening whereas armchair generals need only assess what is likely to happen. Of course, assessments made in the comfort of an office or living room with plenty of time to reflect upon them may be accurate and insightful, but they may also be quite different from the assessments which the same individual might make under the pressure of the battlefield. As will be seen, some methods for measuring SA may not distinguish well between these two kinds of assessments. Yet making the distinction is important. Individuals who can develop accurate reflective SA but not good momentary SA will make poor battlefield commanders. In similar fashion, military information systems that poorly support momentary SA may appear better than they are if metrics used to evaluate them actually measure reflective SA.

Attention and SA. Logically, operators cannot know the state of a situational variable until they have attended to it. For example, pilots cannot know whether there is an enemy aircraft at a certain location unless they aim their radar at that location or attend to some other source of information such as provided by a ground control intercept officer. A useful metaphor for attention is that of a spotlight: attention can be spread over a larger or smaller area, but increasing the area lowers the quality of processed information (Eriksen and Yeh, 1985). Further, Downing's (1988) experiments implied that increasing the number of objects within the same-size attentional beam also reduces processing quality. Thus, when the area to be attended grows larger, or when the number of variables to be attended increases, operators' SA should decline.

Effort and SA. When a task becomes more difficult, whether because the load on attention has increased or for some other reason, performance quality may not decline if sufficient additional effort is put forth. Thus, if maintaining SA becomes more difficult, SA may or may not decline depending upon whether and how much effort is increased. Thus, evaluations of the construct validity of SA metrics should include assessments of effort. When task difficulty increases but effort does not, then a valid measure of SA will decline; on the other hand, if effort does increase, then SA may decline little if at all. Measurement of effort--more commonly referred to as "mental workload"--is a fairly recent and controversial development in psychology (Gopher and Donchin, 1986; Moray, 1979; Ogden, Levine, and Eisner, 1979; Wickens, 1984; Wierwille, 1979; Williges and Wierwille, 1979). Further, the theoretical assumptions underlying much workload measurement research have recently come under attack (Fracker and Wickens, 1989; Hirst and Kalmar, 1987; Navon, 1984; Navon and Miller, 1987). Nevertheless, several practical measures of mental workload have become available (Moray, 1988; O'Donnell and Eggemeier, 1986) and are in wide use. As a result, theoretical controversies notwithstanding, it appears possible to evaluate whether an SA metric responds appropriately to increasing task difficulty and changes in assessed effort.

Content Validity

Content validity refers to the degree that the knowledge or behaviors assessed by a metric represent the knowledge or task domain being measured. Assessing content validity usually involves analyzing the specific knowledge or behaviors relevant to the domain and rendering a judgment as to whether the sampled knowledge or behaviors are in fact representative. In SA measurement, establishing content validity first requires analyzing a given military task in order to determine what kinds of information the operator needs to know. This information, once determined, can then be compared to the information sampled by the SA metric. Content validity would be considered high if all important kinds of information in the domain--and no irrelevant domains of information--are sampled by the metric.

Content validity is specific to different mission domains. For example, an SA metric having high content validity for a tactical air defense mission will likely have low content validity for a strategic bombing mission. Nevertheless, there may be a *situational structure* common to most missions. Fracker (1988) proposed that such a structure might have five levels: goals, organizations, functions, processes, and states. In this representation, situations are viewed as sets of variables whose states can change over time. These dynamic variable states are said to result from the interaction of opposing forces each directing their operations toward specific goals. In order to achieve these goals, each force has organized itself into particular units and assigned to each unit specific functions. The interactions among unit functions, referred to as processes, lead directly to the momentary changes in situation variable states.

Fracker's (1988) five-level situational structure implies that operator SA might differ across levels. For example, operators might be aware of enemy objectives (high-level SA) but uncertain as to what specific actions the enemy has undertaken in order to achieve those objectives (low-level SA). Conversely, operators might know what actions the enemy has undertaken but not know what objective those actions served. A content valid measure of SA, therefore, should not only sample the variables which comprise the situation, but should also sample all five levels of the situational structure.

Criterion Validity

Criterion validity refers to the degree of correlation between the metric and some objective measure that could be used to evaluate the accuracy of a decision based upon the metric. For example, if the SA metric is to be used to select one of several competing cockpit designs for a new fighter aircraft, the criterion might be success in combat.

Establishing criterion validity is usually complicated by the fact that many factors may contribute to the criterion measure. Combat success, for instance, depends not only upon accurate SA but also upon wise decision making and effective response execution. While wise decisions and effective responses are dependent upon accurate SA, possessing the latter is no guarantee that the others will follow. Thus, an otherwise valid measure of SA might appear poor if it is tested on operators who make poor decisions or unskilled responses. This observation implies a dilemma in establishing the criterion validity of SA metrics. If inexperienced or only partially trained operators are included in the study, the correlation between measured SA and the criterion may appear low for reasons that have nothing to do with the SA metric itself. On the other hand, if only experienced and highly trained operators are included, a high correlation may be precluded for purely statistical reasons (restriction of range). Paradoxically, then, criterion validity--which is often the most important form of validity to the user--may be the most difficult to establish and hence the least likely to be assessed.

REVIEW OF SA MEASUREMENT METHODS

Three major approaches to assessing situation awareness are reviewed: explicit, implicit, and subjective rating. The distinction between explicit and implicit measures comes from a distinction made by some psychologists between explicit and implicit forms of memory (see Roediger, 1990, for a readable discussion). Explicit measures require people to self-report material in memory of which they are consciously aware. As a result, such measures are considered subjective in nature--but are distinguished from subjective rating measures, which involve assignment of numerical values to the *quality* of the content of awareness. Unlike explicit measures, implicit measures do nc rely on self-reports of awareness; rather, such measures are derived from task performance. Specifically, SA is inferred from the influence of prior events on task performance (e.g., evading an attacking aircraft, locking on to an enemy target). Thus, implicit measures may be considered objective rather than subjective in nature.

In reviewing each type of metric, the measurement methodology is first described. Then any evidence pertaining to the reliability and validity of the resulting measures is reviewed.

Explicit Measures

If SA is regarded as the information immediately available in conscious awareness, then explicit measures are the most direct way of assessing SA. Two types of explicit measurement methods can be identified: retrospective event recall and concurrent memory probes.

Retrospective Measures

Kibby (1988) and Whitaker and Klein (1988) both used retrospective event recall to assess SA. Kibbe had laboratory subjects perform a radar warning receiver (RWR) monitoring task alone or with a concurrent pursuit tracking task. During the task, five different types of threats appeared on the RWR several times. Following the task, subjects were asked to recall and position threat events along a timeline representing their flight path. In addition, subjects were asked to estimate the number of times each type of threat had occurred. Kibbe found that timeline recall and placement accuracy depended on the type of threat: the more severe the threat, the more accurate its recall. However, accuracy was not affected by whether the concurrent tracking task was performed. Presence or absence of the tracking task did affect the estimate of threat type frequency, however: in the dual-task condition, subjects underestimated the number of threats; in the single-task condition, subjects overestimated threat frequency.

Whitaker and Klein (1988; see also Klein, Calderwood, and Clinton-Cirocco, 1985) took a quite different approach to retrospective SA measurement, using what they called the "Critical Decision Method." Based on Flanagan's (1954) critical incident technique, subjects were asked to recall their step-by-step decisions during a complex real-world task such as planning a military operation. Applying protocol analysis techniques, Whitaker and Klein made a significant observation: subjects seemed to use only immediately available information. Regardless of its importance to task success, information that required more than a cursory search was not sought.

Reliability. No reliability studies of retrospective event recall are known to have been conducted. Kibbe's (1988) timeline recall method could be reliable to the extent that errors in recall are averaged over time and events, however. Regarding Whitaker and Klein's (1988) Critical Decision Method, proprietary scoring and analysis procedures prohibit an assessment of the likelihood that the method could be reliable.

Construct validity. The most serious challenge to the construct validity of retrospective SA measurement is its inability to distinguish between momentary and reflective SA. A growing body of research shows that as the time between an event and its recall increases, people become more likely to recall "facts" about the event that in fact are not true (Loftus, 1979; Loftus and Loftus, 1980). These false recollections appear to be otherwise reasonable inferences drawn from information that people are still able to remember (Carr, 1986). Because progressively more information is forgotten as time goes on, such false inferences increase in frequency as the event becomes more distant. Thus, retrospective recall seems as likely to measure what operators can infer happened (reflective SA) as what they can actually remember having happened (momentary SA).

Besides confounding momentary with reflective SA, retrospective recall may also fail to decline as the load on attention increases. In Kibbe's (1988) experiment, adding tracking to the RWR monitoring task should have diverted attention away from the monitoring task thereby degrading the quality of SA, but adding the tracking task had no effect on timeline placement accuracy. At least three explanations for this failure are possible. First, the failure could have resulted from forgetting: single-task SA may in fact have been more accurate while the monitoring task was performed, but the more accurate information may have been forgotten by the time of recall. Second, Kibbe's subjects may have allocated only residual attention to the tracking task thereby producing no change in the amount of attention allocated to the monitoring task. Unfortunately, this possibility cannot be evaluated because Kibbe did not obtain a single task baseline for the tracking task. Third, subjects may have compensated for the increased difficulty of the task by exerting more effort. Kibbe did not measure mental workload, however, so this possibility cannot be evaluated either.

In spite of the foregoing ambiguity, it is still possible that attention may play a role in retrospective recall because information that attracts more attention is more likely to be recalled later (Logan, 1988; Wyer and Srull, 1986). Thus, the fact that Kibbe's (1988) subjects remembered high threat but not low threat events may indicate that the former received more attention than the latter.

Content validity. Retrospective techniques can achieve a degree of content validity depending upon how well they are structured. Kibbe's (1988) time-line placement technique seems able to measure operator's recall of how variable states changed over time and so may sample both state and process awareness. Whitaker and Klein's (1988) Critical Decision Method may also sample higher levels of SA if operator's give their rationale for doing what they did. Nevertheless, both techniques seem to rely on operators' spontaneous recall of information in order to sample the relevant information domain; in a sense, then, these techniques leave content validity up to the operator.

Criterion validity. In Kibbe's (1988) experiment, a meaningful criterion was subjects' speed and accuracy in detecting and identifying threats as they appeared on the RWR. Unfortunately, she did not report correlations between time-line placement accuracy and the criterion. Nevertheless, a poor correlation may be likely because speed and accuracy on the detection-identification task were affected by threat type whereas placement accuracy was not. Whitaker and Klein (1988) did not report any criterion measures.

Concurrent Measures

The most significant objection to retrospective measures is the confounding of momentary and reflective SA. As discussed, one reason for this confound is the temporal delay between events and their recall. One solution to this problem is to probe memory closer to the time specific events actually occur--during the mission rather than afterwards. Several implementations of such concurrent memory probes have appeared in the recent literature (Endsley, 1989; Fracker, 1991; Fracker and Davis, 1990; Marshak, Kuperman, Ramsey, and Wilson, 1987; Venturino and Kunze, 1989; Wells, Venturino, and Osgood, 1988). The basic idea in most of these implementations is to freeze a simulated mission after some random interval of time, blank the pilot's displays, and ask the pilots to recall certain items of information, such as the locations of enemy aircraft. SA is then quantified as the pilot's error in responding to these queries.

Reliability. Formal studies of memory probe reliability have not been encouraging. Fracker (1991) evaluated the test-retest reliability of memory probes administered on consecutive days to the same subjects under identical experimental conditions. In the experiment, non-pilot subjects performed a simulated combat-like task in which they monitored the positions of friendly, enemy, and neutral aircraft displayed on a computer screen. Periodically, the simulation was frozen and one of the aircraft disappeared. Subjects were either to show where the aircraft had been located (location probe) or to indicate its identity as friend, foe, or neutral (FFN probe). Table 1 shows the reliability (and validity) coefficients averaged across experimental conditions; tests of statistical significance followed Dunlap, Silver, and Bittner's (1986) recommendations. Location probes appeared highly unreliable while FFN probes fared somewhat better, although their reliability was still not impressive. Fracker attributed the generally poor reliability coefficients to idiosyncratic practice effects between sessions. Regarding location probes, Fracker suggested that location error might have been measured with more precision than was psychologically meaningful. Perhaps a more appropriate level of precision would have produced better reliability.

Table 1. *Reliability and validity coefficients from Fracker (1991). Probability of Fisher's z (N = 32) in parentheses.*

	Location Probe	FFN Probe		Envelope
	Error	Accuracy	Latency	Sensitivity
Reliability	.13 (ns)	.49 (.01)	.54 (.005)	.42 (.025)
Correlation w/				
Avoidance Failures	.10 (ns)	-.11 (ns)	.20 (.10)	-.39 (.025)
Kill Probability	.02 (ns)	.10 (ns)	-.29 (.05)	-

In spite of the poor test-retest correlations, other evidence implied that reliability might be better than indicated. Fracker's (1991) two experiments manipulated some of the same factors and observed a high degree of consistency in the memory probe data for each experimental condition across the two experiments. While this consistency across experiments does not formally demonstrate reliability, it does suggest that further research to determine the reliability of memory probes may be justified.

Construct validity. Although memory probes may be less likely than retrospective recall to confound momentary with reflective SA, there are suggestions that the two may still be confounded to some degree. Basic laboratory research has shown that information stored in working memory decays in only a matter of seconds without active rehearsal (Peterson and Peterson, 1959). Thus, there has been concern that pilot SA might decay during the memory probe freezes, particularly SA for information that is probed later rather than earlier during the freeze. In order to respond to probes later in the freeze, operators might then have to rely on reflective SA. To assess the degree of working memory decay, Endsley (1989) had experienced fighter pilots fly simulated combat missions in two experiments. In the first, she manipulated freeze duration and found virtually no increase in error even after delays of six minutes. In the second experiment, she reasoned that memory decay during the freeze would interfere with pilots being able to resume the mission following the freeze. Thus, she varied the number of freezes from 0 to 3 and studied the effect on mission performance (kills and losses). She reported that the number of freezes had no effect on the performance measures.

The divergence of Endsley's (1989) data from well-established laboratory findings demands explanation. Endsley acknowledged that her measures may not have been sufficiently sensitive to detect decay effects, but she felt the most likely explanation lay in differences between her experiments and traditional laboratory tasks. Whereas basic laboratory experiments have typically studied retention of disorganized stimuli such as random sequences of letters or digits, Endsley's experiments involved retention of inherently meaningful tactical information by expert combat pilots. She hypothesized that her pilots relied on information stored in long-term memory in order to respond to the probes and then to resume the mission. While this hypothesis is consistent with most cognitive models of how pilots develop and maintain their SA (Endsley, 1988; Fracker, 1988; see Ericsson and Staszewski, 1989, for a different cognitive approach), it also may render memory probe data ambiguous with regard to whether pilots were actually aware of the probed information prior to the probe (momentary SA). Conceivably, pilots may *not* have been aware of the information prior to the probe; rather, the probe may have served as the stimulus for an inference from knowledge gained through previous experience. In short, the probe may have measured the quality of pilots' reflective rather than momentary SA.

A related difficulty is that the probe procedure may alter pilots' SA. In effect, the probe conveys a message to attend to a specific item of information in the future--and implies a penalty for not attending. As a result, pilots might attend to information that otherwise might have been ignored. Probes might thus shape pilots' SA rather than just measure it. This problem can probably be avoided by ending the mission after the first freeze and never using the same subject again. Such a solution may be impractical, however; the number of pilots available in SA research is usually small, and the need for large amounts of data is usually great.

Unlike working memory decay effects, attention effects have been more supportive of memory probe construct validity. Fracker (1991) manipulated combat intensity by increasing the number of threatening aircraft in the simulation. As noted earlier, this manipulation led to a decrease in SA as measured by location error and FFN accuracy. If an increase in enemy number can be viewed as an increase in the load placed on attention, then the relation between probed SA and attention appears to be confirmed. Other aspects of Fracker's data were not entirely consistent with this conclusion, however. For example, in two of the experiments, subjects sometimes had to monitor an additional information display, but the presence or absence of this additional monitoring task had no effect on the memory probe measures. At present, the reasons for this result are not known.

With respect to effort and SA, the construct validity of memory probes is not clear. Fracker (1991) found that, across experimental conditions, poorer probed SA (i.e., increased location error, decreased FFN accuracy) was accompanied by increased failures to avoid ground threats. One possible explanation is that effort was diverted from the avoidance task to SA maintenance as maintaining SA became more difficult. If this interpretation were correct, then one might expect that probed SA and avoidance failures would be correlated within experimental conditions as well, but the average correlation was small (see Table 1). However, a strong correlation might not be expected if increased allocation of effort to SA maintenance prevented SA from declining. Further, the correlation might also be limited by poor reliability of both memory probes and avoidance failures: reliability of the latter was poor (r = .26).

Content validity. Like retrospective measures, concurrent memory probes can achieve a degree of content validity depending upon how they are structured. Endsley's (1989) work developing SAGAT (Situation Awareness Global Awareness Technique), a sophisticated implementation of memory probes for use in high-fidelity flight simulations, has focused on achieving a high degree of content validity for specific military missions. Nevertheless, while memory probes are particularly useful for sampling the momentary states of various situational variables, it is not clear how they can be used to sample higher levels of SA such as goal or organization awareness.

Endsley (1989) has suggested that pilots can be asked to indicate the future rather than current states of situational variables, but such responses may indicate little beyond pilots' understanding of the immediate processes controlling momentary states. Thus, while the content validity with respect to momentary states and, perhaps, processes may be high, memory probes may possess little potential for content validity at higher levels.

Criterion validity. Of those studies evaluating memory probes, only Fracker (1991) appears to have compared probed SA to a criterion measure of successful mission performance. In Fracker's experiments, subjects controlled an icon representing a friendly aircraft and used it to attack and destroy enemy aircraft. A reasonable measure of mission success, then, is the probability of a kill given an engagement with the enemy. The within-condition correlation between probed SA and kill probability was essentially zero for both location error and FFN accuracy but was statistically significant for FFN probe latency (see Table 1). Again, the poor reliability of probed SA may account for these poor correlations. (Kill probability produced a test-retest reliability coefficient of .48).

Implicit and Surrogate Measures

Explicit measures of SA clearly have liabilities: both their reliability and construct validity are in question. Perhaps for this reason, some researchers have focused on developing implicit measures (Eubanks and Killeen, 1983; Fracker, 1991; Venturino, Hamilton, and Dvorchak, 1989). In implicit measurement, the goal is to determine whether pilots' mission performance has been influenced appropriately by the occurrence of specific events. The most straightforward approach uses signal detection theory to derive an SA metric (Eubanks and Killeen, 1983; Fracker, 1991). In addition, surrogate measures have been proposed which do not directly assess the impact of events on performance but still attempt to use performance as an index of SA (Venturino et al, 1989).

Signal Detection Theoretic (SDT) Measures

Suppose that event X occurs. If pilots are aware of the event's occurrence, then they should respond in one way (a "hit"); but if pilots are unaware that the event occurred, then they should respond in a clearly different way ("miss"). Unfortunately, the interpretation of hits and misses is always complicated by response bias. For example, pilots may be biased to attack other aircraft when they are unsure whether the aircraft is friend or foe. In order to identify and correct for such bias, it is necessary to also measure false alarms (responding as if the event occurred when it did not) and correct rejections (not responding when the event did not occur).

Once these four types of responses have been identified and counted over the course of a mission, there are several methods available for computing the pilots' ability to discriminate occurrence from non-occurrence of the target event, referred to as *sensitivity* (Macmillan and Creelman, 1990). Because sensitivity declines if pilots are unaware of events occurring and increases if they are so aware, the measure provides both an empirical and an intuitively reasonable measure of awareness for a particular kind of event (cf., Hawkins, 1990).

Any discrete measure of performance can be used to measure sensitivity providing that the following three conditions can be satisfied. First, target events as well as the responses to be counted as hits must be unambiguously defined so that the presence and absence of both are clear and countable. Note that continuous measures (e.g., velocity, altitude) can be used if particular changes in the measures can be defined as events or responses (e.g., a sufficiently large decrease in velocity or altitude). Second, when more than one hit response is possible contingent upon which of several alternative forms of an event occurs, the sets of events and responses must both be finite. Third, each alternative form of an event must call for exactly one response, and that response must be unique to that alternative.

In meeting the foregoing three conditions, the main challenge may often be to find response measures that react to the events of interest. Fortunately, for some kinds of events, appropriate measures are not hard to find. Both Eubanks and Killeen (1983) and Fracker (1991) were interested in whether subjects would detect the entry of enemy targets into the subjects' weapon envelope. Eubanks and Killeen studied the performance of Air Force F-4E pilots in simulated air-to-air combat. Hits, misses, false alarms, and correct rejections were defined in terms of whether or not there was an enemy in the envelope, and whether or not pilots fired the weapon.

Reliability. Fracker (1991) reported that the test-retest reliability for envelope sensitivity was similar to that for FFN probes (see Table 1).

Construct validity. A major advantage of envelope sensitivity over explicit measures is that there is little opportunity for momentary SA to be confounded with reflective SA: if envelope sensitivity measures SA at all, it is clearly momentary SA that is measured. Nevertheless, envelope sensitivity may confound momentary SA with other processes that intervene between SA formation and mission success; such processes may include response selection (decision making) and response execution. This possibility may become more likely as the response used to define a "hit" becomes more complex, requiring greater knowledge or skill on the part of the operator. Such unwanted incidental effects may help to explain why some studies have found envelope sensitivity to be a noisy measure insensitive to important experimental manipulations (Wooldridge et al., 1982). Under some circumstances, then, the inference of momentary SA from sensitivity may be invalid if situational factors also affect other intervening processes, a possibility difficult to rule out.

Whether or not sensitivity confounds momentary SA with other factors, Fracker (1991) has found that envelope sensitivity behaves like a measure of SA in at least one respect. Specifically, sensitivity declined as the number of enemy aircraft in the simulation--i.e., the load on the attentional spotlight--increased. The average correlation of sensitivity with avoidance failures (a measure related to mental effort) was about as high as one might expect given their respective reliabilities (see Table 1). Consistent with this correlation, Eubanks and Killeen (1983) found that pilots' envelope sensitivity improved dramatically with training. Assuming that training decreases the amount of mental effort required to perform a task (cf., Schneider and Shiffrin, 1977), this result suggests that sensitivity improves as the demand for effort declines.

Content validity. The most serious challenge to the sensitivity metric may concern its content validity. There are at least three practical problems that may limit the ability of the sensitivity metric to sample the whole content domain of a mission. First, sensitivity can be measured for only a single kind of event. If the researcher is interested in a variety of event types, then each will require its own measure. Thus, the measure of SA will be a set of sensitivity parameters rather than a single parameter. Second, there may not always exist a natural response measure for events that may nonetheless be of interest (e.g., the pilot's awareness of changes in his proximity to the ground). Third, defining non-events so that false alarms and correct rejections can be counted may present a challenge. In simulations, a simple solution is to count the absence of the target event during each program cycle as one non-event. In non-simulated environments, a simple solution may not exist (see Wickens, 1984, for a discussion). In addition to these practical limitations, there is also an important theoretical limitation: the sensitivity metric is based on detections of changes in momentary states. As a result, sensitivity probably cannot be used to assess higher levels of SA such as organization or goal awareness.

Criterion validity. Studies of the criterion validity of the sensitivity metric have not been conducted. In Fracker's (1991) experiments, kill probability was equivalent to the probability of a hit used to calculate sensitivity. As a result, the obtained high correlation between sensitivity and kill probability was both expected and uninformative.

Surrogate Measures

Unlike SDT measures, there is no self-evident logical relation between surrogate measures and SA. The justification for using such measures is purely empirical: they are correlated with an existing measure of SA already believed to be valid. Of course, if a validated measure already exists, there may be little need for another. Still one might desire a measure that is simpler or less costly to obtain than the currently validated one.

Only one attempt to identify surrogate measures is known to have been reported. In a complex study, Venturino, Hamilton, and Dvorchak (1989) measured fire point selection (FPS, the point relative to an enemy target at which pilots launch their missile) during simulated air-to-air combat. The relationship of FPS to subjective self-ratings of SA by the pilots was examined and found to be both non-linear and non-monotonic. As a result, correlation coefficients were not calculated. Nevertheless, the authors felt able to conclude that "extreme or erratic FPS values may be an indicator of poor situation awareness" (p. 4-4).

Reliability. No reliability studies of FPS or any other potential surrogate measures are known to have been conducted.

Validity. The Achilles' heel of surrogate measures is the assumption that one possesses a valid criterion measure to begin with. In Venturino et al.'s (1989) study, however, the assumption is problematic. While pilots' self-ratings of their own SA may sometimes be valid, there is evidence that such is not always the case (discussed below). Venturino et al. were aware of this difficulty and did not base their conclusions on SA ratings alone. Nevertheless, without a valid SA criterion measure, the conclusion that FPS measures SA seems circular: SA is inferred from the measure that it is supposed to explain.

Subjective Rating Measures

Subjective rating measures of SA are by far the easiest to collect and so have proven popular (Arbak, Schwartz, and Kuperman, 1987; Fracker and Davis, 1990; Selcon and Taylor, 1989; Taylor, 1989; Venturino, Hamilton, and Dvorchak, 1989; Ward and Hassoun, 1990). Two classes of rating measures have been used: direct and comparative. In direct ratings, pilots assign a numerical value to their SA during a given mission (or mission segment). While pilots may make these assignments in light of the ratings given to previous missions, the rating technique does not inherently require them to do so (although they may be instructed to do so). In any case, the assigned rating is assumed to have some direct, monotonic relation to the absolute magnitude of SA experienced during the mission. In comparative ratings, pilots compare their SA during one mission to that during another and assign a value to the ratio of one to the other. Thus, in comparative ratings, no attempt is made to determine the location of reported SA with respect to a fixed point on the underlying scale. Rather, one obtains ratio estimates only. For example, values on an underlying scale of 10 and 40 would appear identical to values of 100 and 400.

Direct Ratings

The most common direct rating measures have used Likert scales. For example, Ward and Hassoun (1990) tested pilots' ability to recover from unusual attitudes with three different types of head-up display pitch ladders. Immediately following a trial, pilots were asked whether they agreed with the statement "I experienced no confusion with this pitch ladder configuration and was easily able to recover to straight and level flight." Pilots responded with a number between 1 and 9 indicating their agreement with the statement (1 = "decidedly disagree," 9 = "decidedly agree").

While Ward and Hassoun (1990) used only one rating scale, most researchers have employed multiple scales on the hypothesis that SA is a multi-dimensional construct (Arbak et al., 1987; Selcon and Taylor, 1989; Taylor, 1989; Venturino et al, 1987). Arbak et al. used six rating scales derived from a definition of SA focusing on various elements of air-to-air combat (e.g., friendly locations and actions, enemy locations and actions, available options, and so on). A similar approach appears to have been used by Venturino et al., although those authors did not identify the scales used. Taylor (1989; Selcon and Taylor, 1989) rejected Arbak et al.'s a priori approach to scale construction, opting instead for an empirical approach. Beginning with 44 possible SA dimensions, Taylor used principal components analysis to identify three major factors since incorporated into the Situational Awareness Rating Technique (SART): *attentional demand*, *attentional supply*, and *situational understanding*. Taylor also decomposed these three factors into ten components, but the stability of these components is not currently known (cf., Harmon, 1976).

Reliability. No reliability studies of direct ratings of subjective SA are known to have been conducted.

Construct validity. No coherent theory currently exists either of subjective SA or of how subjective SA might be mapped onto Likert-type rating scales. Consequently, it is difficult to assess just what it is that subjective SA ratings might actually measure. One possibility is that subjective SA ratings are actually confidence ratings; that is, ratings of ones confidence that one knows everything that needs to be known. The usefulness of such confidence ratings probably depends upon how they are related to momentary SA, a relationship that has not yet been explored.

Taylor (1989) has explored the relationship between subjective SA (SART's situational understanding scale) on one hand and attentional load and effort on the other. The measures of load and effort were subjective rather than objective, however (SART's attentional demand and supply scales, respectively). In one experiment, subjective load and effort were positively correlated ($r = .60$) but neither was correlated with subjective SA (r's < .14). This result could mean that as attentional load increased, effort may also have increased in order to maintain SA at a relatively constant level. In a second experiment, subjective load was correlated with subjective effort ($r = .53$) but not subjective SA ($r < .14$); at the same time, subjective effort was correlated with subjective SA ($r = .65$). These results are also sensible; they could indicate that more effort was expended than actually necessary to maintain SA in the face of increasing load. Results in both experiments were apparently consistent with existing theories of situation assessment (Endsley, 1988; Fracker, 1988). Thus, Taylor's (1989) research suggests that SART may indeed possess some degree of construct validity.

Content validity. Content validity has not always been an objective of subjective ratings. Taylor (1989; Selcon and Taylor, 1989) has focused on establishing construct validity with little effort to identify or sample relevant mission content domains. At least one researcher has sought to establish content validity, however: Arbak et al.'s (1987) six rating scales were a deliberate attempt to sample tho content domain of air-to-air combat. The contrast botween Taylor's and Arbak et al.'s research may point to the difficulty of developing rating scales to establish both construct and content validity simultaneously. In principle, such scales could be developed by nesting content-oriented scales within construct-oriented scales (or the other way around). Although such nested scales might prove too complex in practice, their development and evaluation may be a useful direction for future research.

Criterion validity. In spite of their appeal, subjective ratings of SA--when taken alone--confront a major difficulty. While such measures may be able to assess subjects' confidence in their own SA, there is compelling evidence that this confidence is poorly related to measures of mission success. For example, Venturino et al. (1989) reported that pilots who rated their SA as high were as likely to have performed well as poorly. An even more dramatic case has been reported by Ward and Hassoun (1990). Those authors found that the HUD pitch ladder which produced the best subjective SA ratings also produced the greatest percentage of inverted recoveries: pilots believed they were upright when in fact they were upside down! In terms of evaluating that particular display, this outcome was highly informative because it revealed that the display was not just uninformative but was in fact dangerously misleading. These results suggest that an objective assessment of SA lies in the inconsistency between subjective SA ratings and the appropriate performance criteria rather than in the ratings alone. Quantitative assessment of this inconsistency may provide a useful index of SA and may be a fruitful direction for future research.

An alternative approach would be to try and remove the inconsistency between SA ratings and performance criteria. The inconsistency probably arises because pilots do not know that they are unaware of critical information. A procedure to eliminate the inconsistency might be to make pilots aware of task outcome before collecting their ratings. If Ward and Hassoun (1990) had first told pilots whether they were inverted before collecting their ratings, the results would undoubtedly have been quite different. Nevertheless, the "improved" results would have been deceptive in another way: while the ratings would have revealed the poor SA associated with the troublesome display, they would have hidden the fact that the display was actually misleading rather than just uninformative. Thus, what is clear is that subjective SA ratings should not be used alone but should be combined in some way with criterion measures of performance.

Comparative Ratings

Although direct subjective ratings may seem to assess the magnitude of perceived SA, such ratings generally cannot be compared across raters. A pilot who assigns his SA a rating of "9" may mean the same thing as another who assigns her SA a rating of "7." Nevertheless, if one is comparing SA across different missions, such ratings can be compared within subjects if individual subjects are consistent in how they map perceived SA onto the rating scale. Whether subjects are in fact consistent is difficult to evaluate empirically, however. For that reason, Fracker and Davis (1990) proposed a subjective SA scaling technique which both encourages and assesses consistency. Derived from Vidulich's (1989) Subjective Workload Dominance (SWORD) technique (see also Budescu, Zwick, and Rapoport, 1986; Hughes et al, 1990; Lodge, 1981; Saaty, 1977; Ward and Hassoun, 1990), subjects first experience several different experimental conditions and then judge how much better SA in one condition was compared to another, for all possible pairs. The fact that subjects directly compare conditions encourages them to apply the same subjective scale to each condition, and the resulting two-way matrix can be examined to determine the extent to which subjects were in fact inconsistent.

Reliability. No studies of SA comparative rating reliability have yet been conducted.

Construct validity. As with direct subjective rating measures, it may be that comparative ratings are merely an alternative method for assessing operators' confidence in their SA. Nevertheless, Fracker and Davis (1990) provided evidence that such ratings may yet possess a degree of construct validity. Using the combat task from Fracker (1991), the researchers had subjects perform under two levels of combat intensity (Low, High) and two levels of difficulty in identifying objects as friend or foe (Easy, Hard). In addition to paired-comparison ratings of SA, Fracker and Davis also collected Subjective Workload Assessment Technique (SWAT) ratings of mental workload for each of the four experimental conditions (Reid, Shingledecker, and Eggemeier, 1981). SWAT ratings clearly distinguished among the four conditions and ordered them from least to most workload as follows: Low-Easy, Low-Hard, High-Easy, High-Hard. Subjective SA ratings failed to distinguish between the Low-Hard and High-Easy conditions but otherwise provided the same ordering from best to poorest SA. (Within experimental conditions, the correlations between SWAT and SA ratings were virtually zero.) These results may indicate that subjects were able to maintain their SA from the Low-Hard to the High-Easy condition by increasing the amount of mental effort expended. Further, the same pattern found in the SA ratings was also observed in FFN accuracy (the correlation between subjective SA and FFN accuracy across experimental conditions was not strong, however: $r = .35$).

Nevertheless, not all of the evidence from Fracker and Davis' (1990) experiment supported the construct validity of the comparative ratings. The major difficulty was that SWAT ratings dissociated from threat avoidance failures as the experimental conditions increased in difficulty (in the easiest condition, $r = .44$; in the most difficult, $r = -.11$). Because avoidance failures were also a measure of mental workload, this systematic dissociation complicates the interpretation of SWAT as a measure of mental effort (cf., Yeh and Wickens, 1988) and hence the interpretation of the paired-comparison SA ratings. Like direct ratings, then, it seems prudent to avoid relying on comparative ratings alone.

Content validity. In theory, comparative rating scales can be constructed so that they at least appear to possess content validity. To illustrate, suppose that several alternative cockpit displays were being compared to determine which gives pilots the best SA. Two approaches are possible. First, the displays could be compared on each of several dimensions, one dimension at a time (e.g., locations of enamy aircraft, status of enemy anti-aircraft artillery, movements of enemy tank units, locations of friendly ground forces, etc.). If four displays were to be compared on just four such dimensions, pilots would have to make 4 x 6 = 24 separate comparisons; if eight conditions were compared, the number of comparisons would increase to 112. Second, instead of comparing the displays along pre-determined dimensions, empirical dimensions could be extracted using multidimensional scaling methods (Torgerson, 1958). In order to establish four stable dimensions, as many as 30 displays might need to be compared, requiring 435 comparisons. Whether the first or second approach is taken, it is clear that the needed number of comparicons can become quite large very rapidly. As a result, a practical approach to establishing the content validity of comparative ratings seems unlikely.

Criterion validity. In the experiment reported by Fracker and Davis (1990), no systematic relationship between SA comparative ratings and kill probability was found. Both within and across experimental conditions, the correlation between the two was virtually zero. When combined with the poor results obtained with direct rating measures (see above), it appears that subjective SA ratings in general cannot be used to predict criterion performance measures.

DIRECTIONS FOR FUTURE RESEARCH

Development of SA measurement methods has only just begun--and this is evident in the preceding review. Several topics and problems requiring further research still exist and have been identified throughout the discussion. Some of these problems eventually may be solved through continued research on existing SA measures, but new measures doubtless will be needed as well. Although it is not yet clear what those new measures should be, some possibilities suggest themselves and are briefly discussed.

Continuing Research on Existing Measures

Reliability. The reliability of SA metrics will continue to require research, especially because limitations in their validity may well be caused by limited reliability. Explicit measures, particularly location memory probes, appear to be highly unreliable. The reasons for this unreliability need to be explored if the situation is to be improved. In the mean time, probes of location memory should be used with great caution.

Validity. The three categories of SA metrics--explicit, implicit, and subjective ratings--each appears to have its own strengths and weaknesses. In terms of construct validity, explicit measures may fare more poorly than implicit measures because the former seem to confound momentary with reflective SA whereas the latter do not. On the other hand, explicit measures can probably achieve high content validity more easily than can implicit measures. Both the construct and criterion validity of subjective ratings are questionable, but such ratings are usually easier to collect than either their explicit or implicit counterparts. Thus, no one metric seems adequate in and of itself. Perhaps these three classes of methods are complementary, with each providing information not easily available from the other. On the other hand, it may simply be that the collective results of the three methodologies may be no better than any of the methodologies alone. To suggest an analogy, combining a broken thermometer with a broken wind gauge will not provide a more accurate assessment of the weather. Thus, whether explicit, implicit, and subjective ratings are complementary and hence should be used together is a question needing further study.

Research to more clearly establish or improve the construct validity of the various measures continues to be needed. Unfortunately, the cognitive theories that guide tests of construct validity are currently in dispute (Fracker and Wickens, 1989; Hirst and Kalmar, 1987; Navon, 1984). As a result, definitive tests of construct validity may have to await resolution of some of these theoretical controversies. Nevertheless, key issues that need to be examined are (1) the relative contributions of momentary and reflective SA to both concurrent and retrospective explicit measures, (2) the sensitivity of all measures to attentional demand, mental workload, and attention allocation strategies, and (3) the degree to which implicit measures such as envelope sensitivity confound SA with intervening processes such as decision making and response execution. In addition to these issues, considerable work is needed with regard to subjective SA ratings. The most immediate need is for a theory of subjective SA and of how operators go about mapping their perceived SA onto the provided rating scales.

Considerable effort may be needed to improve the content validity of most measures. Existing explicit measures are quite limited in their ability to capture higher levels of operator SA (such as goal awareness). Implicit measures so far have been developed only in the context of simple choices (e.g., whether or not to fire a weapon). Extension of these measures to more complex choice situations seems to be the next logical step in their development. Subjective rating scales have so far focused either on construct or on content validity. If rating scales are to continue playing a role in SA assessment, they should be expanded to achieve both construct and content validity, perhaps in the form of nested rating scales.

Establishing the criterion validity of SA metrics assumes that well-established mission performance criteria exist, an assumption that may not always be met. But where such criteria exist, explicit SA measures in particular have not performed as well as might be hoped. Perhaps this poor performance will improve when the reliability problems are solved. In any case, criterion validity should continue to be a focus in the development of explicit SA assessment. Regarding subjective SA ratings, criterion validity can probably be achieved only by incorporating performance criteria into the measure, perhaps by quantifying the inconsistency between self-assessed SA and the quality of actual mission performance.

Developing New SA Measurement Methods

Sarter and Woods's (1991) have pointed out that new measures of SA are still needed, especially measures that will establish better content validity. For example, there is as yet no good way in which to assess higher levels of SA such as goal or organization awareness. Real-time assessment of these higher levels is difficult to imagine, except possibly for the use of verbal protocols (Sarter and Woods, 1991). A verbal protocol is obtained by having operators verbalize their thoughts as they carry out their missions. These protocols are recorded on tape and later analyzed off line. Retrospective protocols collected after the fact (e.g., Whitaker and Klein, 1987) may also be useful in this regard but would encounter the problems discussed earlier (under Explicit Measures). If either concurrent or retrospective protocols are to be used, new methods of analyzing them may need to be developed in order to reveal the higher levels of SA latent within them.

Further development of subjective rating approaches to content validity might also prove useful. In particular, organizational psychologists have developed subjective rating methods for job analysis (e.g., McCormick, 1976, 1979) that could possibly be adapted to SA assessment. For example, the military aircraft cockpit might be decomposed into individual displays and particular items of information found on those displays. In the case of multi-function displays, a three-level decomposition of displays-pages-information might be needed. Following a mission, pilots might then rate each item of information for a particular page or display. Some pilots might rate how much time they spent attending to each item. Other (or the same) pilots might indicate the importance of each item to the mission. Still others might indicate how difficult the items were to find or use. Items that were critical to mission success but difficult to find or use could point to changes in the displays that would improve SA. Items that were frequently attended but not critical could indicate that the displays are badly formatted, encouraging sub-optimal attention strategies. Finding and correcting such formatting problems would also contribute to better SA.

In summary, much work is still needed before highly reliable, well-validated measures of operator SA will be available. In the meantime, the military services will continue searching for ways to improve--and potential adversaries will continue looking for ways to degrade--friendly combatants' SA.

REFERENCES

Allen, M. J., and Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks-Cole.

Arbak, C. J., Schwartz, N., and Kuperman, G. (1987). Evaluating the panoramic cockpit controls and displays system. Paper presented at the 4th annual symposium on aviation psychology, Columbus, Ohio.

Budescu, D. V., Zwick, R., and Rapoport, A. (1986). A comparison of the eigenvalue method and the geometric mean procedure for ratio scaling. *Applied Psychological Measurement, 10*, 69-78.

Carr, T. H. (1986). Perceiving visual language. In K. R. Boff, L. Kauffman, and J. P. Thomas (Eds.), *Handbook of perception and human performance, Volume II: Cognitive Processes and Performance* (29:1-92). New York: Wiley.

Clausewitz, C. V. (1836/1984). *On war*. Princeton, NJ: Princeton University Press.

Downing, C. J. (1988). Expectancy and visual-spatial attention: Effects of perceptual quality. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 188-202.

Dunlap, W. P., Silver, N. C., and Bittner, A. C. (1986). Estimating reliability with small samples: Increased precision with averaged correlations. *Human Factors, 28*, 685-690.

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society 32nd annual meeting*. Santa Monica, CA: Human Factors Society.

Endsley, M. R. (1989). A methodology for the objective measurement of pilot situation awareness. In *Proceedings of the NATO AGARD Conference on Situational Awareness in Aerospace Operations* (AGARD-CP-478). Springfield, VA: National Technical Information Service.

Ericsson, K. A., and Staszewski, J. J. (1989). Skilled memory and expertise: Mechanisms of Exceptional Performance. In D. Klahr and K. Kotovsky (Eds.), *Complex information processing* (235-267). Hillsdale, NJ: Earlbaum.

Eriksen, C. W., and Yeh, Y. (1985). Attention allocation in the visual field. *Journal of Experimental Psychology: Human Perception and Performance, 11*, 583-597.

Eubanks, J. L., and Killeen, P. R. (1983). An application of signal detection theory to air combat training. *Human Factors, 25*, 449-456.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51*, 327-358.

Fracker, M. L. (1988). A theory of situation assessment: Implications for measuring situation awareness. In *Proceedings of the Human Factors Society 32nd annual meeting* (pp. 102-106). Santa Monica, CA: Human Factors Society.

Fracker, M. L. (1991). *Measures of situation awareness: An experimental evaluation*. Technical Report in preparation. Wright-Patterson AFB OH: Armstrong Laboratory, Crew Systems Directorate.

Fracker, M. L., and Davis, S. A. (1990). Measuring operator situation awareness and mental workload. In *Proceedings of the Fifth Mid-Central Ergonomics/Human Factors Conference*. Dayton, OH: University of Dayton.

Fracker, M. L., and Wickens, C. D. (1989). Resources, confusions, and compatibility in dual-axis tracking: Displays, controls, and dynamics. *Journal of Experimental Psychology: Human Perception and Performance, 15*, 80-96.

Gopher, D., and Donchin, D. (1986). Workload: An examination of the concept. In K. R. Boff, L. Kauffman, and J. P. Thomas (Eds.), *Handbook of perception and human performance, Volume II: Cognitive processes and performance* (41:1-49). New York: Wiley.

Gulliksen, H. (1950). *Theory of mental tests*. New York: Wiley.

Harmon, H. H. (1976). *Modern factor analysis*. Chicago: University of Chicago Press.

Hawkins, H. L., Hillyard, S. A, Luck, S. J., Mouloua, M., Downing, C. J., and Woodward, D. P. (1990). Visual attention modulates signal detectability. *Journal of Experimental Psychology: Human Perception and Performance, 16*, 802-811.

Hirst, W., & Kalmar, D. (1987). Characterizing attentional resources. *Journal of Experimental Psychology: General, 116*, 68-81.

Hughes, E. R., Hassoun, J. A., Ward, G. F., and Rueb, J. D. (1990). *An assessment of selected workload and situation awareness metrics in a part- mission simulation: Final report* (ASD-TR-90-5009).

Wright- Patterson AFB OH: Aeronautical Systems Division.

Kibbe, M. P. (1988). Information transfer from intelligent EW displays. In *Proceedings of the Human Factors Society 32nd annual meeting*. Santa Monica, CA: Human Factors Society.

Klein, G. A., Calderwood, R., and Clinton-Cirocco, A. (1985). *Rapid decision making on the fire ground* (KA-TR-84-41-7). Alexandria, VA: US Army Research Institute.

Lodge, M. (1981). *Magnitude scaling*. Beverly Hills, CA: Sage.

Loftus, E. F. (1979). *Eyewitness testimony*. Cambridge, MA: Harvard University Press.

Loftus, E. F., and Loftus, G. R. (1980). On the permanence on stored information in the human brain. *American Psychologist, 35*, 409-420.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95*, 492-527.

Lord, F. M., and Novick, M. R. (1968). *Statistical theories of mental test scores*. Menlo Park, CA: Addison-Wesley.

Macmillan, N. A., and Creelman, C. D. (1990). Response bias: Characteristics of detection theory, threshold theory, and "non-parametric" indexes. *Psychological Bulletin, 107*, 401-413.

Marshak, W. P., Kuperman, G., Ramsey, E. G., and Wilson, D. (1987). Situation awareness in map displays. In *Proceedings of the Human Factors Society 31st annual meeting* (pp. 533-538). Santa Monica, CA: Human Factors Society.

McCormick, E. J. (1976). Job and task analysis. In M. D. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (651-696). Chicago: Rand-McNally.

McCormick, E. J. (1979). *Job analysis: Methods and applications*. New York: AMACOM.

Moray, N., (Ed.) (1979). *Mental workload*. New York: Plenum.

Moray, N. (1988). Mental workload since 1979. *International Review of Ergonomics, 2*, 123-150.

Murphy, K. R., and Davidshofer, C. O. (1991). *Psychological testing: Principles and applications*. Englewood Cliffs, NJ: Prentice Hall.

Navon, D. (1984). Resources--A theoretical soupstone? *Pyschological Review, 91,* 216-234.

Navon, D., and Miller, J. (1987). The role of outcome conflict in dual-task interference. *Journal of Experimental Psychology: Human Perception and Performance, 13,* 435-448.

O'Donnell, R. D., and Eggemeier, F. T. (1986). Workload assessment methodology. In K. R. Boff, L. Kauffman, and J. P. Thomas (Eds.), *Handbook of perception and human performance, Volume II: Cognitive Processes and Performance* (42:1-49). New York: Wiley.

Ogden, G., Levine, J., and Eisner, E. (1979). Measurement of workload by secondary tasks. *Human Factors, 21,* 529-548.

Peterson,L. R., and Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology, 58,* 193- 198.

Reid, G. B., Shingledecker, C., and Eggemeier, T. (1981). Application of conjoint measurement to workload scale development. In *Proceedings of the Human Factors Society 25th annual meeting.* Santa Monica, CA: Human Factors Society.

Roediger, H. L. (1990). Implicit memory: Retention without remembering. *American Psychologist, 45,* 1043-1056.

Saaty, T. L. (1977). A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology, 15,* 234-281.

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

Schneider, W., and Shiffrin, R. M. (1977). Controlled and automatic human information processing: 1. Detection, search and attention. *Psychological Review, 84,* 1-66.

Selcon, S. J., and Taylor, R. M. (1989). Evaluation of SART as a tool for aircrew systems design. In *Proceedings of the NATO AGARD conference on situational awareness in aerospace operations* (AGARD-CP-478). Springfield, VA: National Technical Information Service.

Taylor, R. M. (1989). Situational awareness rating technique (SART): The development of a tool for aircrew systems design. In *Proceedings of the NATO AGARD conference on situational awareness in aerospace operations* (AGARD-CP-478). Springfield, VA: National Technical Information Service.

Thomas, G. S., Houcke, M. R., and Bell, H. H. (1990). *Training evaluation of air combat simulation* (AFHRL-TR-90-30). Williams AFB AZ: Air Force Human Resources Laboratory, Operational Training Division.

Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: Wiley.

Venturino, M., and Kunze, R. J. (1989). Spatial awareness with a helmet mounted display. In *Proceedings of the Human Factors Society 33rd annual meeting* (pp. 1388-1391). Santa Monica, CA: Human Factors Society.

Venturino, M., Hamilton, W. L., and Dvorchak, S. R. (1989). Performance- based measures of merit for tactical situation awareness. In *Proceedings of the NATO AGARD conference on situational awareness in aerospace operations* (AGARD-CP-478). Springfield, VA: National Technical Information Service.

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).

Ward, G. F., and Hassoun, J. A. (1990). *The effects of head-up display pitch ladder articulation, pitch number location and horizonline length on unusual attitude recoveries for the F-16* (ASD-TR-90-5008). Wright- Patterson AFB OH: Aeronautical Systems Division.

Wells, M. J., Venturino, M., and Osgood, R. K. (1988). Using target replacement performance to measure spatial awareness in a helmet-mounted display. In *Proceedings of the Human Factors Society 32nd annual meeting* (pp. 1429-1433). Santa Monica, CA: Human Factors Society.

Whitaker, L. A., and Klein, G. A. (1988). Situation awareness in the virtual world. In *Proceedings of the eleventh symposium on psychology in the Department of Defense* (USAFA-TR-88-1, pp. 321-325). United States Air Force Academy.

Wickens, C. D. (1984). *Engineering psychology and human performance*. Columbus, OH: Merril.

Wierwille, W. (1979). Physiological measures of aircrew mental workload. *Human Factors, 21*, 575-593.

Williges, R., and Wierwille, W. (1979). Behavioral measures of aircrew mental workload. *Human Factors, 21*, 549-574.

Wooldridge, A. L., Kelly, M. J., Vreuls, D., Obermayer, R. W., Nelson, W. H., and Norman, D. A. (1982). *Air combat maneuvering performance measurement state space analysis* (AFHRL-TR-82-15). Williams AFB, AZ: Air Force Human Resources Laboratory.

Wyer, R. S., Jr., and Srull, T. K. (1986). Human cognition in its social context. *Psychological Review, 93*, 322-359.

Yeh, Y., and Wickens, C. D. (1988). Dissociation of performance and subjective measures of workload. *Human Factors, 30*, 111-120.

U.S. Government Printing Office: 1992 – 648-069/60210