A Review of Situation Awareness Literature Relevant to Pilot Surveillance Functions

John Uhlarik
Doreen A. Comerford

Department of Psychology
Kansas State University
Manhattan, KS 66506

March 2002

Final Report

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

U.S. Department
of Transportation

**Federal Aviation
Administration**

NOTICE

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

Technical Report Documentation Page

1. Report No. DOT/FAA/AM-02/3	2. Government Accession No.	3. Recipient's Catalog No.	
4. Title and Subtitle A Review of Situation Awareness Literature Relevant to Pilot Surveillance Functions		5. Report Date March 2002	
		6. Performing Organization Code	
7. Author(s) Uhlarik, John and Comerford, Doreen A.		8. Performing Organization Report No.	
9. Performing Organization Name and Address Department of Psychology Kansas State University Manhattan, KS 66506		10. Work Unit No. (TRAIS) 98F80691	
		11. Contract or Grant No.	
12. Sponsoring Agency name and Address Office of Aerospace Medicine Federal Aviation Administration 800 Independence Ave., S.W. Washington, DC 20591		13. Type of Report and Period Covered	
		14. Sponsoring Agency Code	

15. Supplemental Notes

This research was performed under task AM-A-00-HRR-514. Thomas E. Nesthus, Ph.D., and Kurt Joseph, Ph.D., were the COTRs for this contracting effort.

16. Abstract

This paper addresses situation awareness (SA) as it relates to surveillance activities in commercial air carriers. The concept of SA in aerospace operations and relevant literature are reviewed and critiqued. Four different approaches that are used to explain SA are examined, and the three major SA measurement techniques and their respective subcategories are described. Research that has explored these techniques is reviewed, and the shortcomings of the measures are discussed in terms of relevant psychology measurement criteria (e.g., validity). Several components of SA also are discussed because they appear to relate to surveillance activities. Research related to these components is reviewed, and a summary that focuses on the current state of SA research is provided, along with an evaluation about the future of SA research and applications.

17. Key Words Situation Awareness, Pilot Surveillance Functions, Measurement	18. Distribution Statement Document is available to the public through the National Technical Information Service; Springfield, VA 22161		
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 24	22. Price

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

ACKNOWLEDGMENTS

TABLE OF CONTENTS

A Review of Situation Awareness Literature Relevant to Pilot Surveillance Functions

1.0 Introduction

The goals of the present paper were to (a) present a general review of the concept of situation awareness (SA), (b) review the methods and issues associated with the measurement of SA, and (c) discuss the SA literature as it relates specifically to surveillance activities in commercial air carriers. The second phase of this research, which is presented in a separate document, identifies and classifies information requirements for an important element of pilot surveillance activities — traffic separation.

Flight instructors and experienced pilots have long held the intuitive notion that successful flight results when a pilot "has the big picture," and conversely when problems arise due to pilot error, it is because some aspect of this picture is missing or incorrect. In the past decade, human factors specialists have attempted to transform this notion into a formal psychological construct to develop both an operational definition (i.e., a definition of the construct in terms of observable behavior) and an experimental paradigm for researching it. An operational definition specifies a construct in terms of empirical units of measurement and allows human factors specialists to make recommendations regarding such issues as: (a) the utility of a novel display (i.e., whether or not a display assists in obtaining an adequate mental model of the relevant information), (b) the content of training (i.e., which type of training facilitates pilots' overall understanding of circumstances), and (c) selection (i.e., in terms of individual difference variables, who is best at obtaining the big picture).

The concept of situation awareness is especially compelling in the operational setting of aviation, which involves the operation and control of a complicated system in dynamic environments. The human has to integrate widely disparate and sometimes inconsistent inter-sensory input (visual, auditory, tactile, vestibular, etc.) with elaborate cognitive models of the machine and the operating environment to control the movement of a vehicle through a medium. The SA construct has also been extended to other domains such as air traffic control (e.g., Endsley & Rodgers, 1994), battlefield management (e.g., Kass, Herschler, & Companion, 1991), medical procedures (e.g., Gaba, Howard, & Small, 1995), and even football

(e.g., Walker & Fisk, 1995). These domains share common characteristics; For example: (a) the environment is often dynamic and information rich; (b) the human may sometimes experience high mental workload; (c) extensive training is often required; (d) the problems are often ill-structured; and (e) time is often constrained.

The impetus and interest in SA have many parallels with the construct of mental workload (cf., Wickens, 1992b). In research on mental workload, what is of interest are the demands that the task(s) impose on the pilot's mental resources. Although that demand is hypothesized to correlate with performance, it does not consistently do so. Like workload, SA is thought to correlate with performance. Workload research can be viewed in three different contexts: (a) predicting task performance based on mental workload, (b) assessing workload imposed by equipment, and (c) assessing workload experienced by the human operator. The same could be said for SA. For instance, like mental workload, SA is a psychological construct that is not directly observable, and there is disagreement regarding an operational definition. A myriad of workload assessment techniques have been proposed, but none satisfactorily meet the criteria, such as sensitivity, diagnosticity, selectivity, unobtrusiveness, bandwidth, and reliability (O'Donnell & Eggemeier, 1986; Wickens, 1992a). These criteria are discussed in the section on measures used to assess SA, but in the present context, many of the lessons learned from the last 25 years of mental workload research are likely to apply to SA.

Many different definitions have surfaced as a consequence of the difficulty of defining SA. This difficulty is demonstrated best by a special issue of Human Factors (Volume 37, No. 1), in which each of nine articles defines SA in a different manner (Baxter & Bass, 1998). The present paper is intended, in part, to provide a primer on the construct of SA and, as such, various conceptualizations and existing measures of SA will be reviewed. The goal in presenting such a review is to provide an understanding of the construct and the various issues surrounding it. After this general review is presented, SA will be discussed as it relates more specifically to pilot surveillance activities.

2.0 Situation Awareness: A Review

2.1 Definitions of SA

The most commonly cited definition of SA is one suggested by Endsley (1995b) who states that "Situation awareness is the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (p. 36). Despite the frequency of its citation, many researchers do not accept this definition of SA. For example, Wickens (1992b) suggests that SA is not limited to the contents of working memory, but it is the ability to mentally *access* relevant information about the evolving circumstances of a flight. Crane (1992) provides a very different conceptualization of SA by focusing on inadequate performance and suggests that SA is synonymous with expert-level performance.

These three definitions provide a sample of the many that exist, and it should be clear from these examples that the conceptualizations of SA are diverse. An exhaustive list of definitions would not be particularly valuable for providing the reader with a detailed understanding of SA. Instead, the next section provides a review of the approaches used to both define and explain SA. An approach is different from a definition in that it is broader than a mere definition and utilizes general models or theories to explain a given psychological construct. The focus on approaches, rather than the specific definitions, should allow one to obtain a general understanding of the SA literature.

2.2 Approaches Used to Define and Explain SA

Four qualitatively different approaches will be addressed in this section:
- use of the i*nformation-processing* model in defining and explaining SA;
- use of the *perception/action cycle* in definitions and explanations of SA;
- equating SA with *expertise*;
- use of SA as a mere *description* of a behavioral phenomenon.

2.2.1 Information-Processing Models

Models of information processing include psychological constructs such as attention and short-term memory. Although these models are meant to describe and explain human information processing, they are also utilized more specifically to understand SA. The most prominent example of the latter use is the approach taken by Endsley (1995b), which is conceptually similar

to the models used to explain human information processing in general. That is, her model of the information-processing mechanisms involved in SA includes such constructs as short-term sensory stores, schemata, and attention. This model is shown in Figure 1. The following excerpt details the components of the information-processing model and illustrates the manner in which Endsley applies all aspects of the information-processing model to SA:

To summarize the key features of SA in this model, a person's SA is restricted by limited attention and working memory capacity. Where they have been developed, long-term memory stores, most likely in the form of schemata and mental models, can largely circumvent these limits by providing for the integration and comprehension of information and the projection of future events (the higher levels of SA), even on the basis of incomplete information and under uncertainty. The use of these models depends on pattern matching between critical cues in the environment and elements in the model. Schemata of prototypical situations may also be associated with scripts to produce single-step retrieval of actions from memory. SA is largely affected by a person's goals and expectations which will influence how attention is directed, how information is perceived, and how it is interpreted. This top-down processing will operate in tandem with bottom-up processing in which salient cues will activate appropriate goals and models. In addition, automaticity may be useful in overcoming attention limits; however, it may leave the individual susceptible to missing novel stimuli that can negatively affect SA (p. 49).

`

Endsley's explanation of SA (1995b) includes three aspects that are distinct from generic information-processing models. First, she suggests that SA consists of three hierarchical phases: Level 1 (i.e., perception of elements in the environment), Level 2 (i.e., comprehension of the current situation), and Level 3 (i.e., projection of future status). For example, imagine a situation in which a pilot is approaching hazardous terrain. This terrain, in Endsley's terms, would be a task factor and represents the state of the environment. If the pilot sees the terrain, the pilot has perceived the element in the current situation (Level 1), and if the pilot recognizes the terrain is hazardous, the pilot has comprehended the situation (Level 2). Furthermore, if the pilot is able to estimate the time at which the aircraft would collide with the terrain and determine when a maneuver is necessary, the pilot has projected the future status of the situation (Level 3).
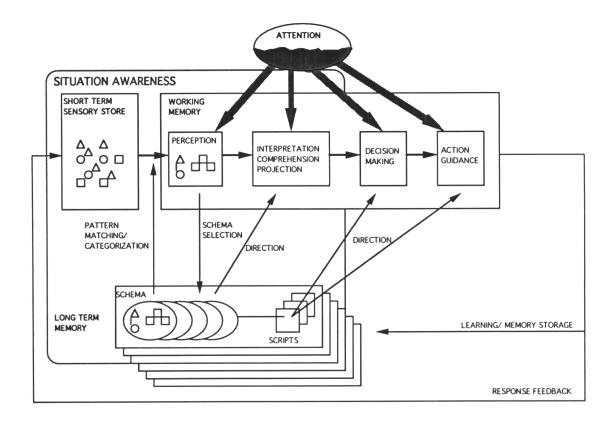
Figure 1. A model of the mechanisms involved in SA (adapted from Endsley, 1995b).

Endsley (1995b) also asserts that SA is separate from the processes used to achieve SA, which is important because her assertion suggests that an operational definition of SA should not include any of the processes involved in the achievement of SA (although her own theoretical definition includes processes such as perceiving, comprehending, and projecting). Thus, Endsley's assertion suggests that the activities performed to achieve SA (e.g., the activities involved in comprehending an event) should not be measured, but rather it is the result of these activities (e.g., whether or not one does comprehend an event) that should be measured. For example, Endsley suggests that the manner in which a pilot comes to be aware of a terrain hazard is not important in the operational measurement of SA. Instead, she suggests SA should be measured by simply assessing whether or not the pilot is aware of such terrain, and (as will be discussed later) the measure she developed presumably measures SA as a product and not a process.

Finally, Endsley suggests (1995b) that a definition of SA should only address a pilot's knowledge regarding dynamic aspects of the environment and should not address all of a pilot's usable knowledge. For example, Endsley and Rodgers (1994) identified the information for which an air traffic controller must have knowledge to obtain SA. As such, they did not include static information like the number of airports in a sector, but they did suggest that a controller must have knowledge of current aircraft positions. Thus, Endsley proposed that a true measure of SA should only assess knowledge regarding aspects of the environment that are dynamic or variable in nature.

Use of the information-processing model to explain SA is potentially problematic for two reasons. First, the information-processing model includes many psychological constructs that are themselves not well-understood (e.g., attention, schemata). Some of these constructs are subject to a great deal of debate and are researched using a wide variety of experimental paradigms. Second, when SA is explained in terms of the information-processing model, the process of achieving SA appears relatively static and finite. Other approaches have been suggested that emphasize the dynamic nature of this process. For example, one approach that emphasizes the dynamic nature of SA is the use of the *perception/action cycle* (Adams, Tenney, & Pew, 1995).

2.2.2 The Perception/Action Cycle

Figure 2 shows that the perception/action cycle consists of three elements: (a) the object (i.e., available information in the external environment); (b) the schema (i.e., internal knowledge that is theoretically structured in an organized manner, developed through training/experience, and is stored in long-term memory when not in use); and (c) exploration (i.e., a search of the environment by the observer). The cycle is hypothesized to work as follows: The object modifies the schema, the schema directs exploration, and exploration leads to sampling of the object. For example, imagine a pilot that is on a familiar route. A river (object) may modify the pilot's current schema in that it may remind him that potentially hazardous terrain is ahead. The activated schema may direct the pilot to explore terrain to the north. When the pilot views the mountain (i.e., samples the object), the schema again would be appropriately modified or attention would simply be redirected. Specifically, the schema would either continue directing the pilot's attention to the hazardous terrain, or if the terrain did not pose a threat, the schema might direct the pilot's attention to other aspects of the environment (e.g., a visual sampling of cockpit displays). As implied by its name, the perception/action cycle suggests that the process of information gathering is cyclical, and the beginning and end of the process are not specified. Therefore, this approach suggests that the process of achieving SA is relatively dynamic.

Adams, Tenney, and Pew (1995) explain SA in terms of the perception/action cycle, but unlike Endsley, they suggest that SA should be conceptualized as both a product and a process. In terms of the perception/action cycle, Adams et al. propose that SA as a product is the state of the currently activated schema, and as a process, SA is the current state of the entire perceptual cycle. In emergency situations, however, they suggest that a more elaborate model is necessary to adequately capture behavior. To explain such circumstances, they expand the perception/action cycle utilizing theory that was developed to understand how individuals comprehend written text. Adams et al. suggest that high-demand situations, such as emergencies are best represented by dividing the schema part of the model into two parts: explicit focus and implicit focus. Explicit focus is essentially equivalent to working memory; implicit focus is synonymous with the entire schema that is activated (where some of the schema is represented in explicit focus). They further suggest that long-term episodic
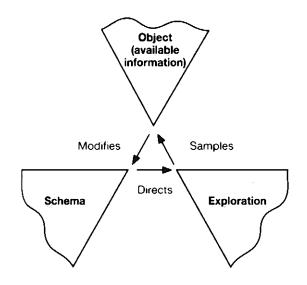

Figure 2. The Perception/Action Cycle taken from Adams, Tenney and Pew (1995).

memory and long-term semantic memory be included in the model. Adams et al. define long-term episodic memory as containing a thorough record of the schemata that have been constructed or activated over the course of a task, and they define semantic memory as containing general knowledge acquired over a lifetime.

There are at least two possible criticisms of Adams' et al. (1995) approach. First, much like the information-processing approach, they include many constructs in their model that are not well-understood (e.g., semantic memory, schemata). Second, their approach provides no suggestion as to how the product (i.e., the state of the active schema) or the process (i.e., the state of the perceptual cycle) of SA can be measured.

Smith and Hancock (1995) also utilize the perception/action cycle to conceptualize and define SA. However, they, at least, imply the manner in which SA should be measured. They define SA as "adaptive, externally directed consciousness" (p. 138). More specifically, they suggest that "adaption" is the process by which the operator uses both knowledge and behavior to achieve goals given the current circumstances and environmental constraints. The phrase "externally directed" suggests that the agent's goal is in the environment rather than in the agent's head, and consciousness is the portion of an agent's knowledge-generating behavior that may be manipulated intentionally.

Smith and Hancock (1995) expand the perception/action cycle by adding a novel element to the cycle: the "invariant." They place the invariant at the center of the cycle and suggest that it produces competent behavior by linking the object, the schema, and exploration, and that it ultimately defines SA. Identification of this invariant (i.e., SA) goes beyond performance, in that they suggest SA is the ability to produce competent performance by appropriately directing consciousness in a dynamic task environment. As a result, they *imply* that SA should not be measured by an evaluation of performance, per se, but rather it should be measured in light of both the competence of the operator (knowledge of goals, rules, etc.) and the current situation. Therefore, their approach is similar to Endsley's in that they acknowledge the importance of the current situation (i.e., dynamic aspects of the task). However, their approach differs somewhat from Endsley's approach in that competence may include a pilot's knowledge of some static elements of a task (e.g., FAA rules such as Instrument Flight Rules).

Although Smith and Hancock (1995) provide the idea of competent performance, their conceptualization is not without criticism. It is questionable whether competent performance, per se, is an adequate measure of SA, given that it may be demonstrated without the operator having SA (i.e., automated systems may be performing tasks, competent performance may be demonstrated purely by coincidence, etc.). Thus, it is more likely that competent performance is a necessary but not sufficient condition for SA.

2.2.3 SA Fused with Models of Decision Making

Crane (1992) claims that coining the term "situation awareness" has led to mixed results in terms of understanding the mechanisms responsible for it, and after reviewing existing cognitive literature, concludes that SA is not a unique psychological construct. Crane focuses on one of many concepts in the decision-making literature and asserts that SA is equivalent to *expertise*, a notion that is somewhat similar to Smith and Hancock's idea of competence (1995). Crane proposes that the decision-making literature is relevant because the behavior of experts has been extensively researched, and he contends that SA is demonstrated by expert-level performance. For example, Crane would simply suggest that a pilot who maneuvers to avoid terrain in an effortless, rapid, and error-free manner has SA. Like Smith and Hancock's notion of competence, one criticism of Crane's approach is that it is quite possible to demonstrate "expert-level" performance without having SA. The separation of SA and performance is an issue that

surfaces quite frequently and will be addressed later in more detail. An additional criticism of Crane's approach is that it also may be difficult to operationally define expert-level performance. While it may be relatively easy to determine if performance is rapid and error-free, it may be difficult to objectively assess if performance is "effortless." Specifically, the problems associated with mental workload measurement surface if SA is measured by degree of effort exerted.

Crane is not alone in his attempt to fuse SA with concepts that have been traditionally associated with judgment and decision making. Federico (1995) suggests that situation assessment may be defined as follows: sizing up the situation, understanding the situation, defining the problem, categorizing the circumstance, constructing a representation of the situation, making a mental model of the circumstance, mentally painting a picture of the situation, or creating an image of the circumstances. The overlap between the construct of SA and situation assessment should be clear from Federico's definition of situation assessment. For example, attainment of SA is often described as having an understanding of the situation or having a mental picture of the situation.

Several researchers apparently have recognized the overlap between the concept of SA and the idea of situation assessment. For example, Wickens, Gordon, and Liu (1998) use the terms SA and situation assessment interchangeably. Further, both Federico (1995) and Fracker (1988) use situation assessment to explain SA. In fact, Federico's research appeared in the special issue of *Human Factors* that was dedicated to SA, and he completely abandoned the term "situation awareness" in favor of "situation assessment." However, like the information-processing model and the perception/action cycle, situation assessment is often discussed in terms of poorly defined psychological constructs. For example, situation assessments are often theorized to be a result of schema-driven processing (e.g., Federico, 1995). That is, situation assessments are thought to be performed based on clusters of knowledge (i.e., schemata) that allow the pilot to categorize events. At present, the poor understanding of schemata raises questions regarding the utility of situation assessment as an alternative to SA.

2.2.4 SA as a Description of a Phenomenon

Flach (1995) suggests that SA should not be used to explain behavior, but it should merely be used as a descriptive label. He makes this proposition based on Underwood's (1957) categorization of psychological concepts. To explain the ideas of both Flach and Underwood, a hypothetical experiment is used, in

which an experimenter wishes to compare terrain avoidance when traditional cockpit displays are used with terrain avoidance when a novel display is added to the cockpit. Underwood suggests that psychological concepts range from Level 1 concepts to Level 5 concepts. A Level 1 concept represents the highest level concept because it does not require a conceptual leap from the objective data. These concepts refer only to the nature of the independent variable, per se. No implication is made regarding the behavior of the subject. The researcher, then, would discuss the hypothetical results in terms of the effects of *Display Type* on performance. On the other end of the continuum, a Level 5 concept represents the lowest level concept because it requires several conceptual leaps from the objective data. These concepts are used very rarely because they are combinations of groups of constructs. As such, there is difficulty in creating an example for the hypothetical experiment.

Flach contends the distinction between Level 2 and Level 3 concepts is particularly relevant to SA. Underwood refers to Level 2 concepts as *phenomenon naming*, and a concept is categorized at Level 2 when the phenomenon is identified but causal processes or conditions are not implied beyond the operations used to define the phenomenon. For example, if pilots maneuver away from hazardous terrain more quickly with a proposed display than with the use of traditional sources of information, then the quicker maneuver might be termed situation awareness. However, the cause for the change in behavior is thought to be the proposed display and no other mechanism or process.

Underwood refers to Level 3 concepts as *causal naming*, and a concept is categorized at Level 3 when a name is applied to a hypothetical process, state, or capacity as a *cause* of observations. In the aforementioned example, the term situation awareness would not be used to mean a change in behavior. Instead, it would be used to describe an intervening process or state, where the proposed display led to attainment of situation awareness (the intervening process), which led to quicker maneuvers.

Flach notes that two problems result when SA is treated as a Level 3 concept. First, he suggests that, if SA is treated as a Level 3 concept, empirical testing is impossible because the construct is conceptualized as being unobservable. Second, as Underwood (1957) initially suggested, Level 3 concepts inevitably lead to circular reasoning. Flach and others (e.g., Baxter & Bass, 1998) recognize the presence of such circularity in the current SA literature. For example, some posit that SA is lost because an operator responds inappro-

priately, and at the same time, some suggest that an operator responds inappropriately because SA was lost. Therefore, as a Level 3 concept, SA theoretically cannot be measured, nor can it be used as an explanatory tool without engaging in circular logic.

Despite criticisms of the manner in which SA is often addressed as a psychological construct, Flach (1995) believes SA is important and suggests that it has value in that it "bounds" the problem. By bounding a problem, SA assists the researcher in two senses. First, SA aids researchers in focusing on relevant variables by requiring researchers to recognize both the objective task situation *and* the mental awareness of the operator. Second, bounding a problem allows researchers to identify similar events that can be placed in categories. Flach proposes that placing events (e.g., erroneous actions) into a category called "loss of SA" might allow the researcher to identify a common feature of these events. For example, the researcher might recognize that a common feature of erroneous actions was the existence of a display with multiple modes (e.g., the flight management system). In such a case, the researcher would be allowed to create *testable* hypotheses regarding the causes of erroneous actions. The researcher might form a hypothesis regarding the effects of modes on human actions. In such a case, both the variable to be manipulated (i.e., modes) and the variable to be measured (i.e., some human action) can be operationally defined because both are directly observable.

In summary, Flach suggests that, although the construct of SA is useful in categorizing events (i.e., as a Level 2 concept), it does not provide utility as a intervening variable (i.e., as a Level 3 concept). One possible criticism of his approach is that it raises the question as to whether or not the term SA is needed at all. Certainly, events could be categorized without the use of the term SA and the issues that surround them.

2.2.5 Summary of Approaches Used to Define and Explain SA

Four qualitatively different approaches were presented and briefly discussed. Table 1 presents a summary of each approach, along with potential problems and criticisms.

2.3 Measures Used to Assess SA

Because there are very different ways to conceptualize SA, there is little surprise in that several somewhat divergent methods are used in assessing SA. This section provides a review of various dependent measures. Some of the measures reviewed here are specifically associated with the theoretical approaches

Table 1. Summary of approaches used to discuss SA and potential problems/criticisms with each.

Approach	Summary	Potential Problems/Criticisms
Use of Information-Processing Models	Traditional, psychological constructs are discussed in terms of their impact on SA (e.g., attention, long-term memory, perception, and automaticity).	• Many of these psychological constructs are themselves not well-understood. • This approach may cause one to conceptualize SA as a static end-state rather than a dynamic process.
Use of Perception-Action Cycle	SA is discussed in terms of the cyclical process of perceiving information in the environment, utilizing pre-existing knowledge structures, and exploring the environment.	• This approach also utilizes several psychological constructs that are themselves not well-understood (e.g., schemata, exploration). • The measurement of SA implied by this approach is unclear at best.
Decision-Making Models	SA is demonstrated by expert-level performance, and SA is equivalent to situation assessment.	• Expert-level performance is a necessary condition for SA, but it is probably *not* a sufficient condition for SA. • There may be difficulties in operationally defining expert-level performance • Models of situation assessment emphasize one psychological construct in particular that is not well-understood (i.e., the schema).
Phenomenon Description	SA should be used as a tool for categorizing situations (i.e., as a Level 2 construct) but should not be used as a psychological construct implying cause and effect (i.e., as a Level 3 construct).	• Why SA is needed to categorize events is questionable.

reviewed above. An understanding of how SA has been operationally defined for empirical assessment should aid both in the theoretical understanding of the construct and make explicit the relative advantages and limitations these measures have in theoretical and practical applications.

Most researchers (e.g., Fracker, 1991; Sarter & Woods, 1995; Vidulich, 1992; Wickens, 1992b) divide the measures of SA into three broad categories: (a) explicit, (b) implicit, and (c) subjective measures of SA. However, these measures also contain subcategories. Table 2 shows the categories and subcategories of SA measures that will be discussed. The potential advantages and disadvantages associated with each measure are discussed and where applicable, examples of each measure are provided, along with relevant research.

First, it is necessary to discuss some general measurement issues and terms as they apply to psychological research. One important issue is validity, which addresses the extent to which a dependent measure actually assesses what it is intended to measure. Four types of validity are relevant for the purposes of this paper:

- *Face Validity* refers to the degree to which a measure intuitively appears to measure the psychological construct in question. Research participants and end-users easily accept measures with face validity. However, readily accepting measures with face validity can be problematic because they do not necessarily measure what they appear to measure.
- *Construct Validity* refers to the degree to which a measure actually assesses the construct that it is intended to assess.
- *Predictive Validity* is the degree to which a measure can predict behavior in real-world settings or tasks.
- *Concurrent Validity* refers to the degree to which a new measure correlates with other existing measures.

In addition to validity, five criteria have been suggested as important for mental workload indices (cf., O'Donnell & Eggemeier, 1986; Wickens, 1992a).

To the extent that there are interesting and important parallels between SA and mental workload, measures of SA also should be critiqued using these criteria. Below, the five criteria are defined, and examples are presented to demonstrate their relevance to measures of SA:

- *Sensitivity* refers to the degree to which a measure distinguishes between differing conditions or states. For example, a sensitive technique would distinguish between levels of SA when the experimenter varied the information available to the participant.
- *Selectivity* is the degree to which a measure is sensitive only to changes in the construct of interest. For example, a measure of SA should be sensitive only to changes in SA and should not be affected by changes in mental workload.
- *Diagnosticity* is the degree to which a measure not only identifies changes but identifies the cause of any variation. In other words, a diagnostic measure would assist in identifying why there were changes in SA.
- *Obtrusiveness* refers to the degree to which a measure interferes with the primary task. For example, a measure of SA should not interfere with piloting duties.
- *Reliability and Bandwidth* refer to the degree to which a measure is consistent and the degree to which a measure can rapidly provide a reliable assessment. For example, if a pilot were tested twice under identical circumstances with an identical understanding of the circumstances (although such a case is quite unlikely), a reliable measure would suggest the pilot had the same amount of SA in both cases. In addition, it is important that a measure of SA can be reliable in dynamic situations where a pilot might be tested several times throughout a flight.

Clearly, these five criteria overlap to some extent with the validity issues, and each of the aforementioned eight issues (validity and measurement criteria) is addressed when or if appropriate.

Table 2. Categories and subcategories of SA measurement.

Categories	Subcategories
Explicit Measures	• Retrospective Measures • Concurrent Measures • Measures Utilizing the Freeze Technique
Implicit Measures	• Global Measures • External Task Measures • Embedded Task Measures
Subjective Measures	• Direct Self-Ratings • Comparative Self-Ratings • Observer Ratings

2.3.1 Explicit Measures of SA

Explicit measures require people to self-report material in memory (Fracker, 1991). For example, pilots may be asked to recall variables associated with the most recent state of the aircraft. As such, the measure might assess whether the pilot was able to correctly recall the aircraft's most recent altitude, speed, location, etc. Several researchers (Endsley, 1995a; Fracker, 1991; Wickens, 1992b) agree that these measures have high construct validity because the data collected is consistent with most theories of SA. In addition, Endsley (1995a) suggests that explicit measures are objective because the data collected can be objectively compared with the true state of affairs (i.e., a normative model of the domain). While Endsley claims they are objective, others suggest that explicit measures are subjective (e.g., Fracker, 1991). The measures may be considered *subjective* because the data are acquired via self-reports rather than some assessment of observable behavior. Therefore, such assessments of SA are likely to be tainted by a participant's bias or preconceptions. Another potential problem with explicit measures is that a normative model of a domain (like aviation) is problematic because the task environment is dynamic (i.e., rapidly changing) and complex enough that it is difficult to understand completely outside of a laboratory setting. Endsley (1995a) and Fracker (1991) suggest that explicit measures can be subcategorized into three types: (a) retrospective measures, (b) concurrent measures, and (c) measures utilizing the freeze technique.

Retrospective measures are utilized *after* a task is completed. These measures require participants to either recall specific events or describe decisions made during an experimental scenario or simulation. Endsley (1995a) suggests that these measures are useful in that they allow participants ample time to respond to questions, but she cautions that the measures may only reliably capture SA for the very end of the task. Fracker (1991) suggests that these measures also may not reveal what actually happened during the task but rather may reveal a participant's retrospective (off-line) *inference* as to what happened. For example, responses could be subject to false recollections or the measure could reflect spuriously high SA because the participant was able to conduct mental operations not possible while actually performing the task.

Concurrent measures, such as verbal protocols, are used during the course of a task. Unlike retrospective measures, these measures assess SA on-line (i.e., while the participant is performing the task). However, Endsley (1995a) suggests that these measures may

have the potential to increase mental workload due to the nature of the measurement task, per se. Both she and Fracker (1991) suggest that such measures may cause participants to act unnaturally by having them attend to information to which they would not normally attend.

Two different types of concurrent measures have been proposed. Verbal protocols are one type (Metalis, 1993; Vidulich, 1992) that essentially requires participants to think aloud. Metalis cautions that verbal protocols may be too obtrusive. A second type of concurrent measure involves the utilization of a confederate who is placed in the task environment and discusses the task with the participant (cf., Metalis, 1993; Sarter & Woods, 1991). Thus, the confederate can probe the participant to determine if the participant is aware of various task-relevant pieces of information. Like all concurrent methods, the use of a confederate may cause the participant to act unnaturally, resulting in the "on-stage" effect in which the participant behaves differently due to the mere presence of a confederate. In addition, there is potential for the confederate to produce systematic bias through verbal as well as non-verbal cues (i.e., "leading the witness"). Metalis suggests that, although using a confederate is likely to be less obtrusive than using verbal protocols, such probing does not completely alleviate the problem of artificially increasing mental workload.

Measures utilizing the **freeze technique** are explicit measures of SA that fall somewhere between retrospective and concurrent measures, since the participant is asked questions mid-task. When using the freeze technique, a simulation is frozen at a particular point in time (usually randomly determined), and the participant is deprived of all task-relevant information (e.g., displays are blanked). At the time of the freeze, the participant is asked to answer task-relevant questions. Endsley (1995a) suggests that these measures are useful because the time-related problems associated with retrospective measures are resolved, and the mental workload issues associated with concurrent measures are eliminated. In addition, she suggests that these measures are practical because, after the appropriate SA requirements are identified, they can be used in any task environment. Specifically, the freeze technique may be used in any domain after the researcher identifies the information of which an operator *should* be aware.

Several shortcomings are associated with measures utilizing the freeze technique. First, Fracker (1991) voices a particular concern regarding the temporal

limits of working memory. Without rehearsal, material is accurately retained in working memory for approximately two seconds, and therefore, Fracker warns that questions asked beyond two seconds following the freeze may be subject to false recollections. Second, like concurrent measures, the freeze technique creates an unnatural environment. Sarter and Woods (1995) contend that the freeze technique is unnatural in that the questions posed may serve as a retrieval cue. They also suggest a participant's responses only divulge what knowledge he/she can demonstrate when asked by the researcher rather than what information the *participant* would have thought was important. Third, Selcon, Taylor, and Koritsas (1991) caution that utilization of the freeze technique involves the assumption that the operator's assessments of the task environment are stored in memory and are accessible. However, in practice, some knowledge may be used without awareness and, therefore, may not be reflected in this technique. Finally, Pritchett, Hansman, and Johnson (1995) question the predictive validity of the freeze technique by suggesting that such techniques only allow researchers to speculate regarding the user's actions given his/her knowledge state. In other words, even if a measure utilizing the freeze technique suggests a participant has high SA, the measure does not provide a way of knowing how the participant would, in fact, *perform*.

A well-known measure utilizing the freeze technique is the Situation Awareness Global Assessment Technique (SAGAT). This measure, developed by Endsley (1995a) specifically for air-to-air tactical combat, is a computerized version of the freeze technique. The SAGAT freezes the simulation at random points in time and queries the pilot with a question chosen randomly from a pre-defined bank of task-relevant questions. As with all measures utilizing the freeze technique, the limits of working memory could pose a problem. However, Endsley (1995a) found that (a) *accuracy* on SAGAT questions was not affected by the amount of time that elapsed after the simulation freeze, and (b) *task performance* was not affected by either the duration or the frequency of freezes. Thus, she concluded that the SAGAT was neither obtrusive nor affected by the limits of working memory. However, it should be noted that the SAGAT generates binomial data (i.e., responses are scored as either correct or incorrect) which, for statistical reasons, requires more data than might be required with other measures.

2.3.2 Implicit Measures of SA

Implicit measures of SA utilize task performance to infer SA. For example, SA might be assessed by computing the deviation of current aircraft heading from the assigned heading. Therefore, implicit measures are different than other types of SA assessments in that the awareness of operators is not assessed directly but is merely implied by their performance. Advantages to using such measures are that they are objective, unobtrusive, and relatively easy to use (Endsley, 1995a; Fracker, 1991; Metalis, 1993). Pritchett et al. (1995) suggest three additional strengths of implicit measures. Specifically, these measures have high predictive validity because they provide information regarding: (a) when and how operators react to real situations where time pressures are present, (b) restrictions on operator behavior that result from training and/or standard procedures, and (c) the operator's confidence in the reliability of information sources (i.e., their willingness to act upon information).

A potential shortcoming of implicit measures is that poor performance may be a result of something other than low SA. For example, a pilot could have high SA but might not perform well due to other factors such as poor response execution. In fact, Venturino, Hamilton, and Dvorchak (1989) conducted a study in which they utilized an implicit measure of SA (i.e., a performance measure) that they called the Pilot Performance Index (PPI). The PPI was the ratio between the number of enemies killed and the number of friendlies killed. Venturino et al. also collected subjective self-ratings of SA. As would be expected, they found that pilots who rated their SA as low had low PPI scores, and pilots who rated their SA as average had average PPI scores. However, they found that pilots who rated their SA as high had PPI scores that were inconsistent, which demonstrates a divergence of performance (i.e., implicit measures of SA) and self-rated SA. Therefore, their study suggests that high SA may be a necessary but not sufficient condition for good performance. The existence of implicit measures also raises the question as to whether the construct of SA is needed at all. That is, if performance is ultimately measured, the utility of SA is suspect.

Endsley (1995a) divides implicit measures into three categories: (a) global measures, (b) external task measures, and (c) embedded task measures. Implicit global measures of SA are simply measures of overall task performance and have the same advantages and disadvantages associated with them as implicit measures in general.

External task measures require the removal of information from a display or the alteration of the information on a display (Endsley, 1995a). The time it takes for the operator to react to the removal or alteration of information is recorded. These measures tend to be too obtrusive and may cause participants to act unnaturally. Endsley also suggests that these measures can be misleading. For example, if the researcher makes an aircraft disappear from an air traffic controller's screen, the controller may know it is gone, but the controller may not demonstrate this knowledge. In such a case, the controller might want to complete other tasks before addressing the problem of the disappearing aircraft.

Embedded task measures assess performance on sub-tasks. For example, Harwood, Barnett, and Wickens (1988) suggest subtasks such as distance estimations, target localizations, or attempted reorientation after being displaced from the flight path during pilot-in-the-loop flight simulations.

Embedded task measures may be helpful in obtaining information regarding the amount of SA a particular display provides about a parameter. However, Endsley (1989) suggests that high SA in one area may result in low SA in another area, and therefore, embedded task measures provide the researcher only partial information regarding SA. In addition, some (Endsley, 1995a; Fracker, 1991) propose that it may sometimes be difficult to ascertain which measure to use for a given situation. As an example, Fracker questions whether an embedded task measure could be developed that truly measures a pilot's awareness of altitude.

Despite their popularity in theoretical reviews, embedded task measures have been used infrequently to measure SA. In order to improve implicit measures, Sarter and Woods (1995) suggest that post-trial debriefings should be used to understand the causes of behavior. In addition, Pritchett et al. (1995) provide three suggestions for the use of implicit measures. They suggest that (a) situations be used in which the participant is forced to engage in actions that are measurable, (b) situations be utilized for which standard procedures mandate a particular response to easily make inferences regarding SA, and (c) situations in which a pilot has little confidence in the information or feels a particular behavior might violate standard procedures should not be ignored.

2.3.3 Subjective Measures of SA

Subjective measures are distinct in that SA is measured either by self-assessment ratings or by the assessments of an observer. In other words, these measures are based solely on the opinion of the participant or the observer. For example, on a given scenario or task, a participant might be asked to use a Likert-type scale ranging from "1" to "7" in rating the amount of SA experienced. Subjective measures of SA are useful due to their ease of implementation, and Metalis (1993) suggests that subjective measures also are practical because they may be used both in simulations and in the actual task environment (e.g., in flight). In addition, these measures are relatively inexpensive to implement. However, Fracker (1991) warns that subjective measures are limited in that they cannot be compared across raters. For example, on a Likert-type scale ranging from "1" to "7," a rating of "4" by one rater may mean something very different than a rating of "4" by another rater.

A taxonomy of three major classes of subjective measures has been developed (Endsley, 1995a; Fracker, 1991) that includes: (a) direct self-ratings, (b) comparative self-ratings, and (c) observer ratings. Direct self-ratings require the participant to rate his/her own SA, as in the example where the participant might be asked to rate the amount of SA experienced on a scale ranging from "1" to "7." Such ratings may be useful in that, theoretically, the participant knows best as to what he or she knows or does not know. However, Endsley warns that participants may have difficulty assessing their own SA during a task, since they are not able to compare their knowledge with the true state of affairs. Thus, the researcher may opt to collect post-task ratings. After the task, the researcher can provide participants with information regarding the true state of affairs, and they can compare the knowledge they had during the task with the true state of affairs. However, Endsley suggests that participants' ratings may be affected by their *performance* on the trial. For example, if a pilot completes a flight successfully, the pilot might assume that SA was quite high when, in fact, it was not. Endsley also warns that, when gathered at the end of the task, direct self-ratings may be prone to rationalizations and overgeneralizations by the participants. Sarter and Woods (1995) also criticize direct self-ratings by contending that they ignore the process of achieving SA and only measure SA as a product.

The selectivity of direct self-ratings can be questioned in that these ratings may actually measure an operator's *confidence* regarding SA rather than SA, per

se. In fact, Endsley (1998) found direct self-ratings to be correlated with participants ratings of both the sufficiency of their SA and their confidence level regarding their SA. However, she notes that, even if direct self-ratings only measure confidence in SA, these measures may be useful. Some behaviors may depend on how aware a person *believes* himself or herself to be. For example, if a pilot does not believe he has high SA, he may choose to scan the instruments a second time rather than make a control action.

The Situation Awareness Rating Technique (SART) is a direct self-rating measure of SA that is more complex than a simple Likert scale. Taylor (1989) developed the SART by eliciting knowledge from pilots and aircrew. Through statistical techniques (i.e., frequencies, principal component analyses, and inter-correlation clustering), he created the 10-D SART, which consists of ten dimensions used to measure SA: (a) Instability of Situation, (b) Complexity of Situation, (c) Variability of Situation, (d) Arousal of Situation, (e) Concentration of Attention, (f) Division of Attention, (g) Spare Mental Capacity, (h) Information Quantity, (i) Information Quality, and (j) Familiarity. Taylor found that these ten dimensions could be further grouped into three overall dimensions, which were named the 3D-SART: (a) *Demands on Attentional Resources* — a combination of Instability of Situation, Complexity of Situation, and Variability of Situation; (b) *Supply of Attentional Resources* — a combination of Arousal of Situation, Concentration of Attention, Division of Attention, and Spare Mental Capacity; and (c) *Understanding of Situation* — a combination of Information Quantity, Information Quality, and Familiarity.

Taylor suggests choosing either (a) a Likert scale, (b) categories (e.g., "low" vs. "high"), or (c) pairwise comparisons as a method of implementing either the 10-D SART or the 3-D SART. One dimension of the 10-D SART, "information quality," will be used to illustrate each of these three options. To use a Likert scale, a pilot would simply be asked to rate a design on its "information quality," where a rating of "1" would represent low information quality and a rating of "7" would represent high information quality. If categories were used, a pilot would simply be asked to rate a design on its "information quality," where a rating of "low" would represent low information quality, and a rating of "high" would represent high information quality. Finally, if pairwise comparisons were used, a pilot would be asked to report whether Display X had higher or lower information quality than Display Y;

Display Y had higher or lower information quality than Display Z; Display X had higher or lower information quality than Display Z; and so on.

SART provides several advantages. First, Selcon et al. (1991) suggest that SART is useful because the scale was developed utilizing aircrew knowledge. Second, Selcon and Taylor (1989) demonstrated that the 3D-SART, which is easier to implement, captures the same information that 10-D SART captures. Specifically, they found that the ten dimensions grouped on the three overall dimensions in a manner similar to that of the original study (i.e., Taylor, 1989). Finally, SART appears to be a relatively sensitive measure.

Endsley (1998) found the SART to be more sensitive than performance measures. Specifically, she found that SART ratings of SA were significantly higher when participants were given an enhanced display, but only one of two performance measures was sensitive to display quality. Selcon and Taylor (1989) found SART to be more sensitive than overall ratings of SA (i.e., where only one number was used to quantify an operator's SA). Both the 3-D SART and the 10-D SART were sensitive to increases in workload, while an overall subjective rating of SA was not sensitive to such increases.

Vidulich (1992) also found the SART to be sensitive in a study that examined workload and expertise. He defined mental workload as the number of objects the participants monitored. In addition, he loosely manipulated expertise by having the "experts" monitor objects that moved in an orderly fashion and had the "non-experts" monitor objects that moved in a random fashion. Consistent with the findings of Selcon and Taylor (1989), Vidulich found that the ratings on SART sub-scales were consistent with the experimental manipulations, and that the sensitivity of SART surpassed a single-scale rating of overall SA.

To summarize research on SART, one study (Endsley, 1998) found SART to be more sensitive than performance measures, and two studies (i.e., Selcon & Taylor, 1989; Vidulich, 1992) demonstrated that SART was more sensitive than an overall rating of SA. However, it should be noted that in a later study, Selcon et al. (1991) found that an overall rating of SA was sensitive to differences in experience while 3-D SART dimensions were not. Therefore, whether SART is a more sensitive measure than an overall rating of SA remains unclear.

The selectivity of SART has been questioned in terms of whether the dimensions of SART measure SA or mental workload (Endsley, 1995a). To address this

question, Selcon et al. (1991) compared the SART with the NASA Task Load Index (NASA TLX). Like SART, the NASA TLX requires participants to rate themselves on several task dimensions, but the NASA TLX was developed to be a measure of mental workload, rather than SA. Selcon et al. compared the NASA TLX with SART by asking pilots to view a simulation and act as if they were flying the mission. The researchers varied the demand of the simulations (low, medium, and high) and whether or not auditory dubbing was present. The dubbing condition, in which auditory information was redundant with visual information, was compared with a condition that included only visual information. Selcon et al. also divided participants by their level of expertise: pilots were either "inexperienced" (60-400 flight hours) or "experienced" (1100-5500 flight hours). Three response measures were used: the 10-D SART, the 3-D SART, and the NASA TLX.

The results indicated that all three response measures were sensitive to differences in levels of demand, but none of the scales were sensitive to differences in auditory dubbing. The most relevant finding was that the NASA TLX produced no differences due to pilot experience, but the 10-D SART scale was somewhat sensitive to experience. Specifically, ratings on the "Familiarity" dimension were different as a function of pilot experience. In addition, for both ratings of "Concentration of Attention" and "Spare Mental Capacity," the effects of experience depended on the level of demand (i.e., there was an interaction between pilot experience and demand for both of these subscales). Thus, the construct validity of SART is open for debate, but the Selcon et al. study suggests SART is selective because it does measure something other than mental workload.

Comparative self-ratings require the participant to *compare* self-assessed SA from one trial to another. Fracker (1991) suggests that such measures are useful because they encourage *within*-participant consistency. However, he also contends that in some situations, the number of comparisons required of the participant can become quite large, and in such cases, these measures may not be very practical.

One example of a comparative self-rating is the SA-SWORD (Situation Awareness-Subjective Workload Dominance Technique) (cf., Vidulich & Hughes, 1991). The SA-SWORD is a modification of the SWORD (Subjective Workload Dominance Technique), which is used in assessing mental workload.

The SA-SWORD is a comparative self-rating tool that requires participants to subjectively rate experienced SA between all possible pairs of potential designs (e.g., comparing various cockpit designs).

There are two potential shortcomings of the SA-SWORD. First, it can only be used in contexts where a within-participants experimental design is used. For example, in the many situations where it is impractical to have participants view more than one potential display, the SA-SWORD, like all comparative self-rating tools, is not an option. In addition, the SA-SWORD, like all subjective measures, does not ensure *between*-participant consistency in ratings of SA. For example, Vidulich & Hughes (1991) found that about half of the participants rated their SA by gauging the amount of information to which they attended, while the other half of the participants rated their SA by gauging the amount of information they thought they had overlooked.

When **observer ratings** are used, an unbiased, neutral expert is asked to observe a participant perform a task and to rate the participant's level of SA. Endsley (1995a) suggests these measures have some utility because, unlike both types of self-ratings, the raters (i.e., observers in this case) *do* have information regarding the true state of affairs. However, a potential drawback of these measures is that the observer cannot know the operator's internal understanding of the situation. For example, Endsley describes a situation in which the operator could be cognizant of a piece of information but does not provide any observable evidence of this knowledge.

2.3.4 Recommendations Regarding the Measurement of SA

Despite their shortcomings, none of the aforementioned measures has been abandoned by human factors researchers and practitioners. The utility, however limited, of the measures discussed here must be recognized until better measures of SA are developed. When measuring SA, there should be an attempt to adhere to the following guidelines: (a) when possible, several measures of SA should be utilized to ensure concurrent validity (Harwood et. al, 1988); (b) scenarios should be lengthy enough to allow participants to become comfortable in the test environment (Sarter & Woods, 1991, p. 54); and (c) as discussed in a previous section, caution needs to be exercised in suggesting SA is the direct cause of behavioral changes (Flach, 1995).

3.0 Situation Awareness as It Relates to Surveillance

From the previous section, it should be apparent that there are many unresolved issues surrounding SA; it is difficult to define, explain, and measure. As such, several researchers have suggested partitioning the construct. For example, Harwood et al. (1988) suggest that SA might consist of such components as spatial awareness, identity awareness, and temporal awareness. Regal, Rogers, and Boucek (1987) regard SA as a broad type of knowledge but also suggest that SA should be examined in terms of its components, such as awareness of environment, awareness of aircraft performance, aircraft systems awareness, and crew awareness.

Several researchers object to the notion of partitioning SA into components. For example, Sarter and Woods (1991) suggest that studies examining the components of SA do not assist in understanding SA as the big picture, and research has suggested that there is some validity to their concern. For example, Entin (1998) used both a global, high-level measure of SA and a more detailed measure of SA. The high-level measure consisted of general questions about the situation (e.g., a question might probe the pilot about the limitations created by relevant geography). The more detailed measure consisted of questions about particular elements of the situation (e.g., a question might inquire about the specific location of the pilot's aircraft). Entin found that the global and detailed measures were only "marginally" correlated early in a mission, and the correlation was essentially nonexistent by the later stages of the mission. Such a finding suggests that overall SA and SA of particular task components may diverge, and care must be taken when components of a task are examined in isolation.

Endsley (1989) also objects to the partitioning of SA because high SA in one area may result in low SA in another area. For example, obtaining awareness of out-the-window information (e.g., weather) might hinder a pilot's awareness of information in the cockpit (e.g., the status of an on-board system). Shively and Goodman (1994) provide support for this concern because they found that display enhancements increased awareness of three task components, had no effect on one task component, and decreased awareness of another task component. These results suggest that SA of particular task components may in fact diverge, and again suggest that care must be taken when components of a task are examined in isolation. Thus, it appears that concerns regarding the partitioning of SA may be well-founded, but it may be the very global nature of the construct that makes partitioning necessary. Because SA, as a global construct, is inherently difficult to define and measure, these problems are magnified in complex tasks like piloting. As was mentioned earlier, reliability and bandwidth are relevant criteria of measurement techniques for both mental workload and SA. Obtaining a reliable estimate rapidly enough so transient changes may be assessed is important. A pilot has to organize numerous activities in a timely manner. The multiple tasks that must be timeshared in a dynamic environment, often with severe temporal constraints, make piloting an aircraft (individually or as part of a group) a very *dynamic* task. Thus, it is doubtful that global measures attempting to capture SA as a *static* or finite product would be able to adequately meet the criteria of bandwidth and reliability.

One goal of the present paper is to examine SA as it relates to *surveillance activities* in the air carrier environment. Surveillance activities are those activities that are continually performed by pilots to gain awareness of potential obstacles and hazards in the external world. Such obstacles include, but are not limited to, other aircraft, terrain, and weather (e.g., turbulence). Surveillance does not require that a pilot be cognizant of all information in the task environment. Rather, to perform surveillance activities well, the pilot need only have high awareness in several specific areas. In this section, the components of SA that are relevant to surveillance activities are identified and defined. In addition, relevant human factors research regarding these components is reviewed.

3.1 Components of SA that Relate to Surveillance

Four components of SA that appear to relate to surveillance are discussed below: (a) environment awareness, (b) spatial awareness, (c) temporal awareness, and (d) navigation awareness. Given the previous discussion of surveillance, it should be clear that surveillance activities would require awareness beyond these four components (e.g., traffic awareness, weather awareness, etc.). To date, however, the literature contains only these four components that appear to be relevant to surveillance activities.

Regal et al. (1987) do not explicitly define **environment awareness**. However, they provide a list that demonstrates the types of knowledge necessary for the commercial pilot to gain awareness of the environment. They suggest that the pilot must be knowledgeable of: (a) weather, (b) windshear, (c) other aircraft, (d) airport conditions, and (e) icing.

Another suggested component of SA that appears relevant is **spatial awareness**, and it includes knowledge of (a) attitude, (b) location relative to terrain, (c) waypoints, navaids, (d) flightpath vector, and (e) speed (Regal et al., 1987). Regal et al. appear to distinguish environment awareness from spatial awareness by suggesting environmental awareness is related to circumstances that occur in the external environment, whereas spatial awareness is related to egocentric spatial orientation. However, Harwood et al. suggest that spatial awareness is achieved when the pilot has knowledge of ownship's location and the spatial relation between relevant objects. Harwood et al.'s definition of spatial awareness is similar to environment awareness as described by Regal et al. At best, the distinction between environment awareness and spatial awareness appears to be unclear.

Harwood et al. define another component of SA, **temporal awareness**, as the pilot's knowledge of events as a mission evolves. Additionally, Wickens (1992b) suggests that temporal awareness is achieved when the pilot knows how much time remains before deadlines.

Several researchers have **suggested navigation** awareness as an important component of SA. For example, Aretz (1991) suggests that navigation awareness is the pilot's ability to answer the question, "Am I where I should be in the world?" More simply, Wickens (1992b) suggests that navigation awareness is achieved when the pilot can answer the following question appropriately: "Where am I with regard to other aircraft, the terrain, and local weather conditions?" Although navigation awareness *is not* easily distinguished from the three components discussed above, it probably includes a combination of spatial awareness and temporal awareness as they apply to activities specifically associated with wayfinding.

3.2 Research Examining the Relevant Components of SA

Three studies specifically address the components of SA related to surveillance. One of these studies examines spatial awareness, and the other two address navigation awareness.

3.2.1 Research Examining Spatial Awareness

Fracker (1989) had participants engage in a simulated air battle by having them view a display on which seven aircraft appeared. Participants controlled one of these aircraft via joystick. Fracker manipulated the identity of the aircraft (i.e., whether they were friend, foe, or neutral) and the number of enemy aircraft (while keeping the total number of aircraft constant). Aircraft identities changed randomly and at random

time intervals. Utilizing the freeze technique, Fracker asked participants to identify (a) the spatial location of one aircraft and (b) the identity of another aircraft. Aircraft were chosen randomly for the test questions.

Although Fracker (1989) also examined another kind of awareness (i.e., knowledge of whether an aircraft was friend, foe, or neutral), what is relevant here is the assessment of spatial awareness, which was defined in terms of the Euclidian deviation of the reported location of an aircraft from the actual location of an aircraft. Fracker found that the spatial awareness of neutrals did not increase when there were less neutrals and concludes that participants coped with increases in demand (i.e., having more enemy aircraft) by sacrificing the attention paid to the low-priority neutrals, rather than sacrificing the attention paid to the higher-priority friendlies. A more general finding was that spatial awareness was highest for those aircraft that might impede task success (i.e., enemy aircraft), somewhat poorer for friendly aircraft, and worst for aircraft that had the least impact on task success (i.e., neutral aircraft). The two findings support a model of limited attentional resources and suggest that components of a task receive attention based on their importance to task success. In other words, spatial awareness of information depended on how essential the information was to the task.

3.2.2 Research Examining Navigation Awareness

Andre, Wickens, Moorman, and Boschelli (1991) investigated the effects of particular displays on navigation awareness. They presented participants with either a planar inside-out display (i.e., a two-dimensional representation with a stationary aircraft), a planar outside-in display (i.e., a two-dimensional representation with a stationary environment), or a perspective outside-in display (i.e., a two-dimensional rendering of three-dimensional space with a stationary environment). Navigation awareness was assessed with four different measures. Two of the measures—the number of pre-determined waypoints participants reached and the accuracy with which participants initiated the appropriate turn after a forced disorientation—were used to represent tasks in which depth and distance judgments were crucial. The other two measures—the proportion of time participants spent controlling pitch and roll simultaneously and the delay between initiation of vertical and lateral control after disorientation—were used to represent cases in which the pilot must integrate tasks. The results suggested that the planar outside-in displays produced the highest navigation awareness when depth

and distance judgment was crucial, while the perspective displays supported processing when integration was necessary.

Aretz (1991) also examined navigation awareness by investigating (a) the importance of mental rotation and triangulation during navigation, (b) the allocation of attentional resources during navigation, and (c) the effectiveness of various map displays. When participants were given questions that required the use of a map (i.e., questions with a world-centered frame), rather than the use of the forward-field-of-view (i.e., an ego-centered frame), response time tended to increase as the aircraft's heading deviated from north. This finding suggests that, to achieve optimal navigation awareness, reference frames must be cognitively aligned. In a dual-task situation, participants appeared to shift from a mental-rotation strategy to a reversal strategy (i.e., saying to themselves "left equals right"), and response time for course changes increased linearly as heading moved away from zero. However, when participants were asked to answer questions *and* simultaneously control the aircraft, the linear trend in response time disappeared. Specifically, given the aforementioned linear trend, participants reacted quicker than would be expected at a 180-degree heading.

Aretz suggested that navigation and flight control compete for limited spatial processing resources. Therefore, he explained this second finding by suggesting that, to free some of the limited, spatial-processing resources, participants used an alternative strategy (in this case the reversal strategy) when available. Finally, Aretz found differences between map displays. Track-up maps resulted in shorter response times to questions regarding course changes in general. However, north-up maps resulted in the identification of more landmarks when participants were questioned regarding the necessary course change for a specified position that was not in their forward-field-of-view. Aretz concluded that the designer must consider what reference frame a navigation task requires before a particular map display is chosen.

4.0 Summary and Implications of Situation Awareness Literature

There have been numerous attempts at developing both adequate definitions and formal models of SA. None of the more widely accepted approaches to defining and explaining SA are without flaws. At the same time, numerous techniques have been suggested for the assessment of SA, and each of these techniques have relative strengths and weakness associated with them. In short, SA, as a formal psychological construct, is both difficult to define and difficult to measure.

Ten years ago, Wickens (1992a) suggested that the Federal Aviation Administration would soon be forced to adopt a mental workload metric as part of the aircraft certification process. Although workload is currently considered in the certification process, present practices require only a cursory evaluation of mental workload by domain experts. Specifically, aircraft are put through an extensive flight test program with FAA pilots and designated engineering representatives. These pilots are asked to fly the aircraft in both normal and abnormal conditions. The mental workload assessment is not one of the numerous formal methods of assessing mental workload. Instead, the assessment is based on the non-scientific opinions of FAA pilots and designated engineering representatives. This circumstance illustrates that years of laboratory research and theory development does not always translate into operational and regulatory consequences.

SA probably will be the focus of future laboratory research with hopes of developing an adequate theory and measurement technique. However, given the parallels of SA and mental workload, the fate of SA probably will be similar to the fate of mental workload. Despite its face validity, there is a strong possibility that SA may not yield practical consequences.

An important caveat regarding SA is that both the term and the concept are often used somewhat indiscriminately as either a psychological state or an implied quality of avionics displays. For example, recent trade journal advertisements have touted a traffic management display as providing "the solution for enhanced situational awareness" (*Global Airspace*, January, 1999, p. 43) and providing "the pilot with situational awareness plus Stormscope data overlaid on an electronic map" (*Global Airspace*, January, 1999, p. 44). Similarly, an aviation writer recently wrote an article entitled, "Enhanced Head-up Symbology Builds Situational Awareness," describing a display as "… improving pilot situational awareness" (*Aviation Week and Space Technology*, April 19, 1999, p. 64). In a different article, the same author suggests that a single, multi-function display, including radar, weather, navigation information, and a ground proximity warning system will "…optimize pilot *situational awareness*" (*Aviation Week and Space Technology,* April 26, 1999, p. 68).

Clearly, the prevailing conventional wisdom is that "more is better" in terms of information in the cockpit, with little concern for allocation of attentional resources or information overload. Further, there are, typically, no performance-based metrics validating such claims. More serious and egregious attributions involving SA occur when "pilot error due to loss of SA" is listed as a cause of accidents (e.g., Bureau Enquetes-Accidents's attribution for the 1994 Roselawn accident, NTSB AAR-96-02). Despite the fact that it is often invoked as an explanation, "pilot error" is not necessarily a root cause of aviation accidents, and using it as an explanation is only exacerbated when SA is included in the mix. Clearly, SA has become an overused cachet. If it is to become an enduring and useful concept, a commonly accepted definition and adequate operational definitions must be developed in the near future.

Some researchers have attempted to concentrate on components of SA in order to make it a more manageable construct. This literature review has identified several dimensions of SA that are specifically related to surveillance. However, no one dimension adequately addresses the knowledge a pilot must have to perform surveillance activities. At the same time, it does not seem likely that a combination of these dimensions would capture the construct that is of interest here. Therefore, concentrating on components of SA has not yet been particularly fruitful.

As part of our current line of research, a cognitive task analysis was undertaken and is described in a subsequent report. This research identifies information requirements that are specifically relevant to surveillance. Once information requirements are identified, assertions may be put forth regarding the knowledge a pilot must possess to perform surveillance activities in an appropriate manner. This kind of research has the potential to reduce or eliminate some of the problems associated with the concept of SA, perhaps sparing it from the same fate as mental workload. SA may have utility in that it is encouraging the resurgence of analyses similar to traditional task analyses with a unique emphasis on the dynamic nature of the task environment. In other words, the idea of SA encourages researchers to think in terms of observable human behavior in light of the environment. Hopefully, the study of SA will force researchers to isolate this complex construct without reference to other fuzzy constructs.

5.0 References

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

Andre, A.D., Wickens, C.D., Moorman, L., & Boschelli, M.M. (1991). Display formatting techniques for improving situation awareness in the aircraft cockpit. *The International Journal of Aviation Psychology*, *1*, 205-18.

Aretz, A.J. (1991). The design of electronic map displays. *Human Factors*, *33*, 85-101.

Baxter, G.D. & Bass, E.J. (1998). Human error revisited: Some lessons for situation awareness. *Proceedings of the Fourth Symposium on Human Interaction with Complex Systems*.

Crane, P.M. (1992). Theories of expertise as models for understanding situation awareness. In *Proceedings of the 13th Annual Symposium on Psychology in the Department of Defense* (pp. 148-52).

Endsley, M.R. (1989). A methodology for the objective measurement of situation awareness. In *Situational Awareness in Aerospace Operations* (pp. 1-9). Copenhagen, Denmark: NATO - AGARD.

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

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

Endsley, M.R. (1998). A comparative analysis of SAGAT and SART for evaluations of situation awareness. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting* (pp. 82-6). Santa Monica, CA: The Human Factors and Ergonomics Society.

Endsley, M.R., & Rodgers, M.D. (1994). *Situation awareness information requirements analysis for en route air traffic control*. (DOT/FAA/AM-94/27). Washington, DC: Federal Aviation Administration, Office of Aviation Medicine.

Entin, E.B. (1998). Measuring situation awareness in an attack helicopter domain: An exploratory investigation. In *Proceedings of the Human Factors Society 42nd Annual Meeting* (pp. 249-53). Santa Monica, CA: The Human Factors and Ergonomics Society.

Federico, P. (1995). Expert and novice recognition of similar situations. *Human Factors*, *37*, 105-22.

Flach, J.M. (1995). Situation awareness: Proceed with caution. *Human Factors*, *37*, 149-57.

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. 102-6). Santa Monica, CA: The Human Factors and Ergonomics Society.

Fracker, M.L. (1989). Attention allocation in situation awareness. In *Proceedings of the Human Factors Society 33rd Annual Meeting* (pp. 1396-1400). Santa Monica, CA: The Human Factors and Ergonomics Society.

Fracker, M.L. (1991). *Measures of situation awareness: Review and future directions*. (AL-TR-1991-0128). Wright-Patterson AFB OH: Armstrong Laboratory, Crew Systems Directorate.

Gaba, D.M., Howard, S.K., & Small, S.D. (1995). Situation awareness in anesthesiology. *Human Factors*, *37*, 20-31.

Harwood, K., Barnett, B., & Wickens, C.D. (1988). Situational awareness: A conceptual and methodological framework. In F.E. McIntire (Ed.) *Proceedings of the 11th Biennial Psychology in the Department of Defense Symposium* (pp. 23-7). Colorado Springs, CO: U.S. Air Force Academy.

Kass, S.J., Herschler, D.A., & Companion, M.A. (1991). Training situational awareness through pattern recognition in a battlefield environment. *Military Psychology, 3*, 105-12.

Metalis, S.A. (1993). Assessment of pilot situational awareness: Measurement via simulation. In *Proceedings of the Human Factors Society 37th Annual Meeting* (pp. 113-7). Santa Monica, CA: The Human Factors and Ergonomics Society.

National Transportation Safety Board. (1996). *In-flight Icing Encounter and Loss of Control Simmons Airlines, d.b.z. American Eagle Flight 4184 Acions de Transport Regional (ATR) Model 72-212*. (Tech. Report NTSB-AAR-96-02). Washington: National Transportation Safety Board.

O'Donnell, R.D., & Eggemeier, F.T. (1986). Workload assessment methodology. In K.R. Boff, Kaufman, & J. Thomas (Eds.). *Handbook of Perception and Human Performance: Vol II: Cognitive Processes and Performance* (Chpt. 42). New York: Wiley.

Pritchett, A.R., Hansman, R.J., & Johnson, E.N. (1995). Use of testable responses for performance-based measurement of situation awareness. In D.J. Garland, & M.R. Endsley (Eds.). *Experimental Analysis and Measurement of Situation Awareness* (pp. 75-81). Daytona Beach, FL: Embry-Riddle Aeronautical University Press.

Regal, D.M., Rogers, W.H., & Boucek, G.P., Jr. (1988). *Situational awareness in the commercial flight deck: Definition, measurement, and enhancement.* (SAE Technical Paper No. 881508). Warrendale, PA: Society of Automotive Engineers.

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

Sarter, N.B., & Woods, D.D. (1995). How in the world did I ever get into that mode? Mode error and awareness in supervisory control. *Human Factors*, *37*, 5-19.

Selcon, S.J. & Taylor, R.M. (1989). Evaluation of the situational awareness rating technique (SART) as a tool for aircrew systems design. In *Proceedings of the AGARD AMP Symposium on Situational Awareness in Aerospace Operations, CP478*. Seuilly-sur Seine: NATO AGARD.

Selcon, S.J., Taylor, R.M., & Koritsas, E. (1991). Workload or situational awareness? TLX vs. SART for aerospace systems design evaluation. In *Proceedings of the Human Factors Society 35th Annual Meeting* (pp. 62-6). Santa Monica, CA: The Human Factors and Ergonomics Society.

Shively, R.J. & Goodman, A.D. (1994). Effects of perceptual augmentation of visual displays: Dissociation performance and situation awareness. In *Proceedings of the Human Factors Society 38th Annual Meeting* (pp. 1271-4). Santa Monica, CA: The Human Factors and Ergonomics Society.

Smith, K. & Hancock, P.A. (1995). Situation awareness is adaptive, externally directed consciousness. *Human Factors*, *37*, 137-48.

Taylor, R.M. (1989). Situational awareness rating technique (SART): The development of a tool for aircrew systems design. In *Proceedings of the AGARD AMP Symposium on Situational Awareness in Aerospace Operations, CP478*. Seuilly-sur Seine: NATO AGARD.

Underwood, B.J. (1957). *Psychological Research*. Englewood Cliffs, NJ: Prentice-Hall.

Venturino, M., Hamilton, W.L., & Dvorchak, S.R. (1989). Performance-based measures of merit for tactical situation awareness. In *Situational Awareness in Aerospace Operations* (pp. 1-5). Copenhagen, Denmark: NATO - AGARD.

Vidulich, M.A. (1992). Measuring situation awareness. In *Proceedings of the Human Factors Society 36th Annual Meeting* (pp. 40-1). Santa Monica, CA: The Human Factors and Ergonomics Society.

Vidulich, M.A. & Hughes, E.R. (1991). Testing a subjective metric of situation awareness. In *Proceedings of the Human Factors Society 35th Annual Meeting* (pp. 1307-11). Santa Monica, CA: The Human Factors and Ergonomics Society.

Walker, N. & Fisk, A (1995, July). Human factors goes to the gridiron. Ergonomics in Design: *The Magazine of Human Factors Applications*, pp. 9-13.

Wickens, C.D. (1992a). *Engineering Psychology and Human Performance* (2nd ed.). New York: Harper Collins.

Wickens, C.D. (1992b, December). Workload and situation awareness: An analogy of history and implications. *Insight*: *The Visual Performance Technical Group Newsletter*, pp. 1-3.

Wickens, C.D., Gordon, S.E., & Liu, Y. (1998). *An Introduction to Human Factors Engineering*. New York: Addison-Wesley.