Scientific Measurement of Situation Awareness in Operational Testing

Elizabeth A. Green
Research Staff Member at the Institute for Defense Analyses

Miriam E. Armstrong
Research Staff Member at the Institute for Defense Analyses

Janna Mantua
Research Staff Member at the Institute for Defense Analyses

Abstract

Situation Awareness (SA) plays a key role in decision making and human performance; higher operator SA is associated with increased operator performance and decreased operator errors. While maintaining or improving "situational awareness" is a common requirement for systems under test, there is no single standardized method or metric for quantifying SA in operational testing (OT). This leads to varied and sometimes suboptimal treatments of SA measurement across programs and test events. This paper introduces Endsley's three-level model of SA in dynamic decision making, a frequently used model of individual SA; reviews trade-offs in some existing measures of SA, and discusses a selection of potential ways in which SA measurement during OT may be improved.

Keywords: Situation Awareness, Human-Systems Interaction, Operational Testing

Introduction

The ITEA Journal of Test and Evaluation, (ISSN 1054-0929), is published four times each year by the International Test and Evaluation Association, 11350 Random Hills Rd, Suite 800, Fairfax, VA 22030 (www.itea.org)

Situation awareness (SA) is relevant to a wide range of operations. Endsley defines SA as an internalized mental model of the current state of the operator's environment (Endsley 1999). However, to the operator, the true definition of SA depends on the goals and decision tasks inherent to operating a system. In other words, SA is not necessarily knowing all information about a situation, but rather knowing the information that is critical to the task at hand and using that information to project forward. When SA is poor, be it because SA is incomplete or inaccurate, the completion of tasks may be compromised and errors may occur (Endsley 1995b). Accordingly, properly quantifying SA within the context of military-relevant system evaluation is a critical endeavor.

Evaluating SA for operators in the military context will become increasingly important as more complex systems and systems-of-systems are developed. The next generation of warfare may take place within the multi-domain operational environment, which involves the integration of systems from a range of domains: space, sea, air, land, cyber, electromagnetic, and others. In such operations, individuals will need to maintain their SA for their current situation, but will likely also have to coordinate with other individuals and teams (which may or may not be co-located). Additionally, there is a broad call for systems to increase use of automation and artificial intelligence (AI). Should these advances come to fruition, we can expect systems will include human teaming, human-machine teaming, human-AI teaming, and increased automation, each of which requires SA to be modeled and measured using even more complex models than those which apply to individual SA. As domains converge and as the use of automation/AI increases, SA will be more difficult to measure, but measurement will be increasingly critical. Accordingly, establishing standards for measuring individual/operator SA during present day – at the current level of system complexity – is necessary and important. Our two primary objectives for this paper are as follows: First, to provide a common definition and model of individual situation awareness for the Test and Evaluation community, and second, to make recommendations for measuring individual SA within operational testing (OT).

Endsley's Three-Level Model of Situation Awareness

There are many published models of situation awareness in the academic literature (e.g., Smith & Hancock, 1995; Sarter & Woods, 1991). Arguably, the most frequently used model of SA represented is Endsley's three-level model of SA in dynamic decision making (Figure 1; Endsley 1995a). This model describes how an individual can achieve and maintain SA in a complex environment. In its basic form, the model shows that SA includes three states of cognitive processing that occur iteratively.

- Level 1 involves the perception of elements in a situation. This level addresses the gathering of information in the environment, such as gathering visual input from a screen or auditory input from a communication channel.

The ITEA Journal of Test and Evaluation, (ISSN 1054-0929), is published four times each year by the International Test and Evaluation Association, 11350 Random Hills Rd, Suite 800, Fairfax, VA 22030 (www.itea.org)

- Level 2 involves comprehension of the current situation. This level addresses the important process of integrating multiple information pieces and determining their relevance to a given operator's goals. Comprehension can be affected by how people combine, interpret, store, and retain information.
- Level 3 involves the projection of future states. This level speaks to an operator's ability to forecast future events and dynamics, projecting from current events to anticipate future events. Accurate projection is important for timely and appropriate decision making and is the hallmark of a skilled expert.

The interaction of perception, comprehension, and projection leads to a decision about how to act, which ultimately leads to an action itself. This action then changes the environmental state, which creates a feedback loop, providing an update to perception and the cascading processes that follow.

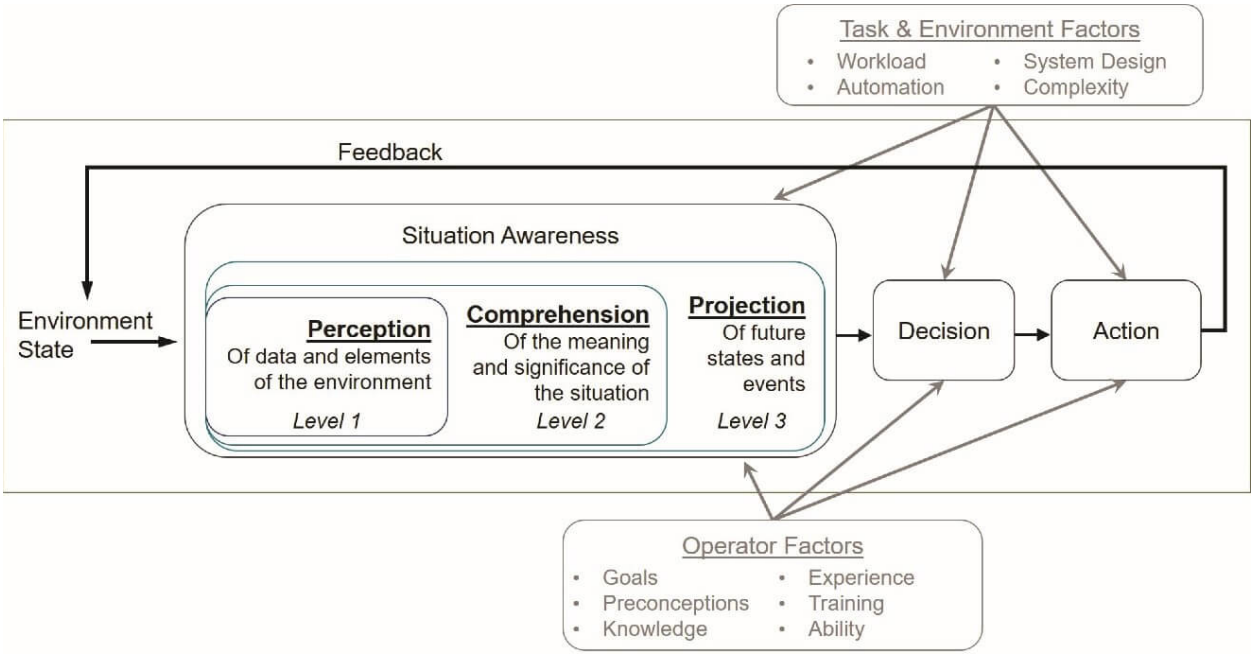

Figure 1. Adaptation of Endsley's Three-Level Model of SA

To demonstrate application of Endsley's SA model in an operational context, think of an air traffic controller (ATC; for more information on this example, see Endsley, Sollenberger, & Stein, 2000). One goal of an ATC is to prevent airborne near-misses or collisions. To achieve this goal, an ATC leverages the three levels of SA. An ATC might use Level 1 SA to perceive information about the airspace and aircraft within it, such as aircraft identification and location; Level 2 comprehension to understand the intended future course of the aircraft and implications of other situational characteristics (e.g., weather); and Level 3 projection to determine whether an aircraft's course will intersect with that of another aircraft or the implications of bad weather. This information is used to make decisions and perform actions to safely guide aircraft from takeoff to landing. This example demonstrates the feedback loop is present: once an aircraft changes location, Level 1 SA is again required

The ITEA Journal of Test and Evaluation, (ISSN 1054-0929), is published four times each year by the International Test and Evaluation Association, 11350 Random Hills Rd, Suite 800, Fairfax, VA 22030 (www.itea.org)

to perceive the updated location. It also demonstrates how deficiencies in any level of SA can lead to a potential failure in task execution.

Endsley's model accommodates incorporation of additional relevant factors (Endsley and Garland 2000). For instance, the efficiency of the SA process is modified by individual factors related to the operator, such as their experience and training. Having prior knowledge or training on a system can improve efficiency at all three levels of SA. In the ATC example, a novice operator may have to carefully search for information (Level 1), integrate multiple pieces of information in order to understand the current state of the airspace (Level 2), and then apply their interpretation of the situation to project future states (Level 3). On the other hand, an expert operator's perceptions may seem almost automated due to their experience; they are able to selectively perceive the most relevant subset of available information and make inferences about other types of information. They are then able to more quickly and efficiently comprehend the current state of the airspace and project and predict future changes in the airspace. However, an expert's efficiency can also cause them to miss or misinterpret information that contradicts their expectations (for example, when an aircraft's type does not match what is present on the ATC's flight strip; Jones, 1997). When one's SA is over reliant on their expert expectations, SA may be reduced (whether the operator knows it or not) and the likelihood of operator errors may increase (Carnino et al. 1988, Klein 1993, Endsley 1995b).

Endsley's model may also be used to account for effects related to task and environmental factors, such as automation, workload, or teaming. Similar to individual factors, task and environmental factors can either facilitate or hinder SA processes. Though individual, task, and environmental factors all increase complexity, they also make Endsley's model more dynamic and adaptable to a wide range of operators and systems. While outside the scope of the current paper, it is important that testers consider individual and task factors which may promote or hinder SA when designing operational tests.

Finally, it is important to note that the relationship between SA and performance is a *probabilistic* link. SA is an intervening variable between stimulus (situation) and response (behavior). That is, having good SA should increase the probability of good decisions and performance, but does not guarantee it. For this reason, accurate, objective SA measurement is crucial; it is ill-advised to assume an operator has good SA based on operator performance alone. A number of studies over the course of several decades have demonstrated this to be true in the military operational setting. For instance, in a study that compared two avionics systems, one old and one new, an objective SA test showed upgrades to the targeting system within the new avionics system improved the operator's ability to find enemy aircraft compared to the legacy system (Endsley 1988). These findings demonstrate the implications of having poor SA: in a real-world scenario, poorer SA within the avionics system could have led to missed enemy aircraft, resulting in catastrophic task failure. A separate study quantified errors committed by an operator who was monitoring

The ITEA Journal of Test and Evaluation, (ISSN 1054-0929), is published four times each year by the International Test and Evaluation Association, 11350 Random Hills Rd, Suite 800, Fairfax, VA 22030 (www.itea.org)

an automated system. The study was conducted because there were concerns that taking the operator "out of the loop" might impact the operator's ability to support it. Objective SA testing found the operator had lower Level 2 SA when they were operating under full automation compared to semi-automation and no automation (Endsley 1995a). That is, the operator's comprehension of the situation was degraded when working with the full-automation system. Again, these results demonstrate the implications of having poor SA, as the operator could not adequately perform his task with reduced Level 2 SA.

Situation Awareness in Operational Testing

Most systems undergo a final OT event prior to fielding. This event is the evaluator's last chance to identify SA-relevant issues before systems are broadly distributed for widespread use.

Within OT, SA measurement practices are inconsistent and rarely objective. This results in poor data and, consequently, yields poor understanding of SA during operations. We believe there are two primary reasons for poor SA measurement during OT. First, there is a lack of agreement on what SA is (and is not) within the T&E community. Second, SA is inherently difficult to measure, particularly with the combat realism required for OT.

To address the first reason, the first aim of this paper is to clarify and standardize the definition of SA for testers. As a starting place for this goal, we've selected Endsley's three-level model of SA in dynamic decision making. In addition to its flexibility, an additional rationale for focusing on this particular model for this paper is that it can be easily aligned with existing military decision-making processes (e.g., Observe, Orient, Decide, Act: Boyd et al., 1996; Military Decision-Making Process: ATTP 5-0.1, 2011). Endsley's model can inform all stages of system development. While evaluator goals and scope of test often differ between test types (e.g., developmental testing, operational testing), these differences should not prevent test designers from ensuring the constructs they measure align across test events. Increasing the utility of information from testing across a system's acquisition lifecycle benefits everyone involved.

A second aim of this paper is to make recommendations for measuring SA within OT. Measuring SA often requires pausing operations and obtaining measures of operator SA in real time, or at a minimum having accurate truth data shortly after a mission or event is completed to compare to operator perceptions. For certain systems, developing SA measures and implementing them for an OT requires a considerable amount of effort. By providing an overview of quantitative measures used to evaluate individual SA, our goal is to promote thought and discussion amongst practitioners tackling this challenging problem.

The ITEA Journal of Test and Evaluation, (ISSN 1054-0929), is published four times each year by the International Test and Evaluation Association, 11350 Random Hills Rd, Suite 800, Fairfax, VA 22030 (www.itea.org)

Considerations When Selecting an SA Measure

Operational testers should evaluate SA for any system expected to impact SA. Choosing appropriate SA measurement methods requires an understanding of available SA metrics and techniques as well as the goals of the test. The next two sections overview philosophical and practical considerations for a selection of existing SA measurement methods as well as how these considerations should inform measurement selection. The following section will provide more detail on specific SA measures and their uses in OT.

Measurement Characteristics

High-quality SA measures are reliable, sensitive, and valid. Reliable measures produce similar results when administered multiple times. Sensitive measures can differentiate between high and low levels of SA. Valid SA measures capture SA specifically rather than some other construct. SA measure validity is typically assessed by comparing the results of a measure to those found using other, well-established measures like operator performance and query measures (discussed further in the next section).

SA measures also differ from each other in that they can be subjective, capturing perceptions of SA; or objective, impartially measuring true SA. Neither subjective nor objective measures are higher quality than the other. Instead, they each capture different information and thus have different purposes.

Measures can be agnostic to test factors or dependent upon test factors. Agnostic SA measures can be administered to an operator in any role regarding any system performing any mission. Agnostic SA measures typically assess perceptions and opinions relevant to SA in multiple types of situations (for example, whether an operator is confident that they observed all critical information). By contrast, SA measures that are dependent upon test factors should be tailored to a specific role, system, and mission. Such measures capture behavior or knowledge specific to a situation (e.g., whether an ATC operator observed a new aircraft entering their airspace). In general, SA measures dependent upon test factors are higher quality than those agnostic to test factors. A practical consideration for the implementation-dependent SA measures is that tailoring them to a new situation requires considerable effort. This effort is likely justifiable when systems are expected to have an impact on (e.g., are expected to improve) operator SA.

The ITEA Journal of Test and Evaluation, (ISSN 1054-0929), is published four times each year by the International Test and Evaluation Association, 11350 Random Hills Rd, Suite 800, Fairfax, VA 22030 (www.itea.org)

Tailoring SA Measures Using Goal-Directed Task Analysis

Goal-Directed Task Analysis (GDTA) identifies and defines a system operator's goals, decisions, and SA requirements within a particular role and domain. GDTA results form a basis for SA measures dependent on test factors (Endsley 1993). A GDTA is typically developed using structured interviews with subject matter experts who are familiar with the system and operator goals. From there, a hierarchy is created in which the cognitive processes that underlie operator goals are identified, and SA requirements are determined. As such, a thorough GDTA will identify all information (Level 1 SA requirements) that people in a given role need for making decisions, ways in which that information needs to be integrated to form situation comprehension (Level 2 SA requirements), and the types of future projections that operators will need to make (Level 3 SA requirements). Testers then use this hierarchical information to tailor SA measures. GDTAs require a significant amount of time and effort to develop, but are crucial for ensuring SA measurements are appropriate for the operator and system.

For some of the SA measures covered in the next section, researchers have already conducted a GDTA (e.g., Strater et al. 2001) and used it to design tailored SA metrics such as survey questions pertinent to a particular role and mission (Matthews and Beal 2002a). In other cases, existing SA measures serve as a technique for translating the results of a GDTA into survey questions, behaviors to record, or queries. For such techniques, testers would be responsible for conducting a GDTA.

Selecting Appropriate Measures

The system under test will determine the appropriate way to measure SA during OT. If a system may impact SA but is not designed to promote SA, test designers may consider including a requirement to maintain SA. In this case, it may be sufficient to measure SA using vetted subjective measures.

Systems designed to promote SA (e.g., most command and control systems and intelligence, surveillance, and reconnaissance systems) should have requirements that state the type of SA that will be improved (e.g., which levels from the Endsley model) and how SA will be improved. To demonstrate improvement, SA should be evaluated using high-quality, objective measures; and operator SA using the system under test should be compared to operator SA using a legacy system. Because high-quality measures typically require tailoring, testers should allow for additional time to prepare appropriate measures.

Measuring Situation Awareness

The following sections overview three types of SA measures: survey measures, behavior-based measures, and query measures. For each type of measure, we will discuss how it is used to assess an operator's SA during a mission, including information on specific scales

The ITEA Journal of Test and Evaluation, (ISSN 1054-0929), is published four times each year by the International Test and Evaluation Association, 11350 Random Hills Rd, Suite 800, Fairfax, VA 22030 (www.itea.org)

or techniques. This information is meant to provide an understanding of why, but not necessarily how, one might use a given SA measure.

Survey Measures

Survey measures provide quantitative ratings of subjective SA. Many surveys are agnostic to test factors, therefore requiring minimal preparation, and can facilitate SA comparisons across different systems or mission types. Surveys can either be self-reports, in which the operator provides their own ratings; or observer ratings, in which someone else, such as a subject matter expert or peer, rates the operator. Operators typically complete self-report surveys after a mission; accordingly, self-report surveys rely on operators' memories of mission events and their SA during those events. Memory biases should be considered in these instances. Observer surveys may be collected during a mission, removing the memory burden placed on self-report surveys. However, observer survey ratings are based on observable SA-related behaviors rather than cognitive SA directly. As such, observer SA ratings are limited: assessment requires the assumption that the observer's interpretation is indicative of the operator's internal SA.

A longstanding and popular self-report SA survey is the Situational Awareness Rating Technique (SART; Taylor 2017) (also referred to as the Situation Awareness Rating Technique). SART comprises three subscales: operator understanding, attentional demand, and attentional supply. Across multiple studies, SART has been found to significantly, but weakly, predict overall mission performance (Bakdash et al. 2022). Because of its focus on attentional supply and demand, a frequent criticism of SART is that it largely measures mental workload rather than SA (Braarud 2021; Jones & Endsley, 2000). As previously mentioned in the introduction, individual factors such as attentional and working memory capacity and workload may impact SA. Measures that capture a factor correlated with SA, but not the rest of the model, are likely to have limited utility. Additionally, a recent meta-analysis suggests that SART is not predictive of objective query-based measures of SA (Endsley 2020).

Other self-report surveys ask operators to self-assess each of the three levels of SA. One such survey, the Mission Awareness Rating Scale (MARS; Matthews and Beal 2002b), was developed for use during infantry missions. A recent meta-analysis indicates that MARS is not predictive of overall mission performance (Bakdash et al. 2022), and in general there is limited evidence toward the validation of any self-report surveys based on the three-level model of SA.

Published observer surveys exist for assessing SA within a small set of domains. The SA Behaviorally Anchored Rating Scale (SABARS; Matthews and Beal 2002a) was initially developed to assess Platoon Leaders' SA during Military Operations on Urbanized Terrain (MOUT) missions. Other observer surveys include The Cranfield SA Scale (C-SAS; Dennehy 1997) and SA Rating Scales (SARS; Waag and Houck 1994). Relatively little research on

The ITEA Journal of Test and Evaluation, (ISSN 1054-0929), is published four times each year by the International Test and Evaluation Association, 11350 Random Hills Rd, Suite 800, Fairfax, VA 22030 (www.itea.org)

these measures has been published, though there is some evidence that SABARS correlates with objective query ratings (Strater et al. 2001) and that SABARS and SARS predict performance (Bakdash et al., 2022). Testers should note that using SABARS outside of the previously validated use cases requires substantial development prior to use, including conducting a GDTA for use in developing survey items. As such, each SABARS-type survey is only as good as the work done to develop it and there is no guarantee that new surveys will retain the predictive properties of those present in academic literature.

SA survey measures, particularly the mission-agnostic self-report measures, are easily implemented in OT. They do not interrupt tasks and thus preserve operational realism during test. Self-report surveys can be quickly administered to operators at the end of missions, though because OT missions are often long and effortful, the total number of surveys should be kept low to reduce the likelihood of survey fatigue and the impact of memory distortion or biases at such a delay should be considered. Further, much research suggests that self-report surveys may not fully or accurately capture SA (Endsley 2020). As such, we caution test designers against relying on survey measures when SA is a critical part of a system under test.

Behavior-Based Measures

Operator behavior during missions can provide indirect measures of SA. These measures include performance measures, task behavior measures, and process indices. Mission performance measures such as total hits or hazard detection are used to infer SA under the assumption that better SA leads to better performance. Task behaviors concern specific task actions within a mission which imply SA; for example, amount of time for operator response to a particularly important event. Process indices provide insight into the processes an operator uses to build SA and can include physiological reactions or transcripts of operators talking through a problem.

Information on operator behaviors can be collected without interfering with an operator's tasks. For example, many systems can record interactions such as when the operator opens a computer window or depresses a brake pedal. Operators can be observed by a tester either directly or from a video recording. Task and process behaviors occur continually throughout a mission, allowing one to assess SA at multiple points within a mission.

Because they rely on observable behavior, behavior-based measures have limitations that are similar to those of observer surveys. Behavior-based measures do not capture SA, only behaviors believed to relate to SA. Additionally, it may be difficult to identify SA at each level using such measures. Because a behavior's relevance to SA depends on an operator's goals within a mission, considerable preparation such as a GDTA is required to select appropriate measures which will indicate SA for various roles and missions.

The ITEA Journal of Test and Evaluation, (ISSN 1054-0929), is published four times each year by the International Test and Evaluation Association, 11350 Random Hills Rd, Suite 800, Fairfax, VA 22030 (www.itea.org)

Physiological process indices include those based on eye movement, cardiovascular activity, and electromagnetic imaging. Physiological measures have the benefit of providing continuous measurement, thus SA assessments could be tied to mission events or assessed across the duration of a mission. However, logistical constraints such as the collection and storage of protected health information, the fact that physiological measurement equipment is often expensive, and that the measurements themselves require technical expertise to develop must be taken into consideration. Further, the relationship between physiological processes and psychological states is not fully understood. As of writing, meta-analysis indicates that eye tracking and cardiovascular measures can be used to assess SA, but further research is needed before concrete conclusions or recommendations may be made (Zhang et al. 2020).

Behavior-based measures present a potentially valuable means of assessing SA within OT, but considerable work is needed to develop that potential. Like surveys, many are unobtrusive and would not disrupt operational realism. However, there are major limitations associated with behavior-based measures. The lack of established measures places a large burden on OT teams wishing to assess SA with behavior-based measures. First, teams must identify behaviors that are relevant to SA and determine how to quantify them (e.g., by marking presence/absence, recording lag time, etc.). Second, teams must either validate their metrics prior to OT or assume the risk that their chosen metrics are not valid and reliable. Finally, even if validated metrics and techniques are available, behavior-based measures are limited in that they indirectly assess SA.

Query Measures

SA can be assessed directly and objectively by querying operators on facts about their situation that are relevant to their mission. Queries pertain to one of the three levels of SA. Depending on the technique, SA is assessed based on percentage of correct answers, time taken to respond, importance of the information, operator's confidence in their response, or some combination of these. Most query-based measurement techniques involve asking operators questions in real or near-real time while conducting a mission. Because it may be disruptive if not dangerous to interrupt a real-life mission, query techniques frequently employ simulators. The operator will conduct a simulated mission inside, for example, a flight simulator or simulated air traffic control room, and testers will query the operator at randomly determined intervals during the mission.

The Situation Awareness Global Assessment Technique (SAGAT; Endsley 1988) and Situation Present Assessment Method (SPAM; Durso et al. 1995) are two widely researched query methods which commonly utilize simulation. SAGAT is an offline measure, meaning that when operators are queried the simulation stops and information from the simulation is removed or restricted. Higher rates of correct answers indicate higher levels of SA. SPAM is an online measure in which operators are alerted that they have a query waiting, then when the operator reaches a low-workload point of the simulation they accept and respond

The ITEA Journal of Test and Evaluation, (ISSN 1054-0929), is published four times each year by the International Test and Evaluation Association, 11350 Random Hills Rd, Suite 800, Fairfax, VA 22030 (www.itea.org)

to the query. SA is assessed based on the time it takes for operators to provide correct answers. Meta-analyses indicate that both SAGAT and SPAM approaches are reliable and predict performance to similar extents (Endsley 2021, Bakdash et al. 2022).

Query measures, particularly well-researched techniques such as SAGAT and SPAM, have an advantage over survey and behavior-based measures because they provide direct and objective assessments of operator SA. Because queries can assess all three levels of SA, they have potential as a wholistic measurement tool for SA. Query measures are more difficult to implement than survey and behavior-based measures. All query measures are dependent upon mission factors; for many domains, a GDTA and design of specific metrics has been conducted (e.g., Endsley 2021), but in all other cases using query measures would require considerable preparatory effort from the testers. Additionally, because they interrupt the task, query measures are not operationally realistic and are therefore not conducive to traditional OT. Implementing query measures in OT would likely require test teams to include additional mini-tests within an OT. Such tests could involve realistic missions conducted in a simulator (e.g., a flight simulator) or conducting a simulation-based excursion. The primary difference between excursion- and simulator-based SA assessment is environmental. Excursion-based evaluations should be short (less than 1 day) and may involve a subset of test participants to enable query data collection from test players during excursion pauses. In developing an excursion, test teams will design a relatively more constrained mission set, allowing for knowledge of ground truth throughout the excursion and a relevant set of query questions for the given mission.

Essential Elements of Information as a Proposed Measure of Situation Awareness

The SA measurement methods described above were derived from the academic literature and have been previously validated. This section differs as it discusses a relatively untested method for using a special information source as a measure of SA. Our description of this potential method does not describe every possible use case. Rather, this section is intended to inspire discussion and further research; we strongly note that this method should be considered preliminary until tested and validated.

During combat operations, soldiers must respond to requests for information by reporting on what they observe in the field. The most critical information requirements are known as essential elements of information (EEIs). EEIs define the commander's priority intelligence requirements regarding the adversary and the environment (Department of Defense 2016). Collection and reporting of EEIs is a well-defined part of military information collection and may serve as an avenue for measuring SA during operations. Information collection activities run on a continuous cycle and depend on the echelon, assets engaged, and the type of operation (Headquarters, Department of the Army 2013). Due to the shared features

The ITEA Journal of Test and Evaluation, (ISSN 1054-0929), is published four times each year by the International Test and Evaluation Association, 11350 Random Hills Rd, Suite 800, Fairfax, VA 22030 (www.itea.org)

of perceiving and characterizing a dynamic environment, these information collection activities closely align with many of the processes described above as SA.

A major advantage of using EEIs to characterize SA is that the SA metrics needed for analysis will arise naturally from mission planning and execution rather than requiring explicit queries. When high command levels (and their subordinates) are included as test players, the mission-specific information collection plans they develop can in turn serve as SA metrics for the test units during operational testing. When these parties are not included in the test, information collection plans can be created prior to test with subject matter expert input and then disseminated to test units. Every level of command issues their own set of intelligence requirements and EEIs which incorporate those of higher-level command.

Information related to EEIs may be collected in real or near-real time from instrumented data, voice- and text-based command and control exchanges, or in some cases, after-action reviews (AARs). Measures that characterize EEIs, including accuracy, completeness, and timeliness of information reports, can be used to assess SA and help evaluators characterize SA at echelons included in a given operational test. Additionally, during longer operational tests, dynamic requesting and reporting of EEIs can provide further insight into changes in SA over the course of the operational test.

The use of EEIs to characterize SA may easily incorporate into an operational test when information-gathering capabilities are part of the system under test. For example, when evaluating sensor-based intelligence, surveillance, and reconnaissance (ISR) systems, target presentation timing and type are often part of the design of experiments. As such, ground truth about targets will be known. In this type of test, EEIs can be defined in terms of what the operators equipped with the system are tasked with observing and the EEI reports can be evaluated by comparing them to ground truth. Some examples of EEI-based measures that test designers might plan to collect for each echelon in a test include:

- Percentage of known EEIs collected. This measure compares the number of EEIs collected to truth data. For example, if there are 50 known EEIs that a sensor operator is asked to collect on, what percentage did they actually report on?
- Percentage of requested EEIs correctly satisfied. This measure evaluates the accuracy of collected EEIs.
- Mean time to disseminate EEI products. In the case of an ISR system, time would begin when an image (from a sensor) is opened or when a request is made for full-motion video and would end when an analyst disseminates a product. Products could be email, electronic transmission, icons sent to a common operating picture, or voice communication. The time element of this measure will be specific to the system under test.

These types of measurements may be more difficult to embed within more dynamic operational tests, such as when assessing systems used by infantry units during free-play, force-on-force engagements. In these situations, EEI products will likely be similar, but test

The ITEA Journal of Test and Evaluation, (ISSN 1054-0929), is published four times each year by the International Test and Evaluation Association, 11350 Random Hills Rd, Suite 800, Fairfax, VA 22030 (www.itea.org)

designers will need to develop a more constrained excursion during the course of the operational test that allows for greater control of ground-truth information.

As previously mentioned, this idea has yet to undergo validation. However, given the close alignment of EEIs with academic models of SA, as well as the existence of related doctrine, measurement and analysis of information collection processes could be an ideal way to learn about SA at each level of operations while preserving a relatively high level of operational realism.

Conclusions and Recommendations

To be adequate, SA measurement methods will need to be tailored to specific operators of given systems. By necessity, "one-size-fits-all" solutions will not suffice. Survey-based SA assessments, which currently comprise the bulk of SA measurement methods used in OT, result in subject feedback that yields an incomplete picture of operator SA and does not correlate strongly with mission performance. These measurement methods are not adequate for current systems and, if nothing changes, will not be sufficient for more complex systems yet to come.

Technological improvements will necessitate increases in the complexity of the warfighters' mission. Examples include making changes to team structures such as integrating human teams with human-machine teams, expansion of command and control (C2) processes toward joint all-domain C2, and integration challenges for multi-domain operations. Increases in operational complexity increase the information needed for warfighters to maintain high SA, and assessing SA will become both increasingly important and difficult to accomplish. As such, the test community needs to start training and preparing methodological transitions now to prepare for future programs and problems.

References

ATTP 5-0.1. 2011. Commander and Staff Officer Guide. Washington, DC: Headquarters, Department of the Army.

Bakdash, Jonathan Z., Laura R. Marusich, Katherine R. Cox, Michael N. Geuss, Erin G. Zaroukian, and Katelyn M. Morris. 2022. "The validity of situation awareness for performance: a meta-analysis." Theoretical Issues in Ergonomics Science 23 (2): 221–244.

Bowers, Clint A., J. Weaver, J. Barnett, and Renée J. Stout. 1998. Empirical validation of the SALIANT methodology. Proceedings of the First Human Factors & Medicine Panel Symposium on Collaborative Crew Performance in Complex Operational Systems.

The ITEA Journal of Test and Evaluation, (ISSN 1054-0929), is published four times each year by the International Test and Evaluation Association, 11350 Random Hills Rd, Suite 800, Fairfax, VA 22030 (www.itea.org)

Boyd, J., C. Spinney, and C. Richards. 1996. "The essence of winning and losing briefing." Slides available at http://www. chetrichards. com/modern_business_strategy/boyd/essence/eowl_frameset. htm.

Braarud, P. Ø. 2021. Investigating the validity of subjective workload rating (NASA TLX) and subjective situation awareness rating (SART) for cognitively complex human-machine work. International Journal of Industrial Ergonomics, 86, 103233.

Carnino, A., E. Idee, J. Larchier Boulanger, and G. Morlat. 1988. "Representational errors: why some may be termed "diabolical"." In Tasks, Errors, and Mental Models, 240–250.

Department of Defense. 2016. Department of Defense Dictionary of Military and Associated Terms. Edited by Department of Defense.

Dennehy, K. 1997. Cranfield situation awareness scale: Users manual. COA Report Number 9702. Bedford, England: Applied Psychology Unit Cranfield University.

Durso, F. T., Truitt, T. R., Hackworth, C., Crutchfield, J., Nikolic, D., Moertl, P., Ohrt, D., & Manning, C. A. 1995. Expertise and chess: A pilot study comparing situation awareness methodologies. Experimental analysis and measurement of situation awareness, 295–303.

Endsley, Mica R. 1988. "Situation awareness global assessment technique (SAGAT)." Proceedings of the IEEE 1988 National Aerospace and Electronics Conference.

Endsley, Mica R. 1993. "A survey of situation awareness requirments in air-to-air combat fighters." International Journal of Aviation Psychology 3 (2): 157–168.

Endsley, Mica R. 1995a. "Measurement of situation awareness in dynamic systems." Human Factors 37 (1): 65–84.

Endsley, Mica R. 1995b. "A taxonomy of situation awareness errors." Human Factors in Aviation Operations 3 (2): 287–292.

Endsley, Mica R. 1999. "Situation awareness in aviation systems." Handbook of Aviation Human Factors, 257–276.

Endsley, Mica R. 2020. "The divergence of objective and subjective situation awareness: A meta-analysis." Journal of Cognitive Engineering and Decision Making 14 (1): 34–53.

Endsley, Mica R. 2021. "A systematic review and meta-analysis of direct objective measures of situation awareness: a comparison of SAGAT and SPAM." Human Factors 63 (1): 124–150.

The ITEA Journal of Test and Evaluation, (ISSN 1054-0929), is published four times each year by the International Test and Evaluation Association, 11350 Random Hills Rd, Suite 800, Fairfax, VA 22030 (www.itea.org)

Endsley, Mica R., and Daniel J Garland. 2000. "Theoretical underpinnings of situation awareness: A critical review." Situation Awareness Analysis and Measurement 1 (1): 3–21.

Endsley, Mica R. & Sollenberger, Randy, and Stein, Earl. 2000. "Situation awareness: A comparison of measures." Proceedings of the Human Performance, Situation Awareness and Automation: User-Centered Design for the New Millennium, Savannah, Georgia.

Fink, AA, and D.A. Major. 2000. "Measuring team situation awareness: A comparison of three techniques." First Human Performance, Situation Awareness and Automation: User-Centered Design for the New Millennium Conference. Savannah, Georgia.

Gawron, Valerie Jane. 2019. Human performance and situation awareness measures: CRC Press.

Hauss, Yorck, and Klaus Eyferth. 2003. "Securing future ATM-concepts' safety by measuring situation awareness in ATC." Aerospace Science and Technology 7 (6): 417–427.

Headquarters, Department of the Army. 2013. Information Collection (FM 3-55). Edited by Department of the Army.

Hogg, David N., Knut Folles, Frode Strand-Volden, and Belén Torralba. 1995. "Development of a situation awareness measure to evaluate advanced alarm systems in nuclear power plant control rooms." Ergonomics 38 (11): 2394–2413.

Jeannott, E., C. Kelly, and D. Thompson. 2003. The development of situation awareness measures in ATM systems. EATMP Report.

Klein, Gary. 1993. "Sources of error in naturalistic decision making tasks." Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Matthews, Michael D., and Scott A. Beal. 2002a. Assessing situation awareness in field training exercises. Military Academy West Point NY Office of Military Psychology and Leadership.

Matthews, Michael D., and Scott A. Beal. 2002b. "A field test of two methods for assessing infantry situation awareness." Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

McGuinness, Barry. 2004. Quantitative analysis of situational awareness (QUASA): Applying signal detection theory to true/false probes and self-ratings. Bae Systems Bristol (United Kingdom) advanced technology centre.

The ITEA Journal of Test and Evaluation, (ISSN 1054-0929), is published four times each year by the International Test and Evaluation Association, 11350 Random Hills Rd, Suite 800, Fairfax, VA 22030 (www.itea.org)

McGuinness, Barry, and Louise Foy. 2000. “A subjective measure of SA: the Crew Awareness Rating Scale (CARS).” Proceedings of the First Human Performance, Situation Awareness, and Automation Conference, Savannah, Georgia.

Muniz, E., R. Stout, C. Bowers, and Eduardo Salas. 1998. “A methodology for measuring team situational awareness: situational awareness linked indicators adapted to novel tasks (SALIANT).” NATO Human Factors and Medicine Panel on Collaborative Crew Performance in Complex Systems, Edinburgh, North Atlantic Treaties Organisation, Neuilly-sur-Seine: 20–24.

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

Smith, K., & Hancock, P. A. 1995. Situation awareness is adaptive, externally directed consciousness. Human factors, 37 (1), 137–148.

Strater, Laura D., Mica R. Endsley, Robert J. Pleban, and Michael D. Matthews. 2001. Measures of platoon leader situation awareness in virtual decision-making exercises. TRW Inc. Fairfax Va Systems And Information Technology Group.

Taylor, Richard M. 2017. “Situational awareness rating technique (SART): The development of a tool for aircrew systems design.” In Situational Awareness, 111–128. Routledge.

Waag, Wayne L., and Michael R. Houck. 1994. “Tools for assessing situational awareness in an operational fighter environment.” Aviation, Space, and Environmental Medicine.

Willems, Ben F., and Michele Heiney. 2001. “Real-time assessment of situation awareness of air traffic control specialists on operational host computer system and display system replacement hardware.” USA/Europe Air Traffic Management R&D Seminar, Santa Fe, USA.

Zhang, Ting, Jing Yang, Nade Liang, Brandon J. Pitts, Kwaku O. Prakah-Asante, Reates Curry, Bradley S. Duerstock, Juan P. Wachs, and Denny Yu. 2020. “Physiological measurements of situation awareness: a systematic review.” Human Factors 0 (0): 1–22.

Author Biographies

Elizabeth A. Green, Ph.D. is a Research Staff Member at the Institute for Defense Analyses. She provides expertise on evaluating human-system interaction during operational testing for the DOD. She provides analytic support to DOT&E action officers assessment of Land Expeditionary Warfare programs. She received a PhD in Experimental Psychology in 2021

The ITEA Journal of Test and Evaluation, (ISSN 1054-0929), is published four times each year by the International Test and Evaluation Association, 11350 Random Hills Rd, Suite 800, Fairfax, VA 22030 (www.itea.org)

from Texas Tech University and a MA in Psychology in 2013 from Marietta College. Her academic research focuses on metacognition, learning, and performance.

Miriam E. Armstrong, Ph.D., is a Research Staff Member at the Institute for Defense Analyses. She provides expertise on evaluating human-system interaction during operational testing for the DOD. Dr. Armstrong received her PhD in Human Factors Psychology in 2021 from Texas Tech University. She is a member of the Human Factors and Ergonomics Society.

Janna Mantua, Ph.D., is a Research Staff Member at the Institute for Defense Analyses. She specializes in influence and deception and is the Co-Chair of the Influence Operations Working Group. She conducts work for the Office of Secretary of Defense that focuses on strategic influence during campaigning. She received her Ph.D. from the University of Massachusetts, Amherst. She is a member of the Irregular Warfare Initiative

The ITEA Journal of Test and Evaluation, (ISSN 1054-0929), is published four times each year by the International Test and Evaluation Association, 11350 Random Hills Rd, Suite 800, Fairfax, VA 22030 (www.itea.org)