A COMPARATIVE ANALYSIS OF SAGAT AND SART FOR EVALUATIONS OF SITUATION AWARENESS

Mica R. Endsley
SA Technologies
Marietta, GA

Stephen J. Selcon, Thomas D. Hardiman, Darryl G. Croft
Defence Evaluation Research Agency
Farnborough, UK

Situation awareness (SA) has become an important criterion for systems evaluation efforts. Several measures of SA have been developed, the most widely used among them being the Situation Awareness Global Assessment Technique (SAGAT) and the Situational Awareness Rating Technique (SART). SAGAT provides an objective measure of SA based on queries during freezes in a simulation. SART provides a subjective rating of SA by operators. This paper presents a direct comparison of the two measures which were used within a display evaluation study. It was found that both SART and SAGAT contributed sensitivity and diagnosticity regarding the effects of the display concept. The SART measure was highly correlated with subjective measures of confidence level, a simple subjective SA measure and a subjective performance measure. The SAGAT and SART measures were not correlated with each other. The implications of these findings for the interpretation of subjective SA measures are discussed as well as advantages and disadvantages of both measurement approaches.

INTRODUCTION

Evaluations of system designs in a wide variety of domains have increasingly begun to include measures of situation awareness. Situation awareness is recognized as a critical, and often elusive, foundation for good decision making in complex and dynamic systems such as aviation, air traffic control, nuclear power, driving and medicine, to name a few. Formally defined, situation awareness refers to "the perception of the elements in the environment within a volume of time and space, the comprehension of thier meaning and the projection of their status in the near future" (Endsley, 1988). Measuring situation awareness during system evaluation posses several advantages (Endsley, 1996):

(1) While performance is always the bottom-line measure of merit for any system design, frequently in the systems of concern, sensitive performance measures are difficult to find. For instance in air traffic control, operational errors happen only very rarely, so the impact of a particular system design concept may be difficult to ascertain during testing. More detailed performance measures are often difficult to interpret. Is it necessarily better or worse if with a particular design aircraft are more closely spaced than with another? Such a finding could mean the controller was more efficient or it could mean the controller was less able to keep aircraft spaced apart. In many complex systems, this difficulty is present.

(2) In order to gain more sensitivity, human factors practitioners have for years attempted to measure workload under the assertion that it will provide more sensitivity in discriminating a good design from one which is harder to use, but which may test okay on performance measures due to increased effort on the part of the operator. Measures of workload can be very useful, however, they only capture half of the picture: How hard the person is working — not what benefit they are gaining for their efforts.

(3) Measures of situation awareness provide an index of how well operators are able to acquire and integrate information in a complex environment where a lot of data may vie for their attention. This is a measure of merit well worth assessing. What matters in these systems is not just how much data we can provide on the many displays available, but how well operators are able to acquire this information (based on their dynamic goals) under operational conditions when faced with many competing sources of information. This factor is a highly useful measure for the evaluation of an integrated system. We often need to know, not just if a given system meets certain display design guidelines, but what such a system buys (or costs) for the operator in terms of improved situation understanding when used in the context of all the other displays in the system.

Over the past decade a number of techniques for measuring situation awareness have been developed. The two most

widely used and tested of these are the Situation Awareness Global Assessment Technique (SAGAT) (Endsley, 1988; Endsley, 1990a) and the Situational Awareness Rating Technique (SART) (Taylor, 1990).

SAGAT

SAGAT is an objective measure of SA. SAGAT employs periodic, randomly-timed freezes in a simulation scenario during which all of the operator's displays are temporarily blanked. At the time of the freeze a series of queries are provided to the operator to assess his or her knowledge of what was happening at the time of the freeze. For instance, queries may ask a military pilot where other aircraft are located, which aircraft pose a threat, ownship airspeed or altitude and whether a given aircraft is hostile. The queries typcially cover SA elements at all three levels of SA (perception, comprehension and projection).

Queries are determined based on an in depth cognitive task analysis that must be conducted for each domain SAGAT is used in. Therefore the technique can be used in many domains, however, the queries must be customized. To date, SA requirements analyses have been conducted for fighter aircraft, bomber aircraft, commercial aircraft, air traffic control, maintenance systems, and nuclear power, allowing versions of SAGAT to be created for each system. Operator's responses to these queries are scored based on what was actually happening in the simulation at the time of each freeze (within operationally determined tolerance zones).

The main advantage of SAGAT is that it allows an objective, unbiased index of SA that assesses operator SA across a wide range of elements that are important for SA in a particular system. The main disadvantage of SAGAT is that it requires freezes in the simulation. Because the freezes are random and cover such a broad spectrum of operator SA requirements, operator's can not prepare for the queries and it has been found that the freezes do not affect performance in the simulations (Endsley, 1995). It has also been frequently asserted that another disadvantage is that the technique relies on memory and thus might not provide a true reflection of operator SA. As the queries are provided immediately during a freeze (which may last from 2 to 5 minutes typically depending on the number of queries provided), however, this does not appear to pose a problem. While queries provided after a simulation (or after performance of some real world task) may be subject to rationalization or generalization in verbal reports (Nisbett and Wilson, 1977), this hazard is most likely not an issue with this technique. Reviews of the literature show that these problems occur primarily when subjects are asked to report how they know something, not what their assessments of the situation are (Dreyfus, 1981; Nisbett and Wilson, 1977). It has also been found that subjects are able to report their assessments for as long as 5 to 6 minutes during SAGAT freezes without memory decay being a problem (Endsley, 1995). This would indicate that the SA of experienced operators performing tasks in a system with high ecological validity (i.e. real task domains and not artificial laboratory tasks) is accessible for verbal report via a fairly stable internal representation. In addition to possessing a high degree of content validity based on the SA requirements analyses used to create the queries, SAGAT has also been found to have predictive validity, predicting operator performance in an air combat task (Endsley, 1990b).

SART

SART is also a highly popular measure of SA. SART provides an assessment of the SA provided by some system based on an operators' subjective opinion. SART has a total of 14 components which were determined through analysis with pilots to be relevant to SA. Operators rate on a series of bipolar scales the degree to which they perceive (1) a demand on operators resources, (2) supply on operator resources and (3) understanding of the situation. These scales are then combined to provide an overall SART score for a given system. SART ratings have been found to be correlated with operator performance in evaluations of cockpit designs (Selcon and Taylor, 1990), and subjective measures of workload (Selcon, Taylor, and Koritsas, 1991).

The main advantages of SART is that it is easy to use and can be administered in a wide range of task types. It does not require customization for different domains and can be used in real world tasks as well as simulations. Potential limitations of SART have been asserted to include (Endsley, 1995): (1) the inability of operators to rate their own SA (without knowing what they don't know or what errors there may be in their own internal representations), (2) the possible influence of performance on their ratings (Operators may make such ratings based on well they think they are doing, as opposed to how good their SA is), and (3) possible confounding with workload issues (supply and demand of attention), while SA may operate as an independent factor from workload in many situations (Endsley, 1993). It has been pointed out, however, that combining SA and workload factors into one scale may provide parsimony in the data collection process (Selcon, et al., 1991).

While there are both advantages and disadvantages to each technique, little data exists regarding their comparative merits. An investigation was conducted within the context of the evaluation of a cockpit display in which both SAGAT and SART measures were assessed simultaneously. By directly comparing operators' subjective assessment of SA via SART to their ability to accurately depict the situation via SAGAT, some of the controversy regarding the use of different types of SA measures can be addressed.

METHOD

An investigation of a system for directly presenting information on threat aircraft capabilities to fighter pilots was conducted. The study examined the use of displayed threat envelopes (as shown in Figure 1) to provide pilots with needed information for replanning around pop-up threats. Such

"explanatory displays" have been shown to be a highly effective means of providing decision support, improving both performance and trust in the system (Fletcher, Shanks, and Selcon, 1996; Selcon, Bunting, Coxell, Lal, and Dudfield, 1995a; Selcon, Smith, Bunting, Irving, and Coxell, 1995b).

In the study, pilots were provided with a part-task mission simulation. The simulator consisted of a cockpit mock-up with a head-down display, an out-of-the-cockpit world display with a HUD overlay, and a stick and throttle. The pilots' task consisted of following a course of waypoints in a low level ingress flight to a target at 450 knots and 2000 feet altitude. Pop-up threats appeared during the flight which pilots were required to avoid, staying as close to their pre-determined flight path as possible. Three aircraft were presented as pop-up threats at the subject's altitude for approximately 30 seconds. Twelve experienced RAF pilots served as subjects during the experiment.

A within-subjects experimental design was used in which each subject was exposed to eight trials in each of two conditions: (1) no envelopes - in which only the threat aircraft symbol and supporting information (aircraft type, speed)were shown, (2) with envelopes - in which the same information was supplemented with a direct graphical display of the threat aircraft's launch success zone (LSZ), as shown in Figure 1. These LSZ envelopes were dynamic based on algorithms that took into account aircraft type, speed, location and relative aspect.

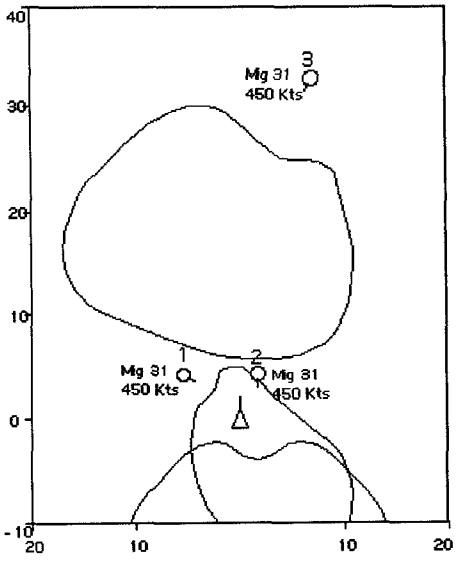

Figure 1. Launch Success Zone (LSZ) Display

Performance was measured in terms of RMS error from the assigned course and total time spent inside the threats' launch success zones. SAGAT and SART were administered during freezes which occurred at random times during the threat avoidance task. In addition, a set of subjective assessments was required of the subjects following each SART administration. These assessments had subjects rate (on a bipolar scale) an overall rating of their SA, the sufficiency of their SA, their confidence level regarding their SA, and their performance. In each condition for each subject, during 2 trials both SART (plus the subjective questions) and SAGAT were administered. During 2 trials only SART (plus the subjective questions) was administered and during 2 trials only SAGAT was administered. During the remaining 2 trials, no SA assessment was done and only performance measures were collected. The order of these administrations was randomized across subjects and conditions.

RESULTS

Performance

Analysis of Variance (ANOVA) was used to evaluate the subjects' performance data. There was no difference between conditions on the RMS flight path error. Subjects' deviation from the prescribed path was not affected by the provision of the threat envelopes. They did spend less time in the threats' launch exposure zones, however, when they were provided with this information, $F(1,11) = 19.962$, $p<.001$. Mean time in the threat zone with the LSZ envelope display was 6.5 seconds as compared to 11.75 seconds without the display.

Situation Awareness

SART. An overall SART score was calculated from the 14 individual SART ratings. The SART rating of SA was found to be significantly higher with the LSZ envelopes than without, $F(1,11) = 12.066$, $p<.01$. In examining the underlying dimensions of SART (understanding, supply of resources, demand on resources), it was found that this result was mainly attributable to differences in subject ratings of understanding. With the LSZ envelope display, subject ratings were higher for understanding ($p<.05$), information quantity ($p<.05$), and information quality ($p<.05$).

SAGAT. The subjects' perception of the situation as recorded via SAGAT was compared to the actual situation at the time of each freeze and an assessment of the accuracy of their SA for each SAGAT query calculated. ANOVAs were performed on each query to examine differences in display conditions on SA. Although there was a subjective impression of higher SA recorded via SART, the SAGAT results show a mixed picture. In terms of Level 1 SA (perception of basic information), subjects had lower SA regarding the location of threat aircraft, $F(1,81) = 3.136$, $p=.08$, with the LSZ envelopes. Most likely the envelopes had a distraction effect and subjects had less SA regarding exact aircraft position. They did however show better SA regarding own heading, $F(1,81) = 5.160$, $p=.026$, and own roll attitude, $F(1,81) = 2.726$, $p=.10$, with the LSZ envelopes. This could be due to lower workload in processing the displays or more likely heading and roll changes were more salient in that the threat envelopes changed considerably with changes in these two variables which occurred when making turns.

In terms of Level 2 SA (understanding of the situation), subjects had a better idea of the imminence of the threats, $F(1,81) = 3.286$, $p=.07$, however they did not have better SA regarding whether the aircraft could launch at them (probably because they could do this fairly well even without the envelopes based on aircraft aspect angle) or of the highest priority threat. (Non-significant trends were in the direction supporting the utility of the LSZ envelopes, however.) Subjects also showed lower SA with the envelopes on the only Level 3 question asked "which aircraft would be a threat if you stayed on your current course?" Subjects were less able to report this information correctly with the envelopes than without, $F(1,81) = 3.641$, $p=.06$. This could reflect less tendency to project ahead with the envelopes (due to the ability to rely more on the display for that information). It might also reflect a tendency for subjects to be more conservative in judging threats without the envelopes (i.e. perhaps they were more likely to be over complacent with the envelopes).

Comparison of SA Measuress. The SART scale was highly correlated ($r^2 = .67$ to $.74$) with the simple subjective SA rating, the evaluation of the sufficiency of one's SA and the subjective rating of confidence level. All of these were also highly intercorrelated ($r^2 = .74$ to $.79$). Of the SART components, the understanding rating was most highly correlated with these factors ($r^2 = .67$ to $.78$). Whatever subjective impression is being tapped by these scales, they appear to draw upon much the same factor. They did vary somewhat, however, in that they were not perfectly correlated. Subjective performance was highly predicted by the subjective SA and SA sufficiency scales ($r^2 = .61$) and less so by a combination of SART and confidence level ($r^2 = .46$).

Using the 48 trials across subjects and conditions in which SART and SAGAT were collected together, a direct analysis was made of how the SART and SAGAT scores compared. First, a component analysis and correlation analysis show that the 13 SAGAT variables collected are fairly independent, agreeing with previous such analyses (Endsley, 1990c). This means that trying to compile SA queries on different situational aspects into one combined SA variable is not supported. (In support of this, a simple SA score added up across queries proved to be of no significance in the envelope vs no envelope comparison). Each SAGAT variable was therefore treated independently in comparing to the SART score. A regression of SAGAT variables on SART was not significant on any component. There was no relationship between the subjective SART rating and any of the SAGAT variables. Examining the SART components (understanding, supply of attention and demand on attention) again there was no correlation with the SAGAT measures. The subjective assessment of SA derived via SART does not appear to be related to the objective measures of SA provided by SAGAT.

DISCUSSION

This study supports the utility of using a test-battery approach for evaluating display concepts. Simply showing reduced time in the threat zones does not provide needed information on why that was the case or on potential downsides associated with the display concept. SART proved to be of use in predicting this performance benefit. In this case, the benefit of the display was found to be associated primarily with improvements in subjective understanding ratings, as opposed to reduced workload ratings (supply & demand of attention). This finding is useful in that SART could be used in actual flight operations to evaluate design concepts when detailed performance measures are not available.

The SAGAT measure provides even further diagnosticity in reagard to changes in SA as a result of the display concept. Although subjects rated their SA as higher with the LSZ envelopes, an analysis of the SAGAT data showed that in actuality while SA may have been higher on some aspects of the situation, it actually was lower in several areas. Subjects had better SA regarding the imminence of the threats and ownship heading and roll. They had poorer SA regarding aircraft location and projecting which aircraft would be a threat in the future. This most likely reflects changes in attention allocation and processing with the new display. Such SA tradeoffs have been noted in previous studies (Endsley, 1995).

Examining the effects of a particular display or technology on SA can help illuminate these issues to the designer, who may wish to employ alternative design concepts or make modifications to deal with any potential SA tradeoffs. For instance, in this display a modification to make the threat symbol more salient (e.g. through color coding or increased intensity) in order to improve SA of aircraft location may be recommended from these results. In addition the effects of displays which directly present high level information (such as the LSZ) on the tendency of pilots to rely on such information to the exclusion of making their own future projections of aircraft states should be further investigated. This effect may be similar to automation complacency effects which have been noted in other studies (Parasuraman, Molloy, and Singh, 1993).

By revealing some of these hidden tradeoffs, the SAGAT measure provides a much greater degree of diagnosticity regarding the effects of the display on pilot SA (and potentially performance) than was available from performance measures alone in the limited simulation testing that is normally conducted in such a study. That is, it reveals SA effects that may be important in the long-run for complex mission performance. It should be noted, however, that because SAGAT scoring is based on binomial data (correct or incorrect), more data is needed to reach a level of statistical significance than might be required with other measures. (This factor accounts for the marginal levels of significance found on several of the tests reported here.) SAGAT is also not currently appropriate for use during actual flight.

CONCLUSIONS

This is one of the first studies to directly compare subjective and objective measures of SA, and it is quite interesting that there was in fact no correlation between these measures. The fact that SART, a subjective measure of SA, was highly correlated with confidence level in SA and subjective performance has interesting implications for SA measurement. It has been previously suggested that this might be the case (Endsley, 1995). This does not mean that subjective SA measures are not useful, however. Such subjective assessments may provide a critical link between SA and performance. That is, a person's *perceived* quality of SA may be important in determining how a person will choose to act on that SA (conservatively if it is low or boldly if it is high), independent of the *actual* quality of that SA.

The fact that the subjective and objective measures of SA showed no correlation, however, also casts some doubt on the validity of subjective SA measurements as an in indication of a person's actual SA. It is possible that the SART ratings were skewed to be more reflective of increases in SA on some factors (e.g. those items for which the SAGAT scores were higher) as opposed to others. If this is the case, however, it shows that subjects may not be aware of other changes in their own SA, such as the SA decrements noted here, that may be induced by a particular design. This indicates that such evaluations should be viewed with caution. As the SART scores were so highly correlated with confidence level and subjective performance, it is recommended that subjective SA ratings be viewed as good indices of these aspects, but perhaps not veridical representations of SA.

ACKNOWLEDGEMENTS

This work was sponsored by the NATO Advisory Group on Aerospace Research & Development (AGARD).

REFERENCES

Dreyfus, S. E. (1981). Formal models vs. human situational understanding: Inherent limitations on the modeling of business expertise (ORC 81-3). Berkeley: Operations Research Center, University of California.

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. Proceedings of the Human Factors Society 32nd Annual Meeting (pp. 97-101). Santa Monica, CA: Human Factors Society.

Endsley, M. R. (1990a). A methodology for the objective measurement of situation awareness. In Situational Awareness in Aerospace Operations (AGARD-CP-478) (pp. 1/1 - 1/9). Neuilly Sur Seine, France: NATO - AGARD.

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

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

Endsley, M. R. (1993). Situation awareness and workload: Flip sides of the same coin. In R. S. Jensen and D. Neumeister (Ed.), Proceedings of the Seventh International Symposium on Aviation Psychology (pp. 906-911). Columbus, OH: Department of Aviation, The Ohio State University.

Endsley, M. R. (1995). Measurement of situation awareness in dynamic systems. Human Factors, 37(1), 65-84.

Endsley, M. R. (1996). Situation Awareness Measurement in Test and Evaluation. In T. G. O'Brien and S. G. Charlton (Eds.), Handbook of Human Factors Testing & Evaluation (pp. 159-180). Mahwah, NJ:LEA

Fletcher, G. C., Shanks, C. R., and Selcon, S. J. (1996). The validation of an explanatory tool for data-fused displays for high-technology aircraft. Proceedings of the Head Mounted Displays . Orlando, FL: SPIE.

Nisbett, R. E., and Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. Psychological Review, 84(3), 231-259.

Parasuraman, R., Molloy, R., and Singh, I. L. (1993). Performance consequences of automation-induced complacency. International Journal of Aviation Psychology, 3(1), 1-23.

Selcon, S. J., Bunting, A. J., Coxell, A. W., Lal, R., and Dudfield, H. J. (1995a). Explaining decision support: An experimental evaluation of an explanatory tool for data-fused displays. Proceedings of the Eighth International Symposium on Aviation Psychology (pp. 92-97). Columbus, OH: The Ohio State University.

Selcon, S. J., Smith, F. J., Bunting, A. J., Irving, M. A., and Coxell, A. W. (1995b). An explanatory tool for decision support on data-fused panaoramic displays. Proceedings of the SPIE International Symposium on Aerospace/Defence Sensing and Control and Dual-use Photonics (pp. 17-21). Orlando, FL: SPIE.

Selcon, S. J., and Taylor, R. M. (1990). Evaluation of the situational awareness rating technique (SART) as a tool for aircrew systems design. In Situational Awareness in Aerospace Operations (AGARD-CP-478) (pp. 5/1 -5/8). Neuilly Sur Seine, France: NATO - AGARD.

Selcon, S. J., Taylor, R. M., and Koritsas, E. (1991). Workload or situational awareness?: TLX vs SART for aerospace systems design evaluation. Proceedings of the Human Factors Society 35th Annual Meeting (pp. 62-66). Santa Monica, CA: Human Factors Society.

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