Developing Operator Capacity Estimates for Supervisory Control of Autonomous Vehicles

M. L. Cummings, Massachusetts Institute of Technology, Cambridge, Massachusetts, and **Stephanie Guerlain,** University of Virginia, Charlottesville, Virginia

Objective: This study examined operators' capacity to successfully reallocate highly autonomous in-flight missiles to time-sensitive targets while performing secondary tasks of varying complexity. **Background:** Regardless of the level of autonomy for unmanned systems, humans will be necessarily involved in the mission planning, higher level operation, and contingency interventions, otherwise known as human supervisory control. As a result, more research is needed that addresses the impact of dynamic decision support systems that support rapid planning and replanning in time-pressured scenarios, particularly on operator workload. **Method:** A dual screen simulation that allows a single operator the ability to monitor and control 8, 12, or 16 missiles through high level replanning was tested on 42 U.S. Navy personnel. **Results:** The most significant finding was that when attempting to control 16 missiles, participants' performance on three separate objective performance metrics and their situation awareness were significantly degraded. **Conclusion:** These results mirror studies of air traffic control that demonstrate a similar decline in performance for controllers managing 17 aircraft as compared with those managing only 10 to 11 aircraft. Moreover, the results suggest that a 70% utilization (percentage busy time) score is a valid threshold for predicting significant performance decay and could be a generalizable metric that can aid in manning predictions. **Application:** This research is relevant to human supervisory control of networked military and commercial unmanned vehicles in the air, on the ground, and on and under the water.

INTRODUCTION

In both the military and commercial sectors, there has been significant research and development in the field of unmanned vehicles, which include remotely piloted vehicles, unmanned air vehicles (UAVs), unmanned air combat vehicles, unmanned underwater vehicles, and a plethora of ground-based robotic systems. Unmanned systems of the future are expected to become more autonomous with improvements in technology and to require less direct human input through manual control. Whereas current UAVs require relatively concentrated human input for flight control, in the future it is likely that the human role for direct flight control will diminish and that the need for supervisory control, including higher level cognitive reasoning, will become much more substantial. Because of the expected future use of clusters of unmanned vehicles that can both respond to human commands as well as operate autonomously, there is a clear need to explore the human requirements for supervisory control of autonomous air vehicles.

In supervisory control, a human operator monitors a complex system state and intermittently executes some level of control on a process, acting through some automated agent (Sheridan, 1992). This paper will detail the results of a research effort designed to explore human supervisory control issues in the management of multiple Tactical Tomahawk missiles. Because of new Global Positioning System retargeting capabilities, Tactical Tomahawk missiles, which were previously fire-and-forget missiles, can now take on missions similar to those of UAVs in that their flight paths can be changed in flight to emergent, time-critical targets. As part of this new capability, Tactical

Address correspondence to Mary L. Cummings, 77 Massachusetts Ave. 33-305, Cambridge, MA 02139; missyc@mit.edu.
HUMAN FACTORS, Vol. 49, No. 1, February 2007, pp. 1–15. Copyright © 2007, Human Factors and Ergonomics Society. All rights reserved.

Downloaded from hfs.sagepub.com at OKLAHOMA STATE UNIV on September 29, 2015

Tomahawk missiles could be stationed in loiter patterns to await further orders, much like what is envisioned for both surveillance and combat UAVs. Because of the similarities between these missions and the need to replan missions and reallocate resources in real time under time pressure, the experimental results reported here are equally relevant to both Tactical Tomahawk missiles and unmanned vehicles.

Human supervisory control of Tactical Tomahawk missiles generally encompasses two primary subtasks: (a) monitoring the missiles in flight, which requires tracking missile progress both in time and distance as well as monitoring subsystem integrity such as navigation and communications; and (b) retargeting missiles when an emergent situation occurs. Emergent situations, for example, can include the appearance of a mobile surface-to-air missile site or a missile system failure that requires redirection of an additional in-flight missile to cover a high-priority target. These tasks of monitoring and redirecting in-flight vehicles are generic to any command and control system that seeks to reallocate resources in real time based on a priority-ranked need.

The U.S. Navy (2002) initially estimated that in the domain of multiple missile control, one operator should control at most four missiles. This number was not empirically determined and was only an initial estimation. The number of missiles a person can effectively monitor and retarget in flight is an important parameter but, as stated, is overly simplistic and does not address all of the complex issues an operator faces in supervisory control. Many factors will affect how an individual performs in a combat rapid real-time resource allocation problem: number of missiles (or UAVs) one person is assigned to monitor, difficulty of decisions, time stress, secondary tasking, situational awareness, available decision support, characteristics of the evolving scenario, and other factors such as fatigue, training, and experience. Because Tactical Tomahawk controllers were all novices in this early stage of domain modeling, design, and testing, the human-in-the-loop experiment detailed in the next section focused on basic performance and workload issues and not on any other metrics that relied on training or on experience (although these would be important parameters to model in more mature systems). The goal of this experiment was to develop better capacity predictions through simulation and experimentation, given increasing numbers of missiles, varying workload levels, and increasing complexity of scenarios.

The original hypothesis for the supervisory control performance measures in this effort was that as the number of missiles and the rate of arrival increased for emergent situations, decision times would increase, the number of incorrect decisions would increase, and situational awareness would be negatively impacted. Through experimental testing, we set out to develop a human supervisory control performance model to empirically provide appropriate manning estimates – for example, to measure the number of missiles that one operator can effectively supervise during a range of representative retargeting scenarios. A preliminary experiment with a previous version of the research test bed described here demonstrated that participants could monitor up to eight missiles (Willis, 2001). With the enhanced decision support capabilities described in the next section, we wanted to examine whether this number could be further increased.

METHODS

Apparatus

A dual screen simulation test bed was developed (Figure 1) that allows a controller to monitor the current status of all in-flight missiles while having a real-time depiction of all potential missile-target pairings in a matrix in the event of the need to possibly redirect missiles in flight. The monitor display represents all missiles and targets in a geospatial display superimposed on a map of the local terrain. The map allows the controller to perceive missile information geographically, and associated time bars allow the controller to perceive important temporal relationships such as missile launch time, time of impact, and time of fuel remaining, all in comparison with the actual time and with each of the other missiles. A detailed discussion of the domain analysis that led to the interface design and the specific details and the nuances of the actual design of the interface are reported elsewhere (Cummings, 2003).

The retargeting display consists of a decision matrix, missile and target information boxes, and a communication box. The decision matrix shows the current status of all retargetable missiles as well as all physically possible missile-target pairings. Retargetable missiles are listed in the left-hand column, and targets in a strike, including

Downloaded from hfs.sagepub.com at OKLAHOMA STATE UNIV on September 29, 2015

Downloaded from hfs.sagepub.com at OKLAHOMA STATE UNIV on September 29, 2015

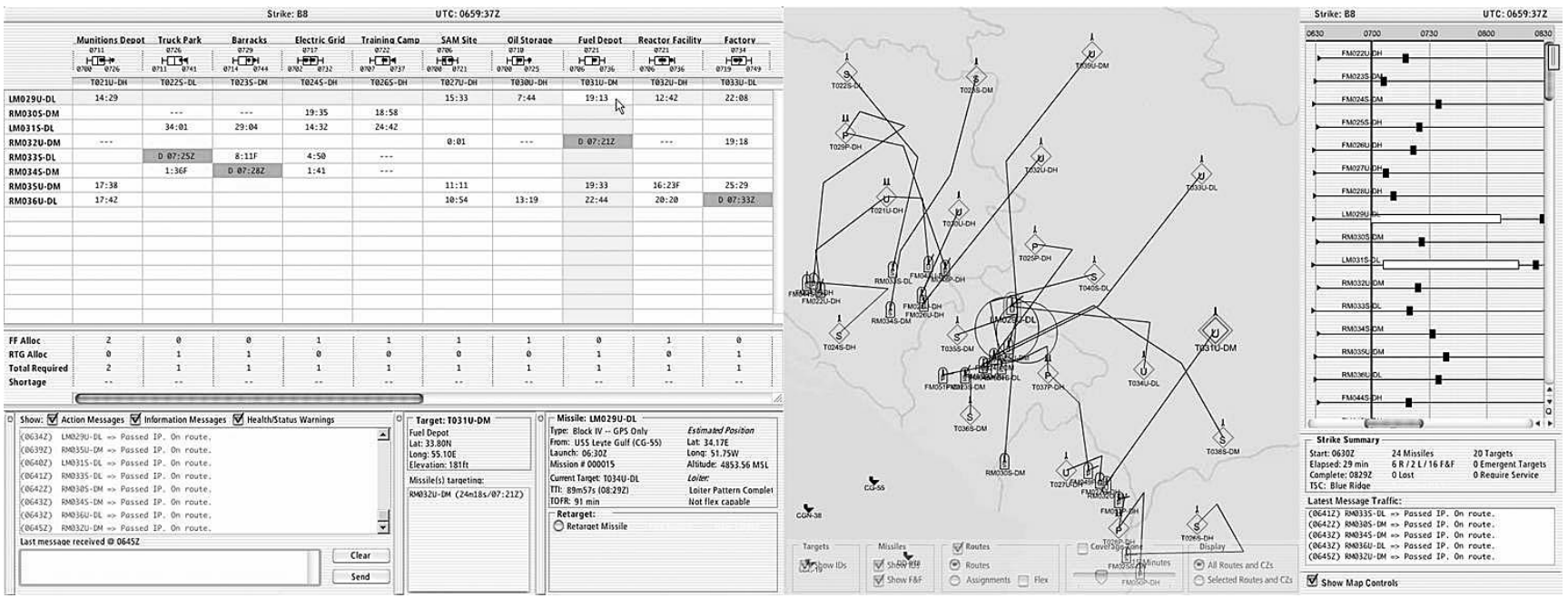

Figure 1. The supervisory control interface for missile management.

emergent targets, are represented across the top. Current missile-target pairs and their associated time on targets are highlighted, and empty cells in the matrix demonstrate that no possible relationship exists between a missile and a target (e.g., due to a wrong warhead type on the missile, or the missile is unable to reach the target). For the possible missile-target pairings, the display shows the user when the missile can get to the target and the time remaining left for the operator to make the decision to retarget.

The other significant display component for the retargeting display is a text-based communication tool, also referred to as the chat box, which resembles current instant messaging programs in widespread use (Figure 2). The popularity of the chat box is not just in mainstream pop culture, as it is already in use today aboard U.S. Navy vessels. In the context of this experiment, operators were "chatting" with superior officers who would disseminate information that required no action as well as query for information that required action on the part of the operator. In addition, the chat box served as an embedded secondary workload measurement tool as it allows for measuring both the accuracy and latency of responses to queries. Traditional secondary tasking can be intrusive and introduce an unrealistic artifact. However, embedded secondary tasks do not fundamentally change the task or task performance and provide more sensitive measurements (Shingledecker, 1987; Tsang & Wilson, 1997; Wickens & Hollands, 2000). Because it is a tool that current U.S. Navy personnel are very familiar with, the chat box provides for ecologic and external validity.

Training and testing was conducted using two dual monitor side-by-side stations as depicted in Figure 1. Each station used a Dell personal computer with two 17-inch (43-cm) color monitors with a screen area of 1024 × 768 pixels, 16-bit high-color resolution, and a Pentium 3 processor. The dual monitors for each system were supported by a MATROX® graphics card. During testing, all user actions, mouse movements, and incoming and outgoing messages on both screens were recorded by software into a text file. In addition, the users' interactions with the retargeting display were recorded using a scan converter and a VCR for later data analysis.

Participants

Forty-two U.S. Navy personnel participated in the experiment. The participant population contained both active duty officers and enlisted personnel who previously worked with the fire-and-forget versions of the Tomahawk missiles, as well as 13 retired U.S. Navy personnel who either had significant experience in operating older versions of the Tomahawk or had experience in strike planning.

Procedure

All participants received approximately 3 hr of training, which included a slide presentation to explain the prototype and two practice sessions, each of which lasted approximately 25 min. During the training, participants interacted with display elements and engaged in retargeting scenarios similar to those that would be seen in test sessions. In addition, participants observed three additional

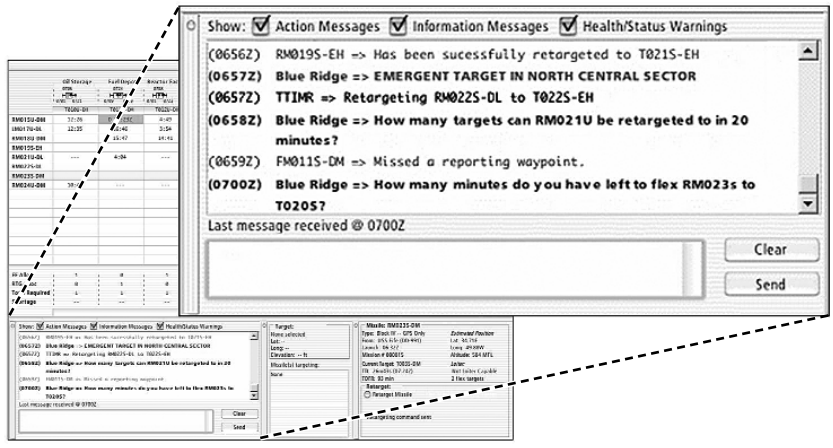

Figure 2. The test-bed chat interface.

Downloaded from hfs.sagepub.com at OKLAHOMA STATE UNIV on September 29, 2015

practice sessions other than their own. The rules for retargeting missiles were explained during all phases of training and reviewed before testing. The rules were relatively simple: no missile paired to a high-priority target should be retargeted, loiter missiles were always preferred if they were candidate missiles, and communications through the chat box were always secondary to the higher priority task of retargeting. The first practice session consisted of a walk-through of all screen elements and a demonstration of the retargeting process. The second practice session mirrored an actual test session, detailed in the next section. The only difference between this second practice session and a test session was the possible pausing that took place in case of mistakes or questions. Participants were thoroughly debriefed, and all questions were addressed.

After training, all participants were tested on two separate experimental sessions, with a break of approximately 25 min between the two. In both test sessions, the participants had approximately 3 min of warm-up in which they monitored missiles and incoming communication messages and were presented with a simple retargeting scenario. After the warm-up period, each participant saw three scenarios: (a) an easy scenario (a single emergent target arrival and only one candidate missile available); (b) a medium-difficulty scenario (a participant had to redirect a missile from its default target to another target in the strike based on the instructions of a "superior officer" simulated as a message coming through the chat box); and (c) a hard retargeting scenario, which consisted of the arrival of two emergent targets 15 s apart. The arrival of the two emergent targets was considered a dual retargeting scenario as both targets were in competition for the same candidate missiles. In all scenarios, there was one correct answer according to the rules. However, because the matrix presents all missiles that can physically reach a target, it was possible for a participant to retarget a missile that was not the best choice but still was viable.

Communication messages were interspersed throughout the test session, including information messages, such as status information of other ships participating in the strike, and health and status messages from missiles. Both information and health and status messages required no response. Participants also received action messages, highlighted in bold, which, by definition, required a response. Six action messages appeared for each participant in every test session; three occurred during monitoring periods and three occurred 30 s after commencing a retargeting scenario. The test sessions concluded when participants retargeted a missile to all emergent targets and were no longer attempting to answer any action messages in the chat box. After each test session, a modified Situation Awareness Global Assessment Technique (SAGAT) test (Endsley, 2000) was administered in the form of a single freeze-frame quiz. In addition, the six action questions provided embedded on-line situation awareness (SA) measurements, which allowed for the collection of SA data without the intrusion of freezes in the simulation. The use of embedded SA measures in the chat box is similar to the Situation Present Assessment Method (SPAM; Durso et al., 1998) as well as to on-line SA probes used in air traffic control research (Endsley, Sollenberger, Nakata, & Stein, 2000).

Experimental Design

There were two primary independent variables of interest in this experiment: (a) retargetable missile level (8, 12, or 16 missiles) and (b) operational tempo (low or high). The three missile levels represented the number of missiles that a participant had to choose from during retargeting scenarios and was included to determine if and when participants would become overloaded because of an increasing number of objects to cognitively process. Participants were randomly assigned to the three missile levels. The second primary independent variable was tempo, which represented the arrival rate of the retargeting problems, simulating either low- or high-tempo operations. In the low-tempo sessions retargeting scenarios were spaced approximately 4 min apart, and in the high-tempo sessions the arrivals were spaced only 2 min apart. Each low- and high-tempo testing session included the three scenarios (easy, medium, or hard), and all participants were tested on all three levels of difficulty, but in each, they saw only one level of missiles. The order of presentation of the fast and slow tempos was counterbalanced across participants.

Several dependent variables were used: decision time, utilization (percentage busy time), and SA. In addition, for each participant an overall performance score was calculated based on decision times, ability to manage the communications, and

Downloaded from hfs.sagepub.com at OKLAHOMA STATE UNIV on September 29, 2015

decision accuracy. (These metrics will be discussed in detail in subsequent sections.) For each of the dependent measures, the experiment was a mixed factorial design in that all participants experienced two test sessions in which they saw both high and low tempos, but for each session they used the same level of missiles. The retargeting scenarios in each of the two tempo sessions were comparable in level of difficulty and presentation of possible correct choices, although they were not exactly the same. Screen objects were modified in name, routings were slightly different, and targets were moved slightly to make it seem as if the problems were different.

A between-subjects secondary factor of session effect was included for all metrics as a measure of experimental internal validity. This factor was represented by "low first" and "high first," which represent whether a participant was tested on the low-tempo or the high-tempo test session first. One recognized threat to internal validity is maturation, in which participants can gain experience or somehow change over the course of an experiment (Adelman, 1991). The session effect was included as a secondary independent factor because participant training was limited and it was possible that learning would take place between the two test sessions and introduce a potential confound. By including session effect as a factor, we could detect any significant learning in the results and measure this threat to internal validity.

For each of the dependent variables, a single covariate of age was used; whereas the use of covariates can lead to a loss of power, a repeated measures design and the addition of a few meaningful covariates helps to increase power (Stevens, 1996). Two other covariates were explored, video game experience and time in the U.S. Navy, but these were not significant and not used in the final analysis. The age covariate was used because the participants ranged in age from 22 to 59 years. In general, the speed of mental processing slows with age, and this slowing accounts for a significant portion of age-related variance in cognitive tasks, more so than any other cognitive mechanism (Zacks, Hasher, & Li, 2000). If age is suspected to be a potential confound, it should be included as a covariate so as not to compromise external validity (Nichols, Rogers, & Fisk, 2003). Covariate assumptions of linearity of regression and homogeneity of regression coefficients were met, with the exception of one secondary independent variable (session effect) for the decision time variable, which will be discussed further.

RESULTS

Given the three levels of missiles (between subjects), the two levels of operational tempo (within subjects), and the two levels of the session effect secondary factor (between subjects), the general linear statistical model was a $3 \times 2 \times 2$ repeated measures linear mixed model. Only two-factor interactions were considered in order to avoid a saturated model and to increase degrees of freedom. One participant was removed from the data pool because standardized scores consistently were greater than 3.29 (Tabachnick & Fidell, 2001). For all the reported results, $\alpha = .05$.

Performance Measures

Two measures were explored to determine the impact of the different levels of missiles and operational tempos on human performance. The first measure examined was decision time for each retargeting event, but because this metric alone does not give a sense of how well participants were able to manage an entire test session, an additional performance measure was used, called the *figure of merit,* which was an aggregate score based on all available time and accuracy measurements.

Decision time analysis. The first performance metric examined was decision time for each retargeting scenario, and as discussed previously, there were three scenarios (easy, medium, and hard). Because the rapid prototype was created in Macromedia Director®, each testing session containing the three scenarios was in effect a movie, and as such it was impossible to randomize the order of the three different scenarios. To determine if the inability to randomize the scenario order would significantly affect the internal validity of the decision time dependent variable, a preliminary experiment was conducted to determine if the order in which participants saw the scenarios produced significantly different results for the decision time dependent variable. The results demonstrated that the order in which participants saw the scenarios was not significant (Cummings, 2003). Thus the confounding effect attributable to the lack of randomization for scenarios for this primary experiment was not considered to be a significant confound.

Because of the additional factor of scenario level of difficulty (a within-subjects factor), the

Downloaded from hfs.sagepub.com at OKLAHOMA STATE UNIV on September 29, 2015

statistical model used for this portion was a 3 × 2 × 2 × 3 repeated measures linear mixed model. Because of deviations from normality, the decision time data were transformed using a square root transformation. Of the main effects, only scenario and missiles were found to be significant: scenario, $F(1, 197) = 202.86$, $p < .001$; missiles, $F(2, 198) = 10.94$, $p < .001$. However, there was one marginal interaction for the Tempo × Session term, $F(1, 195) = 3.80$, $p = .053$, and a significant interaction for the Scenario × Missiles term, $F(4, 130) = 3.19$, $p = .016$. The age covariate was significant, $F(1, 196) = 7.640$, $p = .006$; however, when checking for homogeneity of regression coefficients, it was noted that age interacted with session ($p = .041$). Given that other linearity and homogeneity assumptions were met for the remaining independent variables, and also that the session independent variable was a secondary factor included to test a learning effect, this interaction was not considered confounding.

The significance of the Tempo × Session interaction for decision time is important as it indicates a possible learning effect. Because of the limited training time for participants and the use of a repeated measures design, one potential confound for this experiment was that learning could occur between the two sessions. Repeated measures experimental designs are useful for experiments that seek to measure learning or practice effects; however, if learning is not an explicit measure of the study, it can introduce a possible confound, called a carry-over effect (Keppel, 1982). In the case of the two test sessions, participants' performance could improve as a result of learning between the first and second sessions, which is why the session effect factor was included in the design. One solution to combat the carry-over effect is to counterbalance the order of presentation (Keppel, 1982), which was the case in these experiments. However, this interaction was marginal ($p = .053$), so the possible learning carry-over effect for one combination (high session first) was not a significant experimental confound.

The significant interaction between scenario and number of missiles also required investigation. Figure 3 demonstrates that although the decision times were essentially the same for all missile

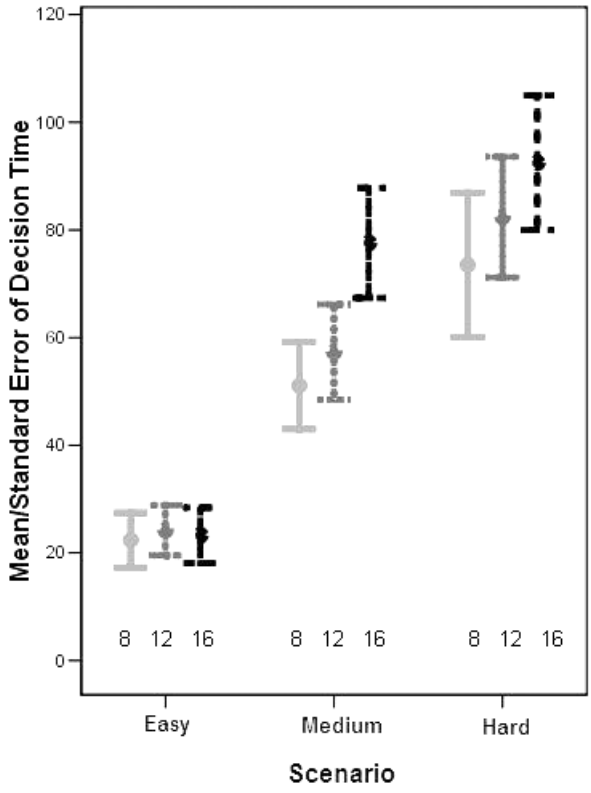

Figure 3. Scenario × Missiles interaction for decision time (in seconds).

Downloaded from hfs.sagepub.com at OKLAHOMA STATE UNIV on September 29, 2015

levels for the easy scenarios, the decision times increased for both the increasing levels of difficulty and the increasing levels of missiles for the medium and hard scenarios. The scenario factor was significant, as expected, as the easy, medium, and hard scenarios were designed to produce different response times and were initially tested to ensure this effect.

Because a primary area of exploration was to investigate the level of missiles that influenced controller performance, the significant main effect for missiles was examined. A Bonferroni comparison between missile levels yielded interesting results. The contrast between the 8- and 12-missile categories was not significant ($p = .268$), but that for both the 8- and 16-missile ($p < .001$) and the 12- and 16-missile comparisons ($p = .015$) was significant.

Overall performance analysis. Because the decision time analysis examined only the performance of participants for specific retargeting situations, not taking into account their answer accuracy or their abilities to attend to both primary and secondary tasking, an overall performance measure was needed. To measure overall performance per session, a figure of merit (FOM) score was developed. Because each test session was composed of three test scenarios (easy, medium, and hard), the FOM represented overall controller performance per test session as defined by the summation of a performance score for each scenario. Each scenario performance score was the product of weighted decision times (the quickest times were rewarded with higher points), answer accuracy (1 for the best answer, 0.5 for a satisficing answer, and 0 for completely wrong), and scenario complexity weight (thus easier scenarios were not weighted as much as difficult scenarios). Answer accuracy in retargeting scenarios included satisficing responses. Satisficing is a form of bounded rationality in which a solution is selected that achieves a "good enough" status. In the context of selecting missiles for retargeting, satisficing occurs when a missile is selected that is not the optimal missile but can achieve the mission nonetheless. In addition, six secondary tasking performance terms were included in the FOM, which incorporated both the time taken to respond to secondary tasking introduced through the chat box action messages as well as answer accuracy. These were also weighted, but on a scale of one-third the weightings for the retargeting scenarios. (The detailed FOM derivation is given in Cummings, 2003.)

Thus the FOM contained nine algebraic terms in two basic categories: three for the weighted retargeting scenario scores and six for the action message secondary tasking scores. Pilot testing was used to determine the correct weights, which were designed so that in the nominal condition, the secondary tasking FOM component would be one third of the weight of the retargeting scenario scores, which indicate primary task performance. Thus, primary task performance was weighted twice as much as secondary task performance. Scores ranged from 67 to 346, with higher scores indicating better overall performance. For the experimental data, the average secondary task component of the FOM was 35.9%, which was commensurate with predictions. However, the secondary task percentage of the total FOM ranged from 7%, if participants performed the retargeting and secondary tasks, up to 76%, if participants neglected to perform either or both of those tasks, and this resulted in lower overall performance scores. For the average FOM of 193, 124 points were attributable to the three retargeting scenario scores and 69 points were attributed to secondary tasking.

The experimental design for the FOM variable was a 3 × 2 × 2 repeated measures linear mixed model. The scenario level of difficulty was omitted because the FOM was an aggregate score that included these levels of difficulty. As in the decision time analysis, the age covariate was used. For the FOM overall performance metric, missile level was significant, $F(2, 71) = 5.08$, $p = .009$, as was tempo, $F(1, 71) = 6.58$, $p = .012$. As in the decision time analysis, the age covariate was significant. There was a single significant higher order interaction, Tempo × Session, $F(1, 71) = 8.14$, $p = .006$. The Bonferroni comparison results for the FOM dependent variable demonstrated the same results as in the decision time analysis. Overall performance was significantly degraded when the missile level increased from 8 to 16 ($p = .028$) and from 12 to 16 ($p = .018$), whereas performance was essentially the same between 8 and 12 missiles ($p = 1.00$). This result demonstrates further evidence that a significant relationship exists among 8, 12, and 16 missiles. Figure 4 demonstrates the relationship among missile level, tempo, and the FOM. There is a clear decrement in performance at the 16-missile level.

Downloaded from hfs.sagepub.com at OKLAHOMA STATE UNIV on September 29, 2015

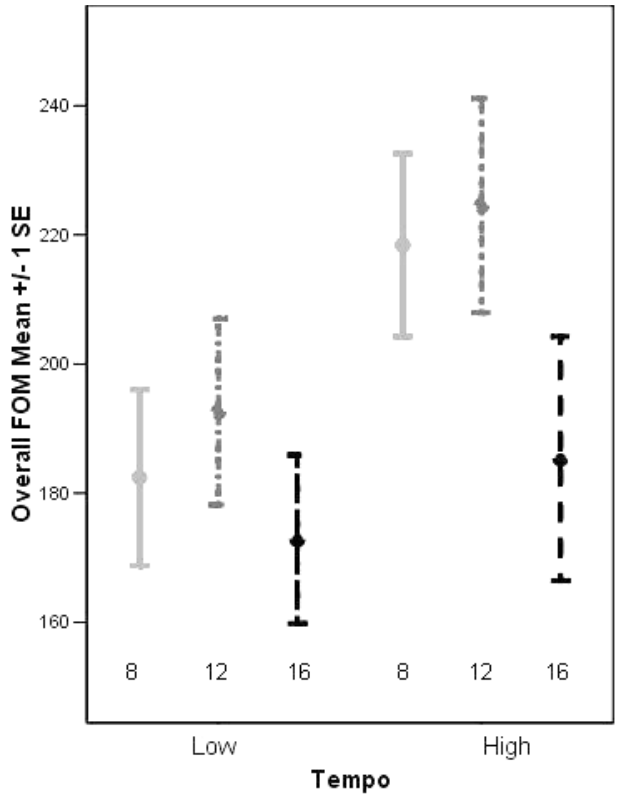

Figure 4. Missile levels versus figure of merit overall performance.

The tempo significant result for the FOM dependent variable was unexpected because the initial hypothesis was that as operational tempo increased, overall performance would decrease. Surprisingly, these experimental results revealed that overall performance increased as tempo increased. However, because there was a significant interaction involving tempo, the significance of the tempo main effect must be interpreted in light of the significant interaction. For the Tempo × Session interaction, if participants saw the low-tempo session first, they did much better overall on the subsequent high-tempo testing session. However, there was essentially no difference in performance between the two if the high-tempo session was seen first. Thus the significance of tempo was confounded most likely by learning, as discussed previously, and is not a conclusive result. This result suggests that in complex testing scenarios that are likely to introduce a learning effect in multiple test sessions, the statistical need to randomize presentation should be balanced against the likely learning effect.

Workload Measurement

When designing a command and control decision support system for rapid real-time replanning of in-flight vehicles, consideration of workload and how changes in system states impact workload and subsequent mission success or failure is paramount. In this experimental analysis of workload, workload was objectively measured through a utilization metric. The utilization metric is defined as the percentage of time the operator is busy, ρ = operator busy time/total operation time. *Busy* was defined as the time the operator was actively engaged in either a retargeting decision or responding to a communication message. The time in both cases was measured from notification of the event to either the successful completion of the retargeting task decision or answering the incoming message. Notification of events was highly salient in that an auditory alarm sounded and, in the case of emergent targets, pop-up windows appeared on both screens. The results from the experiment were expected to illustrate that for given arrival rates

Downloaded from hfs.sagepub.com at OKLAHOMA STATE UNIV on September 29, 2015

and missile levels, operators would experience specific workload levels as expressed through utilization rates. For example, given an arrival rate of 5 targets in 20 min, an operator who can retarget missiles on average in 2 min will experience lower utilization/workload than will an operator who needs 4 min to solve a problem. To measure whether or not the independent variables significantly impacted participant utilization, we used the 3 × 2 × 2 repeated measures linear mixed model including the age covariate. Because of deviations from normality, the utilization data were transformed using a square root transformation.

Tempo, the arrival rate of retargeting scenarios, was significant, $F(1, 62) = 19.67$, $p < .001$, which was an expected result, as was missile level, $F(2, 63) = 7.34$, $p = .001$. The covariate of age was significant, $F(1, 64) = 6.50$, $p = .013$. There were no significant interactions. A Bonferroni comparison between missile levels demonstrated that the utilization dependent variable between 8 and 12 missiles was not significantly different; however, the 16-missile level was significantly different from both the 8- and 12-missile levels (8 vs. 12, $p = 1.00$; 8 vs. 16, $p = .001$; 12 vs. 16, $p = .019$). These results, which mirror those of the decision time and FOM missile comparisons, illustrate that both tempo and missile level significantly impacted the workload of participants, with 16 missiles producing significantly higher workload than that produced by 8 and 12 missiles.

Prior human performance models using steady state queuing models estimate that humans cannot successfully perform at utilization rates greater than 70% (Rouse, 1983; Schmidt, 1978). To determine whether or not this prediction was true for this series of experiments, using the FOM overall performance measure, we performed a Pearson correlation (which measures the linear relationship between two quantitative variables), which revealed a significant relationship between utilization rates and overall performance, $r = –.394$, $p < .001$. As utilization rates increased, overall performance declined. Figure 5, representing the utilization percents from both test sessions for every participant, illustrates that given 10% increases in utilization rates, there is a significant drop in overall performance when utilization rates rise above 70%. This difference in overall performance scores below and above 70% was statistically confirmed with a nonparametric Mann-Whitney test ($p < .001$ for both test scenarios).

Interestingly, for utilization rates below 40%, it appears that overall performance scores also declined; however, no experimental data existed below 33% utilization. As will be discussed in a later section, it is possible that under low-workload situations, situational awareness can decrease, leading to an overall degradation in performance. This analysis confirms that participants do perform significantly worse at utilization rates greater than 70%, but also it is possible that under low

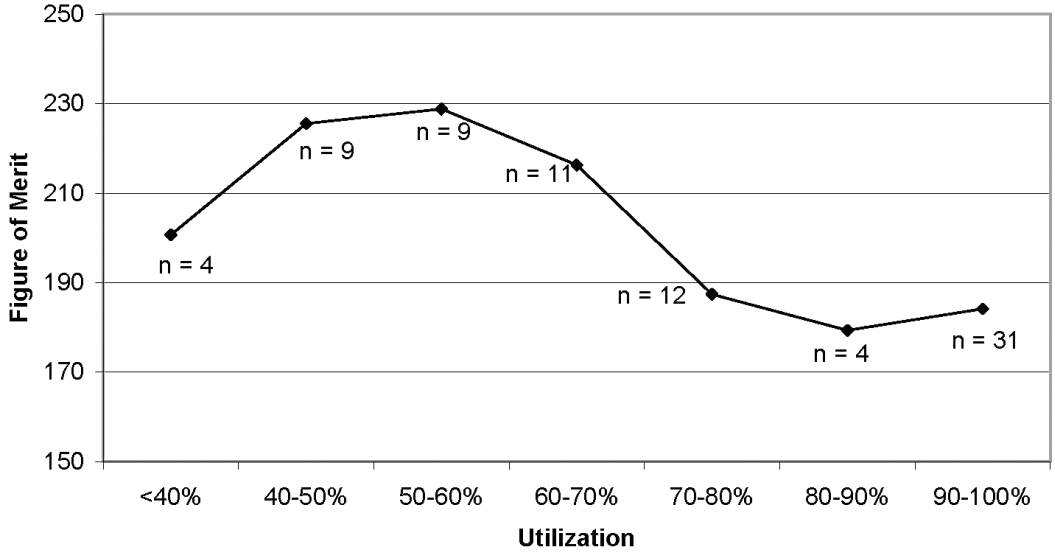

Figure 5. Performance for increasing utilization scores.

Downloaded from hfs.sagepub.com at OKLAHOMA STATE UNIV on September 29, 2015

utilizations, performance will also degrade; however, this degradation could not be tested for significance with an *N* of 13.

Situational Awareness

SA has long been recognized as a critical human factor in military command and control systems. Military command and control centers must attempt to assimilate and reconstruct the battle picture based on information from a variety of sensor sources such as weapons, satellites, and voice communications. In fact, Klein (2000) determined that SA was a key element for those naval personnel engaged in tasks associated with air target tracking aboard U.S. Navy AEGIS cruisers (which are the same platforms that will be responsible for Tactical Tomahawk launches and could control future UAV clusters.) In a domain similar to that of the Tomahawk, poorly designed human interfaces have led to reduced SA as well as degraded mission performance in remotely piloted vehicles (Ruff, Narayanan, & Draper, 2002). Because of the complexity and dynamic nature of the command and control environment, maintenance of SA is considered to be of utmost importance (Santoro & Amerson, 1998).

SA can be measured both through objective and subjective measures. However, conclusions drawn from SA self-ratings are extremely limited and provide insight into judgment processes rather than an objective SA performance measurement (Jones, 2000). In order to avoid any confounds associated with subjective measures, only objective SA measures were used in this effort. As discussed previously, a single freeze-frame quiz administered at the conclusion of each testing session was used to assess SA, in addition to the on-line SA measurements introduced through the chat box. Only response accuracy was considered for SA measurements, as reaction times were indicative of workload, not SA. SA questions introduced through both the on-line probes and the SAGAT quiz targeted comprehension and future projection and thus were testing levels 2 and 3 SA. All participants received practice SA queries, both on line and SAGAT, prior to testing.

Using the SA scores from both the on-line probes and the SAGAT quiz as the dependent variables, we used the 3 × 2 × 2 repeated measures linear mixed model to analyze any possible differences using the same independent variables, which were number of missiles, arrival rate (tempo), and session order. For the on-line probes, there were two significant results: those for the session effect, $F(1, 68) = 8.15$, $p = .002$, and for missiles, $F(1, 68) = 4.57$, $p = .014$. Participants consistently had higher on-line probe SA scores for both test sessions when they experienced the lower workload test session first. This result was not unexpected because the tempo of the second test session was essentially twice that of the first test session, so participants had more time to assess and project correctly during the lower workload session. Thus, this is an example of an established significant learning effect. In addition, the significance of the on-line probe SA missiles factor showed an interesting trend not seen in the other variables. In a Bonferroni comparison, SA as measured by the on-line probes was significantly different only between 12 and 16 missiles ($p = .022$), whereas comparisons between 8 and 12 missiles and between 8 and 16 missiles were not ($p = .056$, a marginally significant effect, and $p = 1.000$, respectively). This result suggests that SA decreased from 12 to 16 missiles but was unchanged between 8 and 16 missiles. Because of the learning effect strongly exhibited for SA as measured by the on-line probes, the significance of the missiles factor is not as compelling as it is for the other dependent variables. However, in a study examining the SA of air traffic controllers tasked to interact with up to 20 aircraft, performance significantly degraded beyond 12 (Endsley & Rogers, 1996).

In a similar study on air traffic control using multiple SA measures, on-line probes did not reveal any significant effects, although the metric was reaction time and not accuracy because participants achieved an overall accuracy of 95% (Endsley et al., 2000). In this Tomahawk study, overall participant accuracy for on-line probes was 77%, and thus this was a viable metric for determining experimental effects. Although the use of an on-line probe as a valid SA measurement tool continues to show mixed results (Durso et al., 1998; Endsley et al., 2000), this study demonstrates that it might show important trends. However, more research is needed to specifically address construct and validity concerns.

In the case of the SA dependent variable represented by the SAGAT questions, there was no significance for either the main effects or any interactions. This lack of significance of main effects for SAGAT questions could be interpreted in different ways. First, the interface design could be

Downloaded from hfs.sagepub.com at OKLAHOMA STATE UNIV on September 29, 2015

effective in promoting SA, so despite increasing workload, SA is not compromised as the workload increases. However, it is likely that exposing the participants to two 20-min scenarios was not long enough to allow a true SA picture to build. In addition, because the SAGAT quiz was administered only twice, it is likely that more measurements would have been needed to capture any significant effect.

The SAGAT method has demonstrated a level of predictive validity that links the SAGAT scores to overall participant performance (Endsley, 2000); however, no such predictive ability has been demonstrated for on-line probes (Endsley et al., 2000). In general, improvement in overall performance is closely linked with improvement in SA (Vidulich, 2000). To see if this relationship between performance and SA held true for these SA scores, both on-line and SAGAT SA scores were correlated with the participants' FOM scores, which represented overall performance. There was no significant correlation with the SAGAT scores; however, the on-line probe SA measures were significantly correlated (Pearson correlation = .432, which was significant at the $p < .001$ level).

DISCUSSION

Table 1 presents the aggregate data analysis results for all the dependent variables and the predetermined main effects. The session effect independent variable was included to determine if a learning effect would confound the results and threaten internal validity. There was only one significant main session effect, the SA on-line probe dependent variable, so this dependent measure was sensitive to the learning that occurred between test sessions. However, there was a significant session effect interaction with the tempo independent variable for the FOM dependent variable and a marginally significant interaction for decision time. In both cases it appears that a learning effect did influence the tempo significance result. This analysis illustrates the importance of analyzing interaction effects before interpreting significant main effects.

The remaining significant effects were not compromised with interactions. The significance of the tempo independent variable for the utilization (percentage busy time) dependent variable, which represented both high and low arrival rates of retargeting problems, was expected. A preliminary discrete event simulation model predicted that the tempo factor would be significant for utilization rates given the 2- and 4-min arrival intervals. In addition, the significance of the scenario independent variable for decision time was expected because the scenarios were designed for different levels of difficulty and, thus, would take longer as complexity increased. As expected, because of the age range of participants, the age covariate was significant for all metrics, which held true except for SA. This result suggests that it is important for the age of the participant pool to closely resemble that of the user population, which for the Tactical Tomahawk community is expected to be 20 to 40 years.

Given the large range in participant ages, from 22 to 59 years, it is important to ensure that the significance among 8, 12, and 16 missiles is not confounded by groupings of older participants within the between-subjects missile factor. Table 2 details the mean age in each missile category as well as the minimum and maximum ages. The finding that controlling 16 missiles, as opposed to 8 and 12 missiles, caused degraded performance is not confounded by age.

The most important finding in the data summary (Table 1) is that the level of missiles was not only significant for three primary performance

TABLE 1: Independent Variable Significance Summary

Factor	Utilization	Decision Time	Figure of Merit	SA On Line
Tempo	*<.001*	.186*	*.012**	.538
Missiles	*<.001*	*<.001*	*.009*	*.014*
Session	.334	.338	.235	*.006*
Scenario	n.a.	*<.001*	n.a.	n.a.
Age (covariate)	*.013*	*.006*	*<.001*	.108

Note. Numbers in italics indicate statistical significance at alpha = .05.

*Interaction detected with the session independent variable.

Downloaded from hfs.sagepub.com at OKLAHOMA STATE UNIV on September 29, 2015

TABLE 2: Missile Factor Age Statistics

	Missiles		
	8	12	16
Mean age	39	36	33
Minimum age	26	24	22
Maximum age	59	57	51

measures (FOM, utilization, and decision time) but also, more importantly, in each of these cases, the 8- and 12-missile levels were statistically no different whereas the 16-missile level produced significantly different results (Table 3).

These results have significant implications for human supervisory control of multiple airborne vehicles. Ruff et al. (2002) determined that at most only four UAVs could be effectively controlled by a human. However, this limitation was primarily attributable to the fact that these four UAVs required significantly more manual human control than do Tactical Tomahawk missiles, which require no manual input and for which the human's role is strictly supervisory. In the case of the Tomahawk missiles and more autonomous UAVs of the future, human intervention is required only for emergencies and higher level mission planning, not for manual control. Thus as systems become more autonomous, humans will be able to individually control a larger number.

Interestingly, in other air traffic control studies investigating workload and performance as a function of increasing numbers of aircraft, similar results have been reported. In a study examining subjective and objective workload of air traffic controllers under free flight conditions (pilot self-separation), participants were presented with a low-workload condition of 10 aircraft on average as well as a high-workload condition of 17 aircraft on average. Just as in the Tomahawk study, participants' objective performance was significantly degraded at the 17-aircraft level as compared with 10 aircraft (Hilburn, Bakker, Pekela, & Parasuraman, 1997). In another study looking at air traffic controller performance in free flight conditions, participants were presented with either 11 or 17 aircraft and, as in the Hilburn et al. (1997) study, performance of participants was significantly degraded at the higher level. Under the moderate traffic conditions of 11 aircraft, conflict detection performance approached 100% but dropped to 50% under the high-traffic condition (Galster, Duley, Masalonis, & Parasuraman, 2001).

TABLE 3: Missile Factor Bonferroni Comparison

Comparison	8 vs. 12	12 vs. 16	8 vs. 16
Utilization	1.000	.019	.001
Decision time	.268	.015	<.001
Figure of merit	1.000	.018	.028

The similarity between the numbers of missiles and commercial aircraft that could be effectively controlled by controllers under free flight conditions is remarkable. During free flight operations, aircraft are operated "autonomously" by the pilots – that is, all navigation decisions are made internal to the aircraft, whereas air traffic controllers monitor the progress of aircraft and are required to intervene only when some unexpected condition (e.g., an emergency) or change in a goal state occurs (e.g., the desire to land at another airport because of weather). This is very similar to the supervision of missiles or autonomous UAVs in that they navigate on their own and the operator needs to intervene only in an emergency or the occurrence of an emergent target, which represents a change in goal state. Thus in two related human supervisory control domains that require rapid decision in real time under time pressure, humans effectively controlled 10 to 12 airborne vehicles, but in both domains performance was significantly degraded at 16 to 17 airborne vehicles. This meta-analysis is preliminary, and this area deserves more research as it suggests some clear design limitations for command and control operators of autonomous systems.

Lastly, another important finding is that this study provides experimental evidence, as predicted by Rouse (1983) and Schmidt (1978), that operator performance beyond utilization rates (percentage busy times) of 70% will significantly degrade. These results appear to be in keeping with the classic Yerkes-Dodson inverted U-shaped function. The original Yerkes-Dodson paper detailed a relationship between stimulus strength and learning (Yerkes & Dodson, 1908); however, a similar relationship was proposed for the effects of arousal level on performance (Hebb, 1955). In this experiment, the arousal level is represented by increasing utilization rates, which reflect increasing stimuli as well as stress. The stimuli in this experiment, defined by increasing numbers of missiles and

Downloaded from hfs.sagepub.com at OKLAHOMA STATE UNIV on September 29, 2015

increasing operational tempo, represent increases in overall mental workload. Participants tended to achieve optimal performance between 50% and 60% busy time (Figure 5), with performance significantly dropping as utilization rates increased and, possibly, dropping under lower utilizations.

This finding provides a tangible design guideline in that systems that task operators beyond the 70% utilization level will operate at a suboptimal level. Whereas the decision times and FOM performance scores are specific to the interface and experimental design, and thus cannot be effectively compared with other interfaces or other unmanned vehicle systems, the utilization dependent variable is not as limited. Because it is based on the amount of time an operator is engaged in control tasks and the 70% ceiling applies across human supervisory control settings, it is a metric that can provide insight for both interfaces and system designs. This is important in that it can be used not only to predict manning levels for a single system but also to provide a generalizable metric that can be compared across automation levels, interface designs, and system architectures such as networks of heterogeneous vehicles.

CONCLUSION

There is significant interest in the Department of Defense to design and build networks of unmanned vehicles that will have the ability to operate autonomously. Not only will these changes affect the military command and control battle space, they will also affect the civilian airspace control picture, as many civilian agencies are considering the use of autonomous vehicles in domains such as cargo flights, border patrol, and other homeland defense purposes. With the influx of higher levels of control autonomy into unmanned vehicles, the nature of human supervisory control is changing and thus introducing new layers of cognitive complexity. Identification of automated decision support issues for unmanned command and control systems is critical because human operators must integrate temporal and spatial elements as well as solve problems, manage assets, and perform contingency planning in a varying workload environment. To this end, this research effort was designed to build a human-in-the-loop simulation platform to mimic the multiple-vehicle supervisory control domain and to investigate basic cognitive limitations in the execution of human supervisory control tasks of autonomous vehicles.

The most significant finding from the examination of the impact of increasing levels of missiles with increasing operational tempo rates was that regardless of operational tempo, participants' performance on three separate objective performance metrics was significantly degraded when controlling 16 missiles, as was their situation awareness. Strikingly, these results are very similar to those of air traffic control studies that demonstrated a similar decline in performance for controllers managing 17 aircraft. These results highlight the need to develop a predictive model for controller capacity that is not based on a single interface design and a single level of automation, as demonstrated in this experiment, but instead can be applied across domains, interfaces, and autonomy levels. Yet another significant finding was that a 70% utilization (percentage busy time) score was a valid threshold for predicting significant performance decay and could be a generalizable metric that can aid in manning predictions as well as address system architecture and decision support design questions.

ACKNOWLEDGMENTS

This research was sponsored by the Naval Surface Warfare Center, Dahlgren Division (NSWC/DD), through a grant from the Office of Naval Research Knowledge and Superiority Future Naval Capabilities Program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSWC/DD or the Office of Naval Research. In addition, we would like to thank Richard Jagacinski, Mica Endsley, and other anonymous reviewers for their insightful comments for both this and follow-on research.

REFERENCES

Adelman, L. (1991). Experiments, quasi-experiments, and case studies: A review of empirical methods for evaluating decision support systems. *IEEE Transactions on Systems, Man, and Cybernetics, 22,* 293–301.

Cummings, M. L. (2003). *Designing decision support systems for revolutionary command and control domains.* Unpublished doctoral dissertation, University of Virginia, Charlottesville.

Durso, F. T., Hackworth, C. A., Truitt, T. R., Crutchfield, J., Nikolic, D., & Manning, C. A. (1998). Situation awareness as a predictor of performance for en route air traffic controllers. *Air Traffic Control Quarterly, 6,* 1–20.

Endsley, M. (2000). Direct measurement of situation awareness: Validity and use of SAGAT. In M. Endsley & D. J. Garland (Eds.), *Situation awareness analysis and measurement* (pp. 73–112). Mahwah, NJ: Erlbaum.

Endsley, M., & Rogers, M. D. (1996). Attention distribution and situation awareness in air traffic control. In *Proceedings of the Human*

Downloaded from hfs.sagepub.com at OKLAHOMA STATE UNIV on September 29, 2015

Factors and Ergonomics Society 40th Annual Meeting (pp. 82–85). Santa Monica, CA: Human Factors and Ergonomics Society.
Endsley, M., Sollenberger, R. L., Nakata, A., & Stein, E. S. (2000). *Situation awareness in air traffic control: Enhanced displays for advanced operations.* Atlantic City, NJ: U.S. Department of Transportation.
Galster, S. M., Duley, J. A., Masalonis, A. J., & Parasuraman, R. (2001). Air traffic controller performance and workload under mature free flight: Conflict detection and resolution of aircraft self-separation. *International Journal of Aviation Psychology, 11*, 71–93.
Hebb, D. O. (1955). Drives and the CNS. *Psychological Review, 62,* 243–254.
Hilburn, B. G., Bakker, M. W. P., Pekela, W. D., & Parasuraman, R. (1997, October). *The effect of free flight on air traffic controller mental workload, monitoring and system performance.* Paper presented at the 10th International Conference of European Aerospace Societies, Amsterdam.
Jones, D. G. (2000). Subjective measures of situation awareness. In M. Endsley & D. J. Garland (Eds.), *Situation awareness analysis and measurement* (pp. 113–128). Mahwah, NJ: Erlbaum.
Keppel, G. (1982). *Design and analysis: A researcher's handbook.* Englewood Cliffs, NJ: Prentice Hall.
Klein, G. (2000). Analysis of situation awareness from critical incident reports. In M. Endsley & D. J. Garland (Eds.), *Situation awareness analysis and measurement* (pp. 51–71). Mahwah, NJ: Erlbaum.
Nichols, T. A., Rogers, W. A., & Fisk, A. D. (2003). Do you know how old your participants are? *Ergonomics in Design, 11*(3), 22–26.
Rouse, W. B. (1983). *Systems engineering models of human-machine interaction*. New York: North Holland.
Ruff, H. A., Narayanan, S., & Draper, M. H. (2002). Human interaction with levels of automation and decision-aid fidelity in the supervisory control of multiple simulated unmanned air vehicles. *Presence, 11,* 335–351.
Santoro, T. P., & Amerson, T. L. (1998, June). *Definition and measurement of situation awareness in the submarine attack center.* Paper presented at the Command and Control Research and Technology Symposium, Monterey, CA.
Schmidt, D. K. (1978). A queuing analysis of the air traffic controller's workload. *IEEE Transactions on Systems, Man, and Cybernetics, 8,* 492–498.
Sheridan, T. B. (1992). *Telerobotics, automation, and human supervisory control*. Cambridge, MA: MIT Press.
Shingledecker, C. A. (1987). In-flight workload assessment using embedded secondary radio communication tasks. In A. Roscoe (Ed.), *The Practical Assessment of Pilot Workload* (AGARDograph No. 282, pp. 11–14). Neuilly sur Seine, France: Advisory Group for Aerospace Research and Development.
Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3rd ed.). Mahwah, NJ: Erlbaum.
Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). New York: HarperCollins.
Tsang, P., & Wilson, G. F. (1997). Mental Workload. In G. Salvendy (Ed.), *Handbook of Human Factors and Ergonomics* (2nd ed., pp. 417–489). New York: John Wiley & Sons, Inc.
U.S. Navy. (2002). *Tomahawk weapons system (Baseline IV) operational requirements document.* Washington, DC: Department of Defense.
Vidulich, M. A. (2000). Testing the sensitivity of situation awareness metrics in interface evaluations. In D. J. Garland (Ed.), *Situation awareness analysis and measurement* (pp. 227–246). Mahwah, NJ: Erlbaum.
Wickens, C. D., & Hollands, J. G. (2000). *Engineering psychology and human performance* (3rd ed.). Upper Saddle River, NJ: Prentice-Hall.
Willis, R. A. (2001). *Tactical Tomahawk weapon control system user interface analysis and design.* Unpublished master's thesis, University of Virginia, Charlottesville.
Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology, 1,* 459–482.
Zacks, R. T., Hasher, L., & Li, K. Z. H. (2000). Human memory. In F. I. M. Craik & T. A. Salthouse (Eds.), *The handbook of aging and cognition* (pp. 293–357). Mahwah, NJ: Erlbaum.

Mary L. "Missy" Cummings is an assistant professor of aeronautics and astronautics and director of the Humans and Automation Laboratory at the Massachusetts Institute of Technology. She received her Ph.D. in systems engineering in 2003 from the University of Virginia.

Stephanie Guerlain is an associate professor of systems and information engineering and director of the Human-Computer Interaction Laboratory at the University of Virginia. She received her Ph.D. in industrial and systems engineering in 1995 from the Ohio State University at Columbus.

Date received: May 12, 2004
Date accepted: November 7, 2005

Downloaded from hfs.sagepub.com at OKLAHOMA STATE UNIV on September 29, 2015