

Overview of the DARPA Augmented Cognition Technical Integration Experiment

Mark St. John

David A. Kobus

Pacific Science & Engineering Group, Inc.

Jeffrey G. Morrison

Space and Naval Warfare System Center, San Diego

Dylan Schmorrow

Defense Advance Research Projects Agency

The Defense Advance Research Projects Agency Augmented Cognition program is developing innovative technologies that will transform human-computer interactions by making information systems adapt to the changing capabilities and limitations of the user. The first phase of the Augmented Cognition program was to empirically assess the ability of various psychophysiological measures to identify changes in human cognitive activity during task performance in real time. This overview describes the empirical results of a Technical Integration Experiment involving the evaluation of 20 psychophysiological measures from 11 different research groups, including functional Near Infrared imaging, continuous and event-related electrical encephalography, pupil dilation, mouse pressure, body posture, heart rate, and galvanic skin response. These "cognitive state gauges" were evaluated on a common, quasi-realistic, military command and control task called the Warship Commander Task. Participants monitored aircraft on a geographical display for their levels of threat and responded to the threatening ones, as they simultaneously monitored ship communications for ship status information. The task involves a combination of perceptual, motor, spatial, auditory, verbal, memory, and decision-making processing. Task load was manipulated by changing the quantity and types of aircraft appearing throughout the primary task and by varying the presence or absence of the secondary verbal-memory task. Eleven of the gauges significantly identified changes in cognitive activity during the task. This overview summarizes the results and examines the prospects for the successful transition of these cognitive state gauges to operational military human-machine systems.

1. INTRODUCTION

Command and control tasks, like many other operational tasks, can vary tremendously in workload from moment to moment. One moment, a user may be understimulated and bored, whereas a moment later the same user may be overwhelmed by multiple competing task demands. Either condition can lead to a variety of performance decrements. Consequently, it would be beneficial if a user's workload could be managed to maintain performance at an optimal level. Furthermore, mental workload may not be a unitary concept. Instead, numerous mental processes may be busy or idle depending on the specific task demands of the moment. If changes in cognitive activity can be measured dynamically in real time, then it may be possible to use this information to manipulate task demands and their modes of presentation to sustain optimal levels of workload, and task performance, for each component mental process.

The Defense Advance Research Projects Agency (DARPA) Augmented Cognition program is an investigation of the feasibility of using psychophysiological measures of cognitive activity to guide the behavior of human-computer interfaces. The goal is to increase the effectiveness of system operators by managing the information presented to them and the tasks assigned to them based on the available cognitive capacity of the operator.

The goal of the first phase of the program was to physically combine multiple psychophysiological sensors and simultaneously detect changes in cognitive activity while participants performed a quasi-realistic military task. This Technical Integration Experiment (TIE) brought together a variety of psychophysiological measures (cognitive state gauges) from different research organizations and evaluated them in a common test environment that manipulated the type and quantity of cognitive activity. The TIE was not intended to serve as the basis for gauge development, per se, although it afforded an opportunity for individual research groups to identify potential improvements and integration possibilities with other gauges. Rather, the goal was to demonstrate the physical combination of multiple gauges on a participant and to evaluate their potential for further development and application to more applied task environments in the second phase of the Augmented Cognition program.

In this study, 20 cognitive state gauges using a wide range of sensor technologies and theoretical approaches were evaluated. Sensor technologies included functional Near Infrared imaging (fNIR), continuous and event-related electrical encephalography (EEG/ERP), eye tracking, pupil dilation, mouse pressure, body posture, heart rate, and galvanic skin response (GSR). Table 1 lists each of the gauges evaluated in this study, type of sensor it used, and the research organization developing the gauge. See the Appendix for more details about each gauge.

The 20 cognitive state gauges were assigned to one of four data collection teams to create suites of gauges that could simultaneously monitor participants as they performed the task. This arrangement was done to (a) assess compatibility issues among the different gauge technologies, (b) allow the direct comparison of results using the different gauges within a team as they assessed the cognitive state changes of the same participants at the same time, yet (c) allow the use of similar sensor technologies that would otherwise compete for access to the same physical locations on test

Table 1: The 20 Gauges and a Summary of Results

Gauge	Sensor Type	Research Group	Team	Task Load Factors			Consistency (Percentage of Participants)
				Number Tracks per Wave (6, 12, 18, 24)	Track Difficulty (Hi and Lo)	Secondary Verbal Task (On and Off)	
fNIR							
fNIR (left)	Blood oxygenation	DrexelU	2	●	○	○	75
fNIR (right)	Blood oxygenation	DrexelU	2	●	○	○	63
EEG-Continuous							
Percentage high vigilance	EEG	ABM	2	●	○	○	63
Probability low vigilance	EEG	ABM	2	●	○	○	75
Executive load	EEG	QinetiQ/UBristol	3	●	●	○	100
EEG-ERP							
Motor effort	ERP-IFF	EGI	1	○	○	○	0
Auditory effort	ERP-Engage sound	EGI	1	○	●	○	0
Loss perception	ERN-Error sounds	Sarnoff/Columbia	4	○	○	●	50
Occular-frontal source	ERP-Comms	UNewMexico	4	●	○	○	100
Synched anterior-posterior	ERP-Comms	UNewMexico	4	○	○	●	100
Visual source	ERP-Comms	UNewMexico	4	○	○	○	100
Arousal							
Arousal meter	Inter-heart beat interval	ClemsonU	1	○	○	○	0
Arousal	GSR	UHawaii	2	○	○	○	0
Arousal	GSR	AnthroTronix	4	○	○	○	17
Physiological							
Head and monitor coupling	Head posture	UPitt/NRL	1	●	○	○	43
Head bracing	Body posture	UPitt/NRL	1	○	○	○	14
Back bracing	Body posture	UPitt/NRL	1	○	○	○	14
Perceptual and motor load	Mouse clicks	UHawaii	4	●	●	○	100
Cognitive difficulty	Mouse pressure	UHawaii	4	●	●	○	100
Index of cognitive activity	Pupil dilation	SDSU	floating	●	○	●	57

Note. fNIR = Near infrared imaging; DrexelU = Drexel University; EEG = electroencephalography; ABM = Advanced Brain Monitoring, Inc.; QinetiQ/UBristol = QinetiQ, Ltd./University of Bristol; ERP = event-related potential; IFF = identify friend or foe; EGI = Electrical Geodesics, Inc.; Sarnoff/Columbia = Sarnoff, Inc./Columbia University; Comms = communication event; ERN = error-related negativity; UNewMexico = University of New Mexico; ClemsonU = Clemson University; CSR = galvanic skin response; UHawaii = University of Hawaii; UPitt/NRL = University of Pittsburgh/Naval Research Laboratory; SDSU = San Diego State University. ● = statistically significant effects ($p < .05$); ● = “marginally” significant effects ($p < .1$); ○ = nonsignificant results. The final column lists the percentage of participants showing a moderate or high correlation ($r > .3$) between gauge values and the Number of Tracks per Wave.

participants. Each team included an EEG sensor, an “arousal” sensor, such as GSR or heart rate, and one or two other gauges whose sensor hardware were physically compatible with each other in terms of their placement on the head and body of participants. The San Deigo State University (SDSU) research group had two eye tracking systems, and they floated among the four teams, adding their pupil dilation measure to the mix for specific data collection sessions. The teaming arrangements are also indicated in Table 1.

The TIE was not the first attempt to combine multiple psychophysiological technologies and measure cognitive activity during a complex task. For example, Fournier, Wilson, and Swain (1999) used a complex personal-computer-based flight simulation to manipulate user workload while measuring cognitive activity using EEG, heart rate, and eye blinks. Smith, Gevins, Brown, Karnik, and Du (2001) used the same task while measuring EEG, and Van Orden, Limbert, Makeig, and Jung (2001) used a mock air warfare target identification and memory task while measuring eye blinks, fixation durations, and mean pupil diameter.

The goals for the TIE were to increase the number of different sensor technologies substantially, demonstrate their effective physical combination on the body of participants, evaluate their ability to detect changes in cognitive activity, and demonstrate their ability to detect these changes in near real time. Achieving these goals required the development of a rich, multitasking environment that was reflective of military command and control tasks. The task had to be simple enough for undergraduates to perform without elaborate training. It had to contain multiple modes of input and multiple stages and types of cognitive processes, and it had to be possible to manipulate the amount of workload and cognitive activity across a wide variety of cognitive processes to give the different cognitive state gauges something to measure.

To fulfill these criteria, the authors developed a personal-computer-based “Warship Commander Task” (WCT; St. John, Kobus, & Morrison, 2002). Participants monitored aircraft on a geographical display for their levels of threat and responded to the threatening ones according to explicit rules of engagement, and they simultaneously monitored ship communications for ship status information. This task was based on previous mock air warfare tasks (Ballas, Heitmeyer, & Perez, 1992; Van Orden et al., 2001), although the pace is faster and the task is more complex in the WCT.

2. METHOD

2.1. Participants

The participants were five men and three women ranging in age from 22 to 47 years ($M = 30.1$ yrs, $SD = 8.6$). Four of the participants were undergraduate students at a local university who were employed by Pacific Science & Engineering Group (PSE) part time, and four were full-time employees of PSE or allied companies. Participants varied in their experience with the task ranging from roughly 2 hr to over 100 hr of practice.

2.2. Task.

The WCT consists of a primary “Airspace Monitoring” task and a secondary “Verbal-Memory” task. The Airspace Monitoring task was designed to manipulate a variety of aspects of cognitive activity including perception, motor activity, memory, attention, and decision making. It consisted of four 15-min scenarios during which aircraft (tracks) entered the display in a series of waves of varying numbers of tracks. There were 12 waves in each scenario, and each wave lasted 75 sec. At the beginning of each wave, tracks appeared from the north (top) of the map at random intervals. The tracks changed direction randomly, but generally headed south (bottom). The tracks moved fairly rapidly, 25% of the tracks in each wave moved 24 pixels per second, and 75% of the tracks in a wave moved 12 pixels per second. The speed of each track in a wave was chosen randomly. At the end of each wave, a bell sounded, all the tracks still present on the map were removed, and participants waited for the next wave to begin.

Tracks initially appeared on the screen as white planes. Participants selected tracks by clicking on them using the mouse. When they selected a track and then clicked on the “IFF” (identify friend or foe) button located beneath the map, the track changed color to blue, red, or yellow. Blue tracks signified friendly aircraft and could be ignored. Red tracks signified enemy aircraft and were to be engaged if they crossed south of the Line of Engagement (LOE). Participants engaged a track by selecting it and then clicking on the “Engage” button located beneath the display. Figure 1 shows a screenshot of the Airspace Monitoring task.

Yellow tracks required more monitoring and processing by participants. Following the IFF, participants had to wait 2 sec before taking further action. This wait introduced a time delay that forced participants to multitask because there were too many pressing tasks to simply wait for 2 sec. After 2 sec, participants needed to

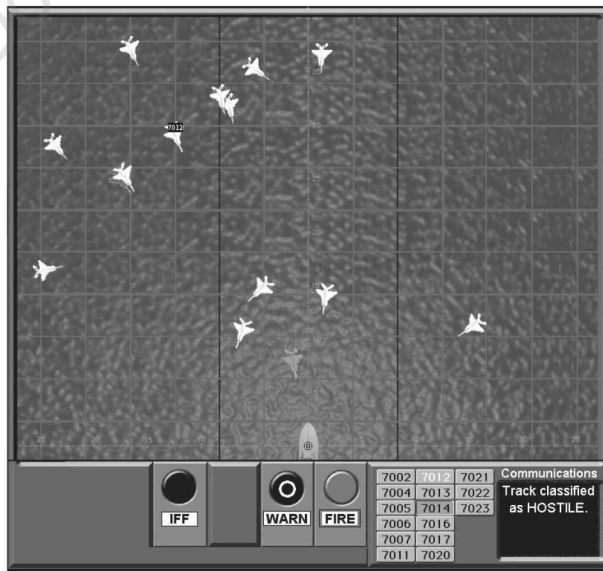


FIGURE 1 Screenshot of the Airspace Monitoring task in the Warship Commander Task.

click on the track's four-digit "track number" located in a list beneath the map which triggered a text message to be displayed adjacent to the track number. The message indicated whether the yellow track was threatening or nonthreatening. This message was removed when participants selected any other track on the map and reappeared whenever participants selected the appropriate track number again. This design was intentional to require participants to memorize which yellow tracks were threatening and which were nonthreatening, because there were no explicit markers on the tracks. This memory burden increased the demands made on the participants.

Nonthreatening tracks could be ignored, but threatening tracks had to be warned and then engaged if they crossed south of the LOE. Tracks could be warned by selecting the track and then clicking on the "Warn" button located beneath the map. Participants needed to observe tracks for 3 sec after providing a warning. If the track continued to head south, participants were to engage the track by selecting it and clicking the Engage button. If the track turned north, participants had to continue to monitor the track and engage the track if it again turned south and crossed the LOE. The 3-sec waiting period introduced another delay that participants had to manage and again forced participants to multitask.

Error tones that were specific to different types of errors were produced whenever participants committed errors. These tones served as feedback to the participants, indicating they were performing inappropriate actions. For example, engaging a yellow track prior to waiting for 3 sec following a warning caused an error sound to play, wherein the engagement did not occur.

A cumulative "game score" for correct actions was displayed at all times beneath the map to motivate participants to obtain a high score. Important actions, such as engaging a track, were given more weight in the score. The score was cumulative across the waves of each scenario.

During the TIE experiment, participants used "keyboard shortcuts" for the IFF, Warn, and Engage buttons rather than the on-screen mouse buttons. Shortcuts were used to reduce the amount of mouse movement, thus reducing motor artifacts in the EEG and fNIR data. The shortcuts were also somewhat faster and easier to use than the on-screen mouse buttons, which reduced the burdens on the participants. However, the task remained difficult at higher levels of task load.

In the Verbal-Memory task, participants periodically received auditory messages regarding the status of three ship systems: current fresh water level, course, and radio channel. The messages were prerecorded using a different voice reporting the state of each ship system, for example, "Fresh water level is five five zero." There were six messages per wave. The first two provided status information, and the third was a query from the captain asking for the status of a system. The fourth and fifth provided more information, and the sixth was another query. As the captain asked a question, four response choices appeared on the left side of the display. Participants used the keyboard numbers (1 through 4) to make their answers. Participants had 10 sec to respond, and they could respond any time after the captain began speaking. Participants received auditory feedback for correct and incorrect answers and time-outs.

Task load, the objective equivalent of cognitive activity and workload, was manipulated in three ways. The “Number of Tracks per Wave” varied the number of tracks that appeared during each wave within a scenario: either 6, 12, 18, or 24 tracks. The more tracks appearing in a wave, the more processing, of all types, a participant had to perform to monitor and respond to each of them. Each scenario contained three repetitions of each level of wave size in the following order (6, 18, 12, 24, 6, 18, 12, 24, 6, 18, 12, 24). Scenarios varied in the individual behavior of each of the tracks and were used to eliminate the possibility of a participant memorizing patterns of track behavior.

“Track Difficulty,” which varied between scenarios, varied the proportion of ambiguously threatening “yellow” tracks appearing within each wave of a scenario. These ambiguously threatening tracks required closer monitoring and more complex processing than other tracks. Two of the four scenarios had a high proportion of yellow tracks within each wave (67%), and the other two scenarios had a low proportion of yellow tracks within each wave (33%).

The third task load factor, which also varied between scenarios, was the presence or absence throughout a scenario of the Verbal-Memory task. Two scenarios were run with the Verbal-Memory task “on,” and two scenarios were run with it “off.”

2.3. Procedure.

Participants were trained and practiced on the WCT prior to the TIE. Each data collection session began by connecting the gauge hardware to a participant. This procedure required from 30 min to an hour, depending on the team and equipment. Some gauges required a short calibration session once the equipment was attached. The four test scenarios required approximately 15 min per scenario for a total of 1 hr to complete all four.

For each team, the WCT was presented on a 17-in. color monitor with a screen resolution of 1024 × 768 pixels.

2.4. Data collection.

The Warship Commander software automatically recorded scenario events, user response times, and errors in real time. The software reported all user and task events in real time to parallel, serial, and Ethernet ports. This real-time reporting allowed experimenters to synchronize (time lock) experiment and user events with events recorded from external devices, such as EEG, eye tracking, and GSR with near-millisecond accuracy.

2.5. Gauges.

The gauges are described in the Appendix. For additional details, see the complete TIE report (St. John, Kobus, & Morrison, 2003).

2.6. Design.

The experiment used a $4 \times 2 \times 2$ repeated measures design, with four levels of Number of Tracks per Wave (6, 12, 18, and 24), two levels of Track Difficulty (Low, High), and two levels of the Verbal-Memory task (On, Off).

3. RESULTS

3.1. Task Performance

To set the stage for evaluating each of the cognitive state gauges, first we confirmed that all three of the task load factors influenced the participants' task performance during the TIE. A number of different response time and error measures were computed for each participant. Here, we report two main measures, RTIFF and Percent Game Score. RTIFF was the mean response time per wave to detect, select with the mouse, and click the IFF button to identify tracks as they first appeared on the screen. The Percent Game Score was the percentage of possible game points attained during each wave of a scenario.

For each scenario, the mean RTIFF for waves of 6, 12, 18, and 24 tracks were computed for each participant. Then, for each team, separate three-way repeated measures Analyses of Variance (ANOVAs) were conducted using the three task load factors of Number of Tracks per Wave (6, 12, 18, and 24), Track Difficulty (high and low), and Verbal-Memory task (on and off). Last, an overall ANOVA was conducted by pooling the data for each participant across teams. For each ANOVA, appropriate Greenhouse–Geisser adjustments to the degrees of freedom were made if sphericity was violated.

All three of the task load factors significantly affected RTIFF—greater numbers of tracks per wave increased RTIFF, higher proportions of difficult tracks in each wave increased RTIFF, and adding the secondary Verbal-Memory task increased RTIFF ($ps < .05$). These effects held true for the overall analysis and each of the four teams (except for team 3 on the verbal-memory factor). Furthermore, in the overall analysis, we tested whether the four levels of wave size (6, 12, 18, and 24) were significantly different from each other. Each increment to wave size significantly increased the RTIFF.

The Percent Game Scores were analyzed by the same method. Again, all three of the task load factors significantly affected the Percent Game Score. Again, these effects held true for the overall analysis and for each of the four teams (except for team 4 on the verbal-memory factor). In the overall analysis, waves of 12, 18, and 24 were significantly different from one another. The overall results are illustrated in Figure 2.

These results indicate that the three task load factors were strong enough to significantly affect task performance. They also indicate that during the TIE, the particular participants and data collection sessions run by each team were sensitive to each of the task load factors (one or both of the performance measures).

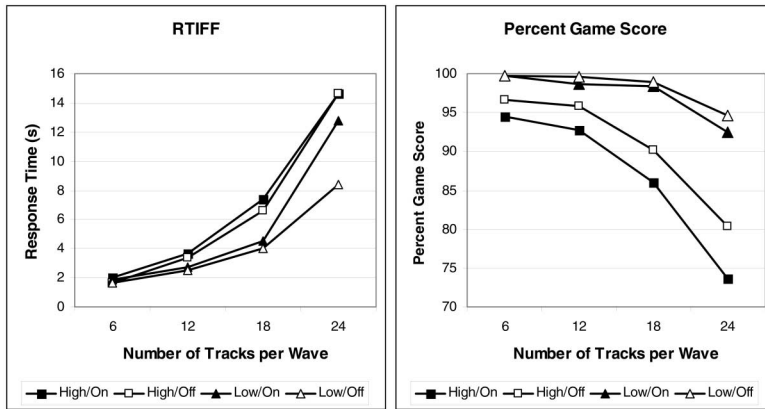


FIGURE 2 Effects on two performance measures of the Number of Tracks per Wave (6, 12, 18, and 24), Track Difficulty (High and Low), and Verbal-Memory task (On and Off).

3.2. Cognitive State Gauge Performance

Turning to the gauges, we evaluated the ability of each gauge to detect changes in cognitive activity as each of the three task load factors varied. Each gauge was analyzed by the same method as the performance measures. Although the within-participant design of this experiment adds a good deal of power to the analysis, it also requires complete data sets from each participant for each experimental session (or the estimation of values for the empty cells). Consequently, incomplete sessions were removed from these analyses. This reduction resulted in small sample sizes for some gauges that limited the conclusions that could be drawn from the analyses. The results are summarized in Table 1.

For each aspect of task load, a filled black circle in a column indicates that the gauge was statistically sensitive to changes in that specific task load factor ($p < .05$). A half-filled black circle indicates that a gauge was “marginally” sensitive to changes in that task load factor ($p < .10$). Given the limitations of sample size and the complexity of the multi-apparatus data collection sessions, as well as the experimental nature of many of the gauges, we felt that reporting these marginal effects was important.

In all, 11 gauges showed statistically significant effects for one or more of the task load factors. Two additional gauges showed promising, marginally significant effects. In drawing conclusions from these results, it is important to understand several points. First, positive results indicate that a gauge was successful at detecting changes in the factors that were manipulated in the task. It is likely that these gauges will be successful in similar fast-paced, command and control-type tasks, and perhaps in other types of tasks as well.

Second, negative results do not necessarily indicate a “failure.” The assessment performed during the TIE involved one task and one context and a relatively small sample size. The data collection environment might have been too noisy for the

gauge, or the small sample size might not have contained sufficient statistical power to reveal the sensitivity of a gauge. Furthermore, due to the rapid development of some gauges, the TIE may have been the first attempt to use them on tasks that differed from those used during their development. There also may have been significant individual differences among participants that require the optimization of various sensor technologies and gauge processing algorithms.

More important, a gauge might be sensitive to some aspects of cognition, but not to the specific cognitive task factors that were manipulated by the WCT. For example, in the WCT, the consequences of error are not severe. Further, participants had limited time to acclimate to the myriad combinations of sensors required for the current state of development of some gauges. As a result, it is reasonable to hypothesize that a gauge that measured the stress induced by severe performance anxiety might not react in the WCT, or be sensitive under the necessary test conditions of the TIE.

In sum, conclusions from these results must be viewed within the context of the TIE test conditions and the test task; generalization to other tasks and other situations must be drawn with care. Nonetheless, given that many of the gauges were very early prototypes that were previously unproven, these results are extremely encouraging.

3.3. Gauge Consistency Across Participants

The analyses discussed earlier address the question of the overall sensitivity of each gauge. A further question is how well, and how consistently, each gauge detects changes in task load for each participant individually: Is a gauge equally sensitive for all participants, or is it sensitive for some participants but not others?

To answer this question, we first computed the correlation between gauge values and the Number of Tracks per Wave for each scenario for every participant. Then, the mean correlation across scenarios was computed for each participant. This correlation provided a measure of gauge sensitivity for each participant. Only the Number of Tracks per Wave factor was examined in this analysis because this factor varied from very low task load to very high task load, and many gauges were able to detect changes in it. Finally, the percentage of participants that showed at least a moderately sized mean correlation was computed. A moderately sized mean correlation was defined to be greater than 0.30. These percentages are listed in the final column of Table I.

Drexel University's fNIR gauge of the left hemisphere was substantially sensitive to changes in task load for all but one participant, although the degree of sensitivity varied from participant to participant. Advanced Brain Monitoring Inc.'s EEG-based vigilance measures showed a similar pattern of high sensitivity to changes in task load for some participants, moderate and more variable sensitivity for other participants, and poor sensitivity for yet other participants. Other gauges, such as QinetiQ's EEG-Based Executive Load gauge and the University of Hawaii's Cognitive Difficulty gauge were consistently highly sensitive. These results are very encouraging, but they also suggest that one limitation to the feasibility of ap-

plying these technologies in operational settings may be differences in gauge sensitivity to particular individuals.

4. CONCLUSIONS

The Augmented Cognition TIE brought together a range of non-invasive technologies that are being developed by a number of independent research groups, and allowed the performance of those gauges to be evaluated in the context of common quasi-realistic command and control tasks.

The experiment was innovative in several ways. First, several of the sensors contained emerging technologies and innovative sensor hardware. For example, the gauge developed by Drexel University uses functional fNIR sensors for detecting changes in cortical blood flow. The EEG system of Advanced Brain Monitoring is wireless and therefore very lightweight and mobile. Electrical Geodesics's 128-electrode EEG sensor net significantly increased the number of sensors that could be practically placed on the head, and it was comfortable enough that users could wear it for hours without complaint. The University of Hawaii's pressure mouse detects motor and cognitive workload from users' hand pressure on a mouse, and the University of Pittsburgh/Naval Research Laboratory's newly developed "posture chair" measures changes in body posture related to changing task demands. Several of these technologies are patented or have patents pending.

A second area of innovation lay in the methods used to compute cognitive state information from the raw sensor data. Many of the gauges use novel analytical methods for turning raw sensor data into meaningful cognitive state gauges. For example, the vigilance gauges developed by Advanced Brain Monitoring and the "Executive Load" gauge developed by QinetiQ/University of Bristol depended on complex decomposition, filtering, and recombination of continuous EEG signals. These methods constitute major advances in sophistication from earlier signal processing methods. The Sarnoff/Columbia "Loss Perception" gauge utilized an innovative adaptive neural network technique to improve its identification capabilities over time. The Index of Cognitive Activity (ICA) gauge developed by San Diego State University takes an innovative twist on an older approach by using complex mathematical procedures to measure high frequency changes in pupil dilation. Several other gauges were developed especially for the Augmented Cognition program and the TIE. Several of these gauges are also in the process of being patented. More detail on each of the gauges is available in the complete TIE report (St. John et al., 2003).

This high degree of innovation in gauge development comes with significant risk in terms of construct validity and validation. For some gauges, their theoretical foundations in neuropsychology and their empirical support are well established and documented. However, a handful of gauges were developed or significantly modified specifically for the TIE, and in several cases, after the TIE data collection was completed. These gauges, as a consequence, are not validated against any other task and their theoretical underpinnings and their relations to established cognitive functions are speculative or unknown at this time.

A third area of innovation lay in the attempts to provide real-time computation of cognitive states. Typically, the complex computations required to turn sensor data into meaningful gauge values are performed after an experiment session is completed. Many of the gauge developers made considerable strides in developing computational methods to allow cognitive state detection for individual participants in real time, or near real time, for the first time. These real-time computations of gauge values constitute a significant advance for the field, which is essential to support the successful manipulation of cognitive state—the goal of Phase II of the Augmented Cognition program.

Several of the gauge researchers noted that the relatively complex and fast-paced WCT was accompanied by a high level of motor activity in the form of mouse movements, key presses, and eye movements. Although this high level of activity made the task more realistic of many command and control-type tasks, it complicated the data analysis for several of the gauges. For example, the muscles around the eyes that control eye movements create electrical artifacts that “contaminate” the brain-based electrical signals that some researchers were attempting to measure. To address this problem, several of the gauge researchers, especially those using EEG sensors, had to substantially extend and modify their sensor-data analysis capabilities to identify and parcel out unwanted muscle artifacts. The detection of changes in task load by several of the EEG gauges point to the success of these efforts. However, the quantity and types of user activities are likely to grow substantially as these gauges are introduced into more applied settings, so there will be a continuing need for improved innovative data analysis techniques and artifact decontamination.

More gauges were sensitive to changes in the Number of Tracks per Wave than to the other task load factors. This finding makes sense because the range of task load was much greater for this factor than for the others—it ranged from very low to nearly overwhelming. Several of the more robust gauges were able to detect intermediate levels of task load for this factor. Additionally, two gauges were significantly sensitive to changes in Track Difficulty, and three gauges were significantly sensitive to the presence or absence of the Verbal-Memory task. Although not predicted by any of the research groups, these different sensitivities suggest that the gauges may be sensitive to different aspects of cognition and task load. In turn, successful augmentation of cognition may require an integrated suite of gauges that are sensitive to different aspects of cognitive activity. For example, some gauges may specifically focus on detecting levels of executive function whereas others focus more on verbal or auditory function.

As a class of gauges, the “arousal” gauges stood out for their inability to detect changes in any of the three task load factors. Because arousal gauges are perhaps the best understood of the gauges used during the TIE, their inability to detect changes in cognitive activity during the WCT is somewhat surprising. These results suggest that there may have been a mismatch between the cognitive states measured by these gauges and the cognitive states elicited by the task, rather than that the gauges themselves were insensitive. It may simply be the case that well-practiced command and control-type tasks do not evoke strong stress responses, and arousal gauges may not be appropriate for measuring changes in

workload in such tasks. However, under more realistic operational conditions, the negative consequences of errors could be profound, and changes in stress levels might be important to detect.

In either case, the ultimate success of arousal-type gauges will depend on their ability to predict changes in participant performance, rather than changes in arousal, *per se*. It is well known that highly trained operators, such as pilots, can be highly aroused or stressed, for example, while landing on an aircraft carrier, with little or no change in their operational performance (e.g., Berkan, 2000; Menza, 2002). It may be that arousal gauges are better suited for monitoring novices during training and for noting how changes in arousal affect human learning. These issues are complex, the research literature is large and varied, and there appear to be many factors that influence the impact of stress on operational performance. More research is required in this area to better understand the relations between task load, stress, and performance outcomes in different types of command and control tasks and different levels of expertise and motivation.

Another class of gauges, the ERP gauges, showed mixed results: some were effective, whereas others were not. Moreover, the development and use of ERP gauges is somewhat problematic in the sense that the user's task must be well understood to identify appropriate task events to measure. It is also necessary to have some means of determining when these events occur during a task. The WCT provided this information to each gauge, but gauges may not have this luxury in real tasks. If these problems can be addressed, then this class of gauge has the potential to measure specific cognitive processes that occur in response to specific task events.

The continuous EEG, fNIR, and ICA gauges, on the other hand, all showed substantial promise for detecting changes in workload. For the TIE, they measured average cognitive activity throughout each wave, but it appears quite possible that they could also measure changes in cognitive activity at much finer time scales. Further, although the EEG gauges, as a group, measured global cognitive functions, such as attention and executive load, there is support for the idea that EEG measures could also be tailored for more specific cognitive processes (e.g., Pleydell-Pearce, Whitecross, & Dickson, 2003).

Beyond the question of detection sensitivity is the question of consistency: Does a gauge consistently and reliably detect changes across trials, across participants, and across experiment conditions? Some gauges proved to be consistently sensitive, whereas other gauges were sensitive for some participants but not others. Still other gauges, unfortunately, were never sensitive. These results, although encouraging, point out that one limitation to the feasibility of applying these technologies in operational settings may be differences in gauge sensitivity to particular individuals.

It is not known at this time what might account for the gauge variability found during the TIE. High variability may have been due to a range of reasons from robustness of the measures, to loose fitting headgear, differences in physiology, and differences in fatigue and distraction during the data collection. Further research may show, for example, that systems must be trained on specific "user-gauge profiles" to control for individual differences, and future improvements in sensor hardware may eliminate some variability. As gauge researchers gain experience

with their innovative gauges and with working in noisier environments, sensitivity and consistency may increase substantially. It may also prove that for certain types of gauges, one suite of sensors may not be universally applicable. Instead, individualized suites of cognitive state gauges may need to be available for different users, or a prescreening process may need to be established to assess the applicability of an augmented cognition system to a specific user performing a specific task in a specific environment.

Finally, the TIE evaluated the practical issue of the ability to combine the sensor hardware into suites of gauges for each team. The most common difficulty arose from the lack of headspace available for multiple sensors and the time required to attach and verify their placement. The development of integrated headgear for multiple sensors should be able to address these concerns. Promising developments include Drexel University's observation that their fNIR sensors on the forehead integrated well with all EEG sensors and SDSU's and ABM's demonstration of integrated EEG/eye-tracking headgear. The introduction of wireless technology for transmitting sensor data to computers is also promising for increasing mobility and reducing weight on the participant. Several developers demonstrated wireless technologies, including ABM, Clemson University, and the University of Pittsburgh/National Research Laboratory.

Improving sensor integration should simultaneously improve the mobility and comfort of sensor suites, which ultimately will be crucial to user acceptance, particularly by war fighters. War fighters cannot be constrained by bulky, uncomfortable equipment that is difficult or tedious to use. Usability will be a critical factor in the successful development of augmented cognition systems in relatively stationary command and control centers, and especially in more mobile environments. Applications where the performer is relatively mobile, such as vehicle operators and foot soldiers, will be orders of magnitude more daunting in their challenges. Many of the gauge and hardware systems are promising in these regards, but this issue will only increase in its importance as the Augmented Cognition program moves forward to more applied settings.

In sum, the TIE results point to the great potential for a number of psychophysiological gauges to sensitively and consistently detect changes in cognitive activity during a relatively complex command and control-type task, and to their practical integration into an effective sensor suite. The goal for Phase II of the Augmented Cognition program will be to take these gauges and incorporate them into systems for demonstrating the manipulation of cognitive states as the basis for augmenting cognition.

REFERENCES

- Ballas, J. A., Heitmeyer, C. L., & Perez, M. A. (1992). *Direct manipulation and intermittent automation in advanced cockpits* (Tech. Rep. No. NRL/FR/5534-92-9375). Washington, DC: Naval Research Laboratory.
- Bastiaansen, M. C., van Berkum, J. J. A., & Hagoort, P. (2002). Syntactic processing modulates the theta rhythm of the human EEG. *Neuroimage*, *17*, 1479-1492.

- Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*, 276–292.
- Berkan, M. M. (2000). Performance decrement under psychological stress. *Human Performance in Extreme Environments*, *5*, 92–97.
- Fournier, L. R., Wilson, G. F., & Swain, C. R. (1999). Electrophysiological, behavioral, and subjective indexes of workload when performing multiple tasks: Manipulations of task difficulty and training. *International Journal of Psychophysiology*, *31*, 129–145.
- Gevins, A., Smith, M. E., McEvoy, L., & Yu, D. (1997). High-resolution EEG mapping of cortical activation related to working memory: Effects of task difficulty, type of processing, and practice. *Cerebral Cortex*, *7*, 374–385.
- Jensen, O., & Tesche, C. D. (2002). Abstract frontal theta activity in humans increases with memory load in working memory task. *European Journal of Neuroscience*, *15*, 1359.
- Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice Hall.
- Levendowski, D. J., Berka, C., Olmstead, R. E., & Jarvik, M. (1999, October). *Correlations between EEG indices of alertness measures of performance and self-reported states while operating a driving simulator*. Paper presented at the 29th annual meeting of the Society for Neuroscience, Miami Beach, FL.
- Levendowski, D. J., Olmstead, R. E., Konstantinovic, Z. R., Berka, C., & Westbrook, P. (2000). Detection of electroencephalographic indices of drowsiness in real time using a multi-level discriminant function analysis. *Sleep*, *23*(Abstract Suppl. 2), A243–A244.
- Levendowski, D. J., Olmstead, R. E., Konstantinovic, Z. R., Davis, G., Lumicao, M. N., et al. (2001). Electroencephalographic indices predict future vulnerability to fatigue induced by sleep deprivation. *Sleep*, *24*(Abstract Suppl.), A243–A244.
- Menza, M. D. (2002, March). The pucker factor. *Approach*, *47*, 8–9. Retrieved June 27, 2003, from <http://www.safetycenter.navy.mil/media/approach/issues/mar02/pucker.htm>
- Mezzacappa, E., Kindlon, D., & Earls, F. (1994). The utility of spectral analytic techniques in the study of the autonomic regulation of beat-to-beat heart rate variability. *International Journal of Methods in Psychiatric Research*, *4*, 29–44.
- St. John, M., Kobus, D. A., Morrison, J. G. (2003). DARPA augmented cognition technical integration experiment (Tech. Rep. No. 1905). San Diego, CA: Space and Naval Warfare System Center.
- Muth, E. R., Koch, K. L., & Stern, R. M. (2000). The significance of autonomic nervous system activity in functional dyspepsia. *Digestive Diseases and Sciences*, *45*, 854–863.
- Picton, T. W., & Hillyard, S. A. (1974). Human auditory evoked potentials: Effects and attention. *Electroencephalography and Clinical Neurophysiology*, *36*, 191–199.
- Pleydell-Pearce, C. W., Whitecross, S. E., & Dickson, B. T. (2003). Multivariate analysis of EEG: Predicting cognition on the basis of frequency decomposition, inter-electrode correlation, coherence, cross phase, and cross power. In *Proceedings of the 36th Annual Hawaii International Conference on System Sciences, USA* (p. 131a). Hawaii: IEEE.
- Rappaport, M., Clifford, J. O., & Winterfield, K. M. (1990). P300 response under active and passive attentional states and uni- and bimodality stimulus presentation conditions. *Journal of Neuropsychiatry and Clinical Neurosciences*, *2*, 399–407.
- Redfern, M. S., Jennings, J. R., Martin, C., & Furman, J. M. (2001). Attention influences sensory integration for postural control older adults. *Gait and Posture*, *14*, 211–216.
- Redfern, M. S., Muller, M. L., Jennings, J. R., & Furman, J. M. (2002). Attentional dynamics in postural control during perturbations in young and older adults. *Journals of Gerontology Series A-Biological Sciences & Medical Sciences*, *57A*, B298–B303.
- Smith, M. E., Gevins, A., Brown, H., Karnik, A., & Du, R. (2001). Monitoring task loading with multivariate EEG measures during complex forms of human–computer interaction. *Human Factors*, *43*, 366–380.

- St. John, M., Kobus, D. A., & Morrison, J. G. (2002). A multi-tasking environment for manipulating and measuring neural correlates of cognitive workload. In J. Persensky, B. Hallbert, & H. Blackman (Eds.), *Proceedings of the 2002 IEEE 7th Conference on Human Factors and Power Plants* (pp. 7.10–7.14). New York: IEEE.
- Tesche, C. D., & Karhu, J. (2000). Theta oscillations index human hippocampal activation during a working memory task. *Proceedings of the National Academy of Sciences, USA*, *18*, 919–924.
- Van Orden, K. F., Limbert, W., Makeig, S., & Jung, T. (2001). Eye activity correlates of workload during a visuospatial memory task. *Human Factors*, *43*, 111–121.
- Villringer, A., & Chance, B. (1997). Non-invasive optical spectroscopy and imaging of human brain function. *Trends in Neuroscience*, *20*, 435–442.

APPENDIX

Gauge Descriptions

Functional Near Infrared imaging (fNIR)—left and right frontal lobes, Drexel University (contact: Scott Bunce). In principle, oxygenated and deoxygenated hemoglobin have characteristic optical properties in the visible and near-infrared light range. Therefore, based on functional optical measurement, concentration changes of these molecules can be measured during functional brain activation. Specifically, the gauge measures hemodynamic responses underneath the forehead, where executive functions, such as attention and working memory, take place. The fNIR brain imaging uses LEDs and photodiodes as the specific sensors to collect data. There is one probe consisting of four LEDs and 10 photodiodes that connect to the participant. There are two measures being used from the data collected at each sensor as an input to the gauge. The first is 3-channel data: NIR wavelengths channel 1, 730 nanometers (nm); channel 2, 850nm; and channel 3, used only for interference. The second is blood oxygenation output calculated using the Modified Beer Lambert Law. The probe placement is fixed on the forehead and the detectors are able to measure the hemodynamic response from 1.5 to 2.0 cm depth in brain tissue (Villringer & Chance, 1997).

Percentage high vigilance, probability low vigilance, Advanced Brain Monitoring, Inc. (contact: Chris Berka). The B-Alert® indexes can quantify changes in vigilance and workload on a second-by-second basis. General arousal level is one of the tonic contributors to the B-Alert indexes, but it is difficult to extract when the task does not manipulate and quantify arousal or control for amount of sleep, time-of-day, or the level of stress and fatigue experienced by the participants during the test sessions. A wireless EEG sensor headset is the specific sensor used to collect data for the B-Alert® EEG indexes. EEG sensors are connected to the participants at Fz, Cz, POz, and mastoids. Electrooculargram (EOG) sensors are placed around the eye. EEG power spectral analysis and computation of the B-Alert® algorithms using regression and discriminant function analyses are being used from the data collected at each sensor (Levendowski et al., 1999, 2000, 2001).

Executive load, QinetiQ, Inc./University of Bristol (contact: Blair Dickson). Changes in the effort required to perform a task are accompanied by changes in the spectral characteristics of EEG recorded across the scalp. In particular, changes in coherence have been demonstrated to provide an index of mental effort. Sensors used to collect data for Executive Load are 14 scalp EEG electrodes and four EOG electrodes. Sensor activity is analyzed in both the time and frequency domain. Measures of EEG activity include spectral analysis, coherence between electrodes, and measures of phase and power between electrodes (Pleydell-Pearce, Whitecross, & Dickson, 2003).

Motor effort, auditory effort, Electrical Geodesics, Inc. (contact: Don Tucker). The theta rhythm has been shown, in both animal and human studies, to be sensitive to cognitive effort. Moreover, the theta rhythm appears to index cortical networks involved in different cognitive tasks (e.g., language demands). A measure of theta averaged .5 sec before and after two "Warship Commander Task" (WCT) events are derived: for Motor Effort, the button press to IFF (identify friend or foe) a track, and for Auditory Effort, the auditory feedback when a track is destroyed. Because the two events represent processing capacity in different domains, the focus of the analysis is on those sensor positions that overlie the somatosensory motor cortex for Motor Effort and over the medial prefrontal cortex for Auditory Effort. Dense-array EEG electrodes (128 channels) are used to collect data. The sensors can be worn comfortably for an indefinite amount of time (Bastiaansen, van Berkum, & Hagoort, 2002; Gevins, Smith, McEvoy, & Yu, 1997).

Loss perception, Sarnoff/Columbia University (contact: Lucas Parra). It is argued that the user's perception of a warning signal along with its negative affect will diminish as task difficulty is increased. Based on prior work on error-related activity, it is hypothesized that differential EEG response to warning signals, as compared to other auditory feedback signals, should represent a measure of the affect associated with loss. Therefore, it is argued that increasing task difficulty should correspond to decreases in intensity of the differential evoked response. Loss Perception consists of 63 channels of EEG, including a few EOG channels used to collect data. Electrodes are placed on the scalp, face, and mastoid. The gauge is based on adaptive linear spatial filtering of the EEG data. The relevant activity is defined as the difference (in time) that is most indicative for a warning signal as compared to control events. The gauge is continually adapting. As more and more events are observed it becomes more and more specific. A practical limitation for the proposed cognitive event detection is that it requires a precise understanding of the task the user is executing. This allows the gauge to identify junctions within the task that will elicit a reproducible evoked response (Picton & Hillyard, 1974; Rappaport, Clifford, & Winterfield, 1990).

Theta power, University of New Mexico (contact: Akaysha Tang). Keeping track of which yellow plane is friend or foe requires working memory. Theta band ac-

tivity within a specified time window in the EEG signal from a specific location within the brain (as opposed to the sensor EEG waves) is associated with increasing working memory demand (Jensen & Tesche, 2002; Tesche & Karhu, 2000).

Arousal Meter, Clemson University (contact: Eric Muth). The Arousal Meter (AM) measures arousal and fatigue. It currently will predict arousal from low (sleep) to active alert. Highly practiced participants in the WCT may have affected the performance of the gauge. Interbeat-intervals, the time between successive heartbeats in milliseconds, is collected at each sensor and used as input to the gauge. Interbeat-intervals are plotted over time and are processed using the Fast-Fourier-Transform (FFT). FFT-derived power is plotted across frequencies to determine the high frequency (HF) peak associated with Peripheral Nervous System activity (between 9 and 30 cycles per min). The mean and standard deviation of the HF peak are continually recalculated. A standardized "arousal" score is derived that drives the AM. Increases in this score are associated with increased autonomic arousal and decreases with decreased autonomic arousal (Mezzacappa, Kindlon, & Earls, 1994; Muth, Koch, & Stern, 2000).

Arousal and stress, University of Hawaii (contact: David Chin). Galvanic skin response (GSR) and infrared oximeter sensors are placed on the participant's toes (two for GSR and one for oximeter). Heart rate from the oximeter and GSR are multiplied and compared to user's calibrated values.

GSR arousal, AnthroTronix, Inc. (contact: Anna Lockherd). The GSR gauge can detect and measure changes in autonomic nervous activity, which is indicative of stress and arousal. However, the GSR gauge cannot distinguish between ANS activity caused by stress and arousal and ANS activity caused by other factors such as fear or embarrassment. A dual-electrode GSR sensor collects GSR skin conductance data. The GSR sensor consists of two electrodes, which are applied to the surface of the skin on the underside of the participant's second and fourth toes. The raw GSR data are averaged over each second. The derivative of the GSR is then calculated from the second-by-second data, and then scaled by 1000.

Head and monitor coupling, head bracing, back bracing, University of Pittsburgh and Naval Research Laboratory (contact: Carey Balaban). Gauges currently being explored include head and monitor coupling, head bracing, and back bracing. The back bracing gauge seems to be sensitive to changes in workload, as measured by changes in tasks. The head and monitor engagement gauge detects a cumulative response to the buildup of tasks pending. In each case, it can produce changes with increments of a single pending task. Further research and analysis will determine which gauges and sensor combinations are the most robust and predictive. The operator's chair is equipped with a 16 × 16 pressure sensor array in the

covers of the seat cushion and back cushion. For head and torso tracking, Flock of Birds sensors are used. One Flock of Bird sensor is attached to the participant's head and one on his or her torso. All other sensors are attached to the operator's chair. Each sensor in the arrays detects pressure and results in a scaled voltage output. Changes in back and seat pressure are detected over time by subtracting consecutive sample collections (sampled four times per second) and then calculating the standard deviation over this set of delta values across the 256 sensors in each array. Head position is determined by the relative position of the Flock of Birds sensors (Redfern, Jennings, Martin, & Furman, 2001; Redfern, Muller, Jennings, & Furman, 2002).

Perceptual and motor load, cognitive difficulty, University of Hawaii (contact: David Chin). The amount of clicking gives a direct indication of the perceptual and motor load in computer tasks that rely heavily on mouse input. A patent-pending pressure mouse is the sensor used to collect data. Due to patent pending, details about the measures being used from the data collected cannot be given. The pressure mouse is also used to measure cognitive difficulty. The waveform of the mouse click produced by users changes when they are thinking. This waveform can be used to judge cognitive difficulty of the task.

Index of Cognitive Activity, San Diego State University (contact: Sandra Marshall). The Index of Cognitive Activity (ICA) measures overall cognitive effort. The ICA is best suited for complex cognitive tasks. It is not sensitive to very simple tasks. The ICA can be displayed in real time, both with the index value and the category level (high, medium, or low). The gauge utilizes two small high-speed cameras to record the size of the pupil in both eyes. Participants are required to wear a headband on which the cameras are mounted. A patented procedure for computing the ICA is applied, yielding an index value for each eye for each second. The procedure uses wavelet analysis to extract the high frequency information from the signal and then applies a statistical threshold (Beatty, 1982; Kahneman, 1973).