TECHNICAL REPORT 1905
December 2003

DARPA Augmented Cognition Technical Integration Experiment (TIE)

SSC San Diego

Approved for public release;
distribution is unlimited.

SSC San Diego

Technical Report 1905
December 2003

DARPA Augmented Cognition Technical Integration Experiment (TIE)

SSC San Diego

Approved for public release;
distribution is unlimited

SPAWAR
Systems Center
San Diego

SSC San Diego
San Diego, CA 92152-5001

SSC SAN DIEGO
San Diego, California 92152-5001

USN
Commanding Officer

.
Executive Director

ADMINISTRATIVE INFORMATION

The work described in this report was performed for the Simulation and Human Systems Technology Division (Code 244) of the Command and Control Department (Code 240) of Space and Naval Warfare Systems Center, San Diego (SSC San Diego) with Pacific Science and Engineering Group, Inc., under contract number N66001-99-D-0050. Funding was provided by the Defense Advanced Research Projects Agency (DARPA), Information Processing Technology Office under program element 0602301E. The DARPA Augmented Cognition program manager was LCDR Dylan Schmorrow.

Released by
SSC SD Principal Investigator

Under authority of
Simulation and Human Systems
Technology Division

This is a work of the United States Government and therefore is not copyrighted. This work may be copied and disseminated without restriction. Many SSC San Diego public release documents are available in electronic format at http://www.spawar.navy.mil/sti/publications/pubs/index.html

LH

ACKNOWLEDGMENTS

No event of this magnitude could have been arranged without the foresight and hard work of numerous organizations and individuals.

Program Management

Defense Advanced Research Projects Agency
LCDR Dylan Schmorrow

Management Support

TLK, Inc.
RADM (ret) Lee Kollmorgen

Strategic Analysis
Amy Kruse, Colby Raley

CACI
Felicia Shands

Technological Integration Experiment

Pacific Science & Engineering Group, Inc.
Mark St. John, David A. Kobus, Michael Quinn, William Walker, Christine Brown, and Martina Savedra

BMH Associates, Inc.
Gary Kollmorgen, Rich Cornwall, David Hamby, and Jeff Harmon

Space and Naval Warfare Systems Center, San Diego
Jeff Morrison

Sonalysts, Inc.
Steve Francis

The Science Advisory Panel
Chair: Jeff Morrison
Panel: LT James Patrey, LCDR Sean Biggerstaff, LT Jeff Alton, and CDR Karl Van Orden

The Cognitive Workload Assessment Developers

Name	Affiliation
Carey Balaban	University of Pittsburgh
Chris Berka	Advanced Brain Monitoring
Scott Bunce	Drexel University
David Chin	University of Hawaii
Joseph Cohn, LT MSC USNR	Naval Research Laboratory

ACKNOWLEDGMENTS (CONTINUED)

The Cognitive Workload Assessment Developers (continued)

Martha Crosby	University of Hawaii
Milenko Cvetinovic	Advanced Brain Monitoring
Jody J. Daniels	Lockheed Martin Advanced Technology Laboratories
Daniel Davenport	Lockheed Martin Advanced Technology Laboratories
Cassandra Davis	San Diego State University
Gene Davis	Advanced Brain Monitoring
Blair Dickson	QinetiQ
Jack Gelfand	Princeton University
Curtis Ikehara	University of Hawaii
Kurtulus Izzetoglu	Drexel University
Frank Keptics	Drexel University
Corinna Lathan	AnthroTronix
Daniel Levendowski	Advanced Brain Monitoring
Anna Lockerd	AnthroTronix
Phan Luu	Electrical Geodesics, Inc.
Sandra Marshall	San Diego State University
Gerry Mayer	Lockheed Martin Advanced Technology Laboratories
Christopher McKinney	University of New Mexico
Eric R. Muth	Clemson University
Banu Onaral	Drexel University
Lucas Parra	Sarnoff Coroporation
Christopher Pleydell-Pearce	University of Bristol
Kambiz Pourrezaei	Drexel University
Rebecca Prichard	Clemson University
Jarad Prinkey	University of Pittsburgh
Mark Redfern	University of Pittsburgh
Tom Renner	Electrical Geodesics, Inc.
Paul Sajda	Columbia University
Andy Swinden	QinetiQ
Akaysha C. Tang	University of New Mexico
Don Tucker	Electrical Geodesics, Inc.
Jack Vice	AnthroTronix
James Weatherhead	Eye Tracking, Inc.
Sharron Whitecross	University of Bristol
Gunay Yurtsever	Drexel University

ABSTRACT

The DARPA Augmented Cognition program is developing innovative technologies that will transform the person–machine interaction by making information systems sensitive to the capabilities and limitations of the human component of the person–machine system. By taking better advantage of individual human capabilities, and being sensitive to human limitations, it is expected that an order of magnitude improvement in system performance can be achieved. There have been many recent advances in the field of Cognitive Science toward understanding human decision-making, and the Augmented Cognition program is taking advantage of them. The technologies developed over the last decade in measuring brain activity and various facets of cognition are serving as the basis for managing the way information is presented to the human operators of complex systems. The Augmented Cognition program will result in demonstrable, quantifiable augmentations to human cognitive ability in realistic operational environments. Towards this goal, the first phase of the Augmented Cognition program was to empirically assess the utility and validity of various psychophysiological measures in dynamically identifying changes in human cognitive activity as decision-makers engaged in cognitive tasks. This report is the culmination of Phase I – *Measuring Cognitive State.* It describes the empirical results of a Technical Integration Experiment (TIE) involving the evaluation of 20 psychophysiologically derived measures (cognitive state gauges) that were developed under Phase I of the Augmented Cognition program. The gauges came from 11 different research groups, and were developed with a variety of theories and scientific backgrounds. The TIE brought these disparate approaches to assessing cognitive state together to be assessed with a common test protocol using a relatively complex cognitive task that was derived from the real world decision-making requirements seen with tactical decision-makers. This task was developed specifically to meet the needs of assessing these very different gauges with necessary empirical controls, yet still maintain the essential character of those tasks from a cognitive perspective as would be found in an operational command and control environment. The results of the TIE assessment discussed in this report concluded that eleven of the gauges successfully identified changes in cognitive activity during the task, and five more gauges showed promise. The report also describes the integration of gauges into suites of gauges to simultaneously measure multiple cognitive indices, and the issues created with sensor technology integration in developing next-generation cognitive state gauges. Additionally, the gauge developers rated the ability of their sensors to integrate with other sensors as fairly high, and most developers reported no problems integrating multiple sensors onto participants. This report summarizes the results found, and attempts to examine the prospects for, and issues that must yet be addressed for, the successful transition of these cognitive state gauges to field-able military person–machine systems in Phase II of the Augmented Cognition program, and beyond.

EXECUTIVE SUMMARY

The Defense Advanced Research Projects Agency (DARPA) Augmented Cognition program is developing technologies capable of extending, by an order of magnitude, the information management capacity of warfighters. This will entail selecting from the myriad of theories and sensor technologies related to the measurement of human cognition developed over the last decade, and marrying them with the many advances in automation and information management. For example, a future C4I system[1] may assign a task to the specific operator having the most unused cognitive capacity, or it may filter information or select the mode or style of its presentation based on a particular operator's available capacity to receive information visually, verbally, or through some other sensory modality.

The objective for the first phase of the Augmented Cognition program, *Measuring Cognitive State,* was to empirically assess the utility and validity of various psychophysiological measures to dynamically identify changes in human cognitive activity during task performance, and explore potential integration and application issues that would need to be addressed during later phases of the program. This report is the culmination of Phase I. It summarizes the empirical results of a Technical Integration Experiment (TIE) that brought together 20 psychophysiological measures (cognitive state gauges) from 11 different research organizations where they were demonstrated and assessed in a common test environment that had the complexity and demand characteristics comparable to those seen by a tactical command decision maker.

The gauges used a wide range of sensor technologies, and they were based on very different, yet sometimes overlapping, theoretical approaches. The sensor technologies included functional Near Infra-Red imaging (fNIR), continuous and event-related electrical encephalography (EEG/ERP), eye tracking and pupil dilation, mouse pressure, body posture, heart rate, and galvanic skin response (GSR). Each of the gauges that was evaluated in the study, the type of sensor it used, and the research organization that developed the gauge are listed in Table A.

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 (1) assess compatibility issues among the different gauge technologies, (2) 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 (3) allow the use of similar sensor technologies that would otherwise compete for access to the same physical locations on test participants.

Coordinating 11 research groups during simultaneous data collection was a major undertaking, and the first attempt to bring so many sensor technologies together at the same time. Over the course of a 4-day period, 3–6 March 2003, all investigators arrived; set up and checked their equipment; configured, calibrated their sensors and algorithms, and conducted the experiment for each participant; and provided a preliminary analysis of their cognitive gauge data.

[1] Command, control, computers, communications, and intelligence

Table A. The 20 gauges evaluated during the TIE.

Gauge	Sensor Type	Developer
fNIR		
fNIR (left)	Blood Oxygenation	DrexelU
fNIR (right)	Blood Oxygenation	DrexelU
EEG-Continuous		
Percent High Vigilance	EEG	ABM
Probability Low Vigilance	EEG	ABM
Executive Load	EEG	QinetiQ
EEG-ERP		
Motor Effort	ERP-IFF	EGI
Auditory Effort	ERP-Engage Sound	EGI
Loss Perception	ERN-Error Sounds	Sarnoff/Columbia
Occular-Frontal Source	ERP-Comms	UNewMexico
Synched Anterior-Posterior	ERP-Comms	UNewMexico
Visual Source	ERP-Comms	UNewMexico
Arousal		
Arousal Meter	Inter-Heart Beat Interval	Clemson U
Arousal	GSR	UHawaii
Arousal	GSR	AnthroTronix
Physiological		
Head-Monitor Coupling	Head Posture	UPitt/NRL
Head Bracing	Body Posture	UPitt/NRL
Back Bracing	Body Posture	UPitt/NRL
Perceptual/Motor Load	Mouse clicks	UHawaii
Cognitive Difficulty	Mouse pressure	UHawaii
Index of Cognitive Activity	Pupil dilation	SDSU

The TIE successfully demonstrated the ability to combine multiple sensors and collect real-time data in an ecologically valid command and control-type decision-making task—which are key requirements of the Augmented Cognition program for transition into Phase II. A key attribute of the TIE was the use of a common experimental test task, under as comparable test conditions as possible. This arrangement allowed us to use a quasi-experimental design[2] for comparing gauges across the data collection teams and for evaluating each of the gauges for their ability to detect changes in cognitive activity as it was manipulated in the common experimental task. The Warship Commander Task (WCT, see Figure A) was designed as a basic analog to a Navy air warfare task. The task was developed to be (1) suitable for use with undergraduate participants, (2) suitable for stimulating as many aspects of cognition as was feasible, and (3) representative of the complex decision-making environments faced by operational warfighters in tactical command centers. Users performed in a series of 15-minute scenarios during which they monitored a varying number of aircraft (tracks) on a display. They evaluated the tracks and determined if and when it was appropriate to warn them, and if necessary, engage them on the basis of explicit rules of engagement. The task was designed to manipulate a variety of aspects of cognitive activity including perception, motor activity, memory, attention, and perceived task load in a semi-realistic command and control-type task.

[2] Though the test conditions were comparable for all the gauge technologies, because the data for all gauges was not collected from the same subjects at the same time, the experimental design is not fully crossed, i.e., it could be argued that uncontrolled factors (e.g., fatigue, time of day, etc.) could confound the interpretation of the results, and preclude the attribution of difference between gauges across the data collection teams to gauges, vice some other factor.

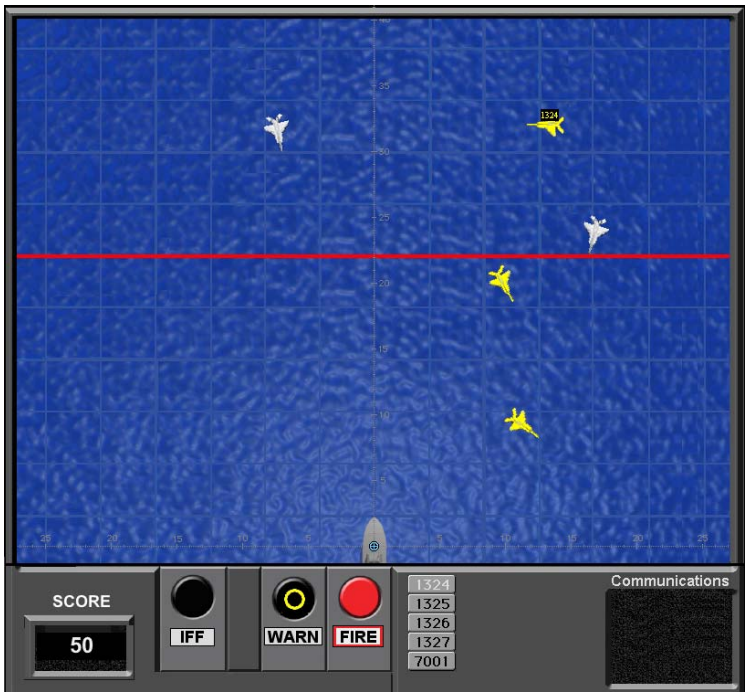


Figure A. Screen shot of the Airspace Monitoring task in Warship Commander.

Figure B provides an illustration of the changing workload demands during the WCT task as perceived by the participants. The pie wedges indicate the proportion of users' activity devoted to each of six dimensions of workload, as defined by the NASA Task Load Index (TLX) (NASA-Ames; Hart & Staveland, 1988)[3]. The left pie chart indicates that during low task load periods of the task, activity on all workload dimensions is low, and users primarily observe and scan the task display. The right pie chart indicates that during high task load periods of the task, temporal and mental demands are high, while other dimensions of workload such as physical demands and frustration remain low, and users have very little time to simply observe the display.[4] The task, however, did not attempt to explicitly manipulate wakefulness/arousal or physical workload, which has implications for the expected diagnosticity of gauges designed to measure those aspects of cognition.

Cognitive activity, or task load, was manipulated through three experimental manipulations during the experiment: (1) Number of Tracks per Wave, which varied from 6 to 24 tracks present on the display during each of 12 waves during the course of each scenario, (2) Track Difficulty, which varied between scenarios according to the proportion of potential threat tracks appearing within every wave (High-67% vs. Low-33%)—which required more actions and decisions than other tracks and were thus more complex, and (3) presence or absence of a concurrent secondary auditory/verbal memory task called the Ship Status Task (on or off), which competed with the primary airspace monitoring task for attentional resources.

---

[3]NASA-Ames. NASA Task Load Index (TLX) V 1.0 Users Manual. Available at http://iac.dtic.mil/hsiac/Products.htm#TLX; Hart, S. G., & Staveland, L. E. (1988). Development of a multi-dimensional workload rating scale: Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human Mental Workload*. Amsterdam, The Netherlands: Elsevier.

[4] The pie wedge proportions are based on a previous task validation study conducted at PSE.

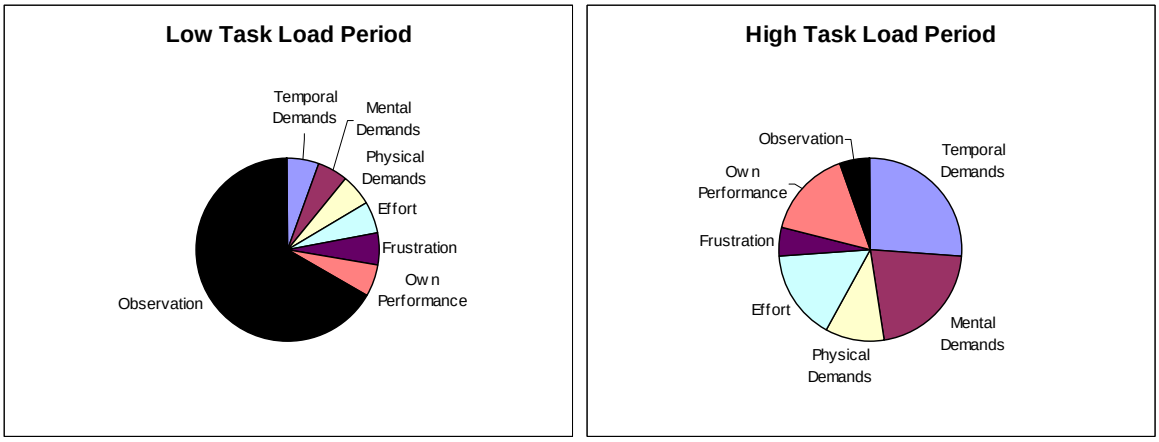


Figure B. Illustration of changing workload demands during the WCT task.

Eleven of the gauges successfully correlated with changes in one or more of the task load factors. Five more gauges showed specific promise for being diagnostic in detecting changes in task load and warrant further development. Since many of the gauges were very early prototypes that were previously unproven, these results are extremely encouraging. 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 similarly successful in tasks that have similar attributes and that are measured under comparable environmental conditions. Specifically, tasks that can be characterized as predominantly involving detection, identification, and memory recall (such as computer-based, fast-paced, command and control-type tasks) and that are presented under similar environmental conditions (such as noise, lighting, and time of day), are likely to show comparable results These gauges may be successful 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. The assessment of such issues was well beyond the scope of the TIE, though this report attempts to explore the issue within the limits of the available data. Consequently, both positive results, and especially, negative results should be interpreted with healthy skepticism.

More importantly, a gauge might be sensitive to 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. The developer appendices address many of these issues in more detail, and the interested reader is encouraged to assess the empirical results for his or herself.[5] In summary, conclusions from these results must be

[5] See Appendix 3, Developer Appendices.

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.

Table B summarizes the overall findings of the experiment. 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 ($p < .05$).[6] A half-filled black circle indicates that a gauge was "*marginally*"[7] sensitive to changes in that task load factor ($p < .10$). A half-filled circle with hash marks indicates that a gauge was "*potentially*" sensitive to changes in that task load factor ($p < .2$). This category was used to indicate gauges that hold some promise for the future. An open circle indicates that a gauge was not sensitive to changes in that task load factor.

The final column of Table B, "Consistency Across Participants," is an indicator of the variability the gauge has shown across the participants. A filled circle indicates a high level of consistency across participants in the degree of sensitivity to changes in task load for that gauge (all participants showed a similar size correlation between gauge value and Number of Tracks per Wave: standard deviation (σ) less than 0.15).[8] In other words, the gauge was equally sensitive (or insensitive) for every participant. A half-filled circle indicates a moderate level of consistency across participants (participants showed moderately different size correlations: $\sigma < 0.30$). An open circle indicates a low level of consistency across participants (participants showed widely different size correlations, $\sigma >$ .30). While some gauges were consistently sensitive for each participant (Hawaii's mouse-based gauges and QinetiQ's EEG-based gauge), the majority of gauges were sensitive for some participants but not others. It will be important, in future development of these gauges, to determine the sources of variability and attempt to control them.

In addition to evaluating the effectiveness of each gauge individually, the TIE evaluated the practical issue of the ability to combine the sensor hardware into useable suites. The gauge "teams" were arranged so that each contained a mix of compatible technologies, although specific assignments were somewhat arbitrary. Following the completion of the TIE, each gauge developer reflected on the challenge of integrating sensor hardware.[9] The developers also rated the ability of their sensor to integrate with other sensors, and specific issues they identified relevant to transition to an operational augmented cognition system. Overall, all developers rated the ease of integration as fairly high, and most developers reported no problems integrating sensors onto participants. For example, the gauges from Clemson University (arousal) and the University of Pittsburgh/National Research Laboratory (head and body posture) were designed to compliment any gauge during the TIE. 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.

---

[6] According to an analysis of variance; see section 4, Results, for details.

[7] Note: the use of the concepts of "*marginally*" and "*potentially*" significant is not consistent with accepted practice in most refereed publications. However, given the intent of this report as a reference for the prospective application of these technologies, the quasi-experimental nature of the experiment design, and the limitations in sample size, the authors feel that it is useful to make these distinctions. We encourage the reader to judge the prospective merits of the various gauges in applying the results to their unique requirements and assess the results accordingly.

[8] See section 4.3, Gauge Consistency,for details.

[9] See section 5, Questionnaire Results and Discussion, for details.

Table B. Summary of Experiment Findings.

Gauge	Sensor Type	Developer	Task Load Factors: Number of Tracks per Wave (6,12,18,24)	Task Load Factors: Track Difficulty (Hi/Lo)	Task Load Factors: Secondary Verbal Task (On/Off)	Consistency Across Participants
fNIR						
fNIR (left)	Blood Oxygenation	DrexelU	●	○	○	◒
fNIR (right)	Blood Oxygenation	DrexelU	●	○	○	◒
EEG-Continuous						
Percent High Vigilance	EEG	ABM	●	⊘	○	◒
Probability Low Vigilance	EEG	ABM	●	○	○	◒
Executive Load	EEG	QinetiQ	●	◒	○	●
EEG-ERP						
Motor Effort	ERP-IFF	EGI	⊘	○	⊘	●
Auditory Effort	ERP-Engage Sound	EGI	○	◒	⊘	◒
Loss Perception	ERN-Error Sounds	Sarnoff/Columbia	⊘	○	●	◒
Occular-Frontal Source	ERP-Comms	UNewMexico	●	○	○	●
Synched Anterior-Posterior	ERP-Comms	UNewMexico	○	○	●	●
Visual Source	ERP-Comms	UNewMexico	○	○	○	●
Arousal						
Arousal Meter	Inter-Heart Beat Interval	Clemson U	○	○	○	●
Arousal	GSR	UHawaii	○	○	○	◒
Arousal	GSR	AnthroTronix	○	○	○	◒
Physiological						
Head-Monitor Coupling	Head Posture	UPitt/NRL	◒	○	○	○
Head Bracing	Body Posture	UPitt/NRL	⊘	○	⊘	◒
Back Bracing	Body Posture	UPitt/NRL	○	⊘	○	◒
Perceptual/Motor Load	Mouse clicks	UHawaii	●	●	○	●
Cognitive Difficulty	Mouse pressure	UHawaii	●	●	○	●
Index of Cognitive Activity	Pupil dilation	SDSU	◒	○	●	○

In summary, the future for integrated, mobile, and comfortable sensor suites appears bright. Again, prospective users of these technologies should look at the detailed descriptions in the full report, and assess prospective transition issues for their proposed application.

The TIE also helped to identify several areas for continuing research as well as development that may prove important to the successful development of an augmented cognition system. One important conceptual issue is the continuing need to define and refine the meaning of each gauge and what it measures. The theoretical constructs being used by researchers from different communities often describe concepts and constructs that significantly overlap. The better these construct definitions, the more accurate, precise, and generalizable each gauge will be. A second important conceptual issue is the need to refine methods for improving the consistency across users with which each gauge measures cognitive activity. There are several approaches for achieving this improvement that are discussed in the conclusion section. A third important conceptual issue is the need to better understand the impact of experience and practice at a task on the ability of each gauge to measure cognitive activity. Some gauges may be more appropriate for different levels of experience. Clearly, the science of cognition would benefit from more comprehensive theories that would provide a common language for both discussing and assessing aspects of cognition in general, and cognitive states in particular. We need to understand how vigilance relates to arousal, how task load relates to work load, as well as what factors limit the ability of decision makers to various kinds of information under different circumstances.

A key practical concern that the TIE experience has highlighted is the need to make the gauge hardware comfortable, mobile, and convenient enough to gain user acceptance. Warfighters cannot be constrained by bulky, uncomfortable equipment that is difficult or tedious to use. Usability is going to be a critical factor in the successful development of augmented cognition systems in relatively stationary command and control center environments, and especially in more mobile environments. Applications where the performer is relatively mobile, such as vehicle operators and soldiers, will be orders of magnitude more daunting in their challenges. Many of the gauge/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. Another practical concern is the need to understand and address potential sources of electro-magnetic frequency (EMF) interference, both between sensors, various bio-amplifiers and with the environmental factors. Several sources of physical and electro-magnetic interference were identified and resolved prior to the TIE. Other interference was noted on an intermittent basis, with no clear source or technical resolution. As we look to the application of these technologies to military environments, it is almost certain that additional sources will appear—operational environments are often noisy and filled with electrical-magnetic interference from many sources. Again, though many improvements in filtering or adapting to this interference have been made, this issue will only grow in importance.

In summary, the TIE results point to the great potential for a number of psychophysiological gauges to sensitively and consistently detect changes in cognitive state (activity) during relatively complex command and control-type tasks and to their practical integration into an effective sensor suite. Phase I of the Augmented Cognition program has achieved its goal of providing a solid foundation for the demonstration of augmented cognition systems. The primary objective of the TIE was to demonstrate the successful integration of multiple psychophysiological gauges to detect changes in cognitive states in real-time. The goal for Phase II will be to take these gauges and incorporate them into systems for demonstrating the manipulation of cognitive states as the basis for augmenting cognition.

CONTENTS

Acknowledgments ..... **i**

Abstract ..... **iii**

Executive Summary ..... **iv**

1. INTRODUCTION **1**

1.1 AUGMENTED COGNITION PROGRAM 1

1.2 TIE EXPERIMENT OVERVIEW 2

1.2.1 Manipulating Cognitive Activity 2

1.2.2 Assessment of Cognitive State Gauges 4

1.2.3 Demonstration of Sensor/Gauge Integration 4

2. METHOD **7**

2.1 PARTICIPANTS 7

2.2 DESIGN 7

2.3 DATA COLLECTION 8

2.4 APPARATUS 9

3. GAUGES AND TEAMING **11**

3.1 TEAMS 11

3.2 TEAM 1 12

3.2.1 Clemson University 13

3.2.2 University of Pittsburgh and Naval Research Laboratory 14

3.2.3 Electrical Geodesics, Inc. 15

3.3 TEAM 2 17

3.3.1 Drexel University 18

3.3.2 Advanced Brain Monitoring 19

3.3.3 University of Hawaii 20

3.4 TEAM 3 22

3.4.1 QinetiQ 23

3.5 TEAM 4 24

3.5.1 Sarnoff Corporation, Princeton University, and Columbia University 25

3.5.2 Lockheed Martin Advanced Technology Laboratories and AnthroTronix 26

3.5.3 University of New Mexico 27

3.6 TEAM SDSU 28

4. RESULTS **29**

4.1 VALIDATING THE TASK LOAD FACTORS 29

4.2 GAUGE EVALUATION 34

4.3 GAUGE CONSISTENCY 49

4.3.1 Clemson University – Arousal Meter 52

4.3.2 University of Pittsburgh/Naval Research Laboratory Head-Monitor Coupling 54

4.3.3 University of Pittsburgh/Naval Research Laboratory – Head Bracing 56
4.3.4 University of Pittsburgh/Naval Research Laboratory – Back Bracing 58
4.3.5 Electrical Geodesics, Inc. – Motor Effort 60
4.3.6 Electrical Geodesics, Inc. – Auditory Effort 62
4.3.7 Drexel University – fNIR (left) 64
4.3.8 Drexel University – fNIR (right) 66
4.3.9 Advanced Brain Monitoring – Percent High Vigilance 68
4.3.10 Advanced Brain Monitoring – Probability Low Vigilance 70
4.3.11 University of Hawaii – Arousal 72
4.3.12 University of Hawaii – Perceptual/Motor Load 74
4.3.13 University of Hawaii – Cognitive Difficulty 76
4.3.14 QinetiQ – Executive Load 78
4.3.15 AnthroTronix – Arousal 80
4.3.16 Sarnoff/Columbia – Loss Perception 82
4.3.17 University of New Mexico – Ocular-Frontal Source 84
4.3.18 University of New Mexico – Synchronized Anterior-Posterior Source 86
4.3.19 University of New Mexico – Visual Source 88
4.3.20 San Diego State University – Index of Cognitive Activity 90

5. QUESTIONNAIRE RESULTS AND DISCUSSION 93

5.1 GAUGE DESCRIPTION QUESTIONAIRE 93
5.2 QUESTIONS AND RESPONSES (TEAM INTEGRATION) 93
5.3 QUESTIONS AND RESPONSES (GAUGE DESCRIPTIONS) 94
5.3.1 Advanced Brain Monitoring, Inc. 101
5.3.2 AnthroTronix, Inc. 102
5.3.3 Clemson University 103
5.3.4 Drexel University 104
5.3.5 Electrical Geodesics, Inc. 106
5.3.6 QinetiQ, Inc 107
5.3.7 San Diego State University 108
5.3.8 Sarnoff 110
5.3.9 University of Pittsburgh and Naval Research Laboratory 111
5.3.10 University of Hawaii 113
5.3.11 University of New Mexico 115
5.4 QUESTIONS AND RESPONSES (SELF-EVALUATION) 116

6. CONCLUSIONS 119

6.1 INNOVATION 119
6.2 ISSUES IN COGNITIVE STATE IDENTIFICATION 121
6.3 TRANSITION AND APPLICATION 124
6.4 SUMMARY 125

APPENDIX 1: WARSHIP COMMANDER TASK ANALYSIS 127

APPENDIX 2: PARTICIPANT COMMENTS 133

APPENDIX 3: DEVELOPER APPENDICES **137**

APPENDIX 4: A COMMENTARY BY ALAN GEVINS AND MICHAEL SMITH **217**

APPENDIX 5: GLOSSARY **221**

FIGURES

1. Conceptual Augmented Cognition architecture. 1
2. Screen shot of the Airspace Monitoring task in Warship Commander. 3
3. A typical AugCog TIE team apparatus layout. Multiple investigators made simultaneous observations of a participant's psychophysiology using a variety of equipment. The participant (sitting) is wearing fNIR sensors, EEG sensors, an eye tracker, toe clips for detecting heart rate, and is using a specially designed " pressure mouse." 11
4. Subjective workload ratings graphed against the number of tracks per wave and split by track difficulty (high or low). 30
5. The mean value of the seven performance measures graphed against the Number of Tracks per Wave and split by Track Difficulty (high and low) and Secondary Verbal Task (on and off). 32
6. The mean response times and percent correct scores on the Secondary Verbal Task as a function of the Number of Tracks per Wave and Track Difficulty (high and low) in the primary task. 34
7. The mean value of each Team 1 gauge graphed against Number of Tracks per Wave, split by Track Difficulty (high and low) and Secondary Verbal Task (on and off). 37
8. The mean value of each Team 2 gauge graphed against Number of Tracks per Wave, split by Track Difficulty (high and low) and Secondary Verbal Task (on and off). 40
9. The mean value of the Team 3 gauge graphed against Number of Tracks per Wave, split by Track Difficulty (high and low) and Secondary Verbal Task (on and off). 41
10. The mean value of Team 4 gauges. 42
11. The mean value of each SDSU gauge graphed against Number of Tracks per Wave, split by Track Difficulty (High and Low) and Secondary Verbal Task (On and Off). 45
12. Clemson Arousal Meter. 53
13. UPitt/NRL Head-Monitor Coupling. 55
14. UPitt/NRL Head Bracing. 57
15. UPitt/NRL Back Bracing 59
16. EGI Motor Effort. 61
17. EGI Auditory Effort. 63
18. Drexel fNIR (left). 65
19. Drexel fNIR (right). 67
20. Advanced Brain Monitoring Percent High Vigilance. 69
21. Advanced Brain Monitoring Probability Low Vigilance. 71
22. UHawaii Arousal. 73
23. UHawaii Perceptual/Motor Load. 75

24. UHawaii Cognitive Difficulty. 77
25. QinetiQ Executive Load. 79
26. AnthroTronix Arousal. 81
27. Sarnoff/Columbia Loss Perception. 83
28. UNew Mexico Ocular-Frontal Source. 85
29. UNew Mexico Synchronized Anterior-Posterior Source. 87
30. UNew Mexico Visual Source. 89
31. SDSU Index of Cognitive Activity. 91

TABLES

1. Participant schedule across the four teams. 7
2. Statistical results of the three task load factors on the WCT performance measures by team and overall. 31
3. Significant interaction results and pairwise comparisons. 33
4. Statistical results of the primary task on performance of the Secondary Verbal Task for each team. 33
5. Statistical results of the three task load factors on each of the Team 1 gauges. 35
6. Statistical results of the three task load factors on each of the Team 2 gauges. 38
7. Statistical results of the three task load factors on the Team 3 gauge. 41
8. Statistical results of the three task load factors on each of the Team 4 gauges. 42
9. Statistical results of the three task load factors on each of the SDSU gauges. 45
10. Effect sizes (η^2) of the statistical results for each specific gauge for each task load factor. 48
11. Variance of the mean correlations for each specific gauge for each task load factor. 51
12. Questions and responses from the CWA developers regarding team integration during the TIE. 95
13. Questions and responses from the CWA developers regarding detailed descriptions of their gauge. 97
14. Questions and responses from the CWA developers (self-evaluation). 117

1. INTRODUCTION

1.1 AUGMENTED COGNITION PROGRAM

The Defense Advanced 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 hope is to increase the effectiveness of combat command and control system operators by managing the information presented and tasks assigned based on the available cognitive capacity of the operator. For example, a C4I (command, control, communications, computers, and intelligence) system may assign a task to the operator with the most unused cognitive capacity or it may filter information or select the mode or style of its presentation based on the operator's available capacity to receive information visually, verbally, or by audible cues.

The anticipated result of the DARPA Augmented Cognition (AugCog) program will be a system to augment the human by providing information in a format more readily assimilated by the human given the state of the human brain and the current tasks being executed by the human. Additionally, it will have the capability to anticipate human task loading (before catastrophic failure) and change information modalities (providing information through a different medium such as aural, spatial, or verbal) or by offloading tasks. Further, the anticipated AugCog systems will augment the human by assisting in task execution through pre-negotiated "crew coordination" either by task sharing between the human and machine or completely offloading certain tasks to machine automation. Figure 1 shows the major components and sub-components of the general AugCog system.

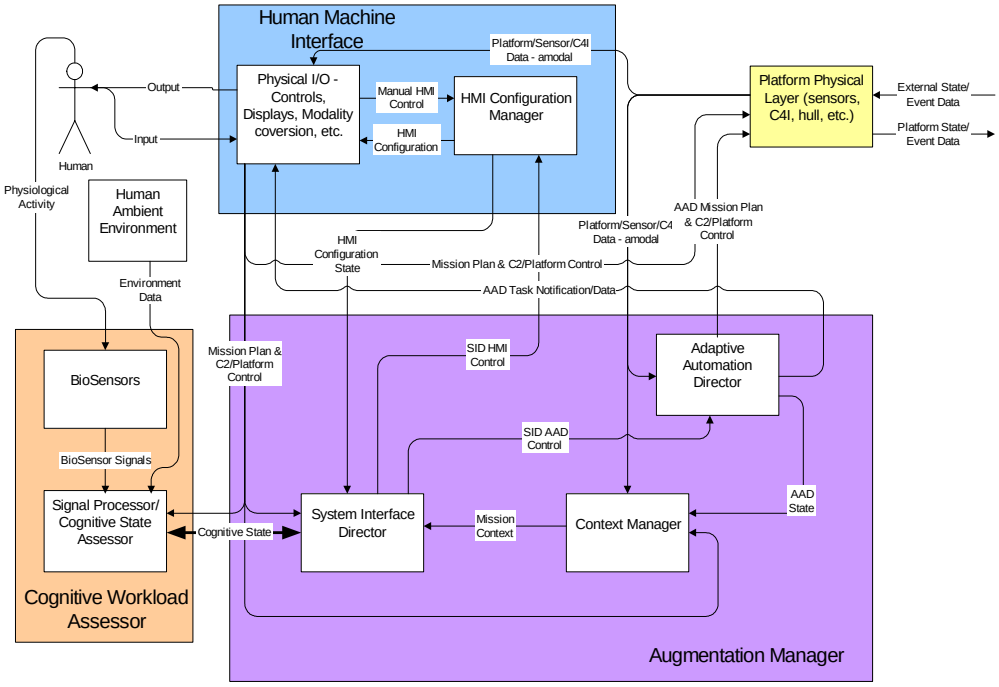


Figure 1. Conceptual Augmented Cognition architecture.

The overall program is divided into several phases. Phase I is a demonstration and evaluation of potential psychophysiological gauges to measure various aspects of cognitive state. The gauges, and

their underlying sensors, measure a variety of physiological and behavioral phenomena and then attempt to identify cognitive state from their basis. A number of the gauges specifically attempt to identify the amount or extent of cognitive workload or activity of a participant. This report is a summary of the results obtained during the Technical Integration Experiment, the final demonstration for Phase I.

Phase II of the Augmented Cognition program will be the application phase in which gauges that demonstrated promise during Phase I will be incorporated into several applied environments. During this phase, gauge outputs will be used to help optimize the human–computer interaction, thereby enhancing operator performance.

1.2 TIE EXPERIMENT OVERVIEW

The Technical Integration Experiment had three goals:

1. Manipulate cognitive activity in a quasi-realistic military task.
2. Assess the ability of each cognitive state gauge to measure the changes in cognitive activity theoretically consistent with that cognitive state.
3. Demonstrate the feasibility of multiple sensor/gauge integration.

1.2.1 Manipulating Cognitive Activity

The AugCog TIE uses an instrumented video game called Warship Commander, which was developed especially for the AugCog program by Pacific Science & Engineering Group (PSE) and SPAWAR Systems Center (SSC San Diego) (St. John, Kobus, & Morrison, 2002),[10] to stimulate the experimental participant. The Warship Commander Task (WCT) was designed to (1) generate the perceptual, memory, verbal, and decision-making demands analogous to those required of a naval tactical decision-maker, yet (2) still be performable by university undergraduate students, and (3) have the necessary software and hardware interface capabilities to support data integration across a variety of measurement systems.

We had numerous criteria for the task environment, summarized as follows:

- Command and control task
- Simple enough for undergraduates to perform
- Engaging for undergraduates
- Multiple channels and modes of input
- Multiple cognitive processes including spatial, verbal, and decision-making
- Multiple "stages" of cognition including information acquisition, analysis, decision selection, and action
- Independent manipulation of workload for each component process
- Amenable to psychophysiological measures of component process workload
- Amenable to modulation of interface and task workload

PSE and SSC San Diego's solution for fulfilling these criteria was a ship-based, quasi-realistic "Warship Commander" task. Warship Commander consists of a primary task of airspace monitoring, plus a secondary task of ship status monitoring. The Airspace Monitoring task requires a combination of spatial, verbal, and decision-making processing, and the ship status task primarily requires

[10] St. John, M., Kobus, D. A., & Morrison, J. G. (2002). A multi-tasking environment for manipulating and measuring neural correlates of cognitive workload. In *Proceedings of the 2002 IEEE 7th Conference on Human Factors and Power Plants*. New York, NY: IEEE. pp 7.10 – 7.14.

auditory-verbal and memory processing. The task, however, did not attempt to manipulate wakefulness-arousal or physical workload.

In the Airspace Monitoring task, displayed in Figure 2, the participant plays the role of a Naval Air Warfare Commander protecting a military convoy located off the screen to the south (bottom of the display). The commander's task is to monitor the airspace and disallow any aircraft determined to be potentially harmful to penetrate south of a "Line of Engagement" (LOE) and pass the ship. The LOE is the horizontal line across the screen as shown in Figure 2. This task involves identifying all aircraft that enter into the airspace, warning threatening aircraft that penetrate the LOE to turn away and leave the airspace within 3 seconds of the warning, and shooting down any threatening aircraft that fail to heed the warning before they can attack own ship or the convoy. A cognitive task analysis of the Warship Commander Task was conducted and is provided in Appendix 1.

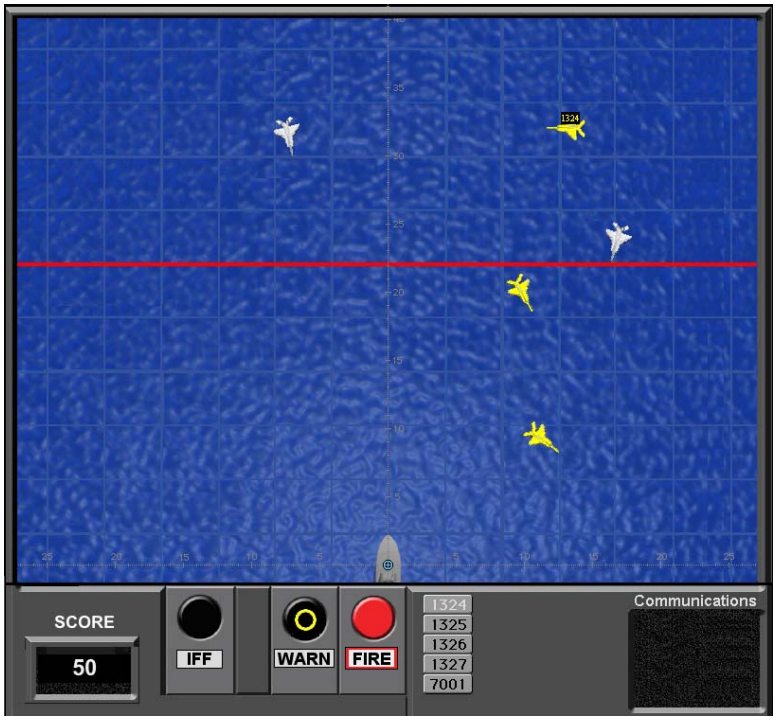


Figure 2. Screen shot of the Airspace Monitoring task in Warship Commander.

Three aspects of task load were manipulated in the experiment:

1. Number of Tracks per Wave – the number of aircraft (tracks) appearing within each consecutive 75-second period of a scenario. There were four levels of this factor: 6, 12, 18 and 24 tracks per wave. Larger waves require more perceptual, motor, decision-making, and memory processing. Our validation and calibration studies indicate that a task load of six aircraft in a 75-second long wave produces a fairly consistent low level of workload for trained users. Larger waves of aircraft produce statistically significant higher subjective workload on participants as measured using the NASA-TLX.
2. Track Difficulty – Some tracks required more work from participants than other tracks. In some scenarios, the proportion of difficult tracks within each wave was low, whereas in other scenarios, the proportion was high. Difficult tracks were initially ambiguous in that there was no way to definitively assess whether they were threatening or neutral. These tracks were assessed only as “Potential Threats,” and were colored yellow on the display. This meant that

additional queries and challenges had to be made for those tracks as they approached own ship and the convoy being protected. Each of these queries/challenges required time to execute, and the results of the queries were not automatically placed on the display when the results were available. Therefore, these tracks placed a significant load on short term memory as well as required awareness of the time required for the different steps necessary before a track could be successfully engaged, creating a significant manipulation of cognitive difficulty and perceived workload.

3. Secondary Verbal Task – The presence or absence of the concurrent Ship Status Task was manipulated. For some scenarios, the Ship Status Task was present, while during others it was absent. When the ship status task was present, participants were asked to listen to a series of audio messages regarding the status of various ship systems. Periodically, the participant was queried about the current status of one of the ship systems. The participant responded by selecting the correct answer from a multiple-choice visual display. The effect of the Ship Status Task was to increase cognitive load by interrupting the primary airspace monitoring task, introduce an audio component to the decision-making, and placed an additional load on short-term memory and resource allocation.

1.2.2 Assessment of Cognitive State Gauges

The gauges were developed at a number of independent laboratories, and they were intended to detect changes in such cognitive states as alertness, vigilance, and cognitive workload, as well as diagnose what task demands are contributing to the cognitive state being measured. The gauges use a wide range of sensor technologies, and they are based on very different theoretical approaches. The sensor technologies include functional Near Infra-Red imaging (fNIR), continuous and event-related electrical encephalography EEG/ERP), eye tracking and pupil dilation, mouse pressure, body posture, heart rate, and galvanic skin response (GSR). Each gauge and research team is described in detail later in this report.

1.2.3 Demonstration of Sensor/Gauge Integration

The objective of the TIE was to bring these gauges together so that they could be compared and evaluated in a common context, and thereby serve as the basis for the DARPA program manager to select which gauges merit continued development and potential integration for U.S. Armed Forces applications in Phase II of the AugCog program.

The scope the TIE included was the demonstration of the Phase I gauges and the showcasing of how each gauge produced measurements to capture ongoing cognitive activity. The TIE was not intended to serve as the basis for gauge development, per se, although we hoped it would in fact afford an opportunity for individual developers to identify potential improvements and integration possibilities. In order to allow different gauges to be compared as directly as possible, it was necessary to instrument participants with as many gauges as practical at one time, as well as test the same participants with each combination of gauges under consistent conditions. The TIE therefore combined several gauges into working teams, and then tested participants on the same WCT scenarios for each team in turn. By using these combinations of gauges, it was possible to cross-validate their results through a repeated measures experimental design and verify to what extent each gauge was responding to cognitive activity. Further, attempts were made to minimize potential confounds due to physical activity, agitation, or arousal that might occur along with decision-making in complex, time-pressured tasks. The experiment design allowed the comparison of one or more emerging cognitive state detection techniques, and comparison of data collected by underlying sensor technologies.

Four separate teams participated in the TIE, each integrating specific prototype devices and their associated gauge outputs. Teams were set up in order to minimize potential sensor interference and maximize the mix of devices available to each team. The four TIE teams and the individual gauges and component technologies are described in section 3. Additional detail is provided in the individual developer appendices.

The TIE was conducted at the facilities of Pacific Science and Engineering (PSE) located at 6310 Greenwich Drive, Suite 200, San Diego, CA 92122. Coordinating 11 research groups during simultaneous data collection was a major undertaking. Over the course of a 4-day period, March 3-6, 2003, all experimenters arrived; set up and checked their equipment; configured, calibrated, and conducted the experiment for each participant; broke down and repacked their equipment; and provided a preliminary analysis of their data. The official data collection experiment was performed following a preliminary experiment (pilot study or Pre-TIE) conducted in February 2003 to reduce many of the technical and procedural issues that may have impacted the running of the TIE. A number of lessons concerning integration of sensors and logistics were learned during the pre-TIE and put to good use at the TIE event.

2. METHOD

2.1 PARTICIPANTS

There were eight official participants plus two additional participants who were available in case of problems. Participants were five males and three females ranging in age from 22 to 47 years (*M* = 30.1 yrs, *SD* = 8.6). Participants varied in their experience with the task, ranging from roughly 2 hours to over 100 hours of practice. Therefore, there was a substantial range of expertise in the task among participants.

Each of the eight participants attempted to visit each of the four teams. During each session, a participant visited a team and performed in a series of four Warship Commander Task (WCT) Scenarios. Table 1 shows the schedule of how participants were assigned to each of the four teams. Numbers in the table reflect the participant ID number. Teams started their data collection runs using participants listed in the first row under each team. After, the first participant completed the session, teams would collect data using the participant listed in the second row and continued down the list for additional sessions.

Table 1. Participant schedule across the four teams.

Team 1	Team 2	Team 3	Team 4
1	2	3	4
5	6	7	8
2	3	4	1
6	7	8	5
3	4	1	2
7	8	5	6
4	1	2	3
8	5	6	7

Across the four performer teams, there were a total of 25 completed sessions (out of 32 possible). Two sessions were excluded because inappropriate task conditions were run on some scenarios (Team 1), and one session was excluded because the participant stopped performing due to pain from the equipment (Team 4). Four sessions were never performed because the teams involved were unable to complete data collection within the timeframe of the TIE event (two for Team 3 and two for Team 4).

2.2 DESIGN

Each data collection session manipulated all three aspects of task load: (1) Number of Tracks per Wave (6, 12, 18, and 24 tracks), (2) Track Difficulty (high or low), and (3) Secondary Auditory/Verbal Memory Task (on or off).

The number of tracks per wave was manipulated within each scenario. Each scenario consisted of 12 waves, and the number of tracks appearing in each wave was manipulated in the following way

across the entire scenario: 6, 18, 12, 24, 6, 18, 12, 24, 6, 18, 12, 24. This arrangement provided three repetitions of each of the four levels of tracks per wave.

Track difficulty was manipulated between scenarios. Yellow tracks require more processing and responses from the participant than other types of tracks. The number of yellow tracks, therefore, could be varied and influence the level of difficulty within a scenario. 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%). For example, at high track difficulty, a wave of six tracks would consist of four yellow tracks, one blue track, and one red track, and a wave of 24 tracks would consist of 16 yellow tracks, four blue tracks, and four red tracks.

The third task load factor was the presence or absence of a secondary verbal task called the Ship Status Task. This factor was also manipulated between scenarios. The task required auditory/verbal processing and memory. Two scenarios were run with the Ship Status Task "on," and two scenarios were run with the Ship Status Task "off."

Therefore, each data collection session was designed to allow developers to collect data during four different scenarios with each participant. Number of tracks varied within each scenario. Task difficulty and Ship Status Task factors varied across scenarios and were crossed to create a 2 x 2 design across the four scenarios performed by a participant during each experiment session. A balanced Latin Square design was used to determine the ordering of scenario conditions within each data collection session.[11] Hence, each of the task load factors was manipulated within each experiment session in a 2 x 2 x 4 repeated measures design.

2.3 DATA COLLECTION

The Warship Commander software automatically recorded scenario events, user response time, and errors in real time. The software computed performance measures for each wave. The wave-by-wave measures were

- RTIFF – the mean time from when tracks appeared on the screen until the participant selected each track and pressed the IFF button.
- RTWarn – the mean time from when tracks crossed the LOE and became eligible for warning until the participant selected each track and pressed the Warn button.
- RTEngage – the mean time from when tracks became eligible for engagement until the participant selected each track and pressed the Engage button.
- Percent Game Score (PctGS) – the percent of total game points for a wave that a participant was able to accumulate.
- Errors of Commission (EC) – Number of errors committed during a wave.
- Errors of Omission (EO) – Number of tasks neglected during a wave.
- Tasks Pending (Pending) – Sum of tasks pending across each second of a wave.

For the Ship Status Task, response times and percent correct responses were also computed and recorded.

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

[11] The balanced Latin Square is a well-known method for arranging the order of conditions within a data collection session. It ensures that experiment conditions are presented in all possible orders.

near-millisecond accuracy. This elaborate data logging and synchronization process allowed experimenters to investigate fine-grained cognitive processing of individual events during the task.

2.4 APPARATUS

Each of the four teams used an identical computer system for presenting the WCT to participants. The computer system included a 17-inch color monitor with a screen resolution of 1024 x 768 pixels, 2.4-GHz processor, 512-MB RAM, 80-GB hard drive, 32-MB VGA/DVI graphics card, sound card with two independent channels, mouse, and keyboard. Each computer system was powered through an IEC 60601-1 compliant isolation transformer.

3. GAUGES AND TEAMING

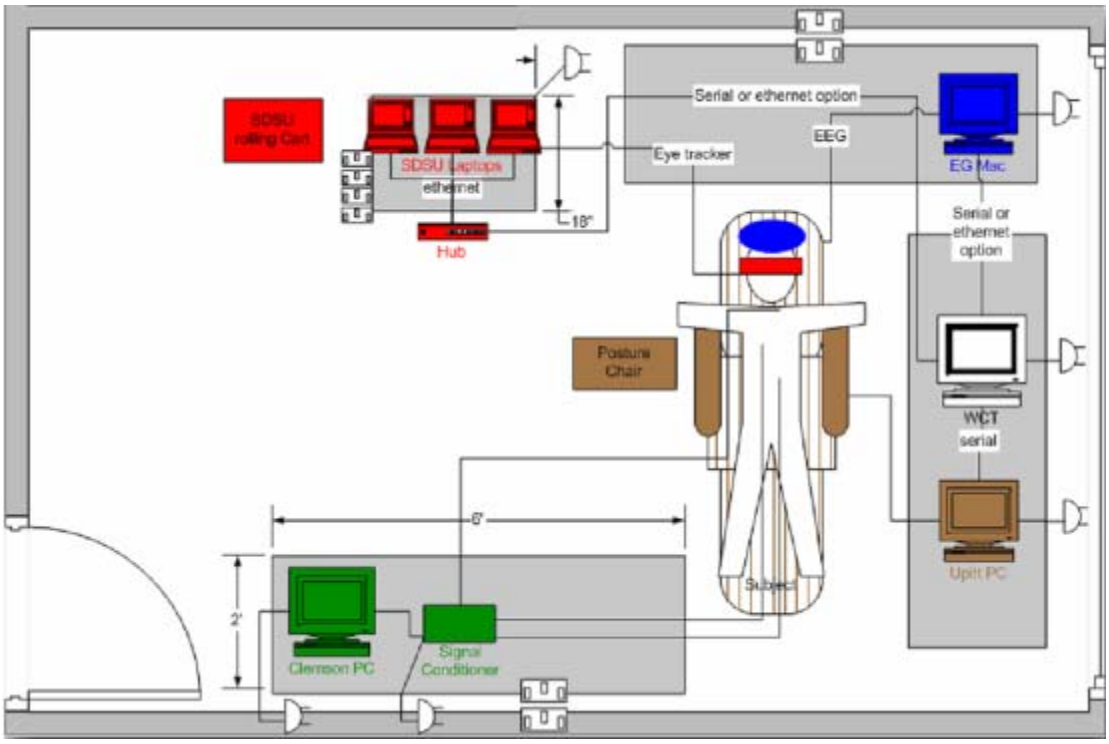
3.1 TEAMS

Although the main goal of the TIE was to correlate gauge responses to the Warship Commander stimuli, a secondary goal was to determine complementary sensor-gauge combinations. Accordingly, investigators were divided into four teams based on a number of criteria. Each team consisted of one EEG investigator and two to three investigators whose measures were hardware compatible with each other. Moreover, team composition reflected electrical compatibility of hardware, researcher preference, and the physical limitations of placing equipment on participants. An example of a team apparatus layout is shown in Figure 3. The San Diego State University (SDSU) eye-tracking group collected data with each team for several data collection sessions. The four teams, and the gauges used by each, are described briefly below.

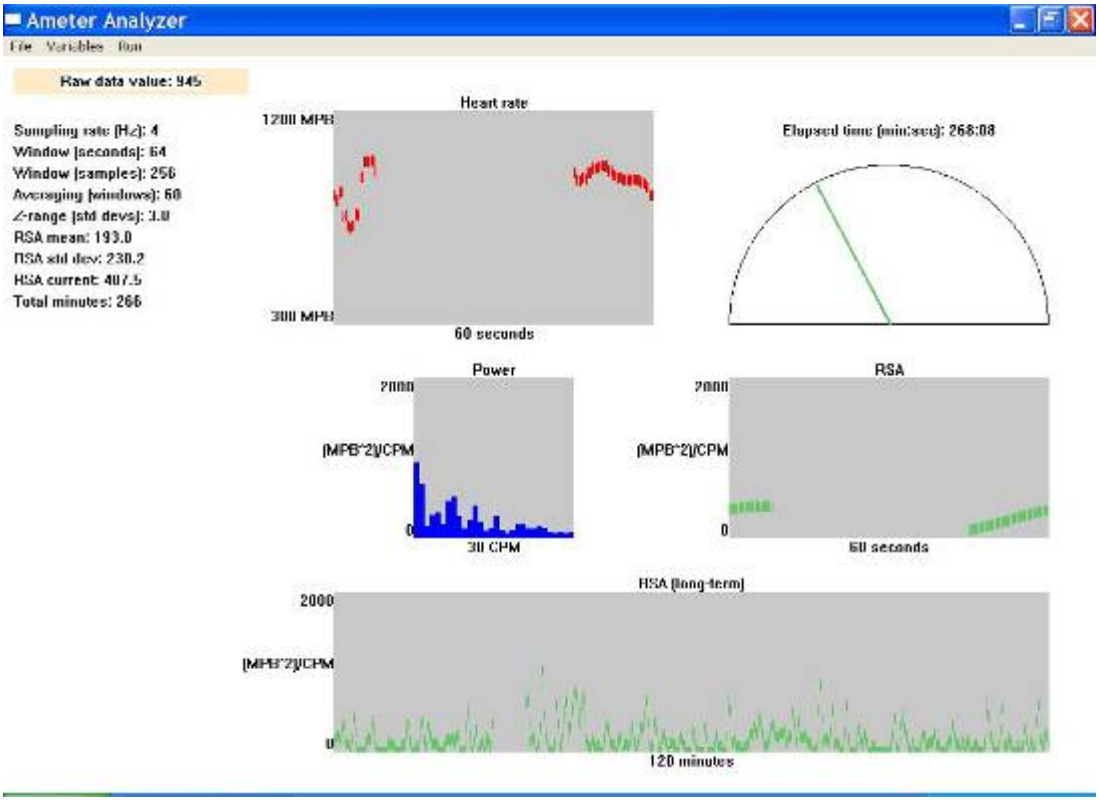
Figure 3. A typical AugCog TIE team apparatus layout. Multiple investigators made simultaneous observations of a participant's psychophysiology using a variety of equipment. The participant (sitting) is wearing fNIR sensors, EEG sensors, an eye tracker, toe clips for detecting heart rate, and is using a specially designed "pressure mouse."

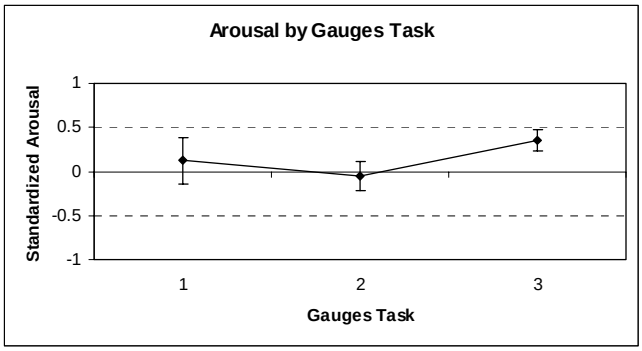
3.2 TEAM 1

Gauge and Technology	Experiment Team	Contact	Organization
Dense array EEG, including alpha, theta, and P3a	3 researchers	Don Tucker	Electrical Geodesics
Assess changes in "immersion" or cognitive engagement (nothing attached to participant)	5 researchers	Carey Balaban	Dept of Otolaryngology, University of Pittsburgh
Arousal Meter (Electrodes to chest)	3 researchers	Eric Muth	Clemson University



3.2.1 Clemson University





Gauge Name: Arousal Meter

Gauge Description

Inter-beat-intervals, the time between successive heartbeats in milliseconds, is collected at each sensor and used as input to the gauge. When state changes are detected, the resolution is second by second, but in reality, real arousal changes may have a time constant of 30 seconds to 1 minute. The potential or practical limit for its sensitivity is 30 seconds to 1 minute based on physiology, not the gauge itself. Arousal and fatigue are the cognitive states the Arousal Meter (AM) measures. The gauge does not predict cognitive state, per se, but will predict arousal. It currently will predict arousal from low (sleep) to active alert. There is work being done to improve the gauge so that it will predict higher states of arousal (e.g., terrified, excited, etc.). There is no post-processing required to calculate

the gauge measurement. However, because of the stage of the gauge development, post-processing is required to reduce data from second-by-second data to wave-by-wave data. Highly practiced participants may have affected the performance of the gauge.

Inter-beat-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 (PNS) activity (between 9 and 30 cycles per minute). The mean and standard deviation of the HF peak are continually re-calculated. A standardized "arousal" score is derived that drives the Arousal Meter (AM). Increases in this score are associated with increased autonomic arousal and decreases with decreased autonomic arousal.

The person to be monitored needs to be connected to the unit via three electrode leads. Two active recording leads (black) are connected; one on the person's right side, just below the collarbone and one the left side just below the left breast. These two electrodes are connected to field effect transistors (Fetrodes) that serve as amplifiers and increase the signal to noise ratio. These leads are positioned to minimize electrode movement and be in line with the major vector of depolarization of the heart. The third lead serves as a reference for signal noise reduction. Actual length necessary for calibration is dependent on whether the gauge is calibrated for specific tasks or if it is calibrated to a person's life. If calibration is task specific, it is likely that a minimum amount of time between 15 to 30 minutes performing the task will adequately calibrate the gauge. If life calibration is required, longer wearing of the gauge is required. Exact time required to reach some sort of calibrated stability remains to be determined through research. The gauge is designed as a long-term wearable device so calibration should not be an issue. There are no constraints placed on the operator when wearing the sensor other than that the wearer cannot get the device wet.

3.2.2 University of Pittsburgh and Naval Research Laboratory

Gauge Names: Head/Monitor Coupling, Head Bracing, Back Bracing

Gauge Descriptions

Gauges currently being explored include head/monitor coupling, head bracing, and back bracing. Further research and analysis will determine which gauges and sensor combinations are the most robust and predictive.

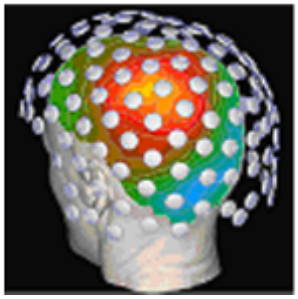
The operator's chair is equipped with a 16 X 16 pressure sensor array in the covers of the seat cushion and back cushion. For head and torso tracking, Flock of Birds sensors are used. One Flock of Birds sensor is attached to the participant's head and one on the torso. All other sensors are attached to the operator's chair. Currently, the head and torso tracking sensors are wired to the Flock of Birds

data collection circuits so the participant's movement is limited to the length of the cable. Additionally, the chair sensors are irrelevant once the subject leaves the seat. While on the seat, however, movement is completely restricted.

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 (if no change in position occurs between two samples, then each sensor shows a delta value of zero and the standard deviation across the 256 sensors is zero). Head position is determined by the relative position of the Flock of Birds sensors (with the reference sensor attached to the back of the chair), and is normalized by subtracting the mean head position for each wave. Head position is calculated for both a-p and m-l dimensions. The head/monitor coupling gauge is the magnitude of the a-p change divided by the variability (as calculated by the root-mean-square value over the same time period).

Currently, the seat cushion and head tracking gauges are treated independently. The true value of the seat cushion sensors will likely be realized in real-world motion environments (at sea, in air, or in a moving ground vehicle). The back bracing gauge seems to be sensitive to changes in workload, as measured by changes in tasks. The head/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. The sensitivity must be determined empirically on a task-by-task basis. The temporal limitation of this gauge is the response rate of automatic behaviors to environmental cues. There are no known limiting factors to the resolution of the gauge. A very high level of predictive value for assessing general arousal is predicted from the gauge based on the conceptual theory underlying posture-mediated state-detection.

3.2.3 Electrical Geodesics, Inc.



Gauge Names: Motor Effort, Auditory Effort

Gauge Descriptions

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). The names given to describe the gauges are (1) motor effort and (2) auditory effort. However, because of the nature of the different events, gauges that assess other cognitive processes can be built.

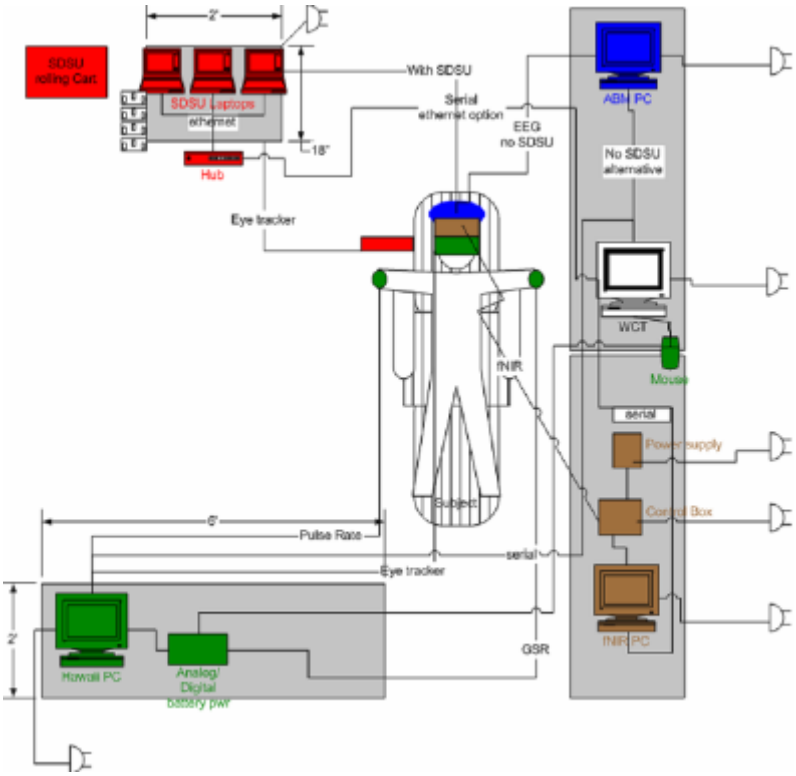
Dense-array EEG electrodes are used to collect data. The sensors are connected to the scalp of the participant. There are 128 EEG channels placed on the participants' head. The entire array takes less than 10 minutes to apply. This time includes amplifier calibration and impedance checking. Each time the sensors are connected, the sensor contact with the scalp needs to be checked. This recalibration requires 1 to 5 minutes to complete. The sensors can be worn comfortably for an indefinite amount of time. However, if the net is worn for more than 3 hours, periodic checking of contact is required. The only constraint placed on the operator wearing the sensor is that the operator must be relatively still.

For the EEG analysis (gauges), a measure of theta averaged 0.5 seconds before and after the WCT event (e.g., KIFF and AHTH) is obtained. Because the KIFF and AHTH events represent processing capacity in different domains, motor and auditory, respectively, the focus on analysis is on those sensor positions that overlie the somatosensory motor cortex for the KIFF (button press to identify a track) event and over the medial prefrontal cortex for the AHTH (feedback sound when a track is destroyed). A limit of the gauges' sensitivity is not foreseen. The current temporal resolution of the gauge in detecting changes in operator state is milliseconds to minutes, depending on computational resources and algorithms. The practical limit of the resolution of the gauge is milliseconds. The aspects of cognitive state/human performance that the gauge measures are working memory, verbal memory, and motor control effort. However, EEG is generally regarded as being quite sensitive to arousal. Currently, minute-by-minute output can be provided.

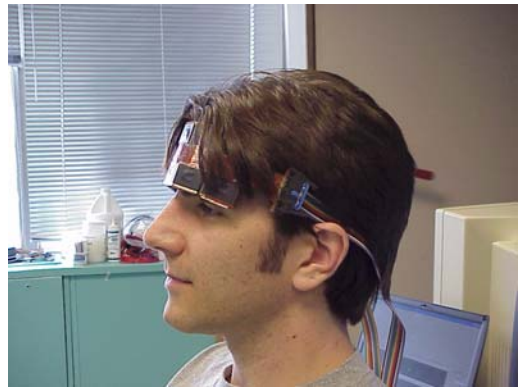
3.3 TEAM 2

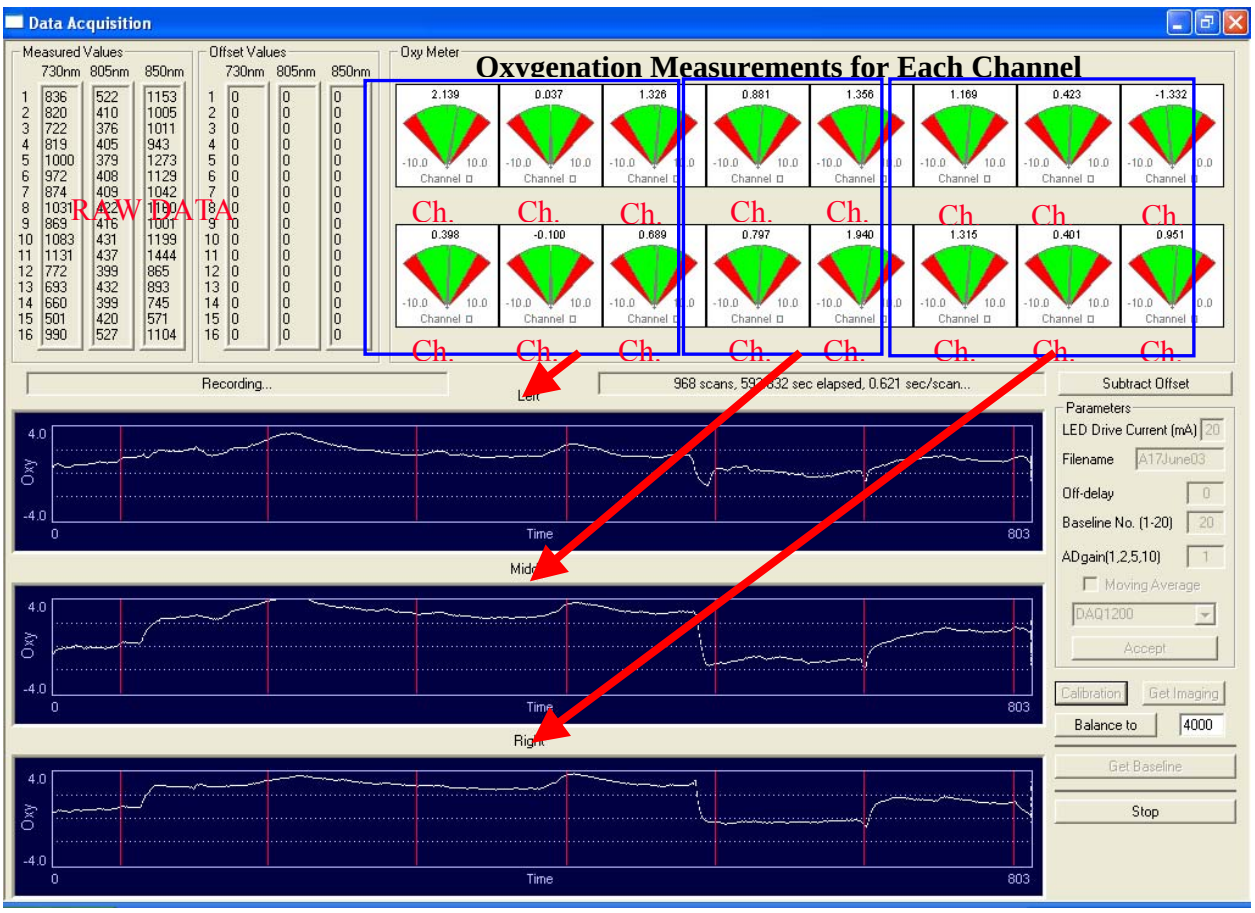
Team 2

Gauge and Technology	Experiment Team	Contact	Organization
fNIR: Measures blood oxygenation and blood volume, during cognitive activity (Device attached to forehead)	5 researchers	Scott Bunce	Drexel University College of Medicine
Global measures of alertness (Electrodes attached to head)	3 researchers	Chris Berka	Advanced Brain Monitoring
Arousal, Perceptual Motor Load and Cognitive Difficulty	3 researchers	David Chin	University of Hawaii



3.3.1 Drexel University





Gauge Name: Functional Near Infrared – Brain Imaging (fNIR) – Left and Right Frontal Lobes

Gauge Description

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. The

fNIR (functional near infrared) 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 Ch. 1 – 730 nm; Ch. 2 – 850nm; and Ch. 3 – used only for interferences. 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 – 2.0 cm depth in brain tissue. The potential or practical limit for its sensitivity is due to the fixed montage of the detectors. The current temporal resolution for the fNIR gauge depends on experimental protocols. In the context of the WCT, the temporal resolution to detect changes in operating state is an average of one-wave point (75 seconds). There are presently two practical limits: (1) 0.5 second due to the gauge, and (2) 4 to 10 seconds due to slow hemodynamic response.

Attention and working memory is the aspect of cognitive state/human performance that the gauge measures. The gauge specifically measures the forehead where executive functions take place. With a full head system this measurement can be extended. The fNIR gauge provides the descriptive information on the participant's general state of arousal in response to task load changes. The current studies will be allowing the prediction of the cognitive load on working memory. The limit of the gauge's ability to describe and/or predict the level of general arousal being measured is due to current probe design and slow hemodynamic responses. For detailed interpretation, there is a need to perform post processing, including the average across the waves and across the channels for better spatial information. Sweat may also interfere with the probe attachment and affect the performance of the gauge. For the current design, the one constraint that is placed on the operator wearing the sensor is head movement. Head movement should be limited as much as it can.

3.3.2 Advanced Brain Monitoring

Gauge Names: Percent High Vigilance, Probability Low Vigilance

Gauge Descriptions

A wireless EEG sensor headset is the specific sensor used to collect data for the B-Alert EEG indices. EEG sensors are connected to the participants at Fz, Cz, POz, and mastoids. EOG sensors are placed around the eye. The operator wearing the headset must restrict excess muscles activity (e.g., chewing gum, dramatic head movements). These data can be filtered, but it is preferable they be kept to a minimum. 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. B-Alert indices can quantify changes in vigilance and workload on a second-by-second basis. There

is no real limit on sensitivity. The current temporal resolution in detecting changes in operator state is on the order of seconds. The practical limit to the resolution is 0.5 seconds.

Vigilance (combination of alertness and attention) is measured and is directly correlated with workload in WCT. B-Alert is approximately 90% accurate when classifying waveforms each second. After implementation of intelligent algorithms based on time series analysis, 98% accuracy should be achievable. General arousal level is one of the tonic contributors to the B-Alert indices, 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. B-Alert has been shown to be highly sensitive to arousal levels in sleep deprivation studies in the lab. Post-processing is conducted in an attempt to filter out contaminated data to salvage as much clean data as possible. Normally, post-processing is used only to tally data from different condition or time periods or to extract event-related EEG. Fatigue, including lack of sleep and circadian cycles (time of day) and the level of expertise on the task significant influences B-Alert.

3.3.3 University of Hawaii

Gauge Name: Arousal and Stress

Gauge Description

GSR and infrared oximeter are the sensors used to collect data. The sensors are placed on the participant's toes (two for GSR and one for oximeter). The participant is immobile with the sensors placed on the toes. However, alternate placement is possible for mobile applications. Heart rate from the oximeter and GSR are multiplied and compared to user's calibrated values. Multiplication is used so that the measures reinforce each other when they agree and cancel each other out when they disagree. The gauge tracks small changes in arousal and stress. The potential limit for its sensitivity is unknown. The current temporal resolution of the gauge in detecting changes in operator state is 2 to 4 seconds. The practical limit to the resolution of the gauge is 1 to 2 seconds.

Arousal and stress are the aspects of cognitive state/human performance the gauge measures. The realistic limit of the gauge's ability to predict the cognitive state being measured is unknown. The participant's general state of arousal is reflected well with the gauge. There is no post-processing required to calculate or interpret the gauge.

Gauge Name: Perceptual and Motor Load

Gauge Description

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. The operator must be in an environment where a computer mouse can be used. Due to patent pending, details about the measures being used from the data collected cannot be given. The gauge seems to track state change very well. The current temporal resolution of the mouse to detect change in operator state is in sub-seconds. The practical limit to the resolution is enough time for a mouse click. The gauge measures perceptual and motor loads and corresponds well to the general state of arousal of the participant.

Gauge Name: Cognitive Difficulty

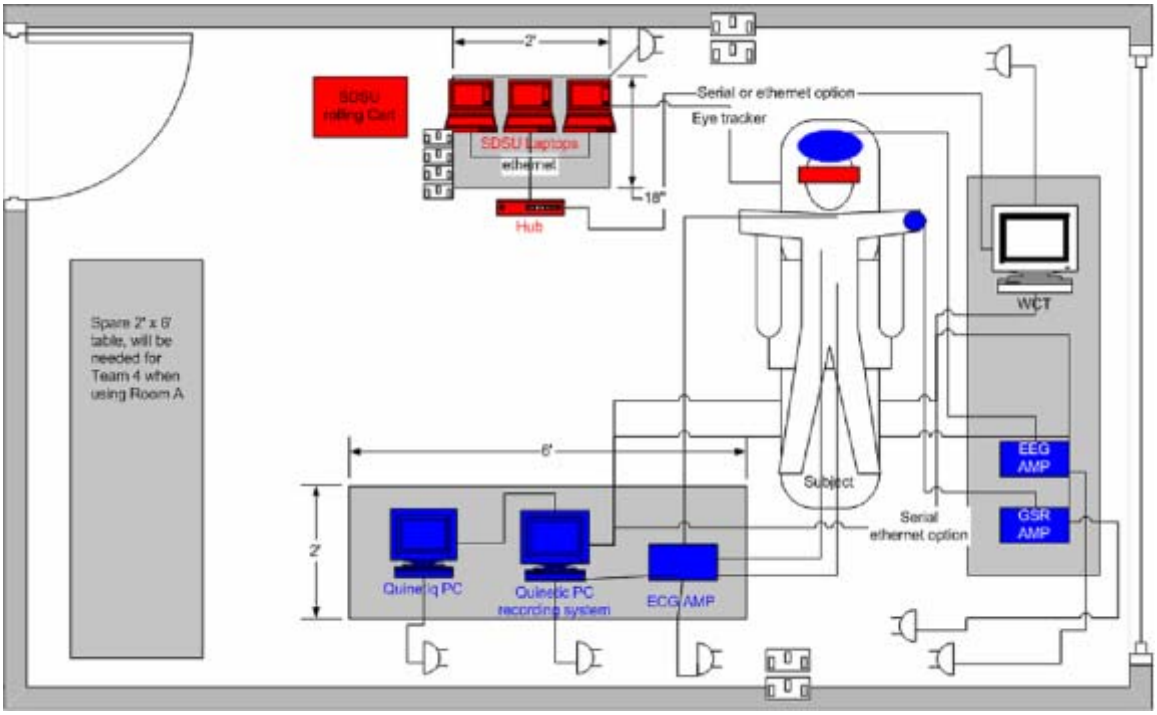
Gauge Description

Gauge description for the Cognitive Difficulty Gauge is similar to the description above, with the following exception: the waveform of the click produced by users changes when they are thinking. This waveform can be used to judge cognitive difficulty of the task.

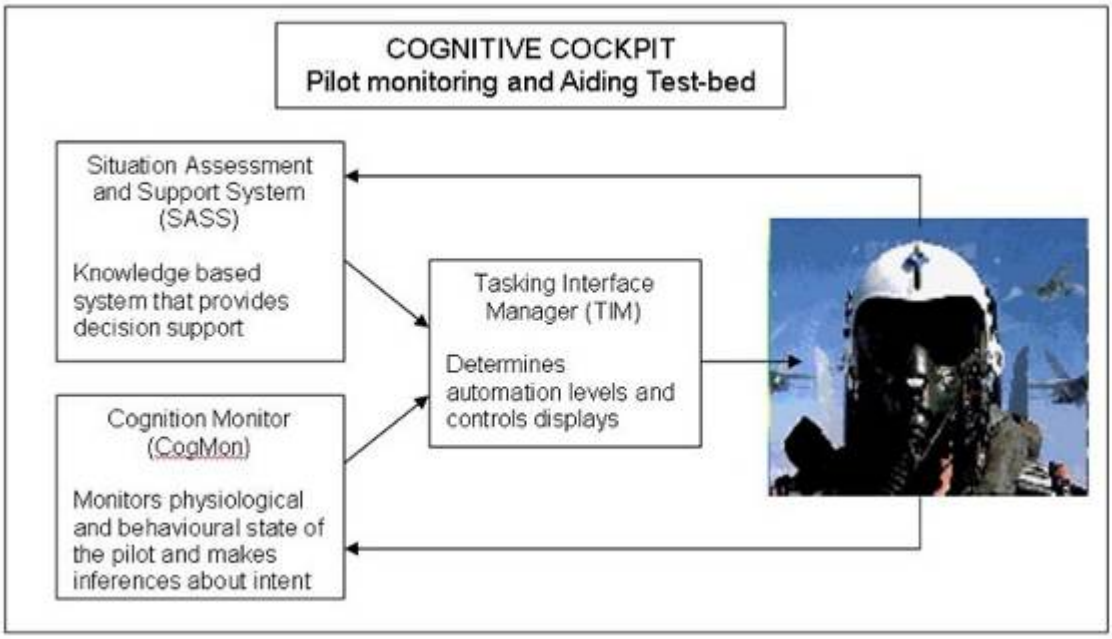
3.4 TEAM 3

Team 3

Gauge and Technology	Experiment Team	Contact	Organization
EEG, ECG, GSR, and EOG (Electrodes to head, face, chest, and hand)	3 researchers	Blair Dickson	QinetiQ, UK



3.4.1 QinetiQ



Gauge Name: Executive Load

Gauge Description

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.

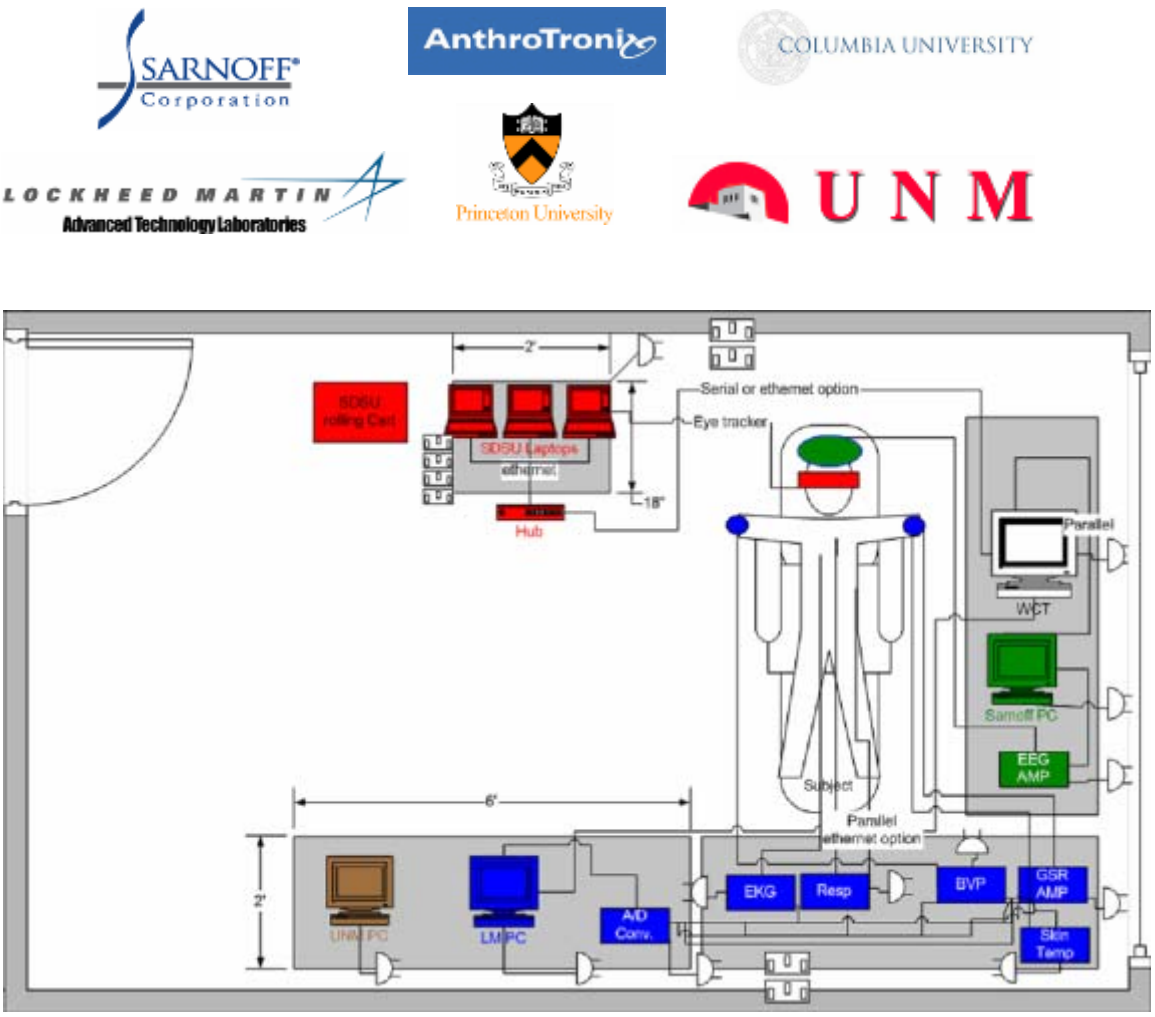
Sensors used to collect data for Executive Load (EL) are 14 scalp EEG electrodes and four EOG electrodes. There are no constraints put on the operator wearing the sensors as long as the effects of any actions/movements/artifacts can be modeled. Currently the hardware is bulky and not wearable, therefore participants must be seated. Sensor activity is analyzed in both the time and frequency domain. Measures of EEG activity include spectral analysis, coherence between electrodes, and measure of phase and power between electrodes.

In the past, sensitivity changes in task difficulty and process type (verbal or spatial) were demonstrated. In the present study, data for executive load are reported. The current temporal resolution in detecting changes is 1.6 seconds. The practical limit to the resolution is the number of points in the FFT; this depends on the windowing employed. This choice would ultimately depend on the application of the technology—the trade off between temporal resolution and accuracy of state elimination given the task being performed.

3.5 TEAM 4

Team 4

Gauge and Technology	Experiment Team	Contact	Organization
EEG, ER correlated measures (Electrode Cap, electrodes on face)	2 researchers	Lucas Parra	Sarnoff Corp
EEG/ERP measures	2 researchers	Akaysha Tang	University of New Mexico Department of Psychology
Measures stress and cognitive load using GSR	2 researchers	Anna Lockerd	AnthroTronix



3.5.1 Sarnoff Corporation, Princeton University, and Columbia University

Gauge Name: Loss Perception

Gauge Description

It is argued that the user's perception of a warning signal along with its negative effect 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 effect associated with loss. Therefore, it is argued that increasing task difficulty should correspond to decreases in intensity of the differential evoked response. The number of errors within a task negatively correlates with the evoked response elicited by warning signals. It is therefore concluded that the proposed EEG measure gives an alternative metric for perceived task difficulty. Available attentional resources modulate the activity of this gauge.

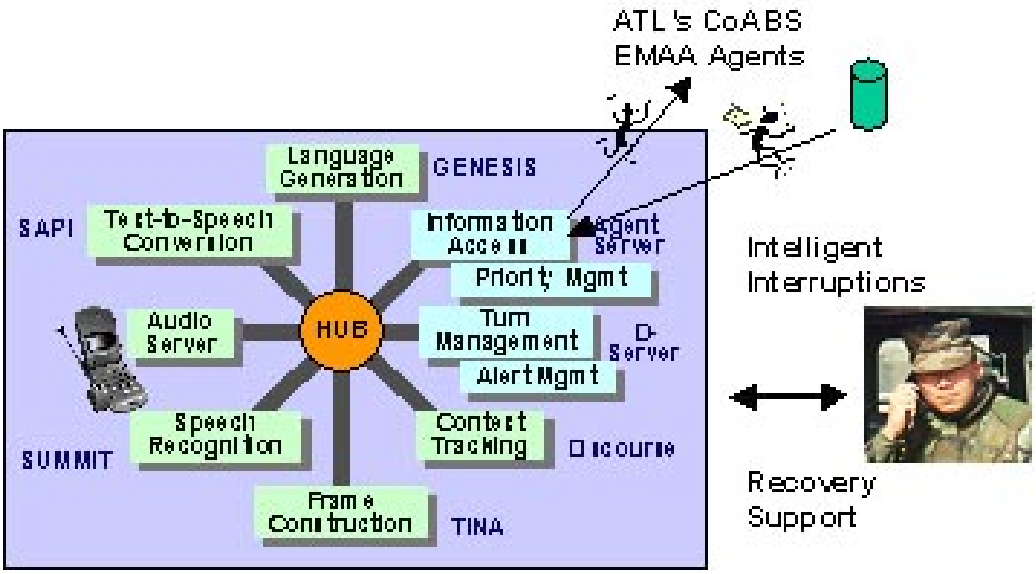
Loss Perception (LP) consists of 63 channels of EEG, including a few EOG channels used to collect data. Electrodes are placed on the scalp, face, and mastoid. The only limitation right now is the wiring to the EEG equipment, which results in a limited range of motion. Speaking is not a problem, but may require modifications to the preprocessing algorithm. It is a consensus among experienced EEG researchers that substantial user motion such as walking or running will introduce substantial motion artifacts. Adaptive linear spatial filtering is the measure being used from the data collected. Spatial filters are continuously adapted. The relevant activity is defined as the difference (in time) that is most indicative for a specific type of event as compared to control events.

The sensitivity of LP, as a measure of state change (between two "cognitive states") provides 75 to 90% correct discrimination when the time of the vent is externally provided with 10 to 50 ms accuracy. The specificity is that the gauge is not meaningful, i.e., non-specific, outside this narrow time window. When considering the average activity across many events, one is limited by the frequency of occurrence of the events. Sensitivity is 100% for a narrowly defined time window relative to external events. The current temporal resolution of the gauge in detecting change is less than 1 second. The potential resolution in time is 100 to 200 ms, and the potential resolution in space is 2 cm.

LP measure is a measure of executive function: self-monitoring or assessment of executed actions. It has potential to measure attentional resources. The 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. In an uncontrolled and flexible environment, it may prove difficult to reliably evoke EEG responses needed to probe the cognitive state. The LP is not related to general state of arousal. If anything, it relates to the immediate affective response to errors. There is no post-processing required to calculate or interpret the gauge measurements. However, the gauge is continually adapting. As more and more events are observed, it becomes more and more specific. Best performance can be obtained after more than 50 events. Therefore post-processing typically yields more relevant results.

3.5.2 Lockheed Martin Advanced Technology Laboratories and AnthroTronix



Gauge Name: GSR Arousal

Gauge Description

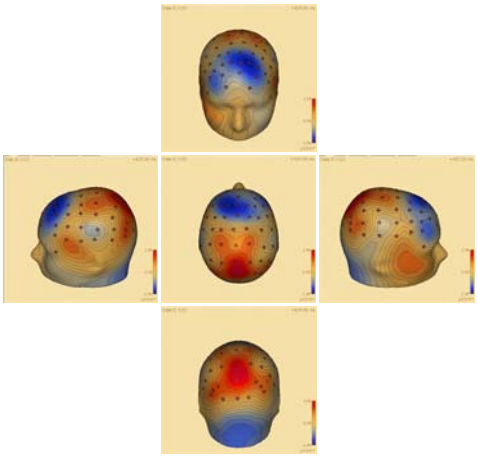
Changes in Galvanic Skin Response (GSR) are indicative of autonomic nervous system activity. Tonic levels of GSR are indicative of a person's general state of arousal, while sudden increases and decreases in GSR are indicative of event-related changes in arousal. By examining raw GSR, a person's tonic levels of GSR are identified, as well as general trends in their arousal state. By examining the rate of change in a person's GSR, it is possible to identify sudden increases and decreases in arousal, and relate those changes to specific events.

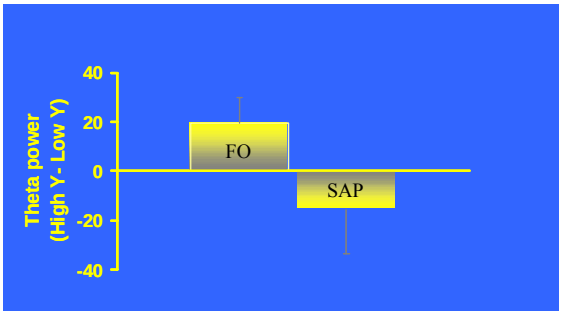
A dual-electrode GSR sensor collects GSR skin conductance data. The GSR sensor consists of two electrodes, which are applied to the surface of the skin on the underside of the participant's second and fourth toes. Participants must refrain from curling toes or applying excessive pressure to the toes fitted with electrodes. Also, participants must restrict arm and facial movement.

The raw GSR data is averaged over each second. The derivative of the GSR is then calculated from the second-by-second data, and then scaled by 1000. The GSR gauge detects changes in state as small as 0.01 micro-mohs with a time resolution of 32 Hz. This is also the practical limit in sensitivity and resolution of the GSR gauge. The GSR gauge can detect and measure changes in autonomic nervous activity, which is indicative of stress/arousal. However, the GSR gauge cannot distinguish between ANS activity caused by stress/arousal and ANS activity caused by other factors

such as fear or embarrassment. By comparing tonic GSR levels measured during the task scenarios and comparing those to tonic levels of GSR during baseline measurements, differences in the participant's general state of arousal is observed. The GSR gauge cannot describe and/or predict the level of general arousal being measured without having a high-quality baseline measurement of each participant's tonic GSR level. Effectiveness of the GSR gauge will be increased by the integration of additional sensors. Wearing multiple sets of the sensors for an extended period of time may cause physical discomfort.

3.5.3 University of New Mexico





Gauge Name: Theta Power (Anterior-Posterior Source, Ocular Source, Visual Source)

Gauge Description

Keeping track of which yellow plane is friend or foe requires working memory. Theta band activity 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.

The EEG sensors from Sarnoff are used to collect data. Theta power changes can be measured from specific functionally distinct brain regions with second-by-second resolution. Spatial resolution is in centimeters. Only when an independent objective measure of working memory is provided could the temporal resolution be determined. The practical limit to the resolution of the gauge is 100 ms.

Working memory is the aspect of cognitive state being measured. The kind of state change that needs to be detected is complete overload or near-complete overload. Theta power in different brain regions is a very good candidate for this. As expressed by many developers during the program, one must first derive a full curve of theta power for each individual to cover the most demanding and the least demanding task conditions. Only with this pre-derived curve can one ask whether a given theta power level measured is indicating a state of overload. The participant's general state of arousal may not be the best measure for physiological arousal using theta power.

3.6 TEAM SDSU

Team SDSU

Gauge and Technology	Experiment Team	Contact	Organization
Index of Cognitive Activity from pupil dilation	3 researchers	Sandra Marshall	Department of Psychology, San Diego State University

Gauge Name: Index of Cognitive Activity (ICA)

Gauge Description

The Index of Cognitive Activity (ICA) uses 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. The movement of the operator is limited due to the computer cable attached to the headband. The operator can move in the chair or can even stand up as long as there is sufficient cable length to do so. The raw pupil signals are obtained from the two cameras. 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. The sensitivity of the gauge as a measure of state change is currently high, medium, and low cognitive effort. The practical limit for its sensitivity is no more than 4 to 6 levels. The current temporal resolution of the gauge in detecting changes in operator state is 1 second as well as the practical limit to the resolution of the gauge.

Overall cognitive effort is the variable being measured by the gauge. The ICA is best suited for complex cognitive tasks. It is not sensitive to very simple tasks. Arousal level is not measured with this gauge. The ICA can be displayed in real time, both with the index value and the category level (high, medium, or low). Both right and left ICA are shown in the current gauge. Fatigue and discomfort may affect the performance of the ICA.

4. RESULTS

4.1 VALIDATING THE TASK LOAD FACTORS

Did each of the three task load factors significantly affect performance on the task?

Ultimately, our goal is to evaluate the ability of each cognitive state gauge to measure changes in the participants' cognitive state and workload. To make this evaluation, though, we must first reconfirm that the three task load factors did in fact meaningfully manipulate participants' workload and task performance. To answer this question, we first briefly report an analysis of participants' subjective ratings of workload from a previous pilot study of the Warship Commander Task. Then we report analyses of how each of the seven WCT performance measures was affected by each of the three task factors during the TIE.

Subjective Workload

The effects of task load on participant's subjective feelings of workload were assessed in a pilot study conducted prior to the TIE. The participants were 14 office personnel who had no previous exposure to the Warship Commander Task. Participants were trained to perform only the primary airspace-monitoring task; the secondary verbal task was not tested. Participants were given 30 minutes of training on the task in which they were guided through the rules of the task and completed several practice scenarios.

Following training, participants were given a series of six test scenarios (three pairs) that were presented in a counter-balanced order across participants. Each scenario contained three waves of tracks with the same number of tracks within each wave (6, 12, or 18).[12] One scenario of each pair contained a high proportion of difficult tracks, and the other scenario in each pair contained a low proportion of difficult tracks. Thus, the experiment used a 3 x 2 repeated measures design, with three levels of number of tracks per wave and two levels of track difficulty. After completing each scenario, participants filled out the NASA Task Load Index Questionnaire (TLX).

A two-way repeated measures Analysis of Variance (ANOVA) of the TLX scores indicated a significant effect of task load for the Number of Tracks per Wave, $F(2, 26) = 13.315$, $p < .05$.[13] Pairwise comparisons revealed significant differences between the 6 to 18 and 12 to 18 tracks, but not between 6 to 12 tracks. These results are shown graphically in Figure 4. The ANOVA results also indicated a significant effect for Track Difficulty, $F(1, 13) = 25.987$, $p < .05$. The participants rated the WCT on the six subscales of the TLX, and determined that the task required primarily "temporal demands" and "effort." In sum, both of the task load factors that were tested, Number of Tracks per Wave and Track Difficulty, were found to produce statistically different levels of subjective workload. Participants felt primarily more temporal demands and more effort at higher levels of task loading.

---

[12] Waves containing 24 tracks were not included in this pilot study.

[13] For readers who are less familiar with statistical tests, the key number is the *p*-value. *P*-values below 0.05 indicate statistically significant results. In more detail, an Analysis of Variance, like all statistical tests, computes the chance that the observed differences between conditions are either real or due to random fluctuations. The *p*-value is the probability that the observed differences are due to random fluctuations. A probability of 0.05 is traditionally taken as the cutoff—indicating that the effect observed would occur by chance only 1 out of 20 times tested. For each repeated measures ANOVA in this report, we tested for sphericity and made appropriate Greenhouse-Geisser adjustments to the degrees of freedom.

Although significant effects were found for the TLX scores, a fourth condition was added to maximize participant's task load. Waves containing 24 tracks were added to provide an "extreme" level of task load during the Pre-TIE and TIE data collections.

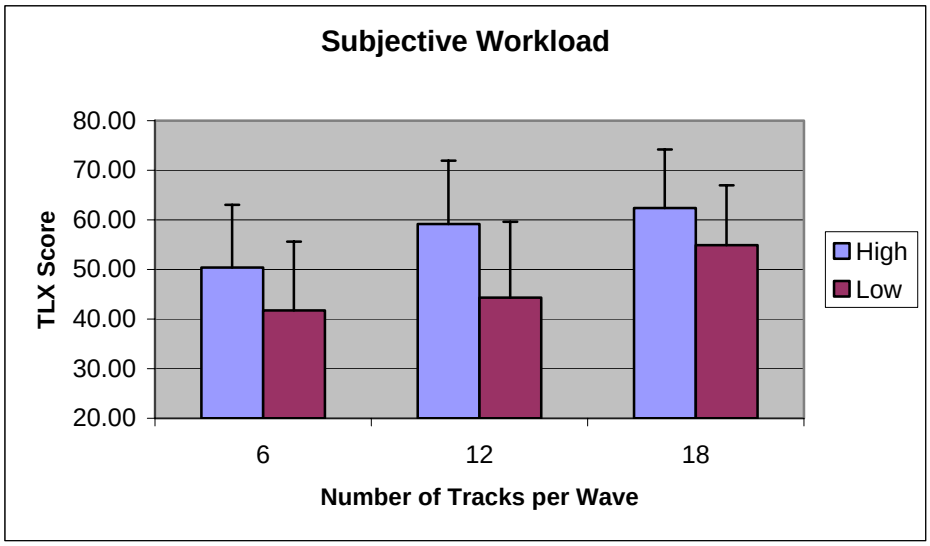


Figure 4. Subjective workload ratings graphed against the number of tracks per wave and split by track difficulty (high or low).

Effects of Task Load on the Primary Airspace Monitoring Task

To analyze data from the TIE, we first collected behavioral data from the WCT software for each of the 25 complete data collection sessions. For clarity, we describe the data analysis for one of the performance measures, RTIFF, but the same analysis was performed for each of the performance measures. First, for each scenario, we computed the mean RTIFF for waves of 6, 12, 18, and 24 tracks. Then, for each team, we computed separate repeated measures ANOVAs using the three task load factors of Number of Tracks per Wave (6, 12, 18, and 24), Track Difficulty (high and low), and Secondary Verbal Task (on and off). Finally, we computed an overall ANOVA by pooling the data for each participant across teams. Table 2 shows the results of these ANOVAs for each team, and overall, for each performance measure. Due to the small sample sizes and exploratory nature of this research, we felt that a more liberal alpha threshold was justified. Therefore, three separate alpha (*p)* levels were used. An alpha value of less than 0.05 indicates that the task load factor significantly affected the performance measure; a value of less then 0.10 was considered to indicate that the task load factor *marginally* affected that performance measure; and a value of 0.20 was used to indicate that the effect may have future *potential* and warrants further research.

Table 2. Statistical results of the three task load factors on the WCT performance measures by team and overall.

Task Load Factors	Team	# prtcpnts	RTIFF df	RTIFF F	RTIFF *p*	RTWarn df	RTWarn F	RTWarn *p*	RTEngage df	RTEngage F	RTEngage *p*	PctGS df	PctGS F	PctGS *p*
Number of Tracks per Wave	1	6	1.084, 5.422	27.1	**0.003**	1.364, 6.818	64.1	**0.000**	1.225, 6.123	44.4	**0.000**	3, 15	27.8	**0.000**
	2	8	1.406, 9.842	59.4	**0.004**	1.493, 10.452	55.8	**0.000**	3, 21	24.1	**0.000**	3, 21	19.2	**0.000**
	3	6	1.363, 6.816	30.2	**0.001**	3, 15	30.2	**0.000**	1.084, 5.422	27.1	**0.003**	3, 15	9.3	**0.001**
	4	5	3, 12	52.9	**0.000**	3, 12	91.8	**0.000**	3, 12	35.2	**0.000**	3, 12	41.0	**0.000**
Track Difficulty	1	6	1, 5	14.9	**0.012**	1, 5	49.7	**0.001**	1, 5	21.9	**0.005**	1, 5	27.9	**0.003**
	2	8	1, 7	22.3	**0.002**	1, 7	69.1	**0.000**	1, 7	27.8	**0.001**	1, 7	48.9	**0.000**
	3	6	1, 5	17.0	**0.009**	1, 5	21.7	**0.006**	1, 5	14.9	**0.012**	1, 5	11.3	**0.020**
	4	5	3, 12	62.2	**0.001**	1, 4	41.8	**0.003**	1, 4	80.7	**0.001**	1, 4	184.6	**0.000**
Secondary Verbal Task	1	6	1, 5	9.3	**0.029**	1, 5	27.2	**0.003**	1, 5	15.5	**0.011**	1, 5	6.7	**0.049**
	2	8	1, 7	6.4	**0.040**	1, 7	25.8	**0.001**	1, 7	16.5	**0.005**	1, 7	13.9	**0.008**
	3	6	1, 5	2.8	0.156	1, 5	24.9	**0.004**	1, 5	9.3	**0.029**	1, 5	19.7	**0.007**
	4	5	3, 12	14	**0.020**	1, 4	6.7	0.060	1, 4	7.9	**0.049**	1, 4	6.6	0.062
# Tracks	Overall	8	1.287, 9.009	66.9	**0.000**	1.129, 7.904	85.4	**0.000**	1.066, 7.461	38.9	**0.000**	1. 095, 7.665	38.8	**0.000**
Difficulty	Overall	8	1, 7	34.8	**0.001**	1, 7	90.0	**0.000**	1, 7	31.8	**0.001**	1, 7	46.1	**0.000**
2nd Verbal	Overall	8	1, 7	19.9	**0.003**	1, 7	41.0	**0.000**	1, 7	21.3	**0.002**	1, 7	15.5	**0.006**

Task Load Factors	Team	# prtcpnts	ECommis df	ECommis F	ECommis *p*	EOmiss df	EOmiss F	EOmiss *p*	TPending df	TPending F	TPending *p*
Number of Tracks per Wave	1	6	3, 15	24.7	**0.000**	1.227, 6.137	36.9	**0.001**	1.152, 5.761	244.4	**0.000**
	2	8	3, 21	33.5	**0.000**	1.454, 10.177	33.7	**0.000**	1.190, 8.328	282.1	**0.000**
	3	6	1.584, 7.921	22.5	**0.001**	1.115, 5.576	13.1	**0.012**	1.157, 5.786	138.5	**0.000**
	4	5	3, 12	13.9	**0.000**	1.147, 4.589	46.6	**0.001**	3, 12	275.7	**0.000**
Track Difficulty	1	6	1, 5	17.7	**0.008**	1, 5	31.3	**0.003**	1, 5	335.9	**0.000**
	2	8	1, 7	17.4	**0.004**	1, 7	37.5	**0.000**	1, 7	297.4	**0.000**
	3	6	1, 5	23.9	**0.005**	1, 5	9.4	**0.028**	1,5	155.2	**0.000**
	4	5	1, 4	1.5	0.283	1, 4	126.3	**0.000**	1, 4	463.9	**0.000**
Secondary Verbal Task	1	6	1, 5	0.4	0.579	1, 5	8.5	**0.033**	1, 5	44.9	**0.001**
	2	8	1, 7	0.1	0.754	1, 7	18.3	**0.004**	1, 7	67.7	**0.000**
	3	6	1, 5	0.0	0.868	1, 5	34.0	**0.002**	1,5	16.0	**0.010**
	4	5	1, 4	0.1	0.806	1, 4	8.7	**0.042**	1, 4	21.7	**0.010**
# Tracks	Overall	8	3, 21	61.6	**0.000**	1.019, 7.136	38.2	**0.000**	1.069, 7.483	285.0	**0.000**
Difficulty	Overall	8	1, 7	37.0	**0.000**	1, 7	35.2	**0.001**	1, 7	430.5	**0.000**
2nd Verbal	Overall	8	1, 7	0.0	0.864	1, 7	21.1	**0.003**	1, 7	57.5	**0.000**

Table 2 also shows that most of the behavioral measures were sensitive to each of the task load factors. Statistically significant effects are highlighted and bolded in the lightest shade of gray (green), *marginally* significant effects are highlighted in darker gray, and *potential* effects are highlighted in the darkest shade of gray. Figure 5 shows graphs of the grand means from the overall analysis for each of the performance measures. From these analyses we can conclude

1. WCT meaningfully manipulates all three task load factors, and
2. Each of the performance measures was sensitive to these factors.

The one exception was for errors of commission. The small numbers of these errors limited the statistical power of this measure, and this limitation may have reduced the reliability of the analysis. We can now confidently evaluate each of the developers' gauges against the task load factors as well as compare each gauge with the performance measures.

Several interactions were also discovered and are summarized in Table 3. Essentially, these interactions tend to indicate that the three task load factors were not strictly additive. Instead, the Track Difficulty and Secondary Task manipulations affected performance more when participants were already experiencing workload from higher Numbers of Tracks per Wave. The pairwise comparisons indicate that for most of the performance measures, each increase in the Number of

Tracks per Wave led to significantly worse performance, either longer response times, lower game scores, or more errors.

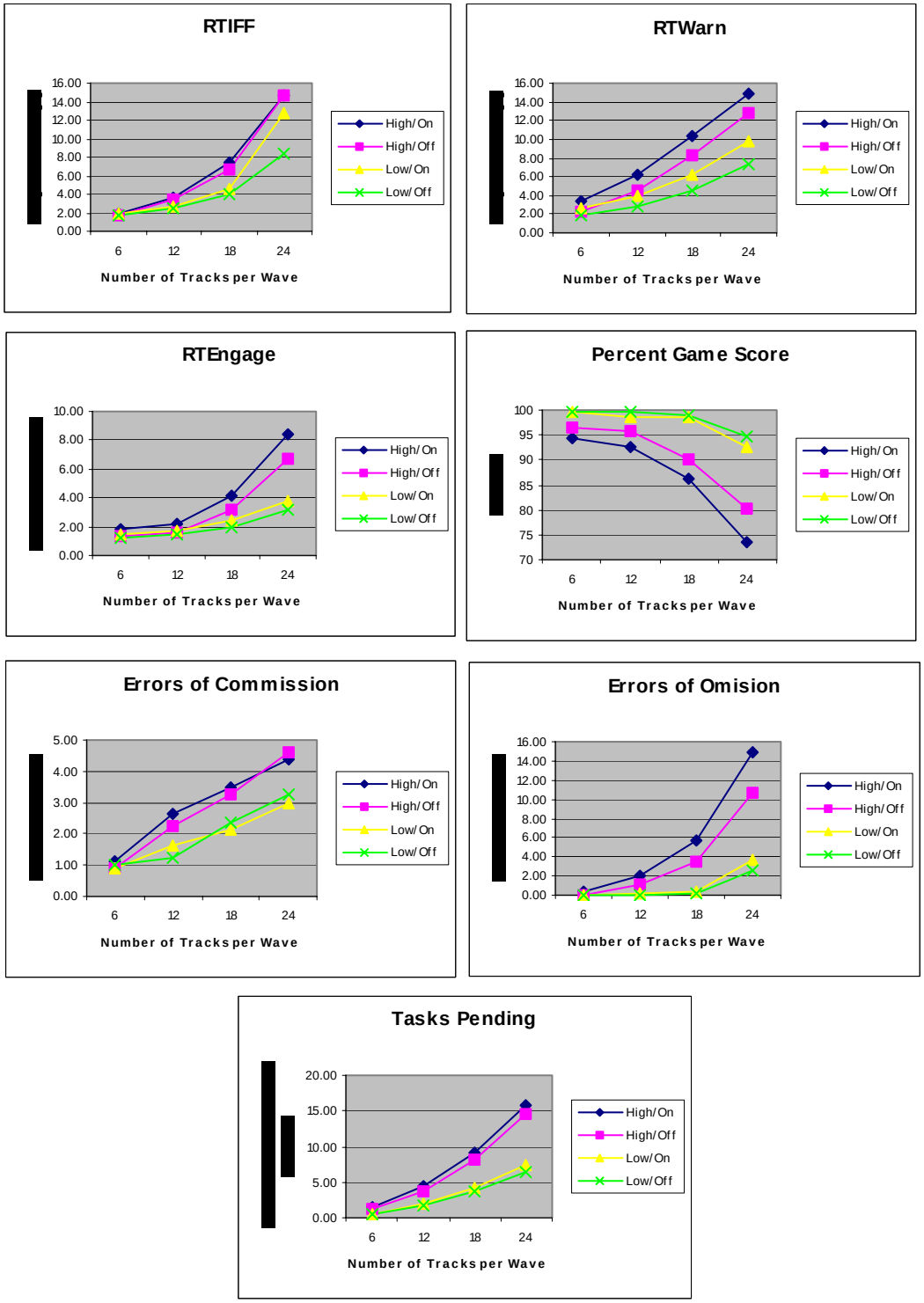


Figure 5. The mean value of the seven performance measures graphed against the Number of Tracks per Wave and split by Track Difficulty (high and low) and Secondary Verbal Task (on and off).

Table 3. Significant interaction results and pairwise comparisons.

Interactions	RTIFF	RTWarn	RTEngage	PctGS	EC	EO	Pending
# Tracks by Track Difficulty	x	x	x	x	x	x	x
# Tracks by Secondary Verbal			x	x		x	x
Track Difficulty by Secondary Verbal		x	x	x		x	x
# Tracks by Track Difficulty by 2nd Verbal						x	
Pairwise Comparisons by Wave							
6<12	x	x	x		x	x	x
6<18	x	x	x	x	x	x	x
6<24	x	x	x	x	x	x	x
12<18	x	x	x	x	x	x	x
12<24	x	x	x	x	x	x	x
18<24	x	x	x	x	x	x	x

Effects of Task Load on the Secondary Ship Status Task

We also analyzed performance on the Secondary Verbal Task and examined how the task load factors of Number of Tracks per Wave and Track Difficulty in the primary task affected performance on the secondary task. Table 4 shows the results of a series of repeated measures ANOVAs for response time and percent correct scores on the Secondary Verbal Task. Figure 6 shows graphs of the effects. The findings were

1. The number of tracks appearing in each wave in the primary task affected performance on the secondary verbal task.
2. Performance on the Secondary Verbal Task was not affected by the Track Difficulty manipulation in the primary task.

Table 4. Statistical results of the primary task on performance of the Secondary Verbal Task for each team.

			Secondary Verbal Task					
		#	RT			PC		
Task Load Factors	**Team**	**prtcpnts**	**df**	**F**	***p***	**df**	**F**	***p***
Number of Tracks per Wave	**1**	6	3, 15	9.7	**0.001**	3, 15	2.66	0.086
	2	8	1.242, 8.694	4.7	0.053	3, 21	3.16	**0.046**
	3	6	3, 15	7.5	**0.003**	3, 15	2.77	0.078
	4	5	3, 9	0.9	0.492	3, 12	1.95	0.176
Track Difficulty	**1**	6	1, 5	0.1	0.742	1, 5	0.14	0.722
	2	8	1, 7	1.3	0.286	1, 7	0.59	0.466
	3	6	1, 5	3.6	0.116	1, 5	0.05	0.840
	4	5	1, 3	0.653*	0.478	1, 4	0.15	0.721
# Tracks	**Overall**	8	1.425, 8.548	13.6	**0.003**	3, 21	4.9	**0.010**
Difficulty	**Overall**	8	1, 5	0.8	0.404	1, 7	0.8	0.394

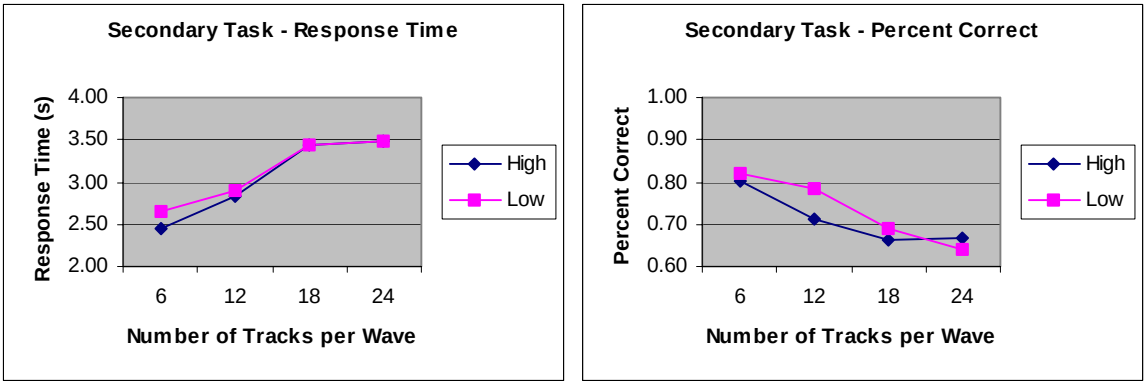


Figure 6. The mean response times and percent correct scores on the Secondary Verbal Task as a function of the Number of Tracks per Wave and Track Difficulty (high and low) in the primary task.

Not every participant performed equally well on the Secondary Verbal Task. Some participants devoted attention to the task and performed well, while other participants appeared to adopt a strategy of essentially ignoring the secondary task and performed poorly—timing out on most questions. One participant clearly used this strategy to an extreme and failed to answer a single question, this participant's data was removed from the analysis. It is not clear why the participant chose, or was unable, to perform the secondary task. However, the secondary task is quite difficul, particularly during waves of 18 or 24 tracks in the primary task. The most important point is that the secondary task had the desired effect of increasing the task load, as shown by the significant impact of the secondary task on performance of the primary task, as shown in Table 2 and Figure 5.

4.2 GAUGE EVALUATION

How well does each gauge detect changes in task load?

In this section, we summarize the evaluation of the sensitivity of each of the cognitive state gauges to the three manipulations of task load. The question we addressed was: Do gauge values reflect changes observed in each of the task load measures of the WCT? Individual CWA developers in Appendix 2 provide more detailed analyses of each specific gauge.

For each gauge, we computed three-way repeated measures ANOVAs. Means for each level of the Number of Tracks per Wave factor were computed for each scenario. These means were submitted to three-way repeated measures ANOVA for Number of Tracks per Wave, Track Difficulty, and Secondary Verbal Task. While 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 this analysis. *This reduction resulted in small sample sizes for some gauges that limit the conclusions we can draw from the analyses*. The results of these analyses are reported in Tables 5 through 9 and Figures 7 through 12.

Team 1

Table 5 gives statistical results of the three task load factors on each of gauges for Team 1.

Table 5. Statistical results of the three task load factors on each of the Team 1 gauges.

Task Load Factors	Clemson (n=7) Arousal Meter			Head-Monitor Coupling (n=5)			UPitt/NRL Head Bracing (n=5)			Back Bracing (n=7)		
	df	F	*p*	df	F	*p*	df	F	*p*	df	F	*p*
# Tracks	3, 18	0.7	0.547	1.055, 4.218	5.0	0.086	3, 12	2.0	0.162	1.249, 7.491	0.5	0.537
Difficulty	1, 6	0.4	0.555	1, 4	0.3	0.608	1, 4	0.9	0.405	1, 6	2.2	0.186
2nd Verbal	1, 6	0.0	0.943	1, 4	0.8	0.423	1, 4	2.9	0.162	1, 6	0.0	0.865

Task Load Factors	EGI (n=5) Motor Effort			Auditory Effort		
	df	F	*p*	df	F	*p*
# Tracks	3, 12	2.0	0.165	1.218, 4.87	0.5	0.528
Difficulty	1, 4	0.0	0.906	1, 4	4.4	0.105
2nd Verbal	1, 4	3.1	0.155	1, 4	3.2	0.147

The Clemson Arousal Meter was not sensitive to the task load factors manipulated in WCT.[14] However, there was a *marginally* significant effect ($p < .10$) for the Track Difficulty X Number of Tracks per Wave interaction. Figure 7 shows that this effect is driven by the gauge values obtained during wave size 18 but not for the other wave sizes. In addition, the Arousal Meter demonstrated a *potentially* significant effect ($p < .20$) for the three-way interaction. Again, it appears that the interaction is primarily driven at the wave size of 18. The High Track Difficulty/Secondary Verbal Tasks (on and off) have higher gauge values than those of the Low Track Difficulty/Secondary Verbal Task values.

> ***Summary:*** *The results indicate that the Arousal Meter was not sensitive to the primary task load incorporated in the TIE. Marginal/potential interactions were found but, due to the high variability in gauge values, they are difficult to interpret. These results suggest that the Arousal Meter may not be best suited for a monitoring type task in which the users are highly experienced. See Appendix 3c for a more detailed analysis and interpretation provided by Clemson University.*

The University of Pittsburgh/Naval Research Laboratory Head-Monitor Coupling gauge was *marginally* sensitive ($p < .1$) to the Number of Tracks per Wave.[15] The Head-Monitor Coupling gauge had a *marginally* significant Number of Tracks per Wave X Secondary Verbal Task interaction. As shown in Figure 7, an observable difference in gauge value is apparent for the 6 and 12 track conditions for both of the Secondary Verbal Task conditions, but differences were not

[14] Clemson reported data from seven participants: p1, 2, 4, 5, 6, 7, and 9, but the raw data set from participant 5 was incomplete. Clemson completed three cells using mean substitution from repeated cells within the same participant and conditions. The repeated measures ANOVA was therefore conducted including the data of all seven participants.

[15] Pittsburgh/NRL reported data from seven participants: p1, 2, 3, 4, 5, 6, and 8. The data sets from participants 5 and 8 were incomplete; therefore the repeated measures ANOVA was conducted with five data sets: p1, 2, 3, 4, and 6. For the Back Bracing gauge, the data sets for participants 5 and 8 were complete; however, for that gauge, all seven data sets were included.

apparent for the 18 and 24 tracks conditions. Furthermore, the Head-Monitor Coupling gauge resulted in a *potentially* significant ($p < .20$) two-way interaction (Track Difficulty X Secondary Verbal Task). Figure 7, shows there is a difference between when the Secondary Verbal Task is on and Track Difficulty. Yet, these differences are not apparent when the Secondary Verbal Task is off. In addition, the Back Bracing gauge had a *marginally* significant Track Difficulty X Number of Tracks per Wave interaction. As shown in Figure 7, a substantial difference in gauge value is apparent during wave size 24 in the high track difficulty condition versus the low track difficulty condition. The remaining wave size levels fail to show this difference. Finally, the Head Bracing gauge demonstrated a *potentially* significant two-way interaction (Secondary Verbal Task X Number of Tracks per Wave). Figure 6, shows larger gauge values when the Secondary Verbal Task was off than all other conditions.

Summary: *All gauges failed to demonstrate statistically significant results to the performance demand of the TIE. However, the number of marginal and potential effects suggests that these results may be related to the low number of participants. The Head-Monitor Coupling gauge did show a marginally significant result to task loading (Number of Tracks per Wave), suggesting an increase in head movement as task load increases. These results are further supported by potentially significant effects found for the Head Bracing gauge for changes in task load and the concurrence of the secondary verbal task. The interactions highlight a couple of interesting relationships such as; higher Head Bracing gauge values being recorded when task requirements were least and the Back Bracing gauge showing potential to discriminate between high and low Task Difficulty conditions at the higher levels of task load, but not at lower levels of task load. These results warrant further investigation. See Appendix 3k for a more detailed analysis and interpretation provided by University of Pittsburgh/Naval Research Laboratory.*

Neither of the Electrical Geodesic measures was sensitive to the Number of Tracks per Wave, but their Auditory Effort gauge was *marginally* sensitive ($p < .10$) to Track Difficulty and the Secondary Verbal Task.[16] In addition, there was a *marginally* significant two-way interaction of Track Difficulty X Secondary Verbal Task. Figure 6 shows that gauge values are largest when the Secondary Verbal Task is on, but not for the remaining conditions. There was also a *potentially* significant ($p < .20$) three-way interaction. This is highlighted by the highest gauge values during the Low Task Difficulty while the Secondary Verbal Task was on, across all wave sizes.

Summary: *Neither of the gauges proposed by EGI demonstrated statistically significant results. However, both gauges did show marginally or potentially significant effects. Due to the low number of participants these results are suggestive and warrant further investigation. See Appendix 3e for a more detailed analysis and interpretation provided by Electrical Geodesic, Inc.*

[16] EGI reported data from five participants: p1, 2, 5, 6, and 8. The data sets from participants 5 and 8 were complete, but were run in an inappropriately configured condition for the secondary verbal task. Nonetheless, these data were retained to increase the sample size. The repeated measures ANOVA was conducted using the data from participants 1, 2, 5, 6, and 8.

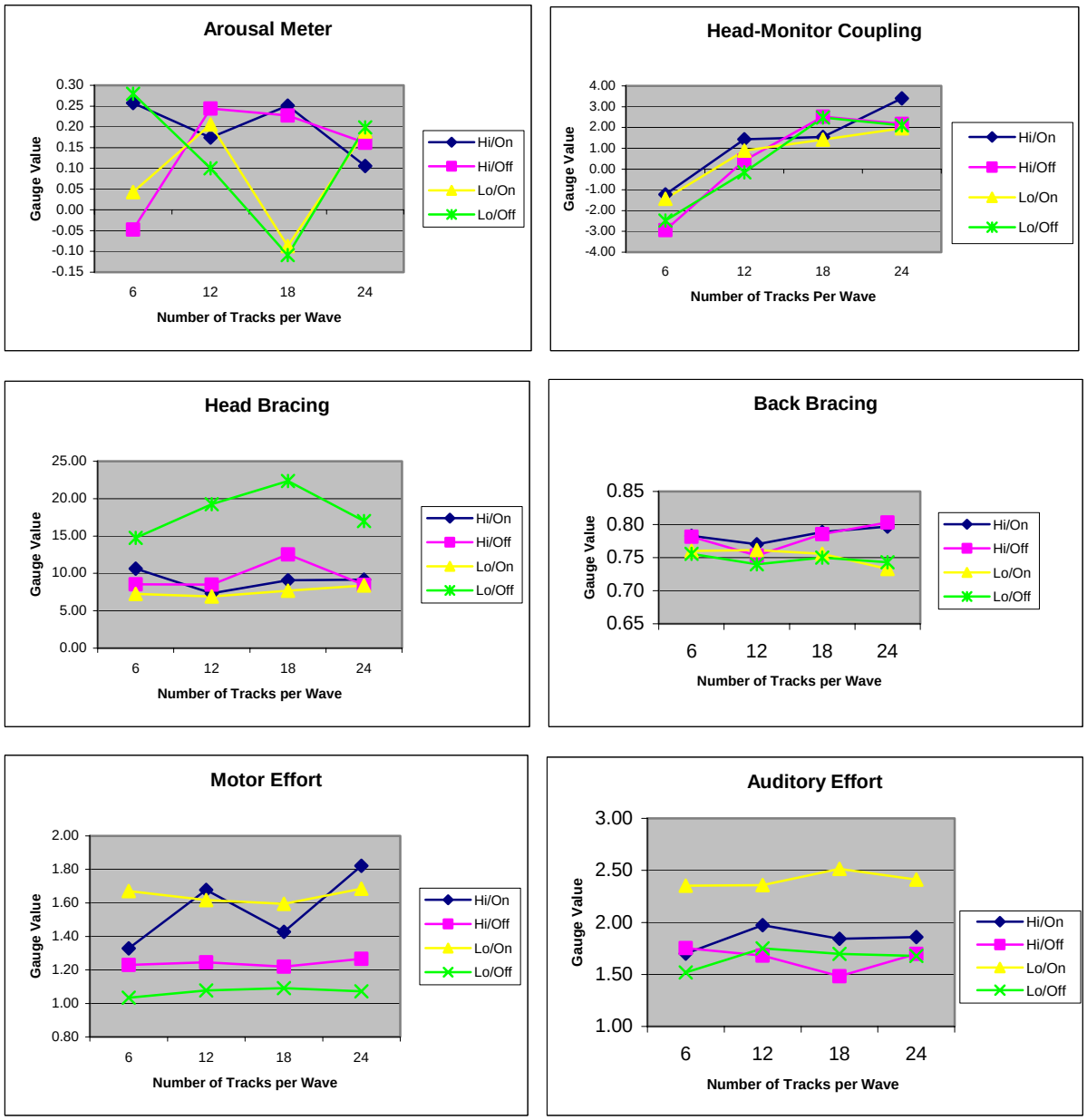


Figure 7. The mean value of each Team 1 gauge graphed against Number of Tracks per Wave, split by Track Difficulty (high and low) and Secondary Verbal Task (on and off).

Team 2

Table 6 gives statistical results of the three task load factors on each of the gauges for Team 2.

Table 6. Statistical results of the three task load factors on each of the Team 2 gauges.

Task Load Factors	Drexel (n=8)						ABM (n=8)					
	fNIR (left)			fNIR (right)			Percent High Vigilance			Probability Low Vigilance		
	df	F	*p*	df	F	*p*	df	F	*p*	df	F	*p*
# Tracks	3, 21	14.9	**0.000**	3, 21	11.7	**0.000**	1.449, 10.145	15.2	**0.002**	1.102, 7.714	11.5	**0.009**
Difficulty	1, 7	0.4	0.544	1, 7	0.3	0.582	1, 7	2.7	0.147	1, 7	1.2	0.301
2nd Verbal	1, 7	1.2	0.316	1, 7	1.1	0.334	1, 7	0.4	0.553	1, 7	0.0	0.851

Task Load Factors	UHawaii (n=8)								
	Arousal			Perceptual/ Motor Load			Cognitive Difficulty		
	df	F	*p*	df	F	*p*	df	F	*p*
# Tracks	1.559. 10916	1.4	0.289	3, 21	416.1	**0.000**	3, 21	602.9	**0.000**
Difficulty	1, 7	0.3	0.626	1, 7	37.7	**0.000**	1, 7	13.7	**0.008**
2nd Verbal	1, 7	1.2	0.316	1, 7	0.4	0.527	1, 7	1.0	0.345

The Drexel University fNIR gauges for both the left and right frontal lobes were significantly sensitive to the Number of Tracks per Wave.[17] The left frontal lobe gauge had a *potentially* significant two-way interaction for Number of Tracks per Wave X Secondary Verbal Task. Figure 8 shows larger differences between gauge values for wave size 6 and 18. These values are lowest for the higher Track Difficulty for wave size 6 and highest for wave size 18. In addition, the right frontal lobe gauge had a *marginally* significant two-way interaction (Number of Tracks per Wave X Secondary Verbal Task). This result is also related to the greater dispersion of gauge values seen at wave size 6 and 18.

> ***Summary:*** *The fNIR left and right gauge were highly sensitive to changes in task load. The interactions are difficult to interpret and are a result of differences found between the scenarios at wave size 6 and 18. Results are promising as a gauge that is sensitive to task load and warrants further evaluation. See Appendix 3d for a more detailed analysis and interpretation provided by Drexel University.*

Advanced Brain Monitoring's Percentage of High Vigilance gauge and Probability of Low Vigilance gauge were also sensitive to the Number of Tracks per Wave. Note that the Probability of Low Vigilance is predicted to *decrease* as task load increases.[18]

> ***Summary:*** *Results indicate that both of the gauges supported by Advanced Brain Monitoring are significantly related to changes in task load. The two gauges complement one another and provide results in the expected direction; an increase on the high vigilance gauge during high task demands and a decrease in the probability of low vigilance. See Appendix 3a for a more detailed analysis and interpretation provided by Advanced Brain Monitoring.*

[17] Drexel reported data for eight participants: p1, 2, 3, 4, 5, 6, 7, and 8. The repeated measures ANOVA was conducted using the data from all eight participants.

[18] ABM reported data for all eight participants plus one extra participant run under a different configuration of sensors. The repeated measure ANOVA was conducted using the data from all eight "official" participants.

The University of Hawaii's Perceptual and Motor Load gauge and Cognitive Difficulty gauge were sensitive to both the Number of Tracks per Wave and to Track Difficulty. Neither gauge was sensitive to the concurrence of the Secondary Verbal Task. Although difficult to see in Figure 8, the Perceptual and Motor Load gauge displayed a significant interaction for Number of Tracks per Wave X Track Difficulty ($p < .05$). The results indicate that the gauge value is significantly greater for the higher Track Difficulty level for wave sizes 6, 12, and 18, but no significant difference in gauge reading during wave size 24. The University of Hawaii Arousal gauge was predicted to be stable across task load manipulations, and it was.[19] The results also indicated a significant Number of Tracks per Wave X Track Difficulty interaction ($p < .05$).

Summary: *Two of the three gauges proposed by the University of Hawaii demonstrated statistically significant gauge changes as the task increased in load. The Perceptual and Motor Load, and the Cognitive difficulty gauge were very sensitive to increases in the Number of Tracks as well as Track Difficulty. Surprisingly, the gauge was sensitive to Track Difficultly changes when task load was low but not when it was high (wave size 24). See Appendix 3i for a more detailed analysis and interpretation provided by University of Hawaii.*

[19] Hawaii reported data for eight participants: p1, 2, 3, 4, 5, 6, 7, and 8. The repeated measures ANOVA was conducted using the data from all eight participants.

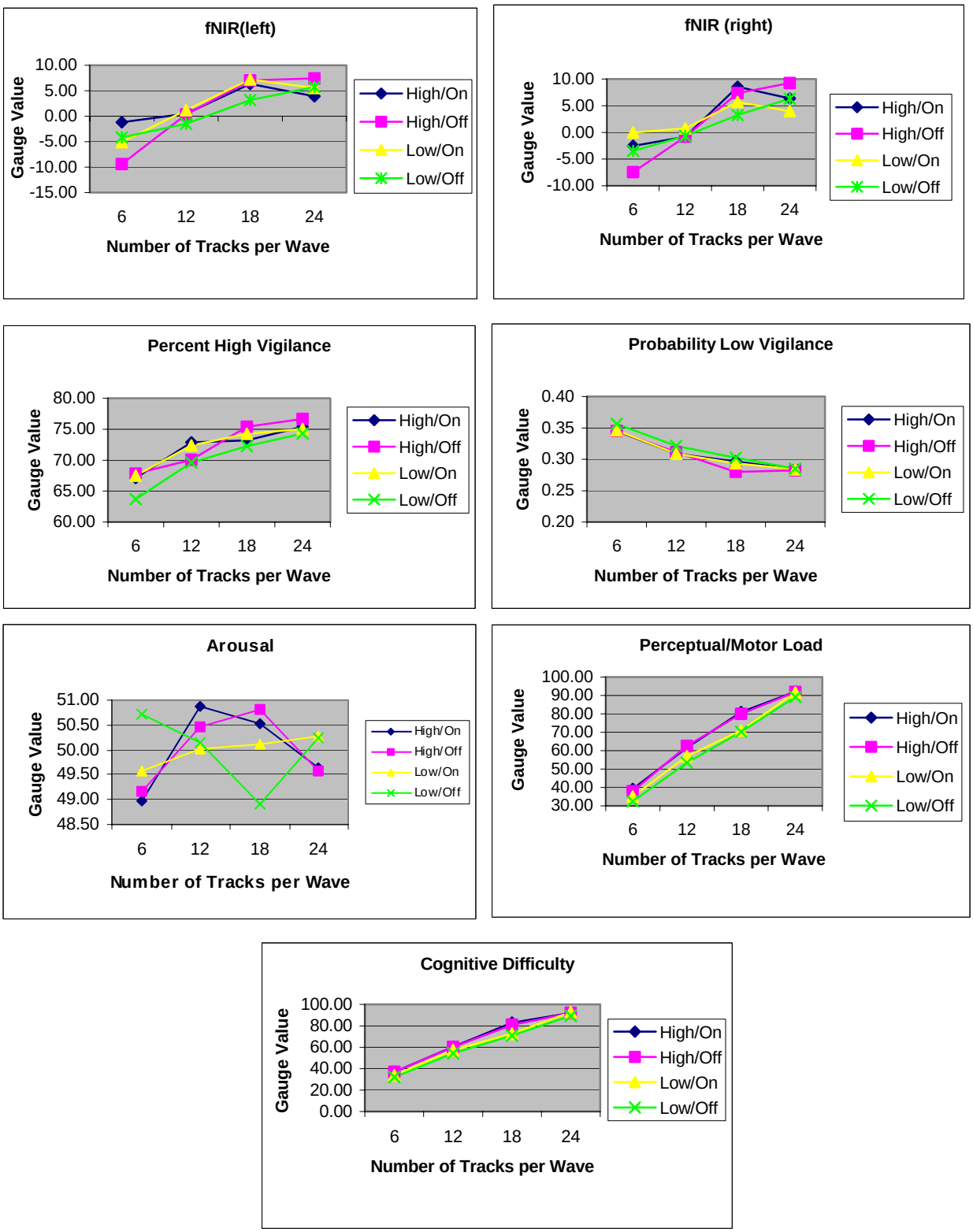


Figure 8. The mean value of each Team 2 gauge graphed against Number of Tracks per Wave, split by Track Difficulty (high and low) and Secondary Verbal Task (on and off).

Team 3

Table 7 gives the statistical results of the three task load factors on the Team 3 gauge.

Table 7. Statistical results of the three task load factors on the Team 3 gauge.

Task Load Factors	QinetiQ (n=6) Executive Load		
	df	F	*p*
# Tracks	3, 15	47.2	**0.000**
Difficulty	1, 5	5.0	0.077
2nd Verbal	1, 5	0.4	0.549

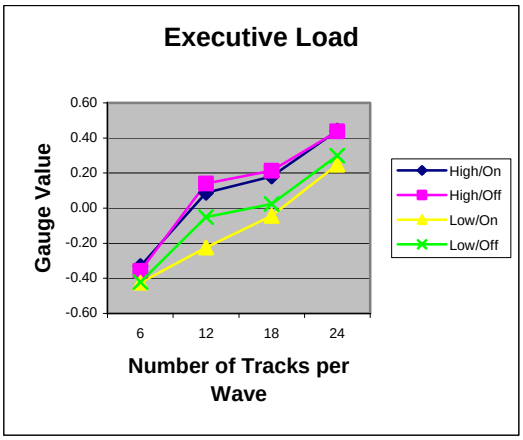


Figure 9. The mean value of the Team 3 gauge graphed against Number of Tracks per Wave, split by Track Difficulty (high and low) and Secondary Verbal Task (on and off).

The QinetiQ Executive Load gauge was significantly sensitive to the Number of Tracks per Wave and *marginally* sensitive to Track Difficulty. The gauge was not sensitive to the concurrence of the Secondary Verbal Task.[20] A significant interaction was found for Track Difficulty X Number of Tracks per Wave ($p < .05$). As shown in Figure 9, a significant difference in gauge values for Track Difficulty during wave size of 12 ($p < .05$). The wave size of 18 had a *marginal* difference, $p < .10$ and the wave size of 24 had a *potential* difference, $p < .20$.

> ***Summary:*** *The Executive Load gauge was significantly sensitive to the Number of Tracks per Wave and marginally sensitive to Track Difficulty. This gauge appears to work best for middle to higher levels of task load. See Appendix 3f for a more detailed analysis and interpretation provided by QinetiQ.*

[20] QinetiQ reported data for six participants: p1, 2, 4, 5, 7, and 8. A repeated measures ANOVA was conducted using the data from all six participants.

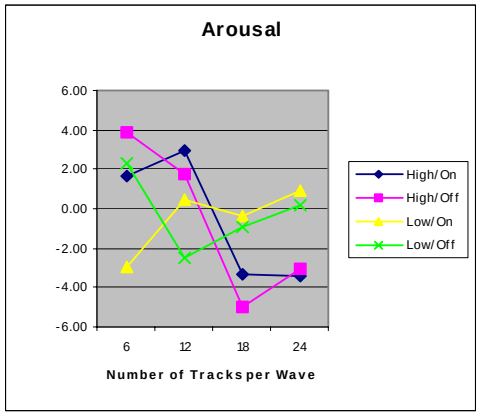
Team 4

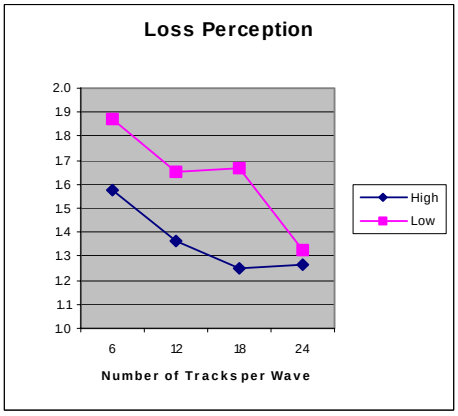
Table 8 gives the statistical results of the three task load factors on each of the gauges for Team 4. Figure 10 shows the mean value of each of the gauges for Team 4.

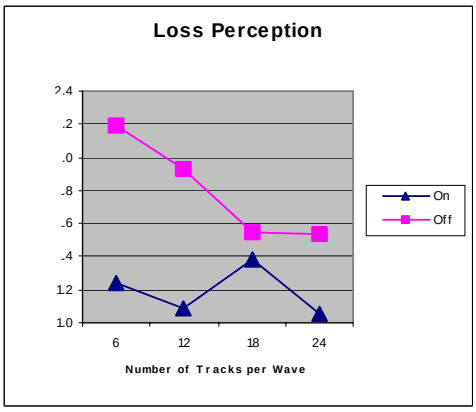
Table 8. Statistical results of the three task load factors on each of the Team 4 gauges.

Task Load Factors	AnthroTronix (n=4)			Sarnoff/Columbia (n=4)		
	df	F	*p*	df	F	*p*
# Tracks	3, 9	1.4	0.307	3, 9	1.9	0.194
Difficulty	1, 3	0.2	0.717	1, 3	0.9	0.402
2nd Verbal	1, 3	0.0	0.905	1, 3	11.3	**0.043**

Task Load Factors	UNewMexico								
	Occular Frontal (n=4)			Scynched Ant.-Post. (n=4)			Visual (n=2)		
	df	F	*p*	df	F	*p*	df	F	*p*
# Tracks	3, 9	5.0	**0.026**	3, 9	1.4	0.296	3, 6	0.6	0.656
Difficulty	1, 3	1.2	0.356	1, 3	1.0	0.400	1, 1	0.7	0.547
2nd Verbal	1, 3	0.0	0.889	1, 3	14.3	**0.032**	1, 2	0.4	0.604

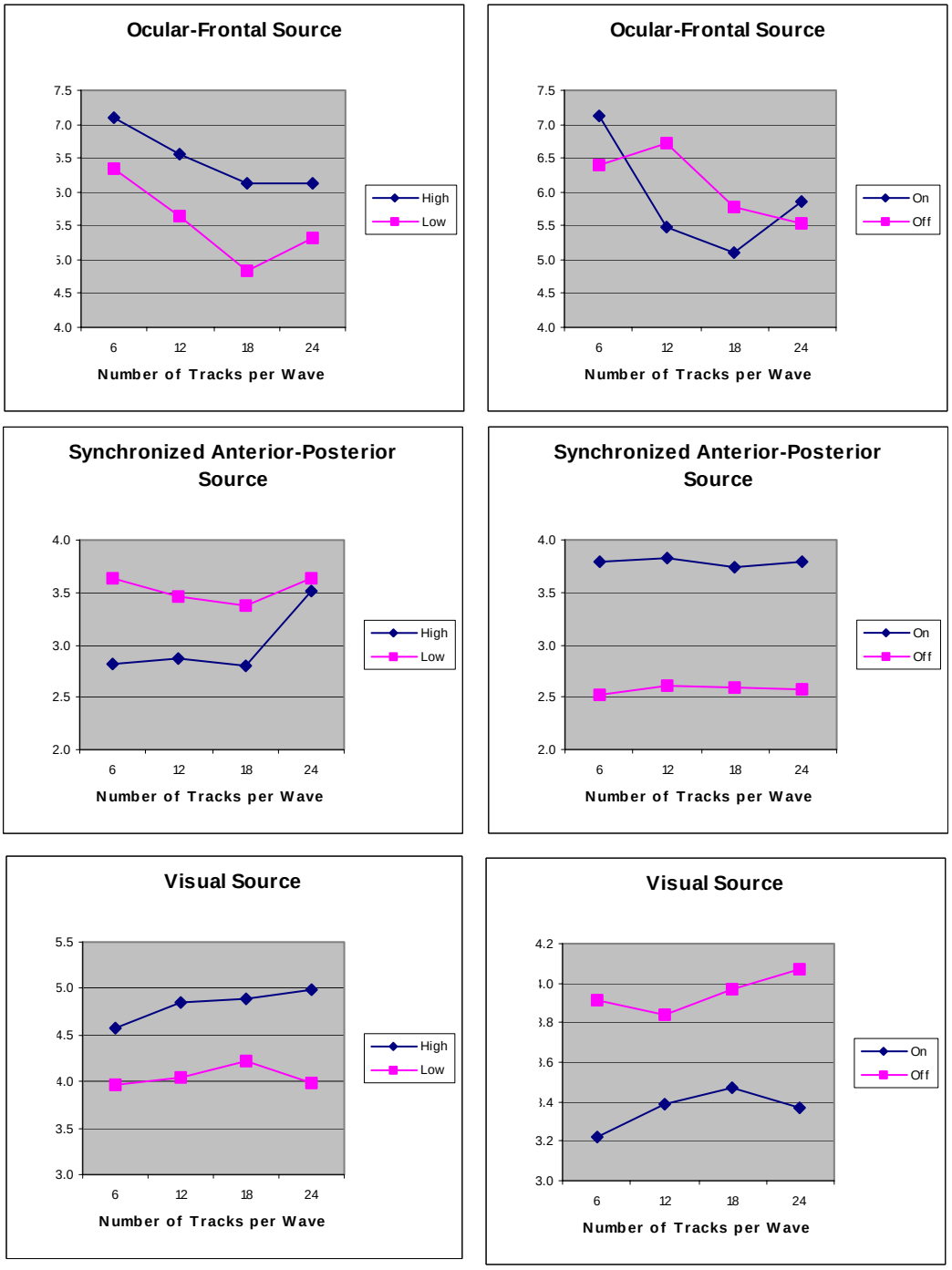






a. AnthroTronix gauge graphed against Number of Tracks per Wave, split by Track Difficulty (High and Low) and Secondary Verbal Task (on and off). The Sarnoff/Columbia gauge is shown as two separate two-way graphs based on the two-way ANOVAs used for this analysis (see text).

Figure 10**.** The mean value of Team 4 gauges.



b. New Mexico's gauges graphed against Number of Tracks per Wave and split by either Track Difficulty (High and Low) or Secondary Verbal Task (On and Off). The data is shown as two separate two-way graphs based on the two-way ANOVAs used for this analysis (see text).

Figure 10. The mean value Team 4 gauges. (continued)

Sarnoff and Columbia's EEG-based Loss Perception gauge was significantly sensitive to the concurrence of the Secondary Verbal Task and is *potentially* significant to Number of Tracks per Wave.[21] Furthermore, there was a *marginally* significant interaction ($p < .10$) for Number of Tracks per Wave X Secondary Verbal Task. These results are shown in Figure 10. However, only for wave size 6 were gauge values significantly different between Secondary Task being either on or off ($p < .05$).

> ***Summary:*** *The Loss Perception gauge was significantly sensitive to the absence/presence of the Secondary Verbal task. Gauge values were significantly greater when the secondary task was off. In addition, when Task Difficulty was high the Loss Perception gauge values were lower than when the Task Difficulty was low. See Appendix 3h for a more detailed analysis and interpretation provided by Sarnoff and Columbia University.*

Anthrotronix's Arousal gauge was not sensitive to any of the three task load factors.[22]

> ***Summary:*** *Similar to other gauges related to changes in general arousal, the Arousal Gauge of Anthrotronix also failed to demonstrate significant changes in gauge values during the task. See Appendix 3b for a more detailed analysis and interpretation provided by Anthrotronix & Lockheed Martin Advanced Technology Laboratory.*

The University of New Mexico's ERP-based Ocular-Frontal Source gauge was significantly sensitive to the Number of Tracks per Wave ($p < .05$), and the Synchronized Anterior-Posterior Source gauge was significantly sensitive to the concurrence of the Secondary Verbal Task ($p < .05$).[23]

---

[21] Sarnoff/Columbia reported data for four participants: p1, 2, 7, and 8. The data set for p7 was incomplete—only two of the four scenarios were conducted. Removing p7, however, would have reduced the sample size to 3. To salvage the data analysis, we conducted two separate two-way repeated measures ANOVAs and pooled the data for the third factor (Number of Tracks per Wave by Secondary Verbal Task and Number of Tracks per Wave by Track Difficulty). Further, since the Sarnoff/Columbia data was based on EEG responses to auditory error sounds in WCT, and since participants made few errors, especially in the lower task load waves (e.g., six tracks), there were missing cells in the data that confounded the ANOVA analyses. To allow the ANOVA to proceed, mean substitution was used to fill empty cells for waves without event markers by taking the average of available data of matching scenario and wave size across participants.

[22] AnthroTronix reported data for six participants: p1, 2, 5, 6, 7, and 8. However, the data sets for p7 and p8 were incomplete. The repeated measures ANOVA was conducted for the remaining four participants' data sets: p1, 2, 5, and 6.

[23] New Mexico reported data from four participants: p1, 2, 5, and 8. However, the data sets from p2 and p5 were incomplete. Removing these data, however, would have reduced the sample size to two. To salvage the data analysis, we conducted two separate two-way repeated measures ANOVAs and pooled the data for the third factor (Number of Tracks per Wave by Secondary Verbal Task and Number of Tracks per Wave by Track Difficulty). This procedure allowed the ANOVAs for the ocular-frontal source gauge and the synchronized anterior-posterior source gauges to be conducted with all four participants, but the ANOVA for the visual source gauge could only be conducted with two participants, making this analysis extremely speculative.

Summary: *Two of the three gauges of University of New Mexico were significantly related to changes in task demands. The Ocular-Frontal gauge was sensitive to changes in task loading (Number of Tracks per Wave,) wherea, the Synchronized Anterior-posterior gauge was sensitive to changes in the concurrence of the Secondary Verbal Task. See Appendix 3j for a more detailed analysis and interpretation provided by University of New Mexico.*

Team SDSU

Table 9 gives the statistical results of the three task load factors for each SDSU gauge. Figure 11 shows the mean value of each SDSU gauges.

Table 9. Statistical results of the three task load factors on each of the SDSU gauges.

Task Load Factors	SDSU (n=7) ICA		
	df	**F**	***p***
# Tracks	3, 4	5.8	0.061
Difficulty	1, 6	0.0	0.964
2nd Verbal	1, 6	9.2	0.023

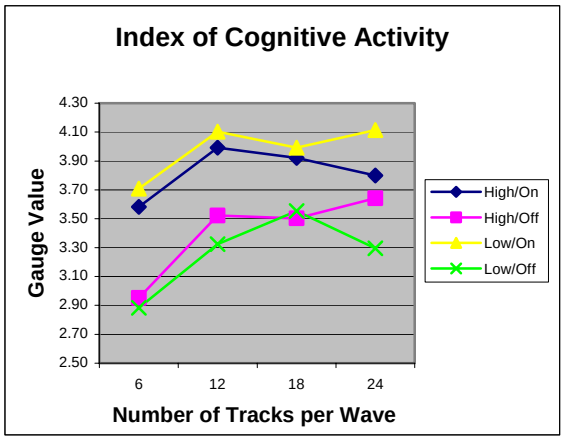


Figure 11. The mean value of each SDSU gauge graphed against Number of Tracks per Wave, split by Track Difficulty (High and Low) and Secondary Verbal Task (On and Off).

SDSU reported substantial levels of electro-magnetic interference (10 Hz) during some sessions. This interference had not occurred previously in their lab or during a pilot study held at the TIE testing site. The interference problem has not been replicated, and its cause has not been determined. It appears to have been limited to the TIE location or other localized variable (Marshall, July 2003, personal communication). SDSU was able to filter some of the interference and report data for seven

participants.[24] This unexpected phenomenon limited the sensitivity of the ICA gauge during the TIE below levels achieved in previous pilot studies. See the SDSU appendix for more information.

The San Diego State University gauge was significantly sensitive to the concurrence of the Secondary Verbal Task, and marginally sensitive to the Number of Tracks per Wave. In addition, there was a *potentially* significant ($p < .20$) two-way interaction for Number of Tracks per Wave X Secondary Verbal Task. Figure 11 shows differences between when the Secondary Verbal Task is either on or off for wave sizes of 6 ($p < .05$) and 24, $p < .05$. For the wave sizes of 12 and 18, there is a *marginal* difference, $p < .10$.

> ***Summary:*** *The Index of Cognitive Activity (ICA) was significantly sensitive to changes related to the Secondary Verbal Task. Gauge values were significantly higher when the Secondary Verbal Task was on. The ICA was also marginally sensitive to changes in the Number of Tracks per Wave. Gauge values increased as the number of Tracks per Wave increased. See Appendix 3g for a more detailed analysis and interpretation provided by San Diego State University.*

Overview of Results

From these analyses, we can conclude that a number of the gauges were sensitive, or *marginally* sensitive, to the three task load factors. Specifically,

1. Eight of the 20 different gauges were significantly sensitive to changes in the number of tracks per wave ($p < .05$):
   a. Drexel's fNIR-based measures for both the left and right hemispheres,
   b. Advanced Brain Monitoring, Inc.'s two EEG-based measures of vigilance,
   c. QinetiQ's EEG-based measure of executive load,
   d. Hawaii's pressure mouse-based perceptual/motor and cognitive difficulty measures, and
   e. New Mexico's ERP-based measure of theta power from an ocular-frontal source.
2. Two gauges were *marginally* sensitive ($p < .1$) to changes in the number of tracks per wave:
   a. Pittsburgh/NRL's body posture measure of head-monitor coupling
   b. SDSU's pupil-based "index of cognitive activity."
3. Three additional gauges were "potentially" sensitive ($p < .2$) to changes in the number of tracks per wave:
   a. Pittsburgh/NRL's body posture measure of head bracing,
   b. Electrical Geodesics, Inc.'s ERP-based measure of motor effort, and
   c. Sarnoff/Columbia's ERP-based measure of "loss perception."
4. Two gauges were significantly sensitive to track difficulty:
   a. Hawaii's two mouse measures.

---

[24] SDSU reported data from 12 participants: from team 1, p1, 4, and 7; from team 2, p1, 2, 4, 6, and 7; from team 3, p4 and 8, and from team 4, p6 and 7. However, SDSU asked that five data sets be excluded due to the extent of EMI. The repeated measures ANOVA was conducted on seven data sets from Team 1, p4; team 2, p2, 6, and 7; team 3, p4, and 8; and team 4, p6. SDSU requested a multivariate analysis of variance.

5. Two gauges were *marginally* sensitive to track difficulty:
   a. QinetiQ's executive load,
   b. EGI's EEG-based auditory effort gauge, and
6. Two additional gauges were *potentially* sensitive to track difficulty:
   a. ABM's high vigilance measure,
   b. Pittsburgh/NRL's back bracing measure.
7. Three gauges were significantly sensitive to the concurrence of the secondary verbal task:
   a. SDSU's pupil-based "index of cognitive activity,"
   b. Sarnoff/Columbia's loss perception gauge, and
   c. New Mexico's ERP-based measure of theta power from synchronized anterior-posterior sources.
8. Three additional gauges were *potentially* sensitive to the concurrence of the secondary verbal task:
   a. EGI's motor effort and auditory effort gauges, and
   b. Pittsburgh/NRL's head bracing measure.

Many of the analyses were performed on very small samples of data. Consequently, both positive results, and especially, negative results should be interpreted with healthy skepticism. The very liberal criterion of $p < .2$ was used to define "*potentially*" significant because of the small sample sizes, the complexity of the data collection—many concurrent gauges and compressed time frame for data collection—and the experimental nature of many of the gauges. In sum, many of the gauges show a good deal of promise.

Table 10 shows the effect size for each gauge and task load factor using the *eta squared* statistic.[25] Given the small sample sizes in this study, solely relying on *p*-values as a measure of sensitivity to identify effects may be problematic insofar as the lack of power may undermine the ability to detect potential effects. Hence, in alignment with the most recent edition of the *American Psychological Association Publication Manual* (2001),[26] we report effect sizes to further illuminate the findings. For this study, given the repeated measures design, the effect size **f** (a standardized measure of the difference between means) as cited in Cohen (1988)[27] can be used for ANOVA and related complex designs. This **f** index can be converted to a measure of magnitude (or strength), and for the purposes of this study, the effect size of interest is the partial eta squared (η^2), which relates to the proportion of variance accounted for by each specific gauge. Cohen makes clear that any taxonomy of small/medium/large effects must take into consideration context and prior strength of effects. Due to the small sample sizes in this study, effects can appear large without becoming statistically significant. Based on this fact, and on an analysis of the distribution of effect sizes in this study, a proportion of 0.40 was set as the definition of a "large" effect. "Large" effect sizes are highlighted in the table.

---

[25] Eta squared is the proportion of the variability in the dependent (measured) variable that can be accounted for by the variation in the independent (manipulated) variable. Thus, the larger the eta squared value the greater the degree to which the variation in the measure is attributed to the different levels of the independent variable. Eta squared is computed by the Analysis of Variance using the ratio of the sum of squares effect/sum of squares total.

[26] American Psychological Association. (2001). *American Psychological Association Publication Manual* (5th Ed.). Washington DC.

[27] Cohen, J. (1988). Statistical power analysis for the behavioral sciences. (2nd Ed.). Hillsdale, NJ: Lawrence Erlbaum.

The effect sizes tell a complementary story to the significance levels from the analyses of variance. The effect sizes can be used to substantiate statistically significant gauges and to help identify non-statistically significant results for gauges that may hold some promise.

Table 10. Effect sizes (η^2) of the statistical results for each specific gauge for each task load factor.

			Task Load Factors		
Gauge	**Sensor Type**	**Performer**	**Number of Tracks per Wave (6,12,18,24)**	**Track Difficulty (Hi/Lo)**	**Secondary Verbal Task (On/Off)**
Team 1					
Arousal Meter	Inter-Heart Beat Interval	Clemson U	0.11	0.06	0.00
Head-Monitor Coupling	Head Posture	UPitt/NRL	**0.55**	0.07	0.17
Head Bracing	Body Posture	UPitt/NRL	0.34	0.18	**0.42**
Back Bracing	Body Posture	UPitt/NRL	0.08	0.27	0.01
Motor Effort	ERP-IFF	EGI	0.34	0.00	**0.43**
Auditory Effort	ERP-Engage Sound	EGI	0.12	**0.52**	**0.45**
Team 2					
fNIR (left)	Blood Oxygenation	DrexelU	**0.68**	0.06	0.14
fNIR (right)	Blood Oxygenation	DrexelU	**0.63**	0.05	0.13
Percent High Vigilance	EEG	ABM	**0.69**	0.28	0.05
Probability Low Vigilance	EEG	ABM	**0.62**	0.15	0.01
Arousal	GSR	UHawaii	0.16	0.04	0.14
Perceptual/Motor Load	Mouse clicks	UHawaii	**0.98**	**0.84**	0.06
Cognitive Difficulty	Mouse pressure	UHawaii	**0.99**	**0.66**	0.13
Team 3					
Executive Load	EEG	QinetiQ	**0.90**	**0.50**	0.08
Team 4					
Arousal	GSR	AnthroTronix	0.32	0.05	0.01
Loss Perception	ERN-Error Sounds	Sarnoff/Columbia	0.39	0.24	**0.79**
Occular-Frontal Source	ERP-Comms	UNewMexico	**0.67**	0.28	0.01
Synched Anterior-Posterior	ERP-Comms	UNewMexico	0.32	0.24	**0.83**
Visual Source	ERP-Comms	UNewMexico	**0.43**	**0.43**	0.16
Team SDSU					
Index of Cognitive Activity	Pupil dilation	SDSU	0.19	0.00	**0.60**

4.3 GAUGE CONSISTENCY

How well and how consistently does each gauge correlate with task load (Number of Tracks per Wave) and the six performance measures, on a wave by wave basis?

ANOVAs in the previous section provide an overview of the performance of each gauge. In addition, we now examine 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? Further, we determine if each gauge is especially sensitive to one or more of the various performance measures: is it especially sensitive to specific aspects of performance such as increasing errors or increasing delays or decreases in task completion? To address these questions, this section is divided into separate subsections devoted to each gauge.

To measure the sensitivity of a gauge to detect task load changes for each participant, we first computed the correlation between the task load and the gauge's value for each wave of a scenario.[28] For this analysis, we only examined the task load factor of number of tracks per wave, since this factor varies from very low task load to very high task load, and many gauges were able to detect changes in it. Second, we computed the mean of the correlations from each of the scenarios that a participant performed during the course of an experiment session. This mean tells us how well the gauge tracked the task load of that participant throughout an experiment session. Third, we computed the mean of the mean correlations. This overall correlation tells us how well, across participants, a gauge tracked changes in task load. Because the sample sizes in the TIE were small, statistical significance for the correlations was rare. Rather than report significance tests, then, we adopted fairly liberal criteria for defining "meaningful" correlation mean correlations. Correlations greater than 0.6 were considered high, correlations greater than 0.3 were considered moderate, and correlations less than 0.3 were considered low. Figure (a) for each gauge shows the mean correlation for each participant and the overall correlation.

Lastly, in order to have a measure of consistency of the correlations across participants, we took the mean correlation for each participant and computed the standard deviation of those means.[29] A consistent gauge would show similar sized correlations for each participant and therefore a small standard deviation. A less consistent gauge would show different size correlations for each participant and therefore a large standard deviation. Of course, a gauge with no sensitivity would be consistent—consistently poor—so both sensitivity and consistency must be considered simultaneously. Consistency values are reported for each gauge and summarized in Table 11.

The second analysis examines how consistently, or reliably, a gauge correlates with each of the performance measures computed from the WCT. Using the performance measures, we can describe a "performance profile" for each participant during a session. We can then compute correlations between a gauge's output values and each performance measure, wave by wave. These correlation coefficients can tell us which aspects of user performance are well correlated with a gauge and which aspects of performance a specific gauge measures best. The mean correlation across participants was graphed in the (b) graph for each gauge.

---

[28] The correlation measures how well a gauge "tracks" task load over the course of a scenario. High positive correlation coefficients (r) indicate that the gauge value increases proportionally as task load increases and decreases proportionally as task load decreases. Correlation coefficients range from 1.0 to -1.0. A correlation of -1.0 would indicate that the gauge consistently moved in the opposite direction as the task load.

[29] The variance (σ) is a description of the distribution of data values around the mean of all values. It is calculated by summing the squared differences between each data point from the mean, divided by the sample size minus one. A small variance indicates a low amount of variability in the measurement (which suggests uniformity).

Lastly, we graphed the mean correlation for each participant for each performance measure in the (c) graph for each gauge. This graph shows the individual variability of correlations across performance measures.

The final column of Table 11 reports the consistency of each gauge for tracking changes in task load across participants. A filled circle indicates a high level of consistency across participants in the degree of sensitivity to changes in task load for that gauge (all participants showed a similar size correlation between gauge value and number of tracks per wave, $\sigma < .15$). A half-filled circle indicates a moderate level of consistency across participants (participants showed moderately different size correlations, $\sigma < .30$). An open circle indicates a low level of consistency across participants (participants showed widely different size correlations, $\sigma > .30$).

Table 11. Variance of the mean correlations for each specific gauge for each task load factor.

Gauge	Sensor Type	Developer	Task Load Factors: Number of Tracks per Wave (6,12,18,24)	Task Load Factors: Track Difficulty (Hi/Lo)	Task Load Factors: Secondary Verbal Task (On/Off)	Consistency Across Participants
fNIR						
fNIR (left)	Blood Oxygenation	DrexelU	●	○	○	◒
fNIR (right)	Blood Oxygenation	DrexelU	●	○	○	◒
EEG-Continuous						
Percent High Vigilance	EEG	ABM	●	◓	○	◒
Probability Low Vigilance	EEG	ABM	●	○	○	◒
Executive Load	EEG	QinetiQ	●	◒	○	●
EEG-ERP						
Motor Effort	ERP-IFF	EGI	◓	○	◓	●
Auditory Effort	ERP-Engage Sound	EGI	○	◒	◓	◒
Loss Perception	ERN-Error Sounds	Sarnoff/Columbia	◓	○	●	◒
Occular-Frontal Source	ERP-Comms	UNewMexico	●	○	○	●
Synched Anterior-Posterior	ERP-Comms	UNewMexico	○	○	●	●
Visual Source	ERP-Comms	UNewMexico	○	○	○	●
Arousal						
Arousal Meter	Inter-Heart Beat Interval	Clemson U	○	○	○	●
Arousal	GSR	UHawaii	○	○	○	◒
Arousal	GSR	AnthroTronix	○	○	○	◒
Physiological						
Head-Monitor Coupling	Head Posture	UPitt/NRL	◒	○	○	○
Head Bracing	Body Posture	UPitt/NRL	◓	○	◓	◒
Back Bracing	Body Posture	UPitt/NRL	○	◓	○	◒
Perceptual/Motor Load	Mouse clicks	UHawaii	●	●	○	●
Cognitive Difficulty	Mouse pressure	UHawaii	●	●	○	●
Index of Cognitive Activity	Pupil dilation	SDSU	◒	○	●	○

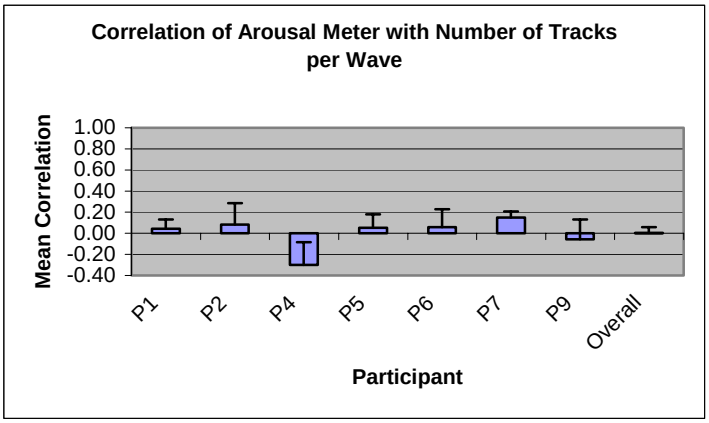
4.3.1 Clemson University – Arousal Meter

Figure 12a shows that the mean correlation for each participant varied from -0.30 to 0.15, with a mean of 0.00. The standard deviation of the mean correlations was 0.15. Thus, the mean size of the correlations was low, but the consistency of the correlations was high ($\sigma = 0.15$).

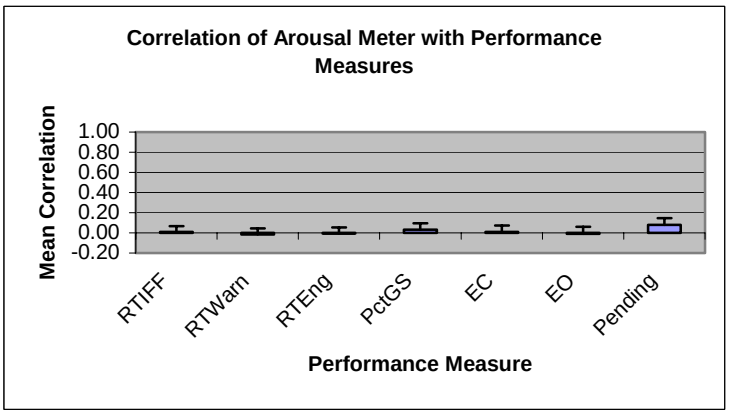
Figure 12b shows the mean correlation between the Arousal Meter gauge readings and the performance measures across participants.

Figure 12c shows the mean correlation between the Arousal Meter gauge readings of participants across the performance each measure. As shown in the graph, there is considerable variability between the participant correlations that range from .30 to -.31.

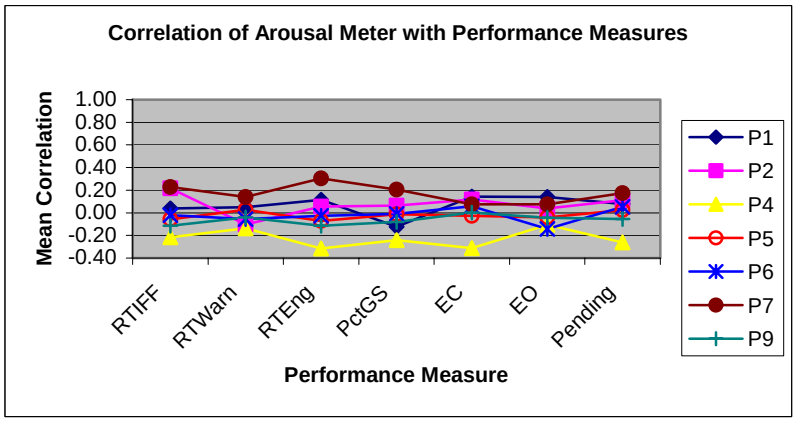
Summary: *The results indicate that the Arousal Meter was not sensitive to task load changes related to Number of Tracks per Wave or any of the performance measures of the WCT. The high consistency indicates that the gauge was not sensitive for any of the participants.*



a. Correlation between gauge value and Number of Tracks per Wave. Error bars are standard errors of the mean.



b. Correlation between gauge value and performance measures.



c. Correlation of gauge value and performance measures for each participant.

Figure 12. Clemson Arousal Meter.

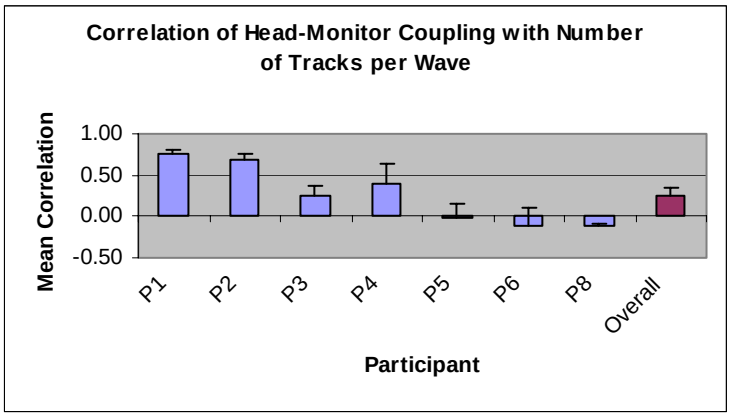
4.3.2 University of Pittsburgh/Naval Research Laboratory Head-Monitor Coupling

Figure 13a shows the mean correlation for each participant varied from -0.12 to 0.77, with a mean of 0.26. The standard deviation of the mean correlations was 0.37. Thus, the mean size of the correlations was low (though just shy of moderate), and the consistency of the correlations was low ($\sigma > 0.30$).

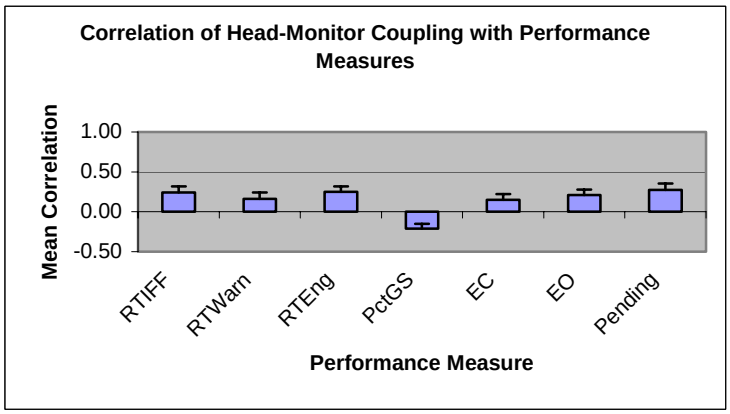
Figure 13b shows the mean correlation between the Head-Monitor Coupling gauge readings and the performance measures across participants. With the exception of the Percent Game Score (PctGS) gauge, which has an expected negative correlation (-.21), correlations range from .15 to .30. The Head-Monitor Coupling gauge had a low correlation with the performance measures.

Figure 13c shows the mean correlations for participants between the Head-Monitor Coupling gauge values with Performance Measures. As can be seen, there are large individual differences in how well the gauge correlates with the various performance measures (ranging from .70 to -.51).

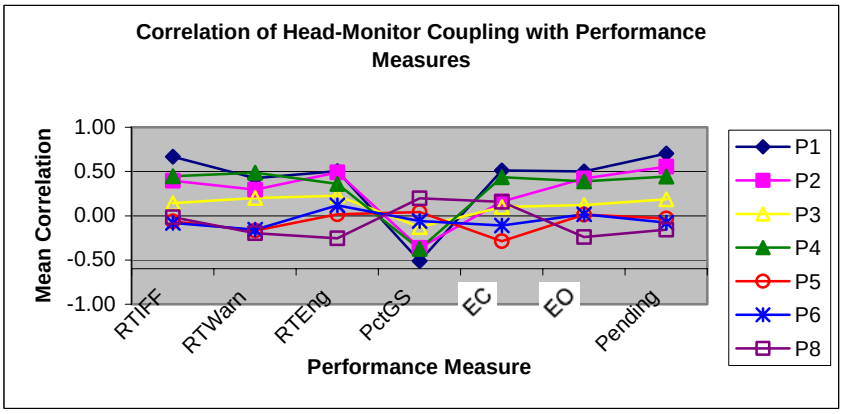
Summary: *The Head-Monitor Coupling gauge was inconsistent in its ability to track changes in task load; it was sensitive for some participants but not for others. Overall, its sensitivity was classified as low. These results indicate the gauge may have potential for predicting changes that occur as task load changes. The gauge also exhibited a moderate relationship with changes in the performance measures, suggesting potential as a predictive tool for variation in participant performance due to task load changes.*



a. Correlation between gauge value and Number of Tracks per Wave.



b. Correlation between gauge value and performance measures.



c. Correlation of gauge value and performance measures for each participant.

Figure 13. UPitt/NRL Head-Monitor Coupling.

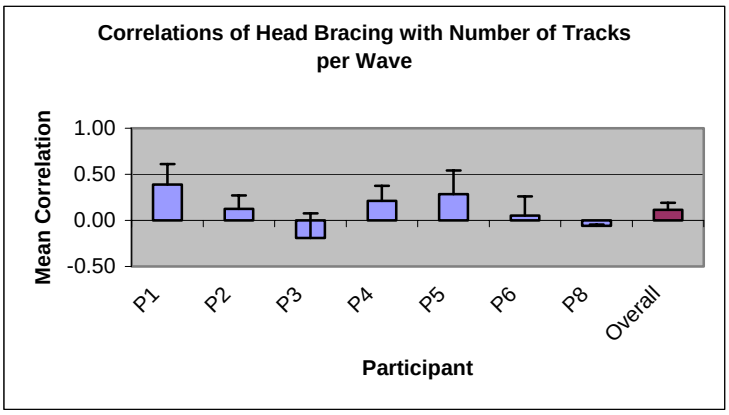
4.3.3 University of Pittsburgh/Naval Research Laboratory – Head Bracing

Figure 14a shows the mean correlation for each participant varied from -0.19 to 0.39, with a mean of 0.12. The standard deviation of the mean correlations was 0.20. Thus, mean size of the correlations was low, and the consistency of the correlations was moderate ($0.15 < \sigma < 0.30$).

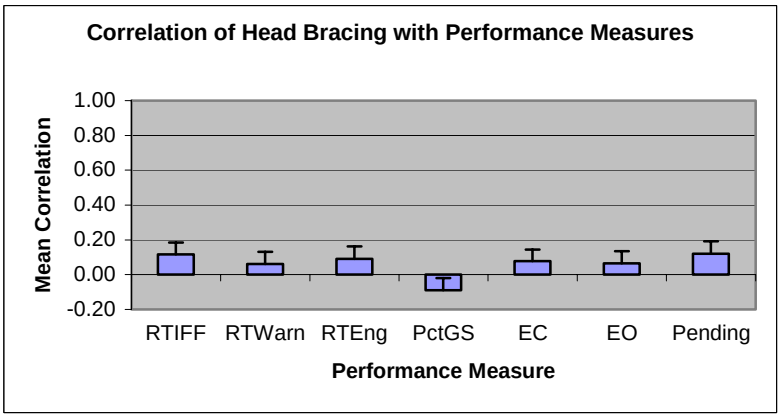
Figure 14b shows the mean correlation between the Head Bracing Monitor gauge readings and the performance measures across participants. The Head Bracing Monitor had very low correlations, -.09 to .12, with each of the performance measures.

Figure 14c shows the correlations between the gauge values and performance measures for each of the participants. The correlations range from a low -.20 to a moderate .40.

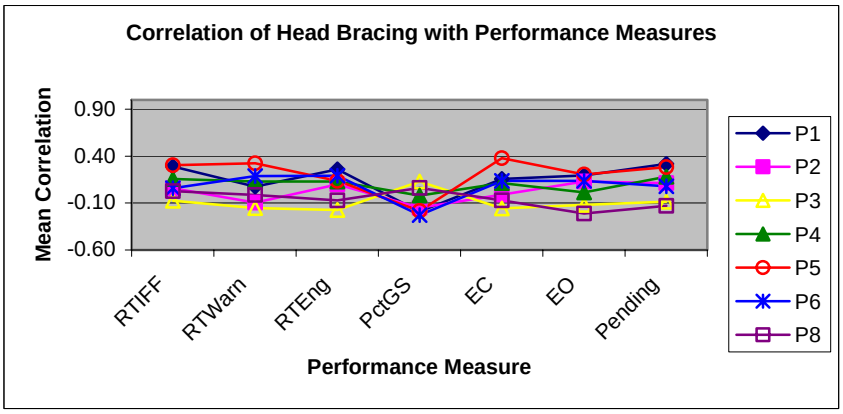
Summary: *The Head Bracing gauge values provided low correlations across performance measures and participants. The consistency of correlations across participants was moderate.*



a. Correlation between gauge value and Number of Tracks per Wave



b. Correlation between gauge value and performance measures.



c. Correlation of gauge value and performance measures for each participant.

Figure 14. UPitt/NRL Head Bracing.

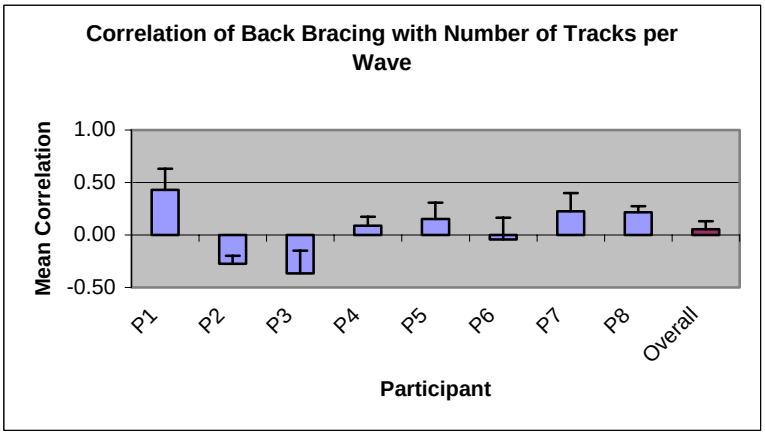
4.3.4 University of Pittsburgh/Naval Research Laboratory – Back Bracing

Figure 15a shows that the mean correlation for each participant varied from -0.37 to 0.43, with a mean of 0.05. The standard deviation of the mean correlations was 0.27. Thus, size of the mean correlation was low, but the consistency of the correlations was moderate ($0.15 < \sigma < 0.30$).

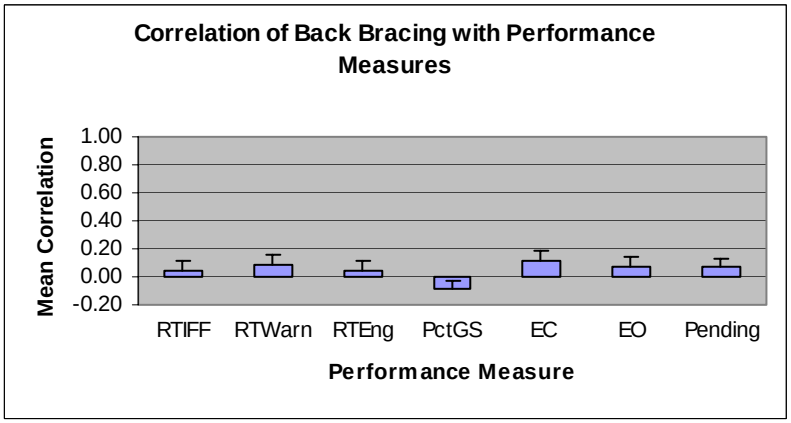
Figure 15b shows the mean correlation between the Back Bracing gauge readings and the performance measures. Correlations range from .11 for Errors of Commission (EC) to -.09 for Percent Game Score (PctGS).

Figure 15c shows the correlation of the gauge values and performance measures for each participant. As can be seen in the graph, there appears to be a great deal of variability in the correlations between participants for a given measure but fairly consistent correlations within a single participant.

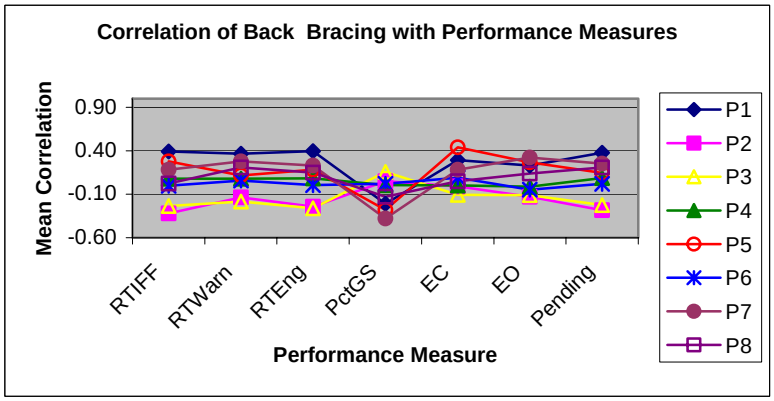
Summary: The Back Bracing gauge demonstrated a low level of sensitivity to task load. The gauge also did not correlate well with any of the performance measures.



a. Correlation between gauge value and Number of Tracks per Wave.



b. Correlation between gauge value and performance measures.



c. Correlation of gauge value and performance measures for each participant.

Figure 15. UPitt/NRL Back Bracing.

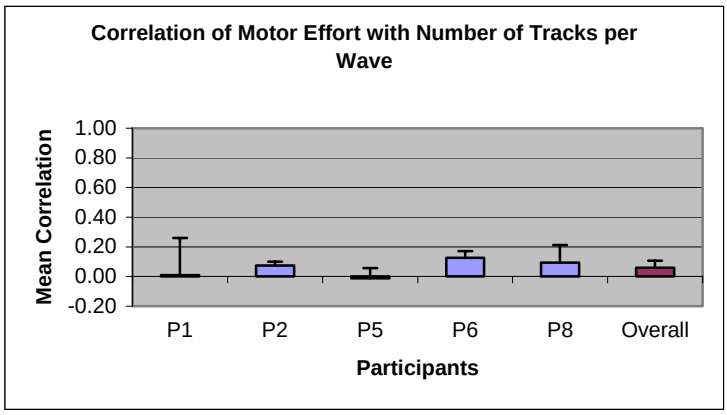
4.3.5 Electrical Geodesics, Inc. – Motor Effort

Figure 16a shows that the mean correlation for each participant varied from -0.01 to 0.13, with a mean of 0.06. The standard deviation of the mean correlations was 0.06. Thus, size of the mean correlation was low, but the consistency of the correlations was high ($\sigma < 0.15$). Correlations were consistent across the participants. Overall, the Motor Effort gauge values were not related to task load changes initiated by the Number of Tracks per Wave.

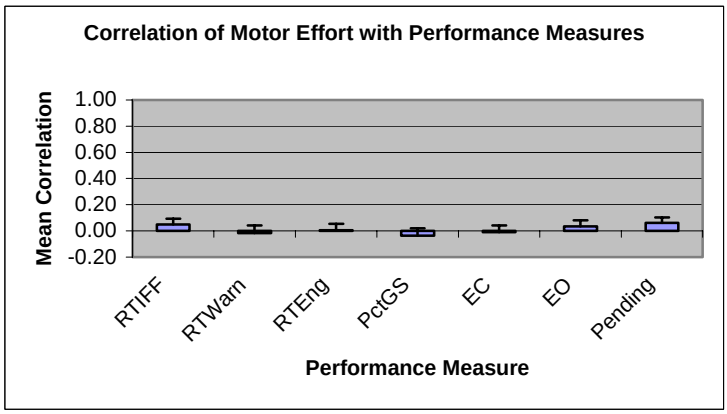
Figure 16b shows the mean correlation between the Motor Effort gauge value and performance measures. Each of the correlations was very low, demonstrating little relationship between the gauge value and the performance measures.

Figure 16c shows the correlation of the Motor Effort gauge reading and performance measures for each participant. Correlations for participants appear to be consistently low across all of the performance measures.

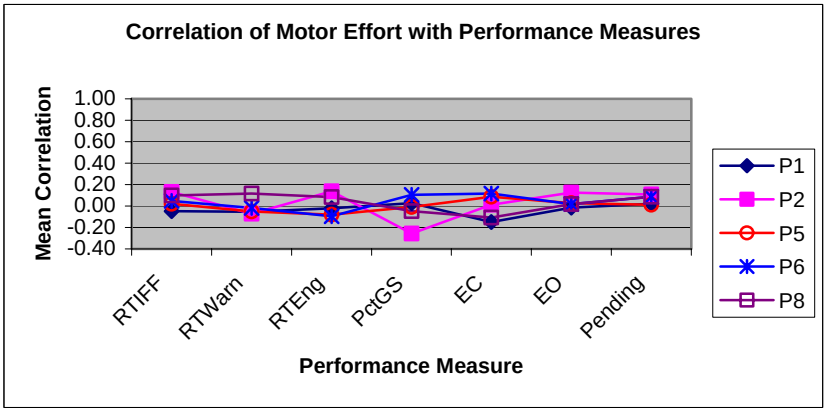
Summary: *The Motor Effort gauge was not sensitive to changes in task load for any participant.*



a. Correlation between gauge value and Number of Tracks per Wave.



b. Correlation between gauge value and performance measures.



c. Correlation of gauge value and performance measures for each participant.

Figure 16. EGI Motor Effort.

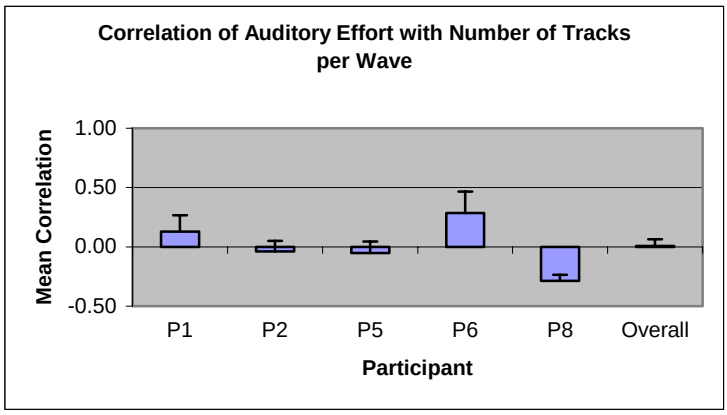
4.3.6 Electrical Geodesics, Inc. – Auditory Effort

Figure 17a shows that the mean correlation for each participant varied from -0.29 to 0.29, with a mean of 0.01. The standard deviation of the mean correlations was 0.21. Thus, size of the mean correlation was low, but the consistency of the correlations was moderate ($0.15 < \sigma < 0.30$). Overall, the Auditory Effort gauge values were not related to task load changes associated with the Number of Tracks per Wave.

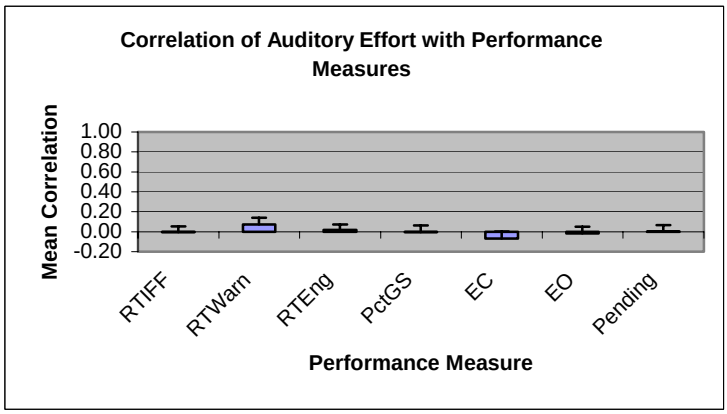
Figure 17b shows the mean correlation between the Auditory Effort gauge readings and the performance measures across participants. The Auditory Effort gauge demonstrated little relationship to any of the performance measures.

Figure 17c shows the correlation of the Auditory Effort gauge and performance measures for each of the participants. The correlations range from low (.02) to moderate (.33).

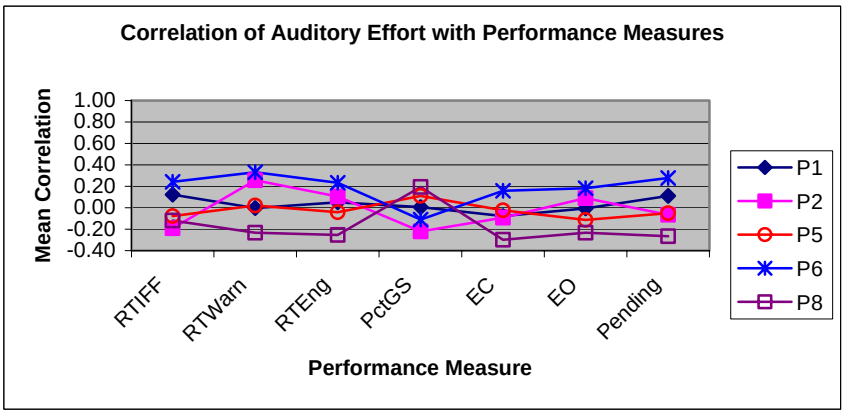
Summary: *The Auditory Effort gauge results indicated little relationship to task load changes. The gauge values were not found to be indicative of changes with any of the performance measures. The auditory effort gauge values were highly variable between participants.*



a. Correlation between gauge value and Number of Tracks per Wave.



b. Correlation between gauge value and performance measures.



c. Correlation of gauge value and performance measures for each participant.

Figure 17. EGI Auditory Effort.

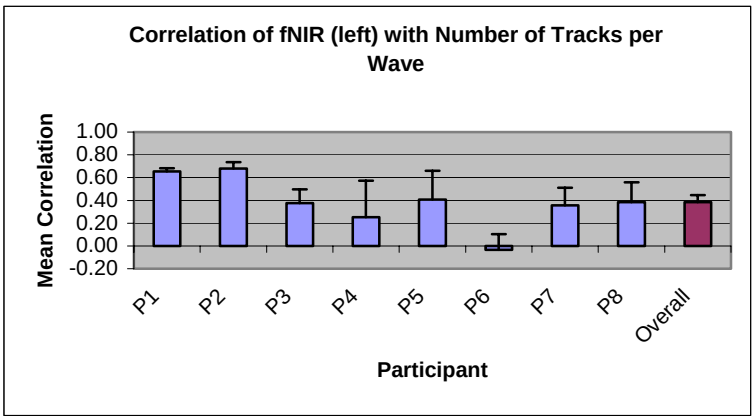
4.3.7 Drexel University – fNIR (left)

Figure 18a shows that the mean correlation for each participant varied from -0.03 to 0.68, with a mean of 0.38. The standard deviation of the mean correlations was 0.22. Thus, size of the mean correlation was moderate, and the consistency of the correlations was moderate ($0.15 < \sigma < 0.30$). Overall, the fNIR (left) gauge values were found to have a moderate relationship to task load changes associated with the Number of Tracks per Wave.

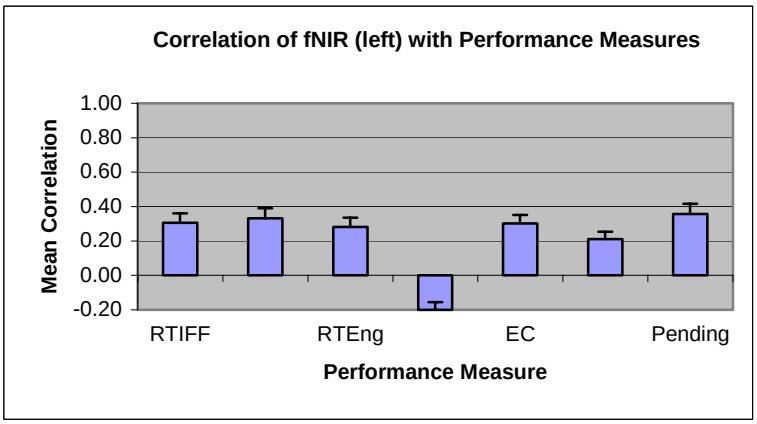
Figure 18b shows the mean correlation between the fNIR (left) gauge readings and the performance measures across participants. fNIR (left) gauge values ranged from .21 to .36, demonstrating a low to moderate correlation with the performance measures.

Figure 18c shows the correlation of the fNIR (left) gauge value and performance measure for each of the participants. Although the variability between participants was high, correlations within an individual for a specific performance measure was fairly consistent.

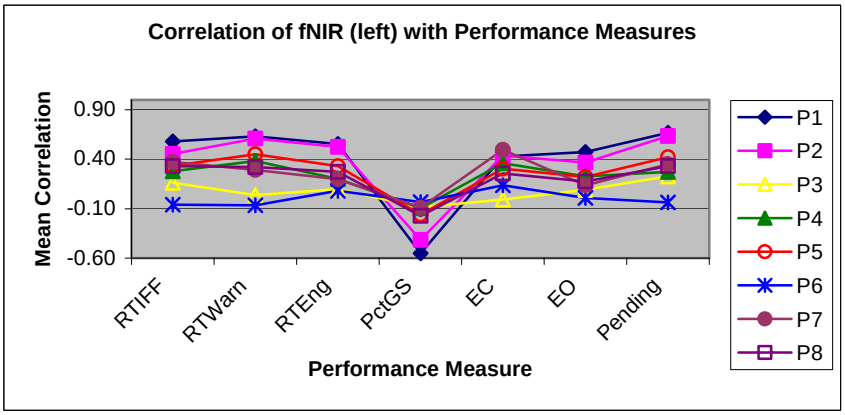
Summary: *The fNIR (left) gauge was found to have a relationship with changes in task load. Results indicate the gauge has potential to predict changes that occur as task load changes. Gauge values were also sensitive to changes in performance associated with task load. Although there was not consistency between participants, there was a fair amount of consistency for individual participants performance in each of the measures.*



a. Correlation between gauge value and Number of Tracks per Wave.



b. Correlation between gauge value and performance measures.



c. Correlation of gauge value and performance measures for each participant.

Figure 18. Drexel fNIR (left).

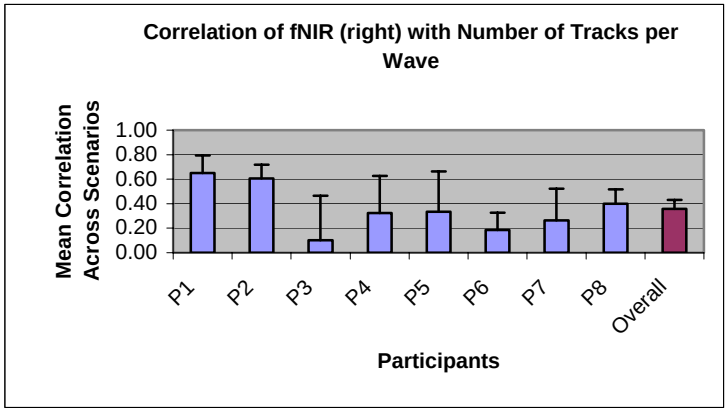
4.3.8 Drexel University – fNIR (right)

Figure 19a shows the mean correlation for each participant varied from 0.10 to 0.65, with a mean of 0.36. The standard deviation of the mean correlations was 0.19. Thus, size of the mean correlation and the consistency of the correlations ($0.15 < \sigma < 0.30$) were moderate.

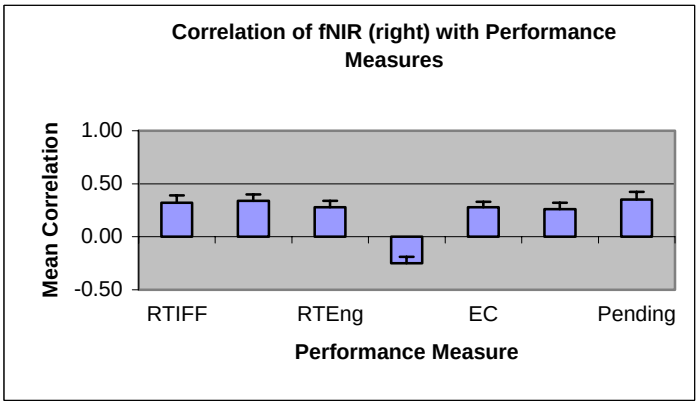
Figure 19b shows the mean correlation between the fNIR (right) gauge value and performance measures across participants. The fNIR (right) gauge moderately correlated with Response Time to Identify Friend or Foe (RTIff), Response Time to Warn (RTWarn) and Tasks Pending (Pending). Correlations for Response Time to Engage (RTEng), Percent Game Score (PctGS), Errors of Commission (EC), and Errors of Omission (EO) were low.

Figure 19c shows the correlation of the fNIR (right) gauge readings and performance measures for each participant. Positive correlations across all participants were found with the exception of participant 3 and excluding Percent Game Score (PctGS).

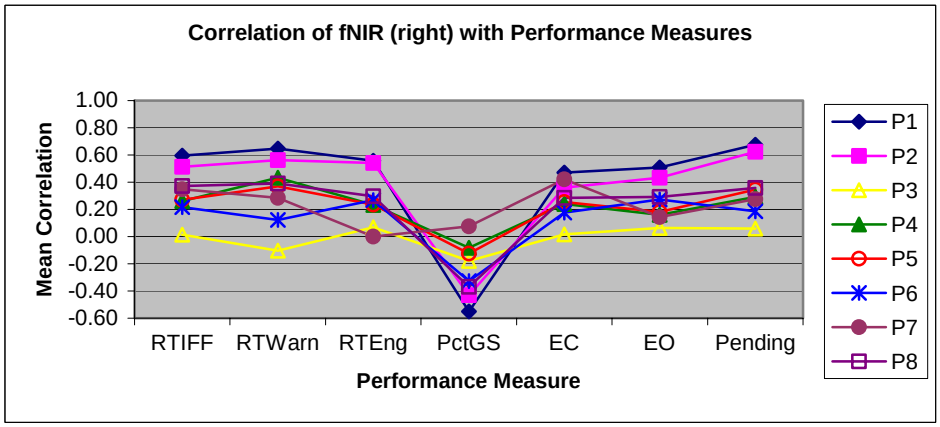
Summary: *The fNIR (right) gauge values were moderately predictive of task load changes related to Number of Tracks per Wave. These results were moderately consistent across each of the participants. Gauge values and Response Time to Identify Friend or Foe, Response Time to Warn and Tasks Pending were found to moderately coincide with changes in the performance measures, indicating the gauge has potential to be predictive of variation in participant performance due to task load changes.*



a. Correlation between gauge value and Number of Tracks per Wave.



b. Correlation between gauge value and performance measures.



c. Correlation of gauge value and performance measures for each participant.

Figure 19. Drexel fNIR (right).

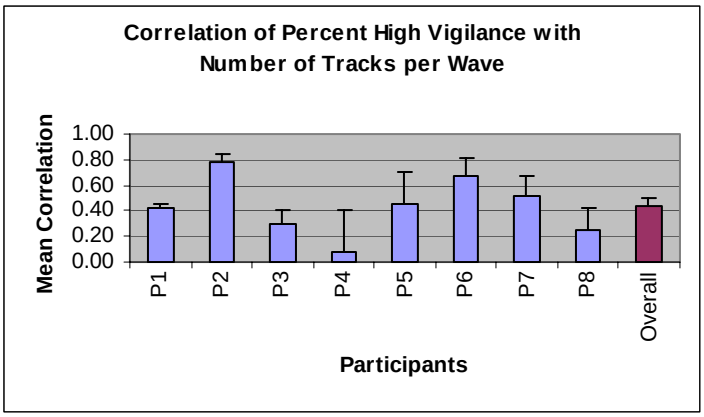
4.3.9 Advanced Brain Monitoring – Percent High Vigilance

Figure 20a shows that the mean correlation for each participant varied from 0.08 to 0.79, with a mean of 0.44. The standard deviation of the mean correlations was 0.23. Thus, size of the mean correlation was moderate, and the consistency of the correlations was moderate ($0.15 < \sigma < 0.30$).

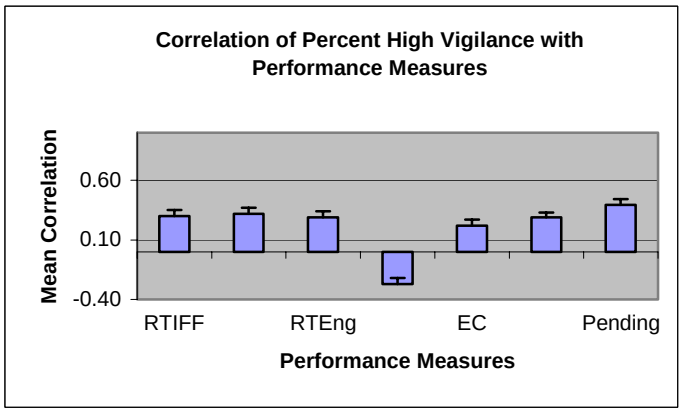
Figure 20b shows the mean correlation for the Percent High Vigilance gauges values and performance measures across participants. The Percent High Vigilance demonstrated a moderate correlation with Response Time to Warn (RTWarn) and Tasks Pending (Pending), Correlations of the Response Time to IFF (RTIff), Response Time to Engage (RTEng), Percent Game Score (PctGS), Errors of Commission (EC) and Errors of Omission (EO) performance measures were low.

Figure 20c shows the correlation of the Percent High Vigilance gauge readings and performance measures for each participant. Although the variability between participants is high, there is some consistency within participants across the performance measures.

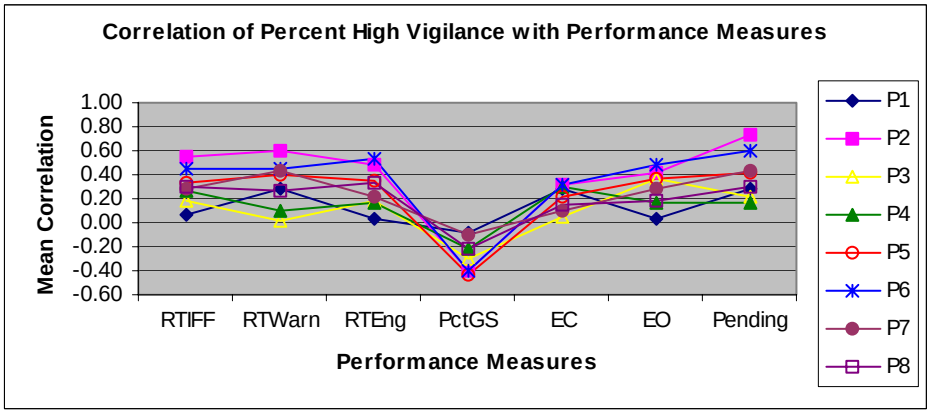
Summary: *The Percent High Vigilance gauge values demonstrated a moderate correlation with the task load changes of Number of Tracks per wave. Gauge values were moderately sensitive to the Response Time to Warn and Tasks Pending changes in participant performance that are associated with task load changes and had low correlations with the other performance measures. These results indicate the gauge may have the ability to predict changes that occur during task performance.*



a. Correlation between gauge value and Number of Tracks per Wave.



b. Correlation between gauge value and performance measures.



c. Correlation of gauge value and performance measures for each participant.

Figure 20. Advanced Brain Monitoring Percent High Vigilance.

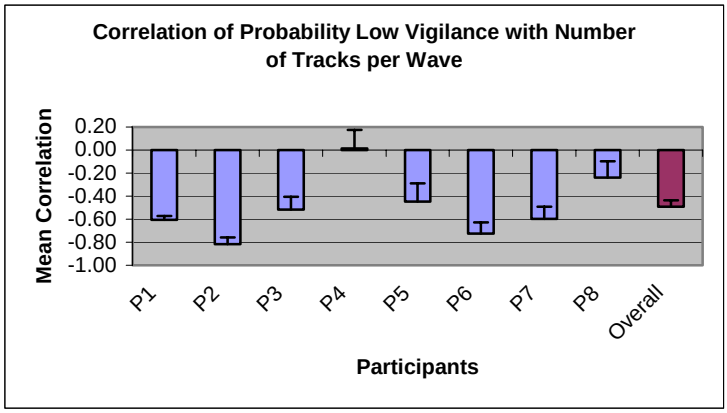
4.3.10 Advanced Brain Monitoring – Probability Low Vigilance

Figure 21a exhibits the mean correlation for each participant varied from -0.82 to 0.01, with a mean of -0.49. The standard deviation of the mean correlations was 0.27. Thus, size of the mean correlation and the consistency of the correlations ($0.15 < \sigma < 0.30$) was moderate.

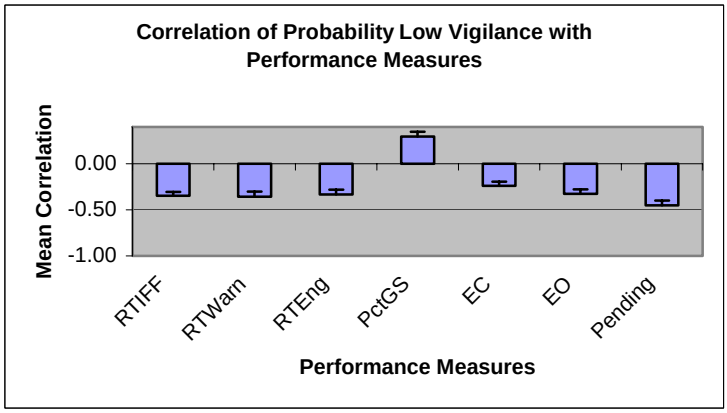
Figure 21b shows the mean correlation of the Probability Low Vigilance gauge values and performance measures. The Probability Low Vigilance gauge moderately correlated with the Response Time to IFF (RTIff), Response Time to Warn (RTWarn), Response Time to Engage (RTEng), Errors of Omission (EO), and Tasks Pending (Pending) performance measures. Correlations for the Percent Game Score (PctGS) and Errors of Commission (EC) measure were low.

Figure 21c shows the correlation of the Probability Low Vigilance gauge values and performance measures for each participant. As can be seen in the graph, Percent Game Score (PctGs) has the highest amount of consistency between participants for the gauge correlations.

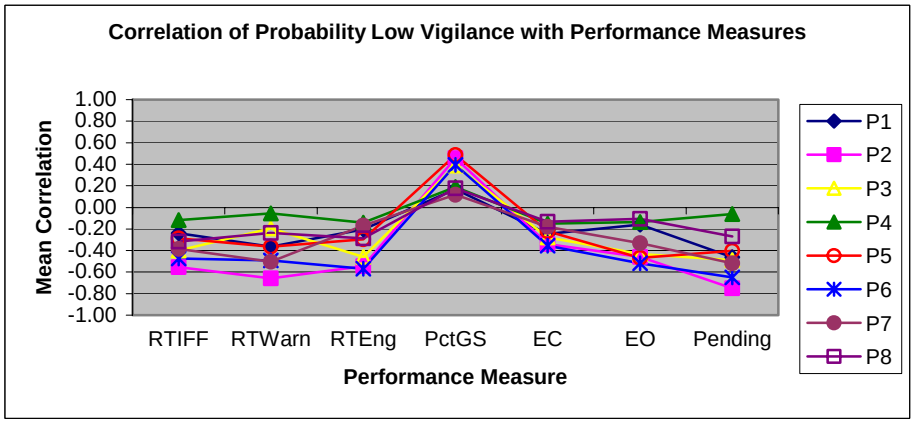
Summary: *The Probability Low Vigilance gauge values were moderately predictive and consistent with task load changes. In addition, the gauge demonstrated sensitivity to participant performance fluctuations due to task load changes, which was moderate for some measures and low for others. These results indicate that the Probability Low Vigilance gauge may have the ability to predict performance changes that occur due to task load changes.*



a. Correlation between gauge value and Number of Tracks per Wave.



b. Correlation between gauge value and performance measures.



c. Correlation of gauge value and performance measures for each participant.

Figure 21. Advanced Brain Monitoring Probability Low Vigilance.

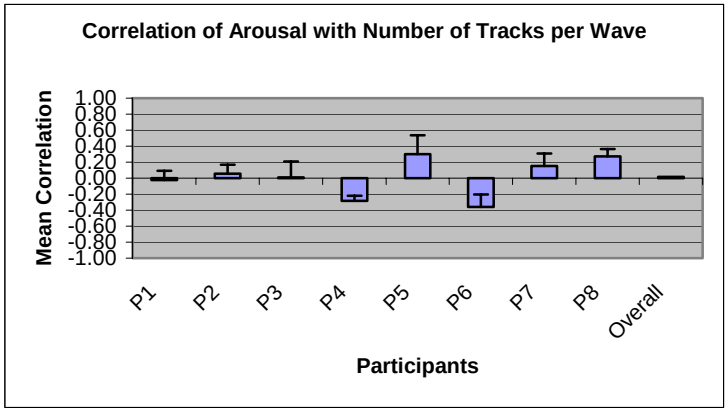
4.3.11 University of Hawaii – Arousal

Figure 22a shows the mean correlation for each participant varied from -0.36 to 0.30, with a mean of 0.02. The standard deviation of the mean correlations was 0.24. Thus, size of the mean correlation was low, but the consistency of the correlations was moderate ($0.15 < \sigma < 0.30$). Overall, the Arousal gauge values were not related to task load changes associated with Number of Tracks per Wave.

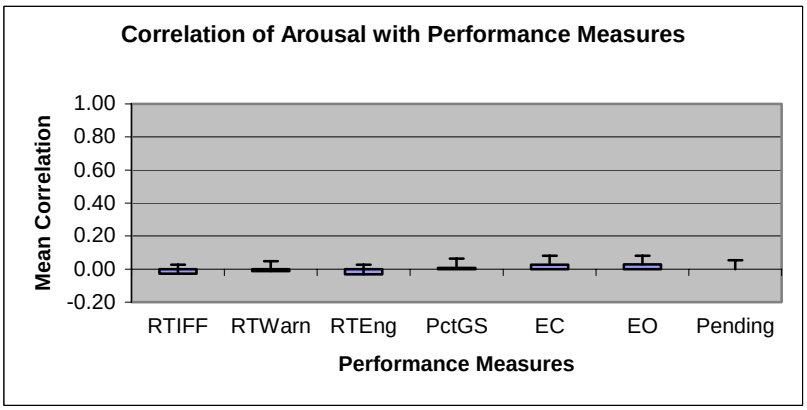
Figure 22b shows the mean correlation between the Arousal gauge readings and the performance measures across participants.

Figure 22c shows the correlation of the Arousal gauge values and performance measures for each participant.

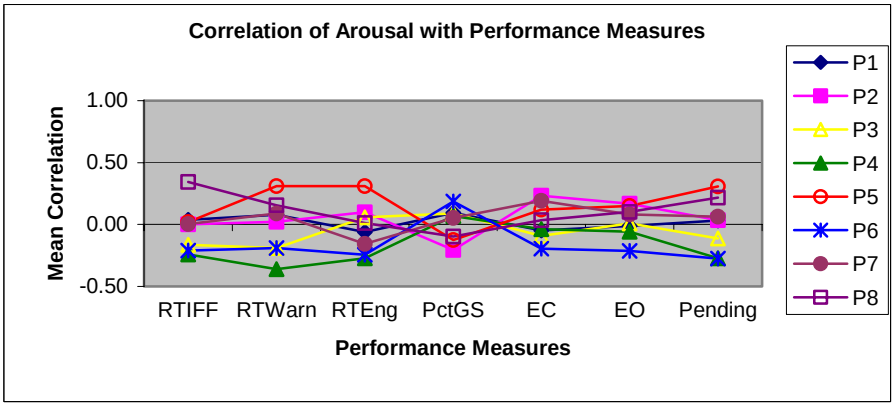
Summary: *The findings indicate the Arousal gauge was not sensitive to the task load changes associated with the Number of Tracks per Wave or the performance measures. The gauge values were highly variable between and within participants.*



a. Correlation between gauge value and Number of Tracks per Wave.



b. Correlation between gauge value and performance measures.



c. Correlation of gauge value and performance measures for each participant.

Figure 22. UHawaii Arousal.

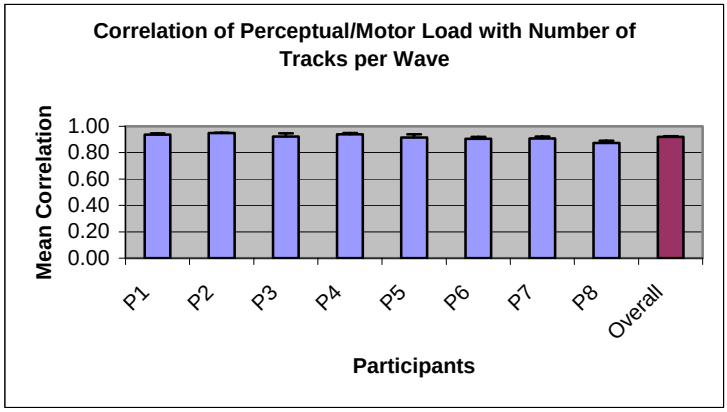
4.3.12 University of Hawaii – Perceptual/Motor Load

Figure 23a shows the mean correlation for each participant varied from 0.87 to 0.95, with a mean of 0.92. The standard deviation of the mean correlations was 0.02. Thus, size of the mean correlation was high, and the consistency of the correlations was high ($\sigma < 0.15$). Very high correlations were found between the Perceptual/Motor Load gauge and Number of Tracks per Wave. Overall, the Perceptual/Motor Load gauge values were related to task load changes associated with number of tracks per wave.

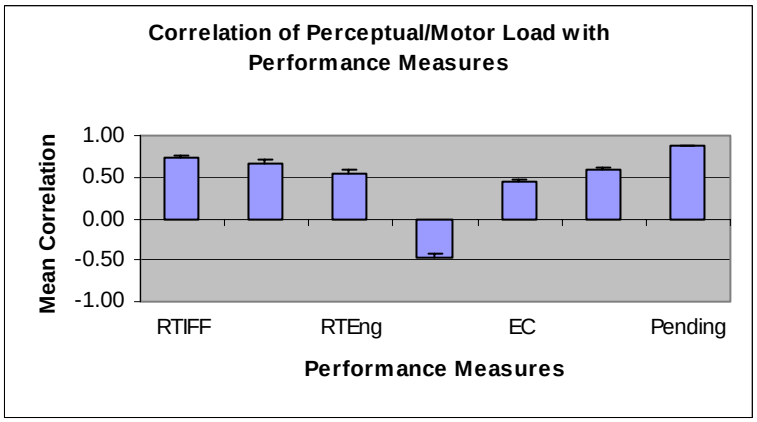
Figure 23b shows the mean correlation of the Perceptual/Motor Load gauge values and performance measures across participants. The Perceptual/Motor Load gauge was found to have a very high positive correlation, with the exception of an expected negative correlation for Percent Game Score (PctGS), with each of the performance measures.

Figure 23c shows the correlation of the Perceptual/Motor Load gauge and performance measures for each participant. Variability between participants is low and there is consistency within an individual for each of the performance measures.

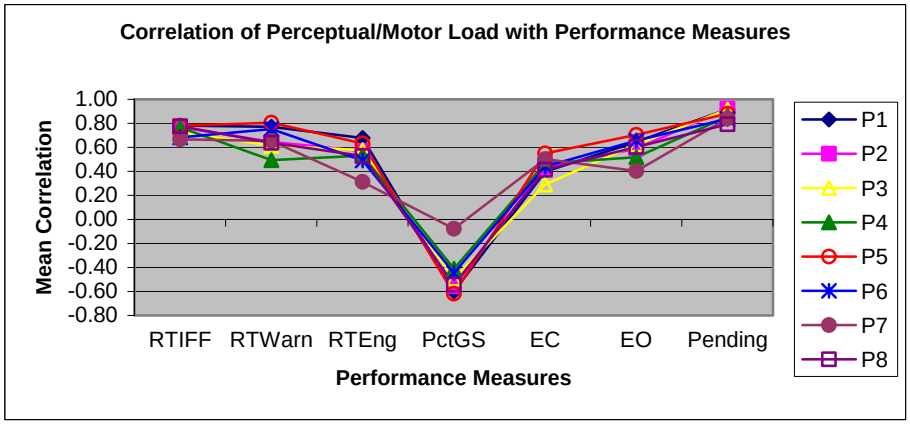
Summary: *The Perceptual/Motor Load gauge values demonstrated a high degree of association to the task load changes, as well as each of the performance measures with a high degree of consistency. Participant performance measures demonstrated consistency both within and between participants.*



a. Correlation between gauge value and Number of Tracks per Wave.



b. Correlation between gauge value and performance measures.



c. Correlation of gauge value and performance measures for each participant.

Figure 23. UHawaii Perceptual/Motor Load.

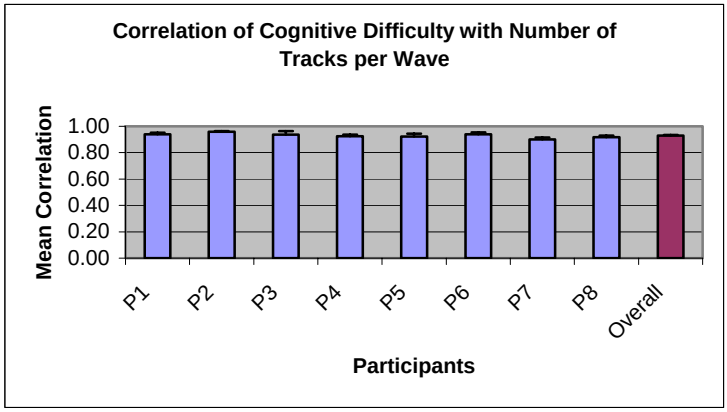
4.3.13 University of Hawaii – Cognitive Difficulty

Figure 24a shows the mean correlation for each participant varied from 0.90 to 0.96, with a mean of 0.93. The standard deviation of the mean correlations was 0.02. Thus, size of the mean correlation was high, and the consistency of the correlations was high ($\sigma < 0.15$). High positive correlations were found between the Cognitive Difficulty gauge and Number of Tracks per Wave. Overall, the Cognitive Difficulty gauge was related to task load changes associated with Number of Tracks per Wave.

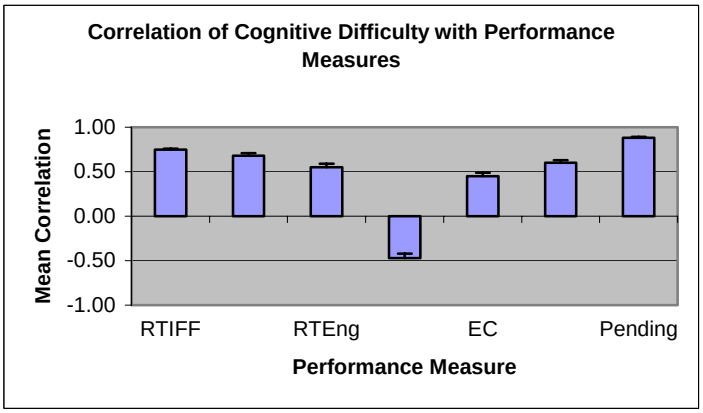
Figure 24b shows the mean correlation between the Cognitive Difficulty gauge readings and performance measures across participants. The Cognitive Difficulty gauge was found to have a high positive correlation with each of the performance measures. Therefore, the gauge values have a statistically significant relationship with the task load changes associated with these performance measures.

Figure 24c shows the correlation of the Cognitive Difficulty gauge readings and performance measures for each participant. Variability is very low between participants across the performance measures.

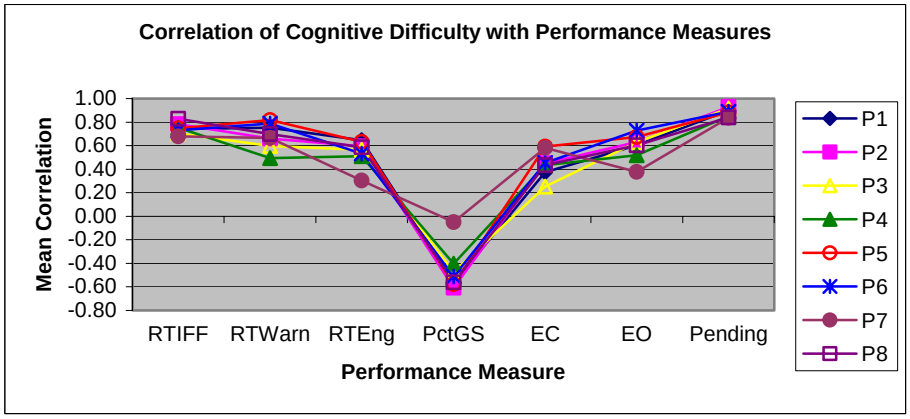
Summary: *Results indicate the Cognitive Difficulty gauge values are significantly sensitive to the task load changes and variation occurring in participant performance for each of the performance measures. The sensitivity of the gauge values is consistent both between participants and within an individual participant's performance across the measures.*



a. Correlation between gauge value and Number of Tracks per Wave.



b. Correlation between gauge value and performance measures.



c. Correlation of gauge value and performance measures for each participant.

Figure 24. UHawaii Cognitive Difficulty**.**

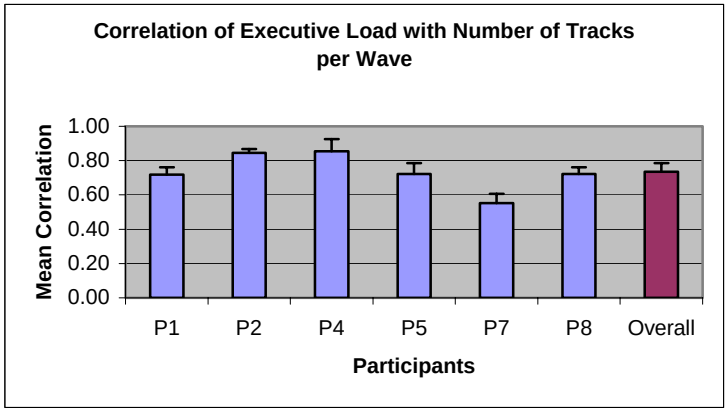
4.3.14 QinetiQ – Executive Load

Figure 25a shows the mean correlation for each participant varied from 0.55 to 0.85, with a mean of 0.74. The standard deviation of the mean correlations was 0.11. Thus, size of the mean correlation was high, and the consistency of the correlations was high ($\sigma < 0.15$). Overall, the Executive Load gauge values are related to task load changes associated with the Number of Tracks per Wave.

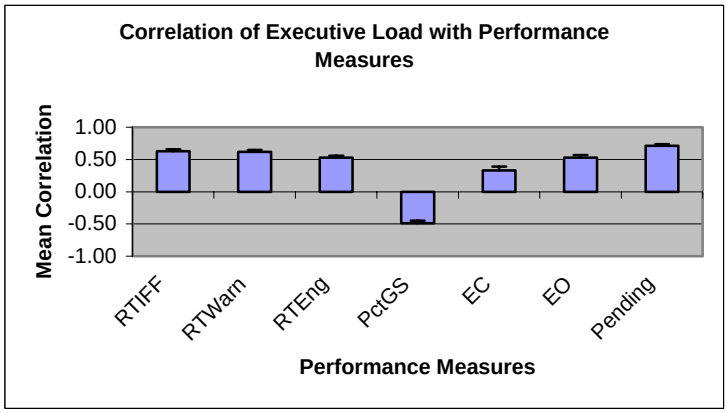
Figure 25b shows the mean correlation between the Executive Load gauge readings and performance measures. The Executive Load gauge significantly correlated with each performance measure.

Figure 25c demonstrates the correlation of the Executive Load gauge values and performance measures for each participant. The highest correlation is found at Pending (.83), which is also the measure with the lowest amount of variability across participants; standard deviation of the mean correlation is .10.

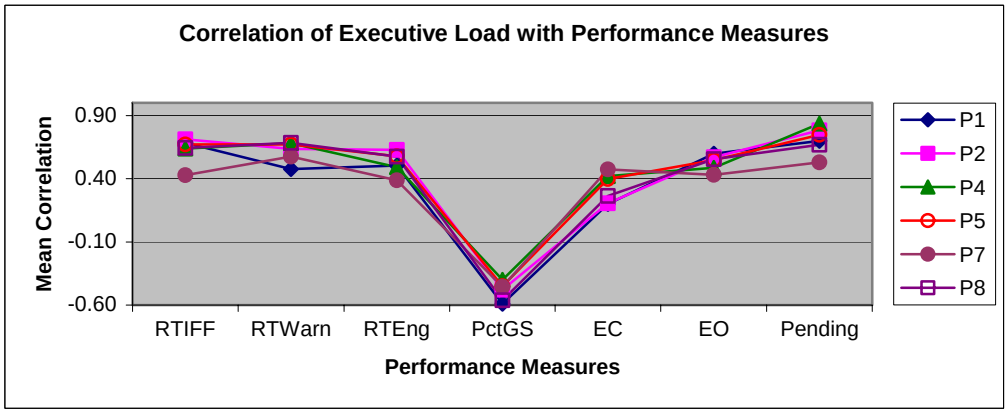
Summary: *The Executive Load gauge values were found to be highly related to task load changes associated with Number of Tracks per Wave. The relationship was high for the Response Time to Identify Friend for Foe, Response Time to Warn and Tasks Pending performance measures and moderate for the Response Time to Engage, Percent Game Score, Errors of Commission and Errors of Omission. Results are consistent both between participants and within an individual participant's performance. These findings indicate that the Executive Load gauge may have a predictive ability to detect changes in task performance.*



a. Correlation between gauge value and Number of Tracks per Wave.



b. Correlation between gauge value and performance measures.



c. Correlation of gauge value and performance measures for each participant.

Figure 25. QinetiQ Executive Load.

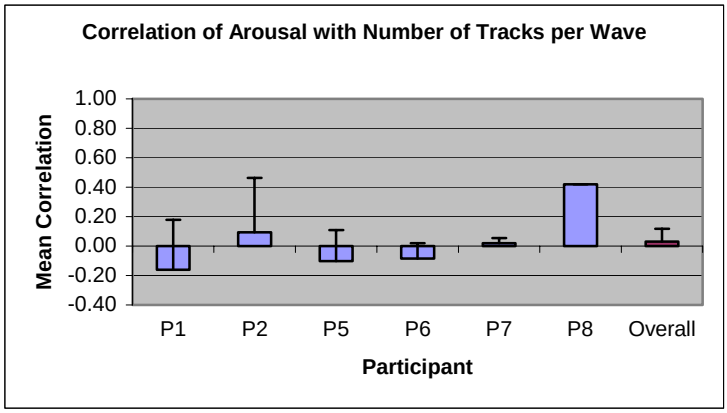
4.3.15 AnthroTronix – Arousal

Figure 26a shows the mean correlation for each participant varied from -0.16 to 0.42, with a mean of 0.03. The standard deviation of the mean correlations was 0.21. Thus, size of the mean correlation was low, but the consistency of the correlations was moderate ($0.15 < \sigma < 0.30$).

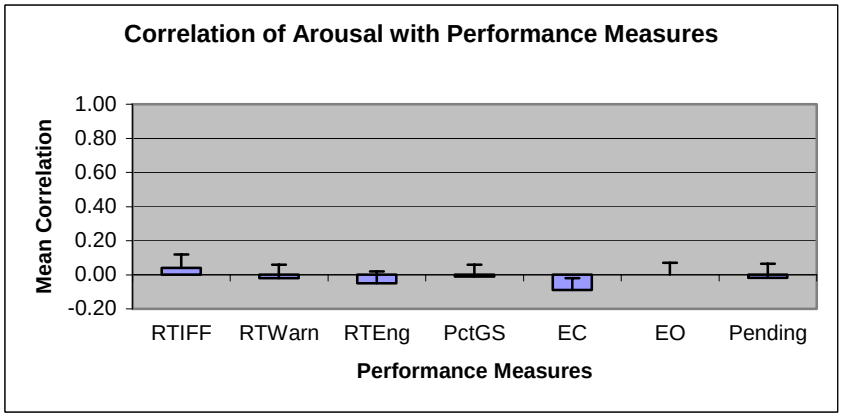
Figure 26b shows the mean correlation between the Arousal gauge readings and performance measures. The Arousal gauge produced low correlations with each of the performance measures.

Figure 26c shows the correlation of the Arousal gauge values and performance measures for each participant. It appears that correlations for participants are low and variability across participants is high.

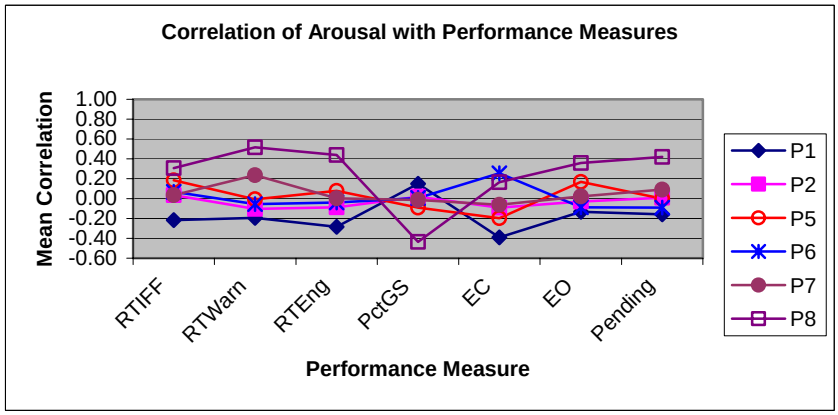
> ***Summary:*** *The Arousal gauge was not found to be sensitive to the Number of Tracks per Wave and performance measure task load changes. These findings were moderately consistent for each of these measures.*



a. Correlation between gauge value and Number of Tracks per Wave.



b. Correlation between gauge value and performance measures.



c. Correlation of gauge value and performance measures for each participant.

Figure 26. AnthroTronix Arousal.

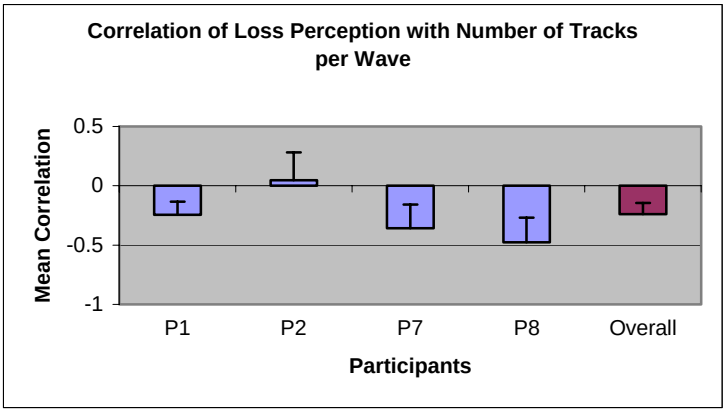
4.3.16 Sarnoff/Columbia – Loss Perception

Figure 27a shows the mean correlation for each participant varied from -0.48 to 0.05, with a mean of -0.26. The standard deviation of the mean correlations was 0.22. Thus, size of the mean correlation was low and the consistency of the correlations ($0.15 < \sigma < 0.30$) was moderate.

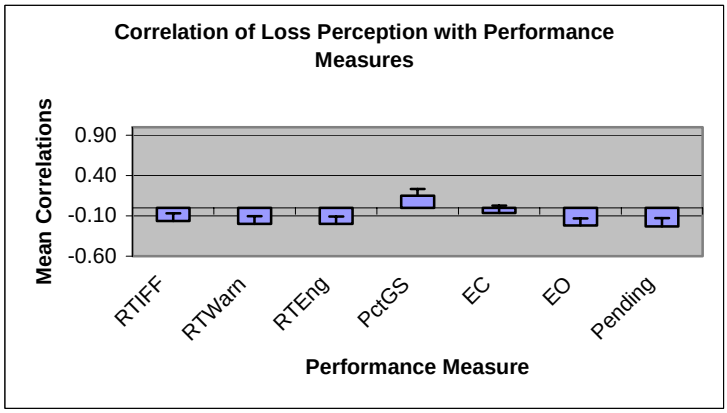
Figure 27b shows the mean correlation between the Loss Perception gauge values and performance measures across participants. The Loss Perception gauge correlated weakly with each of the performance measures.

Figure 27c shows the correlation of the Loss Perception gauge readings and performance measures for each participant. Correlations for the participants for each of the performance measures are low and no clear pattern exists between the participants.

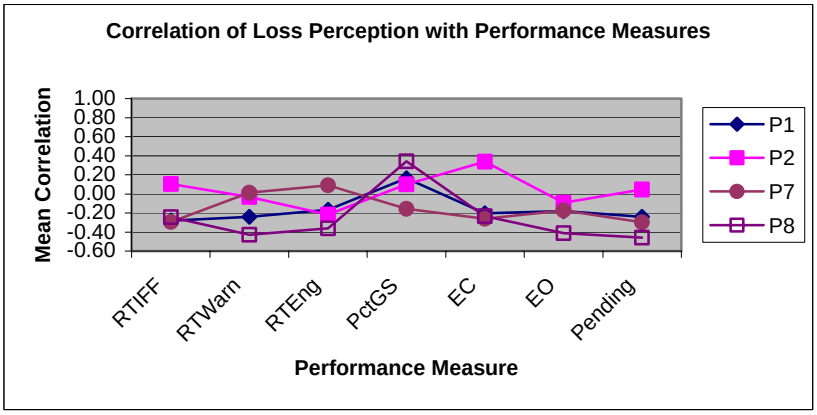
Summary: *The Loss Perception gauge demonstrated a low relationship to the task load changes associated with Number of Tracks per Wave with a moderate degree of consistency. Gauge values had a low sensitivity to changes in participant performance associated with task load. The gauge values were highly variably between and within participants.*



a. Correlation between gauge value and Number of Tracks per Wave.



b. Correlation between gauge value and performance measures.



c. Correlation of gauge value and performance measures for each participant.

Figure 27. Sarnoff/Columbia Loss Perception.

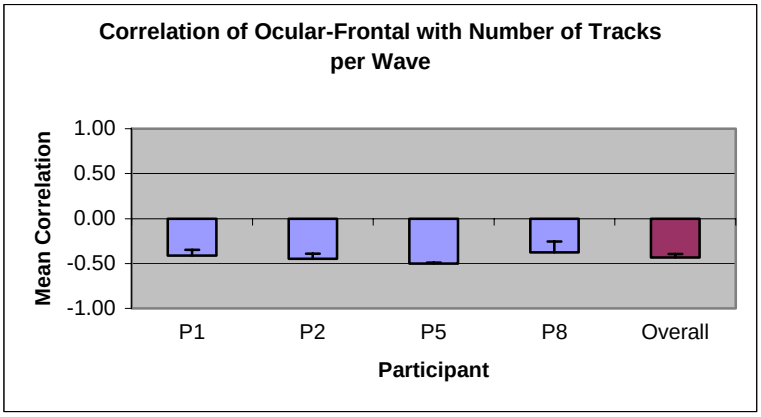
4.3.17 University of New Mexico – Ocular-Frontal Source

Figure 28a shows the mean correlation for each participant and varies from -0.38 to -0.50, with a mean of -0.43. The standard deviation of the mean correlations was 0.05. Thus, size of the mean correlation was moderate, and the consistency of the correlations was high ($\sigma < 0.15$). However, a substantial proportion of the effect may be due to the final wave of 24 tracks in which the gauge value dropped 10-fold. Excluding the 12th wave, the mean correlation drops to -0.32.

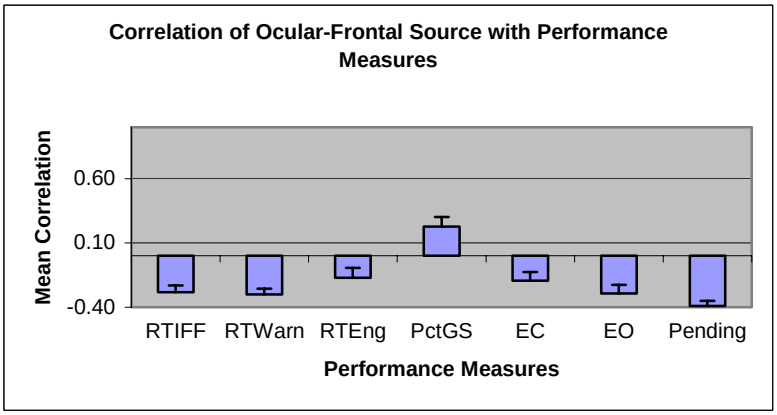
Figure 28b shows the mean correlation between the Ocular-Frontal Source gauge reading and performance measures across participants. The Ocular-Frontal Source gauge moderately correlated with the Tasks Pending (Pending) performance measures and demonstrated low correlations with the remaining six measures.

Figure 28c shows the correlation of the Ocular-Frontal Source gauge readings and performance measures for each participant. The correlations are low to moderate and the results are fairly consistent across participants.

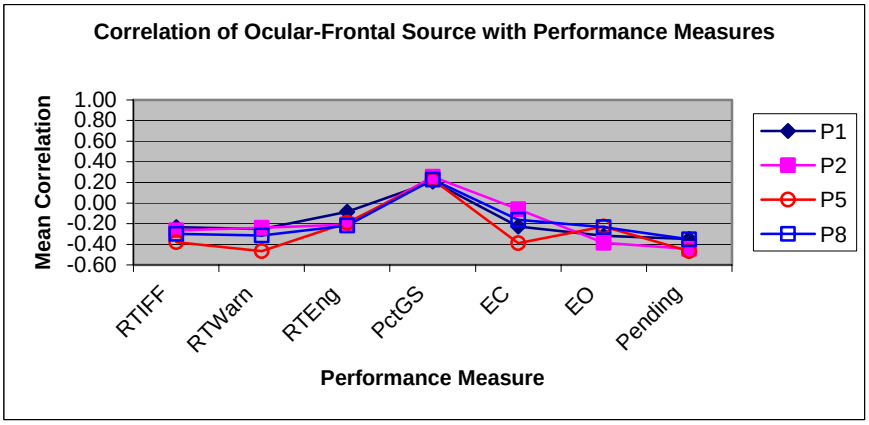
Summary: *The Ocular-Frontal Source gauge was found to be fairly predictive of task load changes related to the Number of Tracks per Wave with consistency for the participants overall. However, the gauge values were not sensitive to changes in participant performance for each of the specific measures. These results suggest that the Ocular-Frontal Source gauge may be able to predict changes occurring in performance in concert with the changes in task load of the WCT.*



a. Correlation between gauge value and Number of Tracks per Wave.



b. Correlation between gauge value and performance measures.



c. Correlation of gauge value and performance measures for each participant.

Figure 28. UNew Mexico Ocular-Frontal Source.

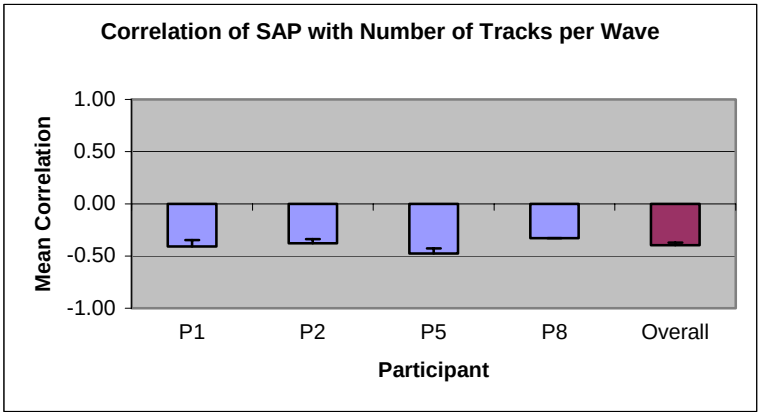
4.3.18 University of New Mexico – Synchronized Anterior-Posterior Source

Figure 29a shows that the mean correlation for each participant varied from -0.33 to -0.48, with a mean of -0.40. The standard deviation of the mean correlations was 0.06. Thus, size of the mean correlation was moderate, and the consistency of the correlations was high ($\sigma < 0.15$). However, a substantial proportion of the effect is due to the final wave of 24 tracks in which the gauge value drops 10-fold. Excluding the 12th wave, the mean correlation drops to 0.0.

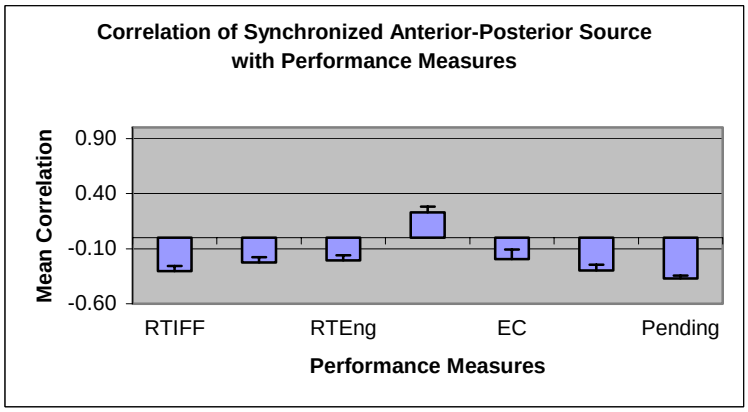
Figure 29b shows the mean correlation between the Synchronized Anterior-Posterior Source gauge readings and performance measures. The Synchronized Anterior-Posterior Source moderately correlated with the Response Time to Identify Friend or Foe (RTIff) and Tasks Pending (Pending) measures. Correlations for the Response Time to Warn (RTWarn), Response Time to Engage (RTEng), Percent Game Score (PctGS), Errors of Commission (EC) and Errors of Omission (EO) performance measures were low.

Figure 29c shows the correlation between the Synchronized Anterior-Posterior Source gauge readings and performance measures for each participant. Although the correlations are low to moderate for each participant, there appears to be a similar pattern across participants for each measure.

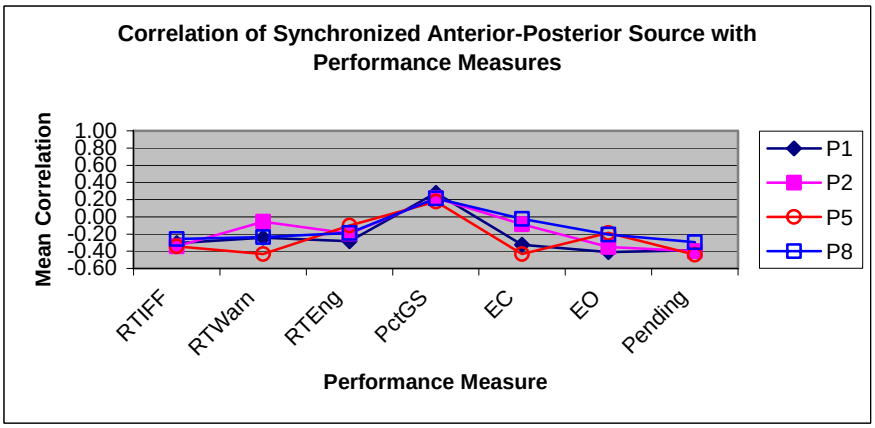
Summary: *The Synchronized Anterior-Posterior Source gauge values were found to be somewhat related to the task load changes associated with Number of Tracks per Wave. The gauge was not sensitive to changes in participant performance associated with task load changes in the WCT.*



a. Correlation between gauge value and Number of Tracks per Wave.



b. Correlation between gauge value and performance measures.



c. Correlation of gauge value and performance measures for each participant.

Figure 29. UNew Mexico Synchronized Anterior-Posterior Source.

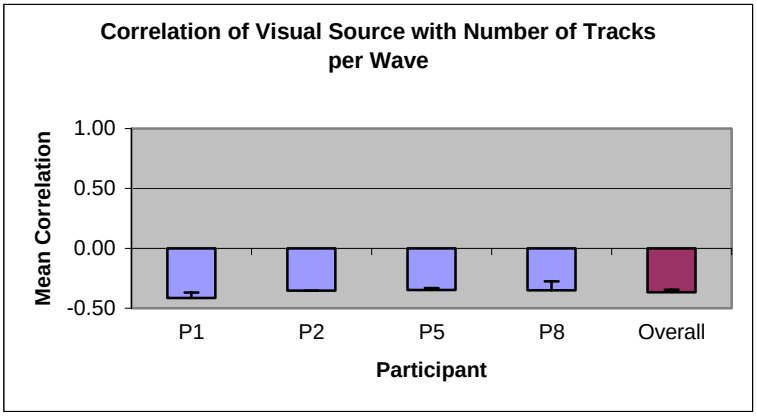
4.3.19 University of New Mexico – Visual Source

Figure 30a shows the mean correlation for each participant varied from -0.35 to -0.41, with a mean of -0.37. The standard deviation of the mean correlations was 0.03. Thus, size of the mean correlation was moderate, and the consistency of the correlations was high ($\sigma < 0.15$). However, a substantial proportion of the effect is due to the final wave of 24 tracks in which the gauge value drops 10-fold. Excluding the 12th wave, the mean correlation drops to 0.11.

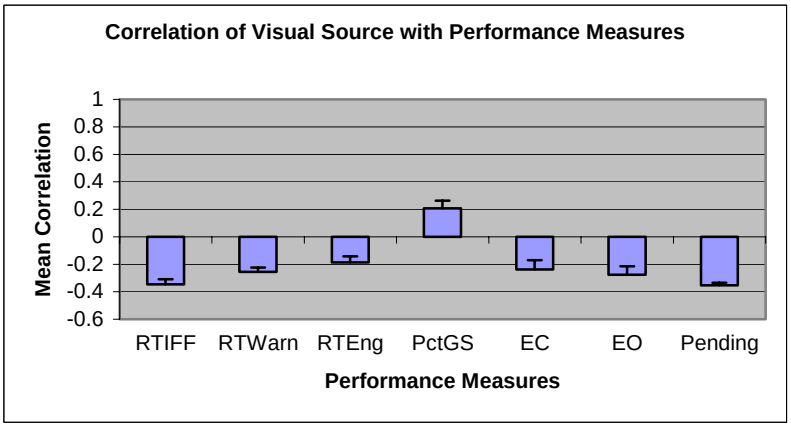
Figure 30b shows the mean correlation between the Visual Source gauge values and performance measures across participants. The Visual Source gauge was moderately correlated with the Response Time to IFF (RTIff) and Tasks Pending (Pending), as well as having low correlations with the Response Time to Warn (RTWarn), Response Time to Engage (RTEng), Percent Game Score (PctGS), Errors of Commission (EC) and Errors of Omission (EO) performance measures.

Figure 30c shows the correlation between the Visual Source gauge values and performance measures for each participant individually. There appears to be some consistency in correlation values across participants for speeded responses but not accuracy related processes.

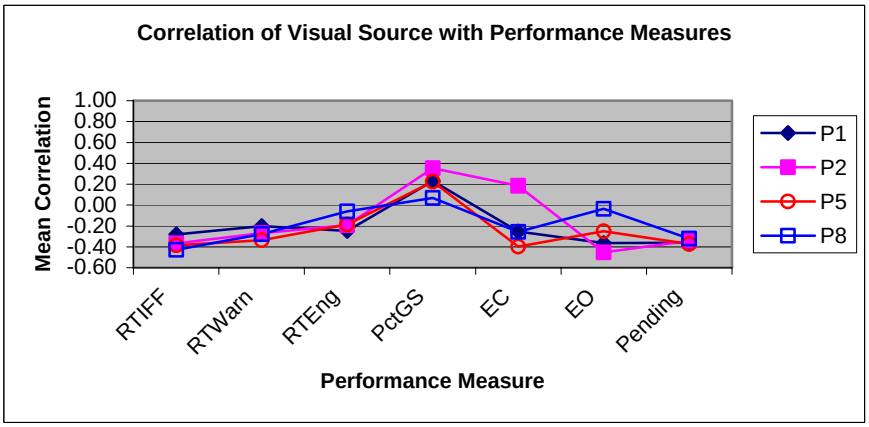
Summary: *The Visual Source gauge values demonstrated a moderate relationship to the Number of Tracks per Wave task load changes. The gauge was also somewhat sensitive to the Response Time to Identify Friend or Foe and Tasks Pending performance measures. In addition, low correlations were found between the gauge values and the Response Time to Warn, Response Time to Engage, Percent Game Score, Errors of Commission, and Errors of Omission performance measures.*



a. Correlation between gauge value and Number of Tracks per Wave.



b. Correlation between gauge value and performance measures.



c. Correlation of gauge value and performance measures for each participant.

Figure 30. UNew Mexico Visual Source.

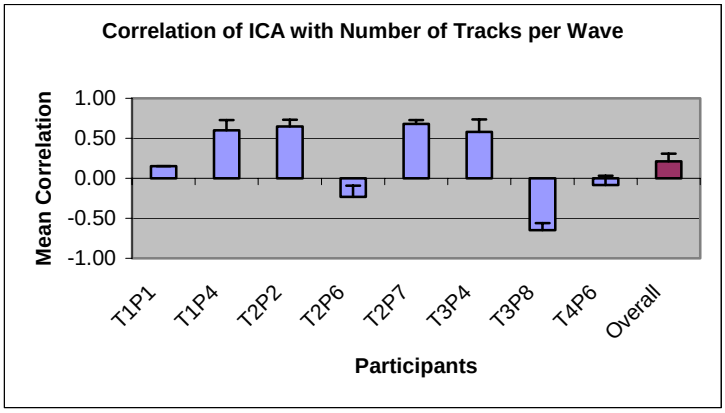
4.3.20 San Diego State University – Index of Cognitive Activity

Figure 31a shows the mean correlation for each participant varied from -0.65 to 0.68, with a mean of 0.21. The standard deviation of the mean correlations was 0.50. Thus, size of the mean correlation was low, and the consistency of the correlations was low ($\sigma > 0.30$).

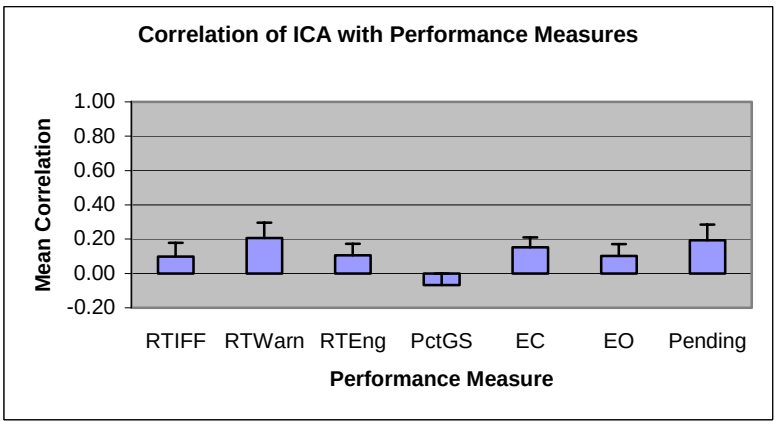
Figure 31b shows the mean correlation between the Index of Cognitive Activity gauge value and performance measures across participants. The Index of Cognitive Activity produced a low correlation with the Response Time to Warn (RTWarn) and Tasks Pending (Pending) performance measures.

Figure 31c shows the correlation between the Index of Cognitive Activity gauge readings and performance measures for each participant. Four of the participants (T1P4, T2P2, T2P7, and T3P4) have similar results across the performance measures.

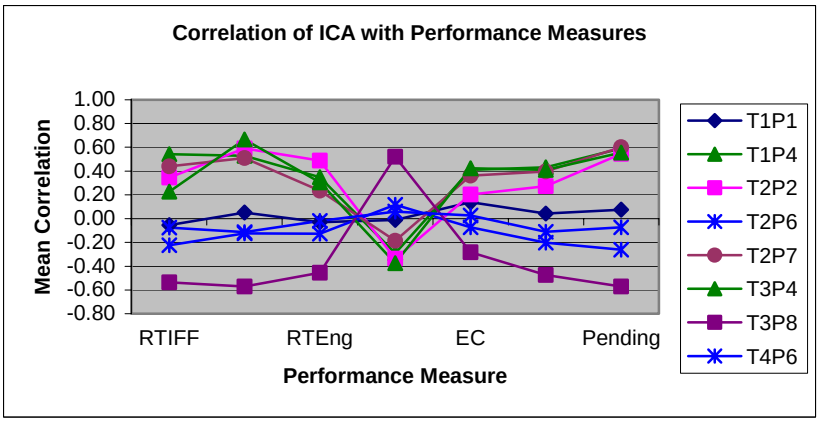
Summary: *The Index of Cognitive Activity gauge appeared to generate large individual differences. The gauge was highly correlated with specific aspects of performance for some participants but demonstrated low correlations with others.*



a. Correlation between gauge value and Number of Tracks per Wave.



b. Correlation between gauge value and performance measures.



c. Correlation of gauge value and performance measures for each participant.

Figure 31. SDSU Index of Cognitive Activity.

5. QUESTIONNAIRE RESULTS AND DISCUSSION

5.1 GAUGE DESCRIPTION QUESTIONAIRE

Each of the gauge developers completed a Gauge Description Questionnaire to provide a description of their gauge and a subjective evaluation of their gauge's performance during the TIE and potential Phase II transition issues.

As developers considered responses to the questionnaire, they were asked to keep in mind the objectives set forth by DARPA for the TIE:

- Demonstrate the Cognitive State Gauges developed under Phase I of the AugCog program.
- Showcase and document potential application of the cognitive state gauges to manipulate cognitive state during Phase II of the AugCog program.
- Show that a number of the cognitive state gauges that have been developed across the AugCog project can be interoperable, and to the degree possible, cross validate them with each other using the same participants at the same time.
- Provide a forum for prototype gauge development teams to meet and exchange data.
- Assess and document maturity, issues, and efficacy of Phase I gauges.

The questionnaire consisted of three parts, each emphasizing a different aspect of a gauge or its performance during the TIE. Part I addressed technical issues encountered during the TIE. These included team integration, such as new procedures for data analysis or sensor deployment learned during the TIE and aspects of the WCT related to gauge feasibility. Part II addressed issues specifically related to gauge descriptions, future applications, advantages/limitations, and key references related to each of the developers' gauges. Part III concentrated on a self-evaluation by developers of their gauges and specific issues related to the transition of their gauges in Phase II.

5.2 QUESTIONS AND RESPONSES (TEAM INTEGRATION)

Each of the gauge developers completed a questionnaire designed to collect information regarding team integration and detailed descriptions of individual gauges developed by the developers. Part I of the questionnaire highlighted issues related to team integration; these are summarized in Table 12. Each of the developers responded to a series of open-ended questions and rated their gauges ability to integrate with other sensors using a 5-point scale.

Overall, the developers rated the ability of their sensor to integrate with other sensors, as fairly high (range = 3.0 to 4.4). In addition, each developer identified the problems experienced during the TIE (see item 2) and rated how likely these problems could be resolved within the next 6 months. Only Electrical Geodesics reported having no problems (one developer failed to respond). The remainder of the developers indicated that the problems would be corrected with a fairly high degree of certainty. University of Hawaii's indication that too much time was required to connect electrodes to a participant and San Diego State University's issues of 10-Hz noise, however, may require longer than 6 months to resolve.

One problem seen consistently by Advanced Brain Monitoring, Qinetiq, Sarnoff, and University of Pittsburgh was interference with other headgear sensors. Participants usually reported head pain when multiple headgear were used. Other problems involved the time to attach sensors and the physical size of the sensors. When considering team sensor-gauge groupings, San Diego State University eye tracker interfered with the results of half of the developers. The problem appears to lie with the combination of the eye tracker headband and underlying EEG sensors; the weight of the

headband can drive the sensors into the skin and create discomfort. Better, more integrated hardware designs should be able to eliminate much of this problem. Additional problems included participant fatigue, largely due to the time required to install and calibrate the myriad instrumentation. Again, better integration of the various technologies can reasonably be expected to resolve this issue.

Table 12 also lists procedure changes identified by each of the developers that were learned due to the nature of the TIE. These procedure changes included improving the order and method of applying the sensors. Techniques learned by other developers were specific to their gauge such as adding new channels and adaptive filters for Drexel University and reducing to a smaller sensor suite for Qinetiq. When asked which sensor would be the most difficult to integrate, five out of the 12 developers had no difficulty in integrating their sensor with other sensors. Most of the difficulty for the remaining seven developers came from the lack of head space available for multiple sensors. The gauges for Clemson University and the University of Pittsburgh were designed to complement any gauge during the TIE. The fNIR sensor and any of the EEG sensors were also complementary to the majority of gauges.

Summary: *The lack of integration of sensor technologies, particularly those that were head-mounted proved a significant issue for the TIE in terms of technology integration, participant discomfort, and fatigue. Future attempts to employ these technologies would be well advised to integrate them into a single device.*

Track Difficulty and Number of Tracks per Wave are common aspects of WCT identified by most of the gauge developers as best demonstrating the feasibility of their gauge. The question concerning limitations of the WCT generated a variety of responses including number of eye movements, motor activity, and a lack of unpredictability.

5.3 QUESTIONS AND RESPONSES (GAUGE DESCRIPTIONS)

Table 13 summarizes the gauge descriptions and is followed by full text descriptions of future applications, advantages/limitations, and key references of each gauge developer. Annotations of "NA" were used to indicate that a question was not applicable, and "NR" to indicate that the developer did not provide a response.

Table 12. Questions and responses from the CWA developers regarding team integration during the TIE.

Gauge	fNIR	EEG-Continuous		EEG-ERP			Arousal			Physiological		
	Drexel University	**Advanced Brain Monitoring, Inc.**	**QinetiQ**	**Electrical Geodesics, Inc.**	**Sarnoff**	**University of New Mexico**	**Clemson University**	**University of Hawaii**	**AnthroTronix, Inc.**	**San Diego State University**	**University of Pittsburgh & Naval Research Lab**	**University of Hawaii**
	FNIR – Left and Right Frontal Lobes	**Percent High Vigilance, Probability Low Vigilance**	**Executive Load**	**Motor Effort & Auditory Effort**	**Loss Perception**	**Theta Power (Anterior-Posterior, Ocular & Visual Source)**	**Arousal Meter**	**Arousal & Stress**	**GSR Arousal**	**ICA**	**Head/Monitor Coupling, Head Bracing, Back Bracing**	**Perceptual & Motor Load**
How well did your sensor integrate with other sensors? Low 1 --- 5 High	4.0	4.0	4.0	4.25	3.0	NA	4.25	4.4	4.1	3.5	3.8	4.4
What problems did you have?	Only one probe; Size-does not fit all	Problems combining multiple headgear	Eye tracker headband with EEG	None	Eye tracker with EEG too painful	NR	No time stamp	Too long to attach electrodes	HR, Respiration rate did not correlate with task	Unable to collect data with Team 1 due to 10hz interference	Head sensors caused interference	Too long to attach electrodes
How likely will problems be resolved in next 6 months? Low 1--- 5 High	4.0	4.1	NR	NA	4.5	NA	4.0	1.3	3.5	2.0	4.1	1.3
Interference during TIE?	Eye tracker sensor interference	Subject fatigue and discomfort	NR	None	60hz interference from unknown source	Eye tracker /EEG interference	None	None	Physical discomfort, EEG cap chest band disconnected EKG	Under investigation	Eye tracker provided some interference	SDSU eye tracker
New procedures learned during the TIE?	Added new channel & adaptive filter	Learned a quick procedure to attach devices	Reducing to a smaller sensor suite	None	Need to adjust the reading of event markers	Heart rate & GSR	Time sampling requirement	None	Least comfortable sensors should be put on last	Order of applying sensors	Use of real-time movement to trigger other gauges	None
Which other sensors complement your gauge?	EEG	FNIR, Pupillometry, Cardio sensors	Modified fNIR/EEG sensors	EKG, Eye tracking	Any mechanical sensor, fNIR	Heart rate & GSR	AM is designed to complement other gauges	EEG	Unable to determine	FNIR and GSR	Posture gauge complements all others	EEG
Which aspects of WCT best demonstrate feasibility of gauge?	Task difficulty & Time to IFF	Workload / Number of tracks	Dynamic changes in workload	Varying levels of workload	Auditory feedback	Request for information	None	Wave difficulty	Onset of new waves of aircraft	Wave size	Events calling for immediate, intensive action	Wave difficulty

Gauge	fNIR	EEG-Continuous		EEG-ERP			Arousal			Physiological		
	Drexel University	**Advanced Brain Monitoring, Inc.**	**QinetiQ**	**Electrical Geodesics, Inc.**	**Sarnoff**	**University of New Mexico**	**Clemson University**	**University of Hawaii**	**AnthroTronix, Inc.**	**San Diego State University**	**University of Pittsburgh & Naval Research Lab**	**University of Hawaii**
	FNIR – Left and Right Frontal Lobes	**Percent High Vigilance, Probability Low Vigilance**	**Executive Load**	**Motor Effort & Auditory Effort**	**Loss Perception**	**Theta Power (Anterior-Posterior, Ocular & Visual Source)**	**Arousal Meter**	**Arousal & Stress**	**GSR Arousal**	**ICA**	**Head/Monitor Coupling, Head Bracing, Back Bracing**	**Perceptual & Motor Load**
Which aspects of WCT limited your gauge?	No pause between waves, Excessive hand movement	Eye movements and motor activity	Eye movements and motor activity	A subjective level of effort would be useful	Infrequent occurrence of auditory feedback	Need an objective measure of working memory	Used expert users	None	Lack of penalties and realism	None	Lacks in unpredictability	None

Table 13. Questions and responses from the CWA developers regarding detailed descriptions of their gauge.

Gauge	fNIR	EEG-Continuous		EEG-ERP			Arousal			Physiological			
	Drexel University	**Advanced Brain Monitoring, Inc.**	**QinetiQ**	**Electrical Geodesics, Inc.**	**Sarnoff**	**University of New Mexico**	**Clemson University**	**University of Hawaii**	**AnthroTronix, Inc.**	**San Diego State University**	**University of Pittsburgh & Naval Research Lab**	**University of Hawaii**	
	fNIR	**Percent High Vigilance, Probability Low Vigilance**	**Executive Load**	**Motor Effort & Auditory Effort**	**Loss Perception**	**Theta Power**	**Arousal Meter**	**Arousal. Stress and Cognitive Difficulty**	**GSR Arousal**	**ICA**	**Head/Monitor Coupling, Head Bracing, Back Bracing, Back Contact**	**Perceptual & Motor Load**	**Cognitive Difficulty**
What is the theory behind your gauge?	See Drexel University Appendix	See ABM Appendix	See QinetiQ Appen- dix	See Electrical Geo- desics Appendix	See ABM Appendix	See University of New Mexico Appendix	See Clemson University Appendix	See University of Hawaii Appendix	See Anthro-Tronix Appendix	See SDSU Appendix	See University of Pittsburgh & Naval Research Lab Appendix	See University of Hawaii Appendix	See University of Hawaii Appendix
What sensors are used to collect data for your gauge?	LEDs, photo-diodes	Wireless EEG sensor headset	EEG & EOG electrodes	Dense-array EEG electrode arrays	63 channel EEG, EOG channels	EEG	Connected to unit via electrode leads	GSR & infrared oximeter	GSR	Two high-speed cameras	Chair with 16X16 pressure sensor arrays	Pressure Mouse	Pressure Mouse
Are sensors connected to participant?	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No
How many and where are the sensors placed?	One probe consisting of 4 LEDs & 10 photodiodes	EEG at Fz, Cz, POz, and mastroids, EOG around eye	14 scalp electrodes, 4 EOG	128 EEG channels put on scalp	Scalp, face, behind ear	NR	2 active recording leads below collar bone, 1 noise reduction lead	3 on toes (2 for GSR, 1 for oximeter)	GSR on toes	Headband on which camera was mounted	One flock of bird sensor on head and torso	Sensors were not connected to participant	Sensors were not connected to participant
How long does it take to connect your sensors?	< 5 min	5-10 min	> 30 min	5-10 min	10-20 min	NR	< 5 min	5-10 min	10-20 min	< 5 min	< 5 min	NR	NR
Do the sensors require calibration? How long does it take?	5-10 min	Yes, 10-20 min	< 5 min	< 5 min	< 5 min	NR	Yes, time depends on task	< 5 min	No	< 5 min	No	NR	NR
Recalibration required? How long does it take?	5-10 min	No	< 5 min	< 5 min	< 5 min	NR	No	No	No	< 5 min	No	No	No

Gauge	fNIR	EEG-Continuous		EEG-ERP			Arousal			Physiological			
	Drexel University	**Advanced Brain Monitoring, Inc.**	**QinetiQ**	**Electrical Geodesics, Inc.**	**Sarnoff**	**University of New Mexico**	**Clemson University**	**University of Hawaii**	**AnthroTronix, Inc.**	**San Diego State University**	**University of Pittsburgh & Naval Research Lab**	**University of Hawaii**	
	fNIR	**Percent High Vigilance, Probability Low Vigilance**	**Executive Load**	**Motor Effort & Auditory Effort**	**Loss Perception**	**Theta Power**	**Arousal Meter**	**Arousal. Stress and Cognitive Difficulty**	**GSR Arousal**	**ICA**	**Head/Monitor Coupling, Head Bracing, Back Bracing, Back Contact**	**Perceptual & Motor Load**	**Cognitive Difficulty**
How long can the sensors be worn comfortably?	> 120 min	> 8 hours	> 120 min	Indefinitely	60-120 min	NR	> 36 hours	NR	1-2 hours	60-120 min	> 120 min	> 120 min	> 120 min
Additional recalibration? How often and how long?	No	No	No	No	No	NR	No	No	No	No	No	No	No
What constraints are placed on the operator wearing your sensor?	Limit head movement	Restrict muscle activity	Participant must remain seated	Limit movement	Limit range of motion	NR	Cannot get device wet	Placement is possible for mobile applications	Minimize toe, arm, & face movement	Limit movement due to cable length	Limit range of motion	Computer mouse must be used	Computer mouse must be used
What measure is being used from the data collected at each sensor?	3 channel data	EEG power spectral analysis	Time & frequency domain, spectral analysis	Theta averaged .5 sec before & after task	Adaptive linear spatial filtering	Raw continuous EEG sensor data	Inter-beat intervals	Heart rate & GSR are multiplied, compared to calibrated values	Averaging over each second, scaling by a factor of 1000	Index value for each eye per second	Seat cushion detects pressure as scaled voltage	Cannot give any details due to pending patent	Cannot give any details due to pending patent
Are measures weighted equally?	No	NR	NR	Yes	No	NA	No	Yes	Only using GSR in gauge calculations	NA	Yes	Single measure	Single measure
What is the sensitivity of your gauge as a measure of state change?	1.5-2.0 cm depth in brain tissue	Quantify changes in vigilance & workload	Changes in executive load	30% of variance can be explained	75-90% correct discrimination	Second by second	Second by second	Tracks small changes in arousal	Detects changes in state at .01 micro- mohs	High, medium, & low cognitive effort	Back bracing sensitive to changes in workload	Tracks load	Tracks load
What is the practical limit for your gauge's sensitivity?	Fixed montage of our detectors	No real limit	NR	To be determined	100% for a defined time window	NR	30 seconds to 1 minutes based on physiology	Unknown	.01 micromohs with a resolution of 32 Hz	4-6 levels	Determined on a task-by-task basis	Unknown	Unknown
What is the current temporal resolution?	One-wave point (75 sec)	Seconds	1.6 sec	Millisecond to minutes	< 1 sec	Requires working memory measure	Minutes	2-4 sec	Maximum of 32 Hz	1 sec	4 times per sec	Sub-second	< 1 sec

Gauge	fNIR	EEG-Continuous		EEG-ERP			Arousal			Physiological			
	Drexel University	**Advanced Brain Monitoring, Inc.**	**QinetiQ**	**Electrical Geodesics, Inc.**	**Sarnoff**	**University of New Mexico**	**Clemson University**	**University of Hawaii**	**AnthroTronix, Inc.**	**San Diego State University**	**University of Pittsburgh & Naval Research Lab**	**University of Hawaii**	
	fNIR	**Percent High Vigilance, Probability Low Vigilance**	**Executive Load**	**Motor Effort & Auditory Effort**	**Loss Perception**	**Theta Power**	**Arousal Meter**	**Arousal. Stress and Cognitive Difficulty**	**GSR Arousal**	**ICA**	**Head/Monitor Coupling, Head Bracing, Back Bracing, Back Contact**	**Perceptual & Motor Load**	**Cognitive Difficulty**
What is the practical limit to the resolution?	Slow brain hemo-dynamic response	0.5 seconds	Number of points in FFT	Millisecond	Time: 100-200 ms, Space: 2 cm	100 ms	30 seconds to 1 minute based on physiology	1-2 sec	Maximum of 32 Hz	1 sec	No known limiting factors	Time needed for a mouse click	Time needed for a mouse click
What aspect of cognitive state does your gauge measure?	Attention & working memory	Vigilance (combination of alertness & attention)	Executive load	Working & verbal memory, motor control effort	Executive function	Working memory	Arousal/ fatigue	Arousal & stress	Changes in ANS activity	Cognitive effort	Posture-mediated state-detection	Perceptual & motor load	Cognitive difficulty task
What is the limit of your gauge's ability to predict the cognitive state being measured?	Can exceed realistic limit with a full head system	90% accuracy when classifying each second	NR	30% of subjective effort ratings	Precise understanding of the task the user is executing	Must detect complete overload or near complete overload	Will predict arousal from low to active alert	Unknown	Cannot distinguish between ANS caused by arousal or other factors	Best suited for complex cognitive tasks	Unknown	Effective only for computer tasks where mouse is being used	Unknown
To what degree is your gauge predictive of the general state of arousal?	Provides descriptive information in response to task load changes	Arousal level is a key contributor to the B-Alert Indices	NR	Not addressed	Not related to arousal	Not related to physiological arousal	Excellently	Reflects arousal well	Comparing GSR levels determines state of arousal	Arousal is not measured	High level of predictive value	Good correspondence	Correlated well
What is the limit of your gauge's ability to predict general arousal?	Spatial limitation & slow hemo-dynamic response	Sensitive to arousal levels in sleep deprivation studies	NR	Close relation between EEG & arousal	Not related to arousal	Not related to physiological arousal	Will predict arousal from low to active alert	Unknown	Requires a quality baseline measurement	Arousal is not measured	High level of predictive value	Unknown	Unknown

Gauge	fNIR	EEG-Continuous		EEG-ERP			Arousal			Physiological			
	Drexel University	**Advanced Brain Monitoring, Inc.**	**QinetiQ**	**Electrical Geodesics, Inc.**	**Sarnoff**	**University of New Mexico**	**Clemson University**	**University of Hawaii**	**AnthroTronix, Inc.**	**San Diego State University**	**University of Pittsburgh & Naval Research Lab**	**University of Hawaii**	
	fNIR	**Percent High Vigilance, Probability Low Vigilance**	**Executive Load**	**Motor Effort & Auditory Effort**	**Loss Perception**	**Theta Power**	**Arousal Meter**	**Arousal. Stress and Cognitive Difficulty**	**GSR Arousal**	**ICA**	**Head/Monitor Coupling, Head Bracing, Back Bracing, Back Contact**	**Perceptual & Motor Load**	**Cognitive Difficulty**
How quickly is the output of your gauge available?	Continually displayed in real-time	Continually displayed in real-time	Continually displayed in real-time	NR	Continually displayed in real-time	NR	Continually displayed in real-time, updated 4 times per second	Continually displayed in real-time	Contin- ually displayed in real-time	Continually displayed in real-time	Continually displayed in real-time	Continually displayed in real-time	Continually displayed in real-time
Is post-processing required to calculate gauge measurement?	Yes, averaging across waves & channels for better spatial information	Yes, only to tally data from different conditions	NR	Filter, detect, & remove artifacts, average data around event	None	SOBI ICA applied to raw data to obtain a spatial filter, domain analysis	Yes, reduces data from second by second to wave by wave	None	Yes, averaging data over each second, scaling by a factor of 1000	No, displays in real time	None	None	None
What factors may affect the performance of your gauge?	Sweat interference with probe attachment	Fatigue, level of expertise on task	NR	None known	60-Hz noise	Highly practiced participants	Highly practiced participants	None	Discomfort from multiple sensors	Fatigue, discomfort	Predictability, replication of task conditions	None	None

5.3.1 Advanced Brain Monitoring, Inc.

Gauge Name: Percent High Vigilance, Probability Low Vigilance

Team Integration During the TIE

In sessions longer than 2 hours, comfort was an issue with multiple headgear (fNIR, EEG, and eye tracking). The primary issue was participant fatigue and discomfort resulting from the layering of three systems on the head. The combination of set-up and run time often had participants complaining of comfort issues. Once the sensor suite was tested during the pre-TIE and finalized for the TIE, Team 2 organized an approach to facilitate multiple-sensor application in a timely manner. Because there was no headspace available for other EEG sensors, it would be most difficult to integrate with other EEG gauges. The fNIR, pupillometry, and any cardiovascular sensors would provide useful data for comparison. Workload levels associated with number of tracks in a wave in the WCT best demonstrated the feasibility of the B-Alert.

The amount of eye movement and muscle artifact increased with increasing number of tracks and contaminated some of the EEG data. Although data loss was less than 5%, the loss was often biased to the highest workload periods, and there were not a sufficient number of epochs per wave to compensate. This was problematic in providing sufficient data to validate the efficacy of the B-Alert indices as workload correlates. This can lead to gauges that are simply measuring sensory processing and/or motor activity, not cognitive workload. Again, these are only concerns in the initial stages of validating a gauge. In the early stages, it was preferable to work with artifact free data and to isolate, where possible, sensory and motor processes from cognitive processes.

Future Application

Additional miniaturization of the B-Alert indices will be expected to change in the future. Interference with the RF must be maintained to optimize gauge performance. Activities that cause excessive movements (jerking head movements, gum chewing) will disturb the quality of the signal.

Advantages/Limitations

Foot Soldier in the Field – Sensor headset could be built directly into field helmets of the soldiers.

Noisy Environments – B-Alert is unaffected by noise.

In a Moving Vehicle – Moving vehicles are fine on smooth terrain; excessive rough movements may cause artifacts.

Multitasking Environment with Multiple Displays and Controls – B-Alert can be used in this situation.

Command and Control Display Operator – B-Alert can be used in this situation.

Key References

See appendix 3b.

5.3.2 AnthroTronix, Inc.

Gauge Name: GSR Arousal

Team Integration During the TIE

Like many of the other developers, it was realized that our GSR gauge was not correlating to event data (wave size, Rtiff, errors, etc.) when averaging over the entire session.

GSR changed significantly over the course of the sessions. Therefore, the GSR was examined more closely to see how GSR was affected by specific task events, and a correlation was found a between event data and the first derivative of the GSR.

Physical discomfort to the participant, caused by other team members' sensors may have interfered with the GSR gauge. It was learned from the TIE that the most comfortable/unobtrusive sensors must be put on the participant first in order to maximize participant comfort. Sensors that are uncomfortable and/or obtrusive must be hooked up to the participant last in order to minimize the length of time that the participant is required to wear them. Due to the fact that only average wave data from most of the other developers was viewed, it is difficult to say which of the other gauges will complement the GSR gauge. Specific events that consistently induce stress, such as the onset of a new wave during the WCT, best demonstrate the feasibility of the GSR gauge. The lack of penalties when the user errors, as well as the lack of a relative feel to the task, led the participants to become less aroused when completing the task, limiting the key capabilities of the GSR gauge.

Future Application

The current footprint of the data acquisition system is no larger than a single laptop computer. It is expected that in the future the footprint will change to an untethered handheld computer. Ambient temperature must be maintained to optimize the performance of the GSR gauge. When applying the GSR gauge to real-world environments, wearability, obtrusiveness, placement and securing of sensors are major challenges that are envisioned.

Advantages/Limitations

Foot Soldier in the Field – GSR data would have to be collected from a site other than the toes to not interfere or cause discomfort when walking long distances as a foot soldier in the field.

Noisy Environments – Not affected by noisy environments.

In a Moving Vehicle – Due to portability, collecting data in a moving vehicle would be limited.

Multitasking Environment with Multiple Displays and Controls – Not affected unless tethered.

Command and Control Display Operator – Not affected unless tethered.

Key References

See appendix 3b.

5.3.3 Clemson University

Gauge Name: Arousal Meter

Team Integration During the TIE

Currently, the arousal meter (AM) hardware and software is not set up to receive or send a time stamp. Therefore, it is impossible to determine specific events related to increases or decreases in arousal. Currently, it only has an internal time stamp for use during the TIE. The AM is designed to measure autonomic arousal. Hence, it was complementary to any gauge that measures workload, working memory allocation, etc. Unfortunately, with highly trained individuals performing a routine task, it was expected and shown that autonomic arousal did not change much. Using only expert participants limited arousal variability.

Future Application

For Phase II, the data acquisition system will eliminate the laptop requirement and the device will wirelessly transmit the data to a base system. When applying AM to real-world environments, having a systems interface director that can handle and use the massive amount of data that the gauge is generating will be a major challenge. Wearing electrodes may limit the utility of the gauge. A material/shirt that acts as the sensor needs to be found.

Advantages/Limitations

Foot Soldier in the Field – The AM is designed to collect data in any environment. That has been one of the target specifications and continues to be; thathis specification should be met in Phase II.

Noisy Environments – The AM is designed to collect data to be used in any environment; that has been one of the target specifications and continues to be. This specification should be met in Phase II.

In a Moving Vehicle – The AM is designed to collect data in any environment; that has been one of the target specifications and continues to be. This specification should be met in Phase II.

Multitasking Environment with Multiple Displays and Controls – The AM is designed to collect data in any environment; that has been one of the target specifications and continues to be. This specification should be met in Phase II.

Command and Control Display Operator – The AM is designed to collect data in any environment; that has been one of the target specifications and continues to be. This specification should be met in Phase II.

Key References

Grossman, P. (1992). Respiratory and cardiac rhythmis as windows to central and autonomic biobehavioral regulation: Selection of window frames, keeping the panes clean and viewing the neural topography. *Biological Psychology*, 34, 131-161.

Katona, P.G. and Jih, F. (1975). Respiratory sinus arrhythmia: noninvasive measure of parasympathetic cardiac control. *Journal of Applied Psychology*, 39, 801-805.

Mezzacappa, E., Kindlon D., and 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.

Muth, E.R., Koch, K.L., and Stern, R.M. (2000). The significance of autonomic nervous system activity in functional dyspepsia. *Digestive Diseases and Sciences*, 45. 854-863.

5.3.4 Drexel University

Gauge Name: Functional Near Infrared – Brain Imaging (fNIR) – Left and Right Frontal Lobes

Team Integration During the TIE

Probe size was a problem experienced during the TIE. In one participant out of eight, there was a fitting problem due to the size of the probe. In addition, there was near infrared interference from the eye tracker sensor that may have interfered with the results of the fNIR gauge. An additional channel in the fNIR was used to help with sensor hardware integration. It captured the effects of ambient light and the interference from other sensors such as the eye tracker. Adaptive filtering was also applied to eliminate interference.

EEG would best complement fNIR because fNIR provides a hemodynamic response and EEG provides information about electrical activity. Also, temporal resolution is better with EEG and spatial resolution is better with fNIR. Hence, integration would give a better temporal and spatial resolution.

Increasing the difficulty level by changing the wave size (number of tracks per wave) and number of yellow tracks (track difficulty) is an aspect of WCT that best demonstrated the feasibility of fNIR as well as tracking of the response time to IFF. WCT limited the gauge's ability because there was not a pause between waves to record new baseline data. In addition, the WCT required extensive hand movements.

Future Application

In the future, wireless broadcast directly from the probe and base station will be PDA. Also, the data acquisition system (control box) and the power supply will be miniaturized. For the current system, the temperature of the environment must be maintained to control for excessive sweating. Because the current design is not yet fully wireless and miniaturized, applying it to real-world environments may not be practical. The current temporal resolution in detecting changes in operating state is an average of one-wave point (75 seconds). The future system will include the single event detection and temporal resolution of 4 to 6seconds.

Advantages/Limitations

Foot Soldier in the Field – Wireless, miniaturized, and ruggedized future generation will be amendable to all situations.

Noisy Environments – Wireless, miniaturized, and ruggedized future generation will be amendable to all situations.

In a Moving Vehicle – Wireless, miniaturized, and ruggedized future generation will be amendable to all situations.

Multitasking Environment with Multiple Displays and Controls – Wireless, miniaturized, and ruggedized future generation will be amendable to all situations.

Command and Control Display Operator – Wireless, miniaturized, and ruggedized future generation will be amendable to all situations.

Key References

Chance, B., Anday, E., Nioka, S., Zhou, S., Hong, L., Worden, K., Li, C., Murray, T., Ovetsky, Y., Pidikiti, and D., Thomas, R. (1998). A novel method for fast imaging of brain function, non-invasively, with light. *Optics Express*, 20, 435-422.

Villringer, A. and Chance, B. (1997). Non-invasive optical spectroscopy and imaging of human brain function. *Trends in Neuroscience*, 20, 435-442.

5.3.5 Electrical Geodesics, Inc.

Gauge Name: Motor Effort and Auditory Effort

Team Integration During the TIE

There were no noted problems during the TIE with team sensor-gauge grouping. There was no anticipation of any difficulties integrating with other technologies. In fact, the EEG was modified to work with the head-mounted eye-tracking device. EKG measures of arousal and eye-tracking data complemented the EEG because they appeared to be controlled by similar cerebral mechanisms. A more subjective measure of effort would have been preferred from the WCT. That is, participants rate the level of effort exerted during each wave.

Future Application

A very small, perhaps even wearable, system within the next 12 months is anticipated. Excessive participant movements are a major challenge when applying the EEG to real-world environments. Tasks that require participants to by physically active may limit the utility of the gauge in applied environments.

Advantages/Limitations

Foot Soldier in the Field – Not applicable.

Noisy Environments – The gauge can be used in situations without additional modification.

In a Moving Vehicle – It is possible to modify the equipment to work in this environment.

Multitasking Environment with Multiple Displays and Controls – The gauge can be used in situations without additional modification.

Command and Control Display Operator – The gauge can be used in situations without additional modification.

Key References

Bastiaansen, M. C., van Berkum, J.J.A., and Hagoort, P. (2002). Syntactic processing modulates the theta rhythm of the human EEG. *Neuroimage*, 17, 1479-1492.

Gevins, A., Smith, M.E., McEvoy, L., and 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.

5.3.6 QinetiQ, Inc

Gauge Name: Executive Load

Team Integration During the TIE

The only problems encountered were those associated with the positioning of the eye tracking headband relative to the EEG electrodes. This resulted in discomfort for the participant and a noisy EEG signal. This problem was resolved by repositioning the headband. Currently, a relatively large sensor suite is being used and has therefore worked to resolve many of the hardware integration issues, such as grounding requirements, over the last few years. Although the fNIR could provide a solution to examining brain activity without the requirement for electrodes, they were the most difficult group to integrate with because the same scalp sites were required. Dynamic changes in workload of the WCT best demonstrated the gauge feasibility. The task itself was very much stimulus driven with appropriate learned responses, such as the amount of high-level decision-making, limited. Difficulty was manipulated by the temporal nature of the task, rather than an increase in the difficulty of any decision. This limited the ability of the gauge to demonstrate key capabilities.

Future Application

The reduction of amplifier size and changing to portable computers will be expected in the future. When applying the gauge to real-world environments, understanding and canceling movement/EM artifacts will be major challenges.

Advantages/Limitations

Foot Soldier in the Field – No response.

Noisy Environments – No response.

In a Moving Vehicle – No response.

Multitasking Environment with Multiple Displays and Controls – No response.

Command and Control Display Operator – No response.

Key References

Pleydell-Pearce, C.W., Whitecross, S.E., and Dickson, B.T. Multivariate Analysis of EEG: Predicting cognition on the basis of frequency decomposition, inter-electrode correlation, coherence, cross phase, and cross power. *Proceedings of the 36th Annual Hawaii International Conference on System Sciences.*

5.3.7 San Diego State University

Gauge Name: Index of Cognitive Activity (ICA)

Team Integration During the TIE

Due to unknown interference at approximately 10 Hz., usable data with Team 1 was not collected. There was a similar problem with one of two participants in Team 4. At this point, it cannot be determined if interference is being picked up or there is an unknown physiological response. Subsequent post-analyses of the Pre-TIE data show that there was a very small amount of interference with the runs in Team 1, but not the large problem experienced at the TIE.

Paying attention to the order in which the sensors are attached was a technique that was learned during the TIE that helped with sensor hardware integration. This made it easier on the participant. Also learned was adding real-time viewing of the raw pupil signal so that it was known immediately if interference was experienced from any source. EOG sensors on the face would be the most difficult to integrate with the gauge. The sensors are often placed where the ICA cameras pick them up. GSR and fNIR complement the eye-tracker because of the hemispheric separation from the fNIR and the GSR measures arousal, which the ICA does not measure. Three other experiments have been completed with WCT and it was found that ICA correlated positively with wave size throughout each study. However, initial results from the TIE are much less conclusive than in previous studies. There are no known aspects of WCT that limit our ability to demonstrate key capabilities of the ICA gauge.

Future Application

A usable off-the-head system (wireless) is a change in hardware expected in the near future. Operator eye problems that make the eye difficult to track (e.g., lazy eye) is one major challenge when applying ICA to real-world environments. A factor that may limit the utility of ICA in applied environments is competition with other gauges placed on the forehead; many sensors need to be attached to that same location.

Advantages/Limitations

Foot Soldier in the Field – Currently, ICA cannot be used with foot soldiers in the field because the gauge depends on computer technology that is not wireless. As the technology changes and/or as remote eye tracking becomes reliable, it may be possible to have the ICA embedded in goggles worn by the foot soldier.

Noisy Environments – It is unknown how the ICA will respond in noisy environments. The ICA has been tested in a number of locations without any interference such as in various labs and simulation environments as well as on ship. However, significant unexplained interference at the TIE was experienced.

In a Moving Vehicle – In a moving vehicle, the chief limitation is the electrical power for the computers that run the eye-tracking programs. The ICA itself could be applied in a moving vehicle that already has an advanced and powerful computer system. Either goggles or a remote camera could acquire the necessary pupil data.

Multitasking Environment with Multiple Displays and Controls – The ICA was developed to measure cognitive activity during multitasking environments. The ICA's efficacy on multiple displays and controls has been shown (e.g., Tactical Decision Making Under Stress (TADMUS) Project, testing in Newport, RI). Eye and head movement do not affect the gauge.

Command and Control Display Operator – The ICA has been used extensively with Command and Control displays such as the DSS (Decision Support Systems) used by COs (Commanding Officers) and TAOs (Tactical Action Officers) in the TADMUS Project as with games like Warship

Commander. At the Naval Postgraduate School, we are currently using the ICA to examine other aspects of decision-making in these environments. The lighting conditions of the CIC do not adversely affect the ICA.

Key References

Kahneman, D. (1973). *Attention & Effort.* Englewood Cliffs, NJ: Prentice Hall.

Beatty, J. (1982). Task-evoked papillary responses, processing load, and the structure of processing resources. *Psychology Bulletin*, 91, 276-292.

5.3.8 Sarnoff

Gauge Name: Loss Perception

Team Integration During the TIE

The head-mounted eye-tracker combined with the EEG electrodes was too painful for most of the participants. In addition, the recordings collected in combination with the eye-tracking equipment had considerable 60-Hz interference. A more quantitative analysis of the deterioration due to 60-Hz noise was difficult given the small number of participants.

Any mechanical sensor (acceleration, posture, remote eye-tracking) would provide complementary information that should in principle not interfere with EEG. In addition, simultaneous recordings of fNIR and EEG are of great interest from a research point of view. While EEG was hypothesized to respond primarily to incoming information in a given cortical area, the hemo-dynamic activity in fNIR would most likely reflect ongoing activity at the given area. Auditory feedback of the WCT is best demonstrated by the feasibility of the Loss Perception gauge. However, the warning signal selected did not occur very frequently and limited the ability to demonstrate key capabilities of the gauge. As a result, the metric could only be sampled infrequently. Other evoked responses modulated by the task may need to be identified.

Future Application

The current footprint of the data acquisition system contains a PC laptop and amplifier connected over a local-area network. In the future, there will be fewer sensors. In practice, five to 10 sensors may be sufficient if properly placed. Algorithms can easily be implemented on general-purpose DSP. Task complexity and/or uncontrolled environments may make it difficult to define a specific "cognitive event" by complicating the information about the timing of an event.

Advantages/Limitations

Foot Soldier in the Field – Difficult because of motion artifacts and uncontrolled environments.

Noisy Environments – Difficult because of motion artifacts and uncontrolled environments.

In a Moving Vehicle – Difficult because of motion artifacts and uncontrolled environments.

Multitasking Environment with Multiple Displays and Controls – Most applicable because the relevant information to interpret EEG on a sub-second basis is available.

Command and Control Display Operator – Most applicable because the relevant information to interpret EEG on a sub-second basis is available.

Key References

Picton, T.W. and Hillyard, S.A., (1974) Human auditory evoked potentials: Effects and attention. *Electroencephalography and Clinical Neurophysiology*, 36, 191-199.

Rappaport, M., Clifford, J.O., and Winterfield, K.M., (1990). P300 response under active and passive attentional states and uni- and bimodality stimulus presentation conditions. *Journal of Neuropsychiatry and Clinical Neurosiciences*, 2, 4, 399-407.

5.3.9 University of Pittsburgh and Naval Research Laboratory

Gauge Name: Head/Monitor Coupling, Head Bracing, Back Bracing

Team Integration During the TIE

During the TIE, attachment of head tracking sensors was sub-optimal due to physical interference by other devices worn on the head. The "hands-on experience" with the other sensors at the TIE demonstrated that there are no insurmountable problems with (1) integrating the sensors with other AugCog sensors and (2) augmenting the approach with additional movement sensors (pressure sensors in a passive restraint (seat belt and harness) system for other platforms). It also became clear that it was completely feasible to use real-time movement data to trigger detection of EEG, eye movement, heart rate, or papillary size events to design new gauges during Phase 2.

It was possible to integrate the posture gauges easily with the sensor requirements of all of the other developers. The posture gauge complements each of the other gauges because it takes its measurements from an independent source (assessing automatic behavioral responses to sensory input/cognitive challenges). As a result, it may provide a source of online, real-time, independent validation of the other gauges.

In the WCT, the posture gauges were able to detect the subtle shifts in seated posture and head position that result from increased attentional perceptual, and/or actionable demands on the user. Thus, events calling for immediate, intensive actions were the easiest for the gauge to detect. The one aspect most lacking in WCT was unpredictability in the timing and difficulty of the waves. Due to the regularity of events, it was impossible to clearly demonstrate the ability of the gauge to detect changes in attention/perception challenges in environments where the participant does not know what task-load is coming or when it will occur.

Future Application

Future advances such as integrating the head tracking data with eye-tracking data to produce accurate gaze determination will require more precise placement of the head tracker. This may be a trivial issue if there is more lead-time to accommodate for the physical needs of each instrument.

Instead of the electronics being mounted under the chair, it is intended to integrate the electronics and computer into the chair back. Further testing and validation in a variety of real-world task environments will need to be conducted to determine if any major challenges when applying the gauge to real-world environments exist.

Advantages/Limitations

Foot Soldier in the Field – Not applicable.

Noisy Environments – Fully applicable and not sensitive to noise.

In a Moving Vehicle – Fully applicable and ideally suited. The gauge was conceived and designed for this environment.

Multitasking Environment with Multiple Displays and Controls – Fully applicable as long as individuals are seated at the displays.

Command and Control Display Operator – Fully applicable as long as individuals are seated at the displays.

Key References

Redfern, M.S., Muller, M.L., Jennings, J.R., and Furman, J.M. (2002). Attentional dynamics in postural control during perturbations in young and older adults. *Journals of Gerontology Series A-Biological Sciences*, 57(8): B298-303.

Redfern, M.S., Jennings, J.R., Martin, C., and Furman, J.M. (2001). Attention influences sensory integration for postural control older adults. *Gait & PostureI,* 14(3): 211-216.

5.3.10 University of Hawaii

Gauge Name: Arousal and Stress, Perceptual and Motor Load, and Cognitive Difficulty

Team Integration During the TIE

It took too long to attach sensors to feet (5 to 10 minutes) and to calibrate the eye tracker sensors (about 10 minutes). Because only one eye tracker can be worn at a time, SDSU eye tracker was alternated with the eye tracker sensors. As a result, there was no integration of the eye sensor data into other gauges. All other gauges would complement the arousal/stress gauges. EEG was especially useful because it adds another source of vigilance to complement the arousal/stress gauge. The wave difficulty of WCT best demonstrated the feasibility of the gauge.

Gauge Name: Arousal & Stress Gauge

Future Application

Miniature sensors with miniaturized wearable transmitters are envisioned for the future. Soldiers will not want to wear sensors on their toes, but may be more receptive to wearing sensors on their fingers. There should be few complaints with wearing the sensors on their forearm.

Advantages/Limitations

Foot Soldier in the Field – Only wear forearm model with miniaturized data collection.

Noisy Environments – The gauge will have no problems in this situation.

In a Moving Vehicle – The gauge will have no problems in this situation.

Multitasking Environment with Multiple Displays and Controls – The gauge will have no problems in this situation.

Command and Control Display Operator – The gauge will have no problems in this situation.

Gauge Name: Perceptual & Motor Load Gauge

Future Application

In the future, the entire analog-to-digital conversion can be done inside of the mouse body itself, leaving a Pressure Mouse externally indistinguishable from a regular computer mouse. Because the gauge is only useful for computer applications that use a mouse as the primary input mode, this may limit the utility of the gauge in applied environments.

Advantages/Limitations

Foot Soldier in the Field – Only if soldier uses a gyro mouse to manipulate a wearable computer.

Noisy Environments – The gauge will have no problems in this situation.

In a Moving Vehicle – Not applicable.

Multitasking Environment with Multiple Displays and Controls – The gauge will have no problems in this situation.

Command and Control Display Operator – The gauge will have no problems in this situation.

Gauge Name: Cognitive Difficulty Gauge

Future Application

In the future, the entire analog-to-digital conversion can be done inside of the mouse body itself, leaving a Pressure Mouse externally indistinguishable from a regular computer mouse. More research needs to be conducted to establish correctness of the gauge in a variety of tasks and with a variety of

users. Because the gauge is only useful for computer applications that use a mouse as the primary input mode, the utility of the gauge may be limited in applied environments.

Advantages/Limitations

Foot Soldier in the Field – Only if soldier uses a gyro mouse to manipulate a wearable computer.

Noisy Environments – The gauge will have no problems in this situation.

In a Moving Vehicle – Not applicable.

Multitasking Environment with Multiple Displays and Controls – The gauge will have no problems in this situation.

Command and Control Display Operator – The gauge will have no problems in this situation.

Key References

http://www.trans4mind.com/psychotechnics/gsr.html

http://www.acumeninc.com/policehr.html

5.3.11 University of New Mexico

Gauge Name: Theta Power

Team Integration During the TIE

From the single run EEG data collected by Sarnoff with SDSU's eye-tracking system, using SOBI ICA, a large number of artifacts associated presumably with variable pressure applied to EEG sensors were found. Heart rate monitoring and GSR measures helped with sensor hardware integration. Anything that sat on top of EEG sensors produced large movement-related artifacts for EEG measurement. Conventional physiological measures such as heart rate and GSR best complemented the gauge.

A primary concern was spatial localization of the brain activation during WCT. A secondary goal was to measure theta power in the frontal cortex as a potential index for working memory and for measuring anterior-posterior communication between the sensory system and the executive function. The focus was on brain activity occurring around event 69, which was depressing the communications button. The task potentially can increase the demand of working memory. However, the output provided does not include some of the critical information that provides an objective measure of working memory. Therefore it was difficult to evaluate whether a gauge for working memory works if an independent and objective working memory variable was not available.

Future Application

General environmental concerns include avoiding large power sources around the participant's head (these power sources must be maintained to optimize the gauge performance.) Also, participants should not move their head in such a way that sensors move relative to the scalp. Sweating and movement-related artifacts are major challenges when applying the gauge to real work environments. The temperature and the requirement of large-scale movement by the operator may limit the utility of the gauge.

Advantages/Limitations

Areas are listed in order of potential application.

Command and Control Display Operator – The gauge is useful in this environment.

Multitasking Environment with Multiple Displays and Controls – The gauge is useful in this environment.

Noisy Environments – The gauge is useful in this environment.

In a Moving Vehicle – The gauge is useful in this environment.

Foot Soldier in the Field – The gauge is useful in this environment.

Key References

Jensen, O., and Tesche, C.D. (2002) Abstract frontal theta activity in humans increases with memory load in working memory task. *Eur J Neuroscience*. Apr;15(8): 1359-9.

Tesche, C.D., and Karhu, J. (2000) Theta oscillations index human hippocampal activation during a working memory task. *Proc Natl Acad Science USA.* Jan;18(2): 919-24.

5.4 QUESTIONS AND RESPONSES (SELF-EVALUATION)

Table 14 gives responses to Part III of the questionnaire. The responses reflect a self-evaluation by CWA developers of their gauges and specific issues related to the transition of their gauges in Phase II. These data are represented in three categories, indicating a high, medium, or low rating for each item. A completely filled in circle indicates a high rating, a half filled in circled indicates a medium rating, and an empty circle indicates a low rating.

Table 14. Questions and responses from the CWA developers (self-evaluation).

● = High ◒ = Medium ○ = Low NR = No Response NA = Not Applicable	Advanced Brain Monitoring, Inc.	AnthroTronix, Inc.	Clemson University	Drexel University	Electrical Geodesics, Inc	San Diego State University	Sarnoff	University of Hawaii	University of Pittsburgh & Naval Research Laboratory	QinetiQ	University of New Mexico
Ease of connecting to subjects	●	○	●	◒	●	◒	○	●	●	○	NR
Ease of connecting to subjects in 3-5 yrs	●	●	NA	●	●	●	◒	●	●	◒	NR
Subject comfort	◒	○	●	◒	●	◒	○	●	●	◒	NR
Subject comfort in 3-5 yrs	●	●	NA	●	●	●	◒	●	●	●	NR
Footprint size	◒	◒	●	◒	◒	◒	○	●	◒	○	NR
Footprint size in 3-5 yrs	●	●	NA	●	◒	●	◒	●	●	●	NR
Sensitivity: Ability to detect small changes in “state”	●	◒	◒	◒	◒	◒	◒	●	◒	◒	NR
Sensitivity: Ability to detect small changes in “state” in 3-5 yrs	●	◒	NA	●	●	●	●	●	●	●	NR
Construct validity	◒	◒	●	◒	●	◒	●	●	◒	◒	NR
Construct validity in 3-5 yrs	●	◒	NA	●	●	●	●	●	●	●	NR
Day-to-day reliability	◒	◒	●	◒	●	◒	○	●	●	◒	NR
Day-to-day reliability in 3-5 yrs	●	◒	NA	●	●	●	●	●	●	●	NR
Resistance to external noise	◒	◒	●	◒	●	◒	○	●	●	○	NR
Resistance to external noise in 3-5 yrs	●	●	NA	●	●	●	◒	●	●	●	NR
Real time capability	●	●	●	◒	◒	●	●	●	◒	●	NR
Real time capability in 3-5 yrs	●	●	NA	●	●	●	●	●	●	●	NR

● = High

◒ = Medium

○ = Low

NR = No Response

NA = Not Applicable

	Advanced Brain Monitoring, Inc.	AnthroTronix, Inc.	Clemson University	Drexel University	Electrical Geodesics, Inc	San Diego State University	Sarnoff	University of Hawaii	University of Pittsburgh & Naval Research Laboratory	QinetiQ	University of New Mexico
Allow operator mobility	◒	◒	◒	◒	◒	◒	○	○	◒	○	NR
Allow operator mobility in 3-5 yrs	●	●	NR	●	●	●	◒	●	◒	◒	NR
Predictive ability	◒	●	◒	○	◒	◒	◒	●	◒	◒	NR
Predictive ability in 3-5 yrs	●	●	NR	●	●	●	◒	●	●	●	NR
Fit into Phase II Architecture	◒	●	●	◒	◒	●	◒	●	●	●	NR
Fit into Phase II Architecture in 3-5 yrs	●	●	NR	●	●	●	●	●	●	●	NR
Overall real-world application	◒	●	●	◒	●	●	○	●	◒	●	NR
Overall real-world application in 3-5 yrs	●	●	NR	●	●	●	◒	●	●	●	NR

6. CONCLUSIONS

The Augmented Cognition Technical Integration Experiment was an ambitious demonstration and evaluation of psychophysiological measures of cognitive activity. It brought together a range of sensor technologies developed by a number of independent researchers to serve as cognitive state gauges, and allowed the performance of those gauges to be evaluated in the context of common quasi-realistic command and control task. There were also a number of issues identified that will be relevant to future augmented technology assessments and the successful transition of these technologies.

6.1 INNOVATION

The experiment was innovative in several ways. First, the sensors used by many of the teams were state-of-the-art and contained emerging technologies and innovative sensor hardware not previously available. For example, the gauge developed by Drexel University used functional near infrared sensors for detecting changes in cortical blood flow—an emerging technology that is in the early stages of development. Other sensor technologies, while more established, were innovative in other ways, either by virtue of their hardware design and implementation, processing algorithms, technical approach, or theoretical underpinnings. For example, the EEG system of Advanced Brain Monitoring is wireless and therefore very lightweight and mobile. Electrical Geodesics' 128-electrode EEG sensor net significantly increased the number of sensors that could be practically placed on the head, thereby increasing the ability to localize neuronal signals to specific regions of the brain. In addition, the net was comfortable enough that users could wear it for hours without complaint. Although not the primary purpose of this report, issues such as comfort and usability clearly need to be considered when evaluating these technologies for future applications and transition to operational systems. Other examples of innovative technology include the University of Hawaii's pressure mouse for detecting arousal and workload from users' hand pressure on a mouse, and the University of Pittsburgh/Naval Research Laboratory's newly developed "posture chair," which measured 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 raw sensor data. Many of the gauges employed 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 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 used 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.

This high degree of innovation in gauge development came 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 (see section 5 and the gauge developer appendices for details on each gauge). 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 relationships to established cognitive functions are speculative or unknown. For the purposes of the DARPA Augmented Cognition program, these limitations are considered acceptable because these gauge developers may now return to their laboratories to replicate their findings, validate their new gauges against other tasks, and connect them to a more solid theoretical foundation.

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 (seconds), or near real-time (minutes), for the first time.[30] Further, several alternative schemes were designed if not implemented to allow both gauge and task performance data to be fused in near real 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.

A fourth area of innovation lay in the simultaneous data collection from multiple sensors. It is our belief that the successful augmentation of human cognition will not come from a single sensor technology or cognitive state gauge, but rather will result from the hybrid integration of several gauges using both physiological and objective data derived from the tasks being performed by the operator(s). To make these combinations possible, several hardware, software, and sensor interactions were discovered and addressed. For example, many of the sensors required contact with the users' head. Some sensors, such as EEG, measure electrical activity on the scalp. Other sensors, such as eye tracking, are mounted on the head. Consequently, headspace had to be coordinated and shared. Sensor types that could be attached to other body parts, such as heart rate, GSR, and body posture, were desirable in this regard, since they left the head free for other sensors. Two of the systems used infrared light (e.g., the eye-tracking cameras and the fNIR diodes), which initially caused interference between the systems. Additionally, the real-time nature of the gauges required careful coordination and synchronization between task events and the readings of each gauge. These issues were identified, and methods were developed to overcome or eliminate problems, through a series of increasingly sophisticated pilot studies prior to the TIE.

The fifth area of innovation involved the use of a relatively complex task. Much of the early development of cognitive state gauges has been conducted using very simple tasks in carefully controlled laboratory environments that allow clear manipulation of single factors and straightforward interpretation of results. This simplification has helped researchers demonstrate the possibility of using psychophysiological measures as indices of cognitive states, but provided little basis for assessing the extension of those findings to other task domains. Though the use of more complex tasks, as well as the use of multiple sensor technologies, is increasing (see for example, Fournier, Wilson, & Swain, 1999[31]; Smith, Gevins, Brown, Karnik, & Du, 2001[32]; and Van Orden,

---

[30] Berka, C., Advanced Brain Monitoring (personal communication, March 2003).

[31] 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.

[32] 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.

Limbert, Makeig, & Jung, 2001[33]), the WCT represented a dramatic increase in task complexity for many of the research groups and their techniques and gauges. Many of the gauge developers were able to meet this challenge by adapting their gauges to the task and demonstrating significant and large effects related to changes in task load and cognitive activity. The relatively complex and fast-paced WCT brings the detection of cognitive states an important step closer to real-world settings.

Several of the gauge developers noted that this increase in task complexity was accompanied by a large increase in motor activity in the form of mouse movements, key presses, and eye movements. Though this high level of activity made the task more realistic, 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. Simpler, standardized tasks commonly used in basic EEG research control for this problem by keeping eye movements to a minimum. This simplification increases the sensitivity and reliability of the EEG measures but greatly reduces the potential for applying these gauges and their data analysis algorithms to real-world tasks. To address the increased user activity found in the Warship Commander Task, gauge developers, 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. If nothing else, participation in the TIE provided a basis for many of these researchers to reassess, and if necessary, redirect their research and gauge algorithms to accommodate the demands of more realistic task environments.

6.2 ISSUES IN COGNITIVE STATE IDENTIFICATION

The Warship Commander Task manipulated three aspects of task load: Number of Tracks per Wave, Track Difficulty (cognitive and procedural complexity), and presence or absence of the Secondary Verbal Memory Task. Of necessity, all three aspects of task load simultaneously manipulated a number of cognitive and perceptual/motor factors so that a whole variety of cognitive state gauges could be applied to the task with a high probability that as many gauges as possible would detect changes in different cognitive states. Each of the task load factors was shown to significantly affect the participants' response times and accuracy during the TIE. Nonetheless, not all aspects of user cognition and workload were specifically manipulated by the WCT. For instance, the amount and complexity of relatively high-level decision-making was limited. Therefore, the assessment of any gauges claiming to measure higher order cognitive processing was limited. Further, since all participants had at least 1.5 hours of practice on the task, and some had far more practice, frustration and stress were not systematically manipulated by the WCT. However, there was stress in the form of substantial time pressure during the higher episodes of task load, as indicated by participants' subjective reports and the performance measures of response time errors and task completion. Further, the demands placed on the participants was such that fatigue was a significant factor over the course of the day, but due to the requirements of the test protocol, there was no way to systematically assess the various gauges in terms of their ability to detect changes in such factors as fatigue across participants or data collection teams.

---

[33] 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.

Eleven of the 20 gauges used during the TIE successfully detected changes in at least one of the three task load factors ($p < .05$). Five additional gauges were *marginally* sensitive ($p < .10$) or *potentially* sensitive ($p < .20$) to changes in task load. Given the high degree of innovation required to participate in the TIE, and the number of new and experimental gauges, the quantity of statistically significant results is impressive, and the number of marginally, or even potentially, significant results is encouraging. Larger studies, with additional participants, as well as further refinements to sensor and gauge technology may be able to quickly and dramatically boost gauge sensitivity.

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 manipulation for task load was much greater for this factor than for any 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 Secondary Verbal Memory Task. Although not predicted by any of the developers, these different sensitivities suggest that the gauges may be sensitive to different aspects of cognition and task load, again suggesting that the successful augmentation of cognition may require an integrated array of gauges tailored to the unique task demands of different application environments. For example, some gauges may specifically focus on detecting levels of executive function while others focus more on verbal or auditory function. The WCT was not designed to explicitly differentiate these aspects of cognition; rather, it was designed to demonstrate the effectiveness of many different gauges in a common complex task. Such differentiation, of course, is an important and ongoing project for many of the gauge developers (see the gauge developer appendices for examples) as they look to the manipulation of cognitive state and optimization of human performance in complex systems during Phase II of the Augmented Cognition program and beyond. Clearly, a lesson learned for the developers of augmented cognition systems is to conduct sufficient cognitive task analyses and differential gauge analyses so that they can identify where and when cognitive demands come from and prescribe an appropriate augmentation.

As a class of gauges, the "arousal" gauges stood out for their inability to detect changes in any of the three task load factors. Since 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, or simply that the gauges themselves were insensitive. As noted above, the WCT does not explicitly manipulate stress, arousal, or physical activity other than in terms of mouse and eye movements. Several of the gauge developers suggested that the introduction of stronger negative consequences for errors committed during the task might have produced more measurable stress changes. For example, game score deductions and loud audio error alerts might have created more stress, especially during high task load periods of the task. It may also be the case that well-practiced command and control-type tasks simply do not evoke strong stress responses, and arousal gauges may not be appropriate for measuring changes in workload in such tasks. However, under operational conditions, the negative consequences of errors can be profound, and changes in stress levels may be important to detect. Therefore, we do not recommend eliminating this class of cognitive state gauges at this time.

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 level of arousal, or operational performance.[34] It may be that arousal gauges are better suited for monitoring novices during training and 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 relationships 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, while others were not. The development and use of ERP gauges is somewhat problematic in 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 the 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 gauges has the potential to measure specific cognitive processing occurring during a task.

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. 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.[35]

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? When gauges were sensitive to the experimental conditions manipulated in the TIE, we were able to assess gauge consistency across trials and participants. For example, Drexel's fNIR gauge of the left hemisphere was considerably sensitive to changes in task load for all but one participant, though the degree of sensitivity varied from participant to participant. ABM'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 other participants. Other gauges, such as QinetiQ's EEG-based Executive Load gauge and Hawaii's Cognitive Difficulty gauge were highly sensitive for each participant. These results are very encouraging, but they also suggest that one limitation to the feasibility of applying these technologies in operational settings may be differences in gauge sensitivity between individuals.

It is not known at this time what might account for the gauge variability. High variability may be related to a range of factors from robustness of the measure, 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" in order to control for individual differences. Future improvements in sensor hardware may also eliminate some problems. As gauge developers gain experience with their innovative gauges and with working in noisier environments, sensitivity and consistency may increase substantially. Future work may also show that for certain types of gauges, one suite of sensors may not be universally applicable to all

---

[34] Berkan, M. M. (2000). Performance decrement under psychological stress. Human Performance in Extreme Environments, 5, 92-97. Menza, Lt. M.D. (2002, March). The pucker factor. *Approach*. Retrieved June 27, 2003, from http://www.safetycenter.navy.mil/media/approach/issues/mar02/pucker.htm

[35] 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. *Proceedings of the 36th Annual Hawaii International Conference on System Sciences.*

operators performing a task. Instead, individualized suites of cognitive state gauges may need to be available for different users, or a pre-screening 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. This issue represents a potentially critical area of research as the quest for augmented cognition moves forward.

6.3 TRANSITION AND APPLICATION

As the Augmented Cognition program transitions into Phase II and beyond, there remain some conceptual and practical issues that will take on added significance. First among the conceptual issues is the continuing need to define or refine the psychological construct of each gauge. The developers will need to determine for themselves: What cognitive process does the gauge really measure? The definition of cognitive state is central to understanding how a gauge will generalize across tasks. Refining the psychological construct of each gauge should also improve the sensitivity and precision of the gauges themselves. As the gauges become better defined, their meaning will become more precise, and their application to different tasks will become more clear.

A second conceptual issue is the need to refine the consistency of each gauge across users. As discussed above, one strategy, employed by several gauge developers during the TIE, was to obtain a set of baseline measures on each participant prior to the actual experiment. Typically, these baseline measures are obtained from very simple psychometric tasks that underlie many complex tasks. This base-lining strategy dovetails with refining the psychological construct of a gauge, since the baseline tasks may be simpler or more pure versions of the processes measured by a gauge. Another strategy for refining the consistency of a gauge is to develop a better understanding of the underlying physiology and dynamics of the cognition measured by a gauge and their variability across users. This strategy is necessarily the product of long-term development and will require substantial applied cognitive neuropsychological research.

A third conceptual issue is the need to understand how experience with a task influences gauge measurements. For example, novice users may suffer stress in task situations that expert users find mundane. It is also quite possible that experience will result in decision-making strategy shifts that will change the amount and/or type of cognitive processes used to perform a task, and thereby diminish or remove the specific processes measured by a gauge. Some individual differences in gauge consistency may be due to strategy changes that occur as users become more adept at a task. Addressing this issue will require an understanding of how the results of a gauge change with experience and how strategies for a particular task shift with experience.

There are also practical matters to consider as gauges are transitioned to more applied tasks. Issues include sensor and gauge robustness, mobility, and wearability. Over the long-term, it is likely that many of the gauges will evolve into small, highly mobile, and minimally disruptive sensor and processing systems. In the short-term, however, there are clear differences among the gauges in their applicability to real-world settings. The successful application of augmented cognition will require gauge technologies that are acceptable to the user communities. Cognitive gauge sensors must be comfortable and convenient to wear for extended periods of time, with users who may move about rooms, if not a battlefield. Sensors cannot interfere with whatever mobility is required for performing all tasks the wearer might be called on to do under any number of operating conditions. The sensors cannot interfere with the users' vision or hearing or with other equipment around or on the user, or they will not be accepted by the user population. Complications may also arise from interactions with other tools and equipment. For example, in many applied settings, users wear headphones or microphones that require mounting on the users' heads—the same parts of the body that are used by many of the cognitive state gauges.

Other aspects of the working environment may interfere with gauge measurements. For example electro-magnetic emissions may interfere with gauge sensors. Similarly, many tasks require users to speak and move to an extent that may interfere with accurate gauge measurements. As discussed above, EEG systems are typically quite sensitive to muscle movement. Since eye muscles lie close to the forehead, they can cause significant interference for gauges that measure forebrain activity. This problem may be particularly significant for tasks that require extensive visual scanning. Similarly, jaw and tongue movements that occur during speaking can interfere with brain measures. While analytical methods for measuring and factoring out this interference is progressing, it continues to pose some problems for some gauges, and therefore the application developer using these gauges must be aware of the unique task requirements of the application environment and the potential mitigating factors that would impact the utility of specific cognitive gauge technologies. Recall that mouse and eye movements were reduced for the TIE by using keyboard shortcuts to reduce interference from the electrical signals produced by muscle contractions. The success of gauges during the TIE indicates that many of these issues can be addressed satisfactorily, but increases in user activity will pose new problems and exacerbate old ones.

In summary, the Augmented Cognition Technology Integration Experiment leads to the following recommendations:

1. Develop sound cognitive constructs for each gauge early in the development process. Ideally there should be an explicit theory that underlies the construct(s), and which will provide a clear basis for predicting and interpreting results.
2. Define explicit requirements for what the gauge needs to do to drive the desired augmentation—define quantitative and qualitative exit criteria. Map the gauge to cognitive task requirements in terms of multiple parameters.
   a. What do you need the gauge to tell you?
   b. Under what conditions must the gauge work?
   c. What do you intend to manipulate based on the gauge?
   d. How stable/consistent is that gauge likely to be under the conditions you expect to use it?
   e. Is the gauge likely to be sensitive in the operational environment?
   f. How much time lag can you accept from the gauge? How reactive can your augmentation be? Is there a way to use the gauge results to be proactive?
   g. How might experience or changes in task strategy affect gauge reliability?
   h. Will users accept the gauge?
3. Consider employing multiple cognitive gauges to address possible individual differences in the sensitivity of individual gauges.
4. Address the integration of gauge technologies and the work environment early in the development process. Design for usability.

6.4 SUMMARY

In summary, the TIE involved a substantial degree of innovation across a range of issues and across a range of technologies. Many of the gauges hold a good deal of promise, both in terms of sensitivity and consistency, for detecting and identifying changes in cognitive state in near real time. Given the degree of innovation seen at the TIE, it seems likely that much more innovation, in terms of sensor hardware, software, and procedures, is possible. In this sense, the TIE represents an important, and critical, demonstration of the potential for real-time cognitive state identification in applied settings.

Phase I of the Augmented Cognition program has achieved its goal of providing a solid foundation for the demonstration of augmented cognition systems. Phase II of the Augmented Cognition program will address the challenges of defining augmentation strategies for complex tasks performed in realistic environments, and demonstrating that cognition can be manipulated based on appropriate cognitive state gauges.

APPENDIX 1: WARSHIP COMMANDER TASK ANALYSIS

The goal of this task analysis is to identify cognitive processes involved in the Warship Commander task that can be manipulated in the task and measured either behaviorally or neurologically.

There are numerous methods for breaking down the task that may prove valuable in different ways. Consequently, two different methods are reported here. The first method uses a standard break down of cognition into information processing stages (e.g., Parasuraman, Sheridan, & Wickens, 2000[36]): information acquisition, information analysis, decision selection, and action implementation. The steps of the Warship Commander task for these stages are categorized in Section I.

A second method for breaking down the task is to write a production system that could perform the task to a reasonable degree of match with users' performance and strategies. A production system seems particularly appropriate for the Warship Commander task because the task involves a great deal of situation awareness and reaction to situations. A production system for the task is described in Section II.

SECTION I – STAGES OF COGNITION

Information Acquisition

Detect new tracks
Observe their colors
Activate communication window
Count time
ID time
Warn time

Information Analysis

Color
Position (e.g. below LOE, lowest, close by)
Communication (reading ID)
Time (counting seconds for ID time warn time)

Decision Selection

Which track to handle next?
ID?
Warn?
Engage?

Action

IFF, Communication, Warn, Engage

[36] Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans, 30,* 286-297.

SECTION II – PRODUCTION SYSTEM

Table 1 presents a production system for performing the Warship Commander task, both the air warfare primary task and the ship status secondary task. This production system is meant to capture the strategies and performance of typical users to a reasonable level of accuracy. It is not meant to capture every nuance of strategy, just the basic activity. Table 2 organizes the conditions and actions from the productions into three separate lists: perceptual non-spatial processes such as recognizing a color, perceptual spatial processes such as recognizing that a track is south of the Line of Engagement (LOE), and mental processes such as recalling the memorized assessment of a track.

Productions are, essentially, condition-action statements. If their condition side is satisfied, their action side will be executed. In a production system, there is a repeating cycle of activity. On each cycle, the productions are checked to see if the conditions of any productions are satisfied. Productions are checked in order of priority, and the first production found to have satisfied conditions is the production that is executed. Once a production has been executed, the cycle repeats. With these simple rules in mind, the Warship Commander production system is easily understood.

The productions in Table 1 are listed in descending priority, meaning that Engage Red has the highest priority. If the conditions of this production are satisfied, this production is the one to execute regardless of whichever other productions are also satisfied. Consequently, whenever there is a red track below the LOE, the production system will engage that track regardless of whatever else is happening. This strategy seems to roughly correspond with users' performance on the task.

The second priority production is to engage a yellow-hostile track. There are several conditions that must be satisfied: the track must be yellow, a perceptual judgment; the track must be below the LOE, a spatial-perceptual judgment; the track must be continuing southbound, a spatial-perceptual judgment; the track must be hostile, a mental judgment that requires recalling the assessment of that track previously found in the communications window; and finally the track must have been warned at least three seconds ago, another mental judgment that requires recalling that the track was warned and that at least three seconds have elapsed.

The third priority production is to wait and do nothing for a cycle if short-term memory has become overloaded with information about too many concurrent tracks. This production places a strategic limit on the number of tracks that can be handled simultaneously. Typically, it is observed that users concurrently handle only two or at most three tracks. Handling more tracks concurrently would appear to place too great a burden on remembering the state of each track: who is hostile, who has been warned, and how long ago each was warned.

The fourth priority production is to warn yellow-hostile tracks that pass south of the LOE. In addition to the perceptual conditions of being yellow and south of the LOE, there are three mental conditions: not previously warned, assessed to be hostile, and at least two seconds having elapsed since the track was ID'ed. Executing this production entails warning the track, setting a mental note to that effect, and starting a mental timer.

The fifth priority production is to assess a yellow-unknown track. This production determines whether the track is hostile or friendly and sets a mental note to that effect. Users tend not to assess yellow tracks as soon as they are discovered. Rather, users tend to wait until the tracks reach the LOE, where they can be warned and engaged. Assessing yellow tracks above the LOE places a burden on short-term memory without any immediate use, and is typically avoided by

users. Subtler strategies include assessing a yellow track just prior to its crossing the LOE and choosing to work on either the lowest tracks on the screen, or alternatively, tracks close to where the user is currently working.

The sixth priority production is to ID any white unknown track that appears on the screen. Even though this is the first action that users take in any run of the task, it clearly begins to defer to the higher priority actions as the task progresses.

The seventh and lowest priority production is simply to scan the display for new white tracks. This production executes when no other production is satisfied.

The three ship status productions work similarly. When the air warfare task and the ship status task are performed concurrently, their productions must be integrated into a single prioritized list. If users give the ship status task highest priority, then its productions will appear at the top of the list, and they will be executed in preference to any of the air warfare productions. It appears that users more or less do give the ship status task highest priority, but that the rehearsal production falls lower in the list, perhaps even below ID White.

Errors can occur in any of the productions' conditions or actions. Users can fail to check a condition and execute a production even when that condition is not satisfied. Users can miss-evaluate a condition and execute a production on an inappropriate track. Users can also incorrectly execute an action. Perhaps most frequently, users can forget a mental value and have to repeat an assessment or warning or perform an inappropriate action.

In Table 1, the mental conditions and actions are written bold. The conditions and actions are organized by category in Table 2.

Table 1. Production systems for the Warship Commander task: Air warfare and ship status.

Air Warfare

1. Engage Red

if: <red> <below LOE>
then: engage track

2. Engage Yellow-Hostile

if: <yellow> <below LOE> <Southbound>
<hostile> <warn time greater than 3s>
then: engage track

3. Wait

if: **<memory load greater than maximum>**
then: **wait one cycle**

4. Warn Yellow-Hostile

if: <yellow> <below LOE>
<not warned> <hostile> <ID-new time greater than 2s>
then: warn track,
set warned

5. Assess Yellow Track

if: <yellow> <below or near LOE> <lowest or closest yellow>
<not assess-yellowed>
then:comms,
set (hostile or friendly), set assess-yellowed

6. ID White

if: <white>
then: IFF

7. Detect Track

if: <other priority productions are satisfied >
then: scan for new/white tracks

Table 1. (continued)

Ship Status

1. Answer Query

if: <query>
then: understand query, **recall value,** find answer number, press key

2. Encode Statement

if: <statement>
then: understand message, **memorize value**

3. Rehearse

if: <other ship status productions are satisfied >
then: **rehearse all system values**

Table 2. Types of conditions and actions in the productions.

Perceptual NonSpatial	**Perceptual Spatial**	**Mental**
<yellow>	<below LOE>	**<hostile>**
<white>	<Southbound>	**<warn time greater than 3s>**
		<memory load greater than maximum
		<not warned>
		<ID-new time greater than 2s>
		<not assess-yellowed>
engage track		**wait one cycle**
warn track		**set warned**
Comms		**set (hostile or friendly)**
IFF		**set assess-yellowed**
	scan for new/white tracks	
<query>		
<statement>		
understand query		**recall value**
find answer number		**memorize value**
press key		**rehearse values**
understand message		

APPENDIX 2: PARTICIPANT COMMENTS

At the end of participation in the TIE all participants were asked to provide comments regarding their experience with each of the teams. The following are all of the participant comments. Some of the remarks have been edited for clarification. Direct quotes of the participants are in quotation marks. Developers may find these comments useful in the future development of their gauges.

TEAM 1

- "The web-like sensor (EEG) was the most comfortable of all devices placed on my head. The water was a nice change from the messy gel. After five minutes or so, I barely noticed it was on."
- "Very friendly, funny and very informative. They explained every step and the equipment was very comfortable."
- "Total time was two hours." I liked that there were only three investigators in the room at a time while setting me up and that each piece of equipment used was explained to me. "I appreciated the female investigator taking me to the restroom to put electrodes on my chest, lower rib cage, and other areas. All investigators were gentle during setting up and removal of equipment. The team was very quick and organized. Data collected was explained and shown to me. Overall, I had a good experience."
- "The team was very pleasant. I was asked if I needed anything and was comfortable."
- "Wonderful. I would have another 4 trials. Tom and Phan were always concerned with how I felt."
- "Team one's task performance was fine. "Made me feel relaxed and comfortable."
- "Total time was one and a half hours." "I liked that each piece of equipment used was explained to me. I also liked that one organization was in the room at a time when setting up. The room was very spacious. All investigators were very gentle when setting up equipment. The posture chair wasn't ready when we started the experiment. When the chair was ready, the experiment went on uninterrupted. All investigators asked how I was doing several times. The team was very quick and organized. Investigators of the Dense Array EEG were not present when the experiment was complete. SDSU investigators helped me remove the EEG cap. Overall, I had a good experience."

TEAM 2

- "Combined with the head tracking device, the brain sensor pad on my forehead [fNIR] became bothersome after the third [scenario]. Other than a couple of itchy places on my head during the experiment, my vision was not affected by the eye tracking equipment."
- "Team 2 was nice, it just took awhile to set up and get situated." "The waffle [fNIR] pressed on my forehead so hard that it still hurts the next day. They were accommodating and any time I experienced discomfort they were quick to help."
- "Team members were coming in and out of the room during testing and it was very distracting. "Eye tracker [U of Hawaii's] was uncomfortable." I had to look thru a piece of glass with one of my eyes. It made the screen half one color and half another. "Waffle hurts".
- "A bit distracting as people walked in and out of the room." They were very kind and concerned for my well-being. "They worked well together."
- "Team two performance was fine. "They made me feel relaxed and comfortable."

- “Total time was one and a half hours”. “All investigators were very gentle when setting up equipment. The team was very quick and organized.” All of the investigators asked how I was doing several times. “When I did feel discomfort, the investigators came to my aid immediately. When the experiment was complete, investigators guided me in taking off some of the equipment that might hurt me if they did it themselves. I really appreciated this.”

TEAM 3

- “None of the equipment bothered me throughout the experiment. I thought that the sensor gloves placed on my hand were inventive. The team was friendly. The only complaint I had was the scraping of dead cells on my scalp with the cotton swab to get a good contact with the [EEG] sensors.”
- The team was very easy to work with. “Equipment was comfortable.”
- “The best team to run for! Made me extremely comfortable. Excellent teamwork.”
- The performance of team three was fine. “They made me feel relaxed and comfortable.”

TEAM 4

- I did not find any of the equipment to be bothersome. “The team was great. However, the calibration time added some irritation to the wait time.”
- I had no practice trial, I inquired about this on the first wave and was told not to worry about it. While the neural cap [EEG] was being applied the word “whoops” was said a few times without explanation. “The guy who said he had trouble getting out of bed and could make breakfast on time, forgot to plug in the data cable to record on the search task”. The infrared mouse had trouble functioning and was “sticky” throughout the first two trials. Electrodes from the EISG were accidentally ripped off of my chest. “When starting trial one we had to break so something could be done that was said “should have been done before I got there”. The team did not seem to work well together.
- The performance of team four’s tasks were fine. “They made me feel relaxed and comfortable.”
- “Total time was three hours. Two and a half hours to set up.”
  - University of New Mexico: “The EEG/ERP cap was not put on gently.” When the cap was removed, the electrodes on my face and chin were still attached to either the cap or another device, which pulled at my face. “I stopped the experiment after only completing two scenarios because I felt discomfort and I was fatigued. When I quit the experiment, I felt UNM was very disappointed that I stopped.” The investigators tried to push another “quick” five-minute experiment after I had already quit. “The female investigator said that she could do the experiment and was going to after I was done. Later I saw her with the equipment on.” Two of the investigators were very apologetic and showed their concern with a small gift.
  - San Diego State University: “The eye tracker pushed on electrodes against the sides of my head. All investigators reminded me that if I felt discomfort, I could stop at any time. All investigators were very concerned with my comfort level and were very apologetic when I stopped the experiment.”
  - Lockheed Martin: “Equipment for this organization was put on last and they had several problems. This equipment could have been put on first before any headgear was added. Instead, I sat with the EEG/ERP cap and eye tracker until they found the problem”. The female investigator who put on the equipment did not take me into another room when they had to attach electrodes to my chest and abdomen underneath by blouse. The entire team

was present during this time. When I stopped the experiment, the investigators were not around to remove their equipment, therefore the SDSU investigators helped me. “I left the experiment with several electrodes attached to my chest because I didn’t realize I had several on my chest and abdomen.

- Overall: “The room was very little and there were too many people in it at a time. There were cords/wires all over the floor. The team did not work together. There was lots of talking during the set up and much confusion was expressed between teams. There was a lot of plugging and unplugging of wires that could have been done before I was set up. The order of equipment used could have made it easier on me if Lockheed went first, followed by UNM, and lastly SDSU. I left the experiment very unhappy. Overall, I had a horrible experience and I would not return as a participant for UNM.”

APPENDIX 3: DEVELOPER APPENDICES

In addition to the analysis performed in Section 4, each individual CWA developers submitted a summary for the data collected and analyzed from their specific gauge or gauges during the TIE. These summaries are provided in a free format allowing each developer to choose a format that best presents their report. These individual reports are presented alphabetically in the following sections.

Appendix 3a	Advanced Brain Monitoring, Inc.
Appendix 3b	Anthrotronix
Appendix 3c	Clemson University
Appendix 3d	Drexel University
Appendix 3e	Electrical Geodesics, Inc.
Appendix 3f	Qinetiq
Appendix 3g	San Diego State University
Appendix 3h	Sarnoff
Appendix 3i	University of Hawaii
Appendix 3j	University of New Mexico
Appendix 3k	University of Pittsburgh

3A ADVANCED BRAIN MONITORING, INC.

EEG Indices of Workload in the Warship Commander Test Augmented Cognition Technical Integration Experiment

Introduction

The research goal was to determine whether quantification of the EEG in real-time using the B-Alert™ indices of alertness would correlate with cognitive workload as manipulated by the Warship Commander Test (WCT). This evaluation was conducted with data acquired during the Technical Integration Experiment (TIE), where multiple sensors were employed to detect a variety of psycho physiological parameters during WCT, and in three preliminary studies. The B-Alert model system classifies each second of EEG on the alertness-drowsiness continuum and was developed with the intention of providing early warnings of the onset of drowsiness. Classifications are obtained using a discriminant function analysis (DFA) derived from a large normative database and fitted to each individual's unique EEG patterns with data acquired from three baseline conditions. The B-Alert system was validated in sleep deprivation studies with performance in a driving simulator [1], accuracy and reaction time during a psychomotor vigilance task, behavioral evidence as measured by cessation of finger tapping, visually scored observations of facial signs of drowsiness (eye closures, head nods) and responses to a subjective sleepiness questionnaire [2, 3]. B-Alert was also independently validated by visual inspection of the EEG signals by two board-certified sleep specialists [2, 4]. Analysis of the B-Alert indices during 44 hours of sleep deprivation revealed that changes in the EEG can predict performance deficits [4] and confirmed previous reports [5, 6] that individuals differ in their vulnerability to sleep deprivation.

Although the B-Alert system was not specifically designed to assess workload, previous studies revealed that highly engaging or difficult tasks induce higher overall levels of vigilance as measured by B-Alert. In addition, other investigators have reported validation of EEG measures of workload that reflected differences in task-related cognitive resource allocation, task mastery and task overload [7-11]. The EEG variables employed in these models (e.g. alpha suppression, beta/alpha plus theta) reflect changes in vigilance, providing support for the potential utility of the B-Alert model in assessing workload.

Methods

Acquisition Hardware System: The ABM sensor headset acquires six channels of EEG or EOG using either a unipolar or bipolar montage. Data are sampled at 256 samples/second with a band pass of 0.5 Hz and 65 Hz (at 3dB attenuation) obtained digitally with Sigma-Delta A/D converters. The RF link is frequency-modulated to transmit at a rate of 57 kBaud in the 915 MHz ISM band. When utilized in the bi-directional mode, the firmware allows the host computer to initiate impedance monitoring of the sensors, select the transmission channel (so two or more headsets can be used in the same room), monitor battery power of the headset, and retransmit dropped packets. The RF sub-system includes a micro-controller, 3-volt power supply, battery recharging circuitry and connector, two LED function indicators and an onboard RF antenna. The standard hardware montage includes bipolar recordings from Fz-POz and Cz-POz for the B-Alert system, unipolar recordings from Fz, Cz and POz referenced to linked mastoids for Event Related Potentials analysis, and a bipolar configuration for HEOG and VEOG.

Artifact Identification and Decontamination, and Signal Processing: The B-Alert system automatically detects and decontaminates data points associated with amplifier saturation or dropped data packets from the RF transmission. Three sets of filtered EEG data are then derived using a 0.5 Hz 256th order high-pass FIR filter, a 4 Hz 640th order FIR high-pass filter and a 7 Hz IIR low-pass filter. In order to obtain faster computations, both high-pass filters are realized by subtracting the output of the corresponding low-pass filter from the original signal. Identification of eye blinks in the EEG without the use of a reference EOG channel is achieved by filtering the fast component of the EEG with a 7 Hz IIR low-pass filter, applying cross-correlation analysis to the filtered signal using the positive half of a 40 μV 0.1875 Hz sine wave as the reference, and applying thresholds to the outputs from the cross-correlation analysis. Once eye blinks have been recognized and saturation removed, the 0.5 Hz high-pass filtered EEG signal from each channel is further decontaminated. The data points corresponding to the range between the start and end of each eye blink are replaced by the same data points after application of the 2.5 Hz filter. Spikes and excursions are then identified based on changes in amplitude over defined ranges. In each of the EEG channels, all data points contaminated with saturation, spikes, and excursions are replaced with zero values.

Decontaminated EEG is segmented into overlapping 256 data-point windows called overlays. An epoch consists of three consecutive overlays. Fast-Fourier transform is applied to each overlay of the decontaminated EEG signal multiplied by the Kaiser window ($\alpha = 6.0$) to compute the power spectral densities (PSD). The PSD values are adjusted to take into account zero values inserted for contaminated data points. The PSD between 70 and 128 Hz is used to detect EMG artifact. Overlays with excessive EMG artifact or with fewer than 128 data points are rejected. The remaining overlays are then averaged to derive PSD for each epoch with a 50% overlapping window. Epochs with two or more overlays with EMG or missing data are classified as invalid. PSD values are derived for each one Hz bin from 3 Hz to 40 Hz and the EEG bands from 3 to 30 Hz and 3 to 40 Hz.

Classification Model: The four-class B-Alert classification model was developed using a database of 150 healthy subjects with data from three 5-minute baseline conditions (i.e., finger-tapping eyes open (EO) and eyes closed (EC) and 3-choice psychomotor vigilance task (PVT)) and sleepy epochs selected from sleep-deprivation data. Five variables were computed for each one Hz bin between 5 and 40 Hz (5 variables x 36 bins) for each channel: the logged PSD, the relative power compared to the total power between 3 and 40 Hz, and the z-scores for each one-Hz bin compared to the means and standard deviations from the three baseline conditions. The recognition of fast blinks in the epoch was also used as a predictive variable. A total of 361 variables were available for each epoch based on a two-channel classification model (i.e., 180 variables each for FzPOz and CzPOz, and fast-blink).

The variables from each artifact free epoch for the three baseline conditions plus sleepy epochs from subjects with available data were submitted to step-wise analysis to select those variables most predictive in a four class model (PVT = high vigilance, EO = low vigilance, EC = high alpha, and Sleepy = high theta). A total of 22 variables were selected. The most predictive variables were: the z-score of the 10 Hz bin from CzPOz relative to the PVT task (partial $r^2 = 0.46$), the presence of a fast blink (partial $r^2 = 0.09$), and the z-score of 11 Hz from CzPOz relative to eyes closed (partial $r^2 = 0.06$). The approach developed for the B-Alert system was to utilize population data to establish the underlying model and then refine the DFA by adjusting for individual differences in the EEG using data from the three baseline conditions. Although baseline data could be readily acquired for development of the classification models for new individuals, sleep data could not be obtained a priori. Rather, the mean values of all variables from the three baseline conditions (180 variables x 2 channels x 3 conditions) for all subjects in the database were submitted to multiple simple linear regressions to derive equations to predict the DFA coefficients for the "sleepy or high theta" classification for the each of the 22 predictive variables. Matrices were then derived using the

above-mentioned analyses in order to fit the four class model to the individual and compute the probability of classification into each of the four output classes on a second-by-second basis off-line or in real-time.

To classify each one-second epoch of EEG, the DFA generates values representing the probabilities associated with each of the four classes. The final class assigned to each epoch is the one with the highest probability. One of the research goals was to provide a second-by-second measure of workload that could be displayed in real-time and ultimately used as input to a closed-loop intelligent system. Because previous work revealed that highly engaging and difficult tasks resulted in a majority of epochs classified as High Vigilance (HV), the probability values associated with HV and Low Vigilance (LV) were also investigated as potentially more informative for real-time assessment.

Preliminary Studies

Study 1: EEG was acquired from eight healthy subjects between 8:30 AM -1:30 PM. Continuous EEG (Fz, Cz, POz referenced to mastoids and FzPOz and CzPOz-differential) and EOG recordings were acquired with the sensor headset. Three five-minute baseline recordings were acquired during eyes open, eyes closed, and a 3-choice psychomotor vigilance test. Subjects were provided a minimum of 35 minutes of practice on the WCT and then completed three sessions with increasing difficulty (three waves of 6, 12 and 18 tracks). EEG B-Alert classifications (i.e., HV, LV, high alpha activity, and drowsy), reaction times (Time to IFF) and game scores (% of total possible points) were computed for each of the three difficulty levels. The mean probabilities associated with the HV and LV classes were also computed as investigational variables.

Results: Repeated measures ANOVA across the three WCT workload conditions revealed an increasing percentage of high vigilance classifications as a result of increasing workload with a significant main effect for workload ($F = 5.204$, $p = 0.02$) (Figure 1a). Comparisons between the easy, moderate and hard levels of the WCT revealed a significant difference between easy vs. hard ($p = 0.02$) and marginally significant moderate vs. hard ($p = 0.06$). Easy vs. moderate did not reach significance. The correlations between B-Alert HV classifications and the WCT reaction time.

Subj No.	RT %HV	%Score %HV	RT %Score
509	1.00	-0.99	-0.98
510	0.84	-0.94	-0.98
512	0.96	-0.97	-1.00
522	0.96	-1.00	-0.98
525	0.87	-0.68	-0.95
528	0.11	0.82	-0.48
529	0.70	-0.92	-0.93
530	0.60	-0.85	-0.93

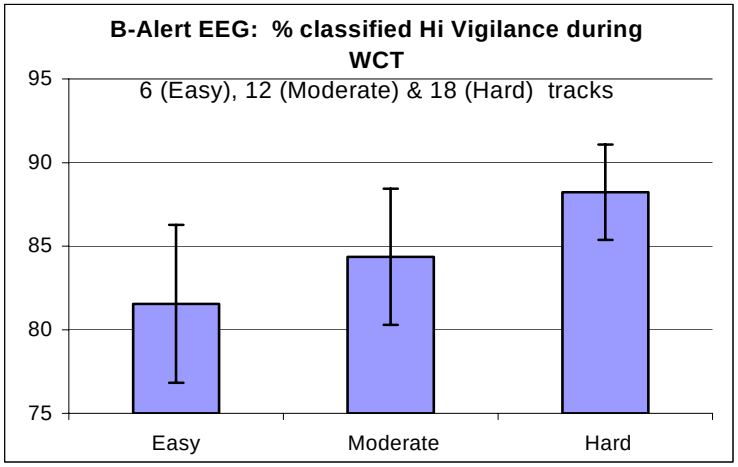


Figure 1. a. B-Alert % high vigilance during WCT 6, 12 and 18 tracks b. individual correlations WCT performance

(WCT-RT) and game scores (WCT-%score) for individual subjects are presented in Figure 1b. The correlations between HV and at least one of WCT performance measures was $\geq$ 0.85 in seven of the eight subjects. For subject 528, the correlations between HV and the WCT performance measures were poor.

Study 2: To evaluate the B-Alert workload indices without the sensory and motor confounds associated with workload levels in the WCT, a 3-level cognitive task developed by Klaus Mathiak at the University of Tubingen, Germany was evaluated in ten healthy subjects. In this task, both the stimuli and motor demands are kept constant during three levels of increasing task difficulty (easy, moderate, hard). For each of 250 trials, single integers between 1 and 8 are presented with a 1.6-second ISI. The order of the digit presentation is identical for each of the levels of difficulty to maintain consistent visual inputs. For level 1 (easy), subjects are instructed to press the space bar with the index finger of both hands only when they see the number 5. In level 2 (moderate), subjects respond only after any 3 consecutive even numbers and for level 3 (hard) they respond only to a number the <u>same</u> as the number <u>two</u> trials previous (2-back task).

Results: Repeated measures ANOVA across the three difficulty levels revealed significant increases in the reaction time ($F = 6.3$, $p < 0.02$) and decreases in the percentage of correct responses ($F = 21.6$, $p < 0.001$), confirming the actual increase in the levels of task difficulty. The mean percentage of B-Alert high vigilance classifications for each of the three difficulty levels is illustrated in Figure 2a. Repeated measures ANOVA across the three levels of difficulty revealed an increasing percentage of high vigilance classifications as a result of increasing difficulty with a significant main effect ($F = 24.4$, $p < 0.001$). Comparisons between the easy, moderate and hard levels revealed significant differences between easy vs. hard and moderate vs. hard ($p < 0.001$). Easy vs. moderate did not reach significance. The correlations between B-Alert HV classifications and the reaction time and percent correct responses are presented by subject in Table 1. The correlations between HV and at least one of performance measures was $\geq$ 0.85 in all ten subjects.

Table 1. Correlations between B-Alert HV classifications and the reaction time and percent correct responses

Subj.	B-Alert HV % correct	HV % errors	HV RT
521	- .93	.93	- .48
522	- .97	.99	.52
523	- .99	.99	.82
524	NA	NA	.90
525	- .91	.91	.93
526	- .92	.92	.99
528	- .85	.85	.79
530	*	.87	.99
531	- .94	.87	.89
532	*	.91	.52
NA no data available			
* no correlation - performed 100% at all 3 levels			

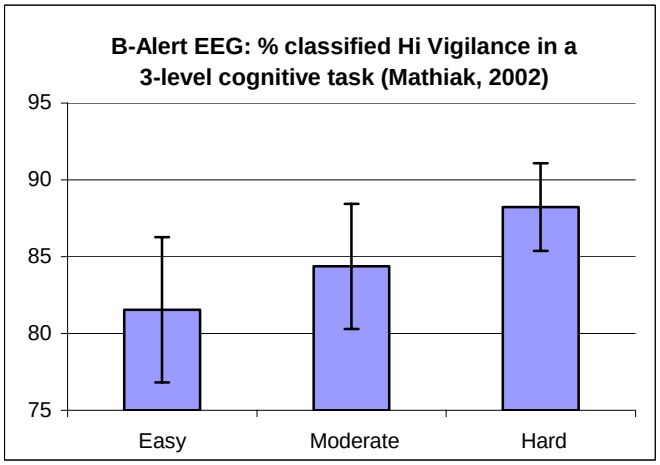


Figure 2. a. B-Alert % high vigilance in 3-level cognitive task
b. individual correlations with performance

The percent high vigilance increased with increasing level of task difficulty. Subjective reports on task difficulty and performance results confirmed the B-Alert data. The results suggest that the B-Alert indices are not increasing as a function of changes in sensory or motor demands in WCT.

Study 3: EEG and WCT data were acquired from four subjects with increasing levels of WCT training: 35 minutes, 5 hours, 10 hours training, and over 200 hours of training respectively. The subjects with 10 and 200 hours completed a fourth difficulty level (24 tracks). Figure 3 illustrates that

the relationship between the B-Alert percent high vigilance and WCT workload level (number tracks) remained consistent within subjects, however, the overall high vigilance percentages decreased dramatically as a result of training. This suggests that as subjects gain expertise, the level of vigilance is modulated to meet task demands.

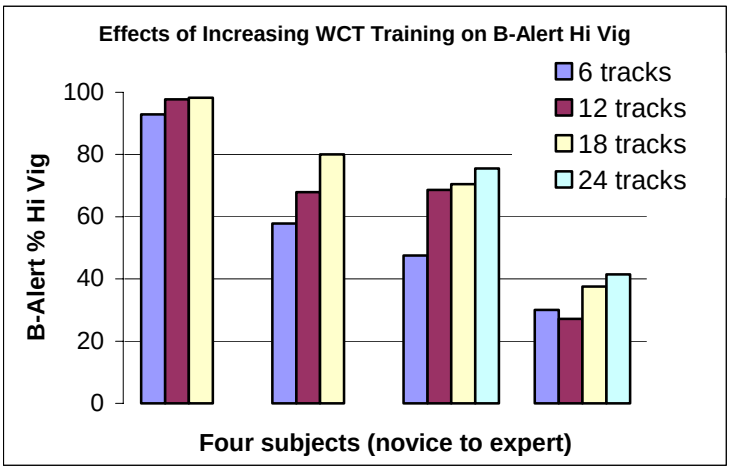


Figure 3. B-Alert % high vigilance in WCT 6, 12, 18, 24 tracks in 4 subjects with increasing levels of WCT training

Conclusions from three preliminary studies: The percentage of high vigilance, measured by the B-Alert EEG indices, is correlated with level of workload during WCT and Mathiak's cognitive task. The probabilities assigned to HV and LV for each epoch may be useful in providing more discrete measures of vigilance and workload on a second-by-second basis.

TIE Results

Nine subjects completed the TIE protocols (see TIE plan). The level of WCT expertise across subjects varied from one hour to over 300 hours of training. EEG was acquired and analyzed as described for the previous studies. The percentage of B-Alert high vigilance was submitted to repeated measures ANOVAs for each of the four scenarios. As predicted, main effects for number of tracks were significant for all scenarios (all $p < .005$). Inspection of the overall means suggested that there were no differences in B-Alert EEG for comparisons between low/ high or audio/no audio scenarios as a result of the high levels of variability across subjects.

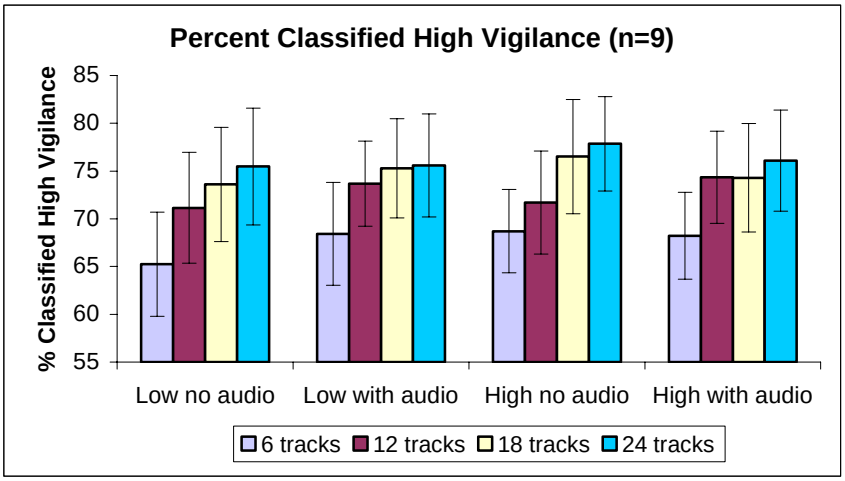


Figure 4. B-Alert % high vigilance for all scenarios of WCT

A comparison of the probabilities associated with HV and LV revealed that both within- and between-subject variability was lower for the LV probability values. The mean LV probabilities for each of the four scenarios and four workload levels (number of tracks) are illustrated in Figure 5. **It is important to note that the LV probabilities are inversely correlated with workload.** LV probability was also selected as the best variable to be included as the second-by-second values in the data summaries and for use in a preliminary investigation of the responsiveness of the B-Alert indices to the introduction of auditory messages and queries in the WCT.

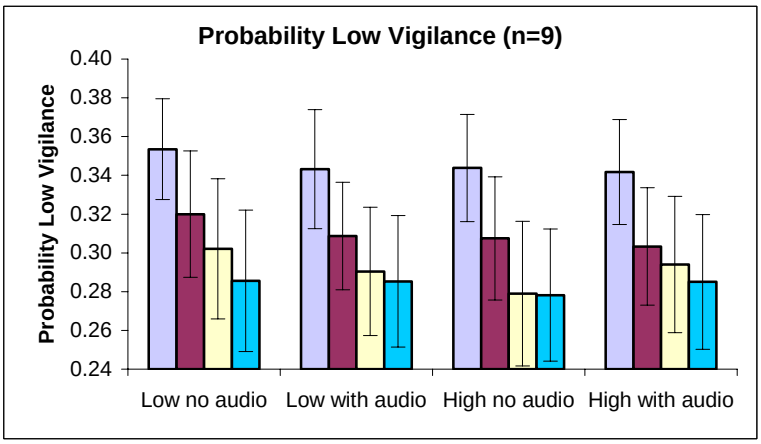


Figure 5. Probability Low Vigilance for all scenarios of WCT

Study 3 revealed that level of WCT expertise significantly influences B-Alert results. Differences in WCT training in the TIE subjects (1 hour to 300 hours) may account for some of the variability. To investigate this hypothesis, the ANOVA was computed again with a grouping variable for subjects divided into three groups based on amount of WCT training: ≤ 3 hours, 3 – 8 hours and 9 – 300

hours. Despite the small sample size, the results revealed a significant interaction ($F=2.51$, $p < 0.05$) between level of expertise and number of tracks, confirming the results reported in Study 3.

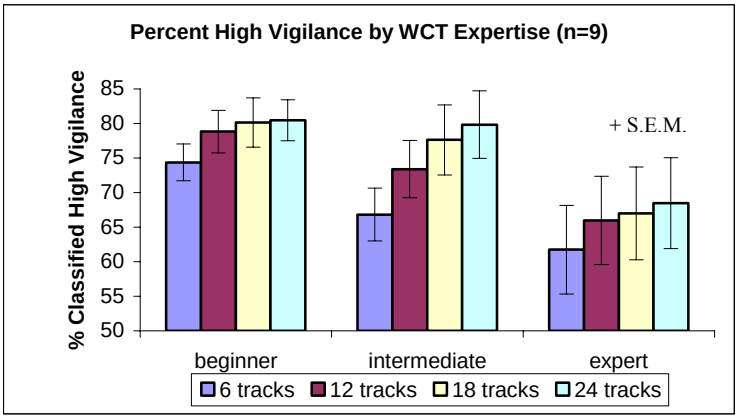


Figure 6. WCT training effects on B-Alert % high vigilance

The wave-by-wave correlations between B-Alert indices and WCT performance measures (Rtiff, game scores, etc.) were not as robust as those previously obtained. Individual results for each scenario were highly variable with correlations ranging from $r = 0.1 – 0.9$. One contributing factor may have been that the WCT generated an increasing amount of artifact (from eye movement and muscle activity) as a function of increasing number of tracks. Although the automated decontamination algorithms are designed to extract contaminants from the EEG, it was impossible to avoid some data loss due to contamination particularly during the 18- and 24-track tasks. This is a potential weakness of any EEG-based gauge and argues in favor of building redundancy into any suite of sensors.

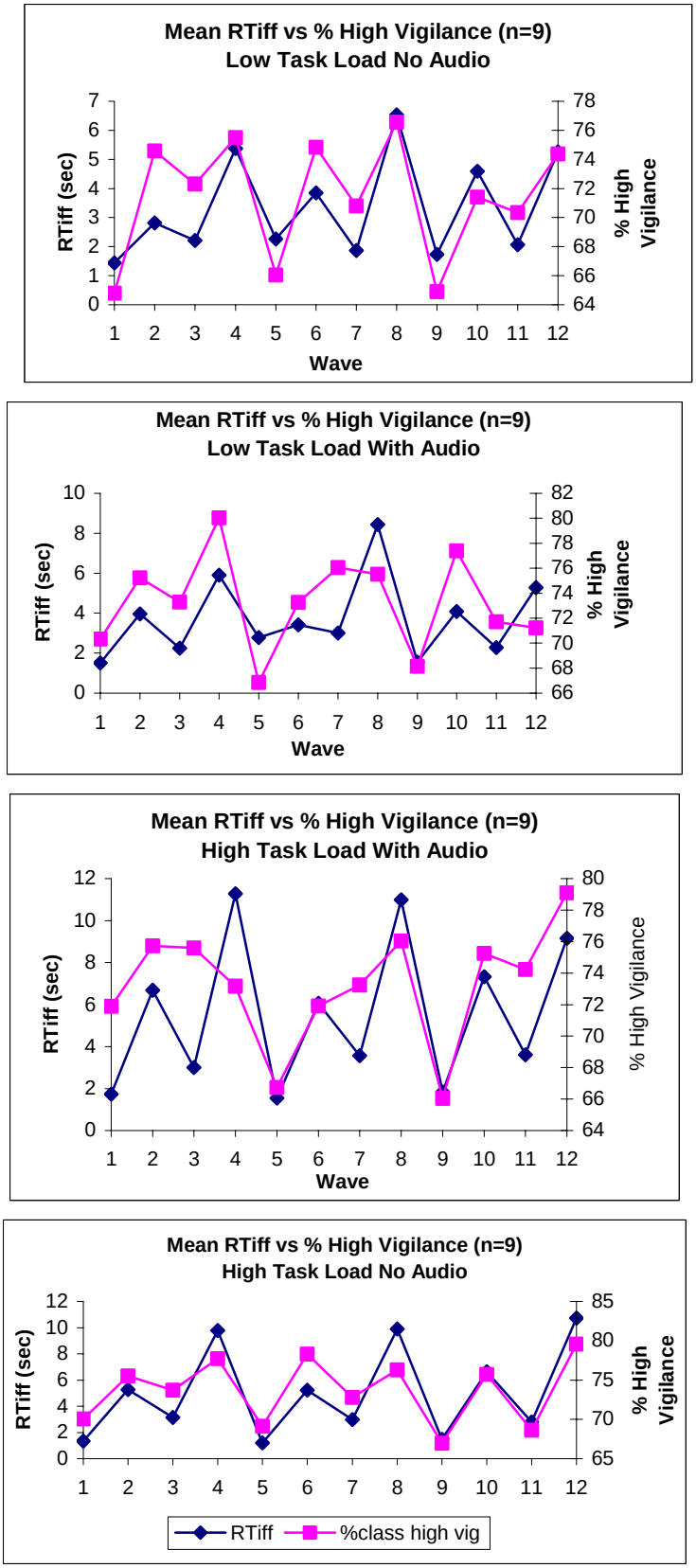


Figure 7. Rtiff vs. B-Alert % high vigilance for each of the WCT scenarios.

An investigational analysis was conducted to assess changes in the B-Alert indices in response to the auditory messages and requests. The probability of low vigilance was averaged for two seconds before and three seconds after each message and request. Mean data are presented in Figure 8, suggesting a shift from low to high vigilance for both messages and requests.

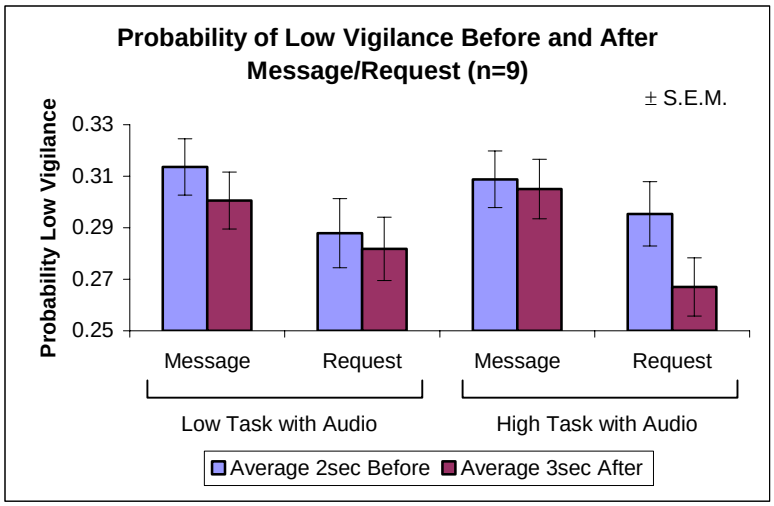


Figure 8. B-Alert LowVig Probability before & after messages & Requests.

Literature Cited

1. Levendowski, D.J., et al. Correlations between EEG Indices of Alertness Measures of Performance and Self-Reported States while Operating a Driving Simulator. In 29th Annual Meeting, Society for Neuroscience. 1999.
2. Levendowski, D.J., et al., Detection of Electroencephalographic Indices of Drowsiness in Real time using a Multi-Level Discriminant Function Analysis. Sleep, 2000. 23(Abstract Supplement #2): p. A243-A244.
3. Levendowski, D.J., et al., Electroencephalographic indices predict future vulnerability to fatigue induced by sleep deprivation. Sleep, 2001. 24(Abstract Supplement): p. A243-A244.
4. Mitler, M.M., et al., Validation of automated EEG quantification of alertness: methods for early identification of individuals most susceptible to sleep deprivation. Sleep, 2002. 25(Abstract Supplement): p. A147-A148.
5. Balkin, T.J. Sleep Deprivation Research at WRAIR. In DARPA Workshop. 2001. Las Vegas, NV.
6. Doran, S.M., H.P. Van Dongen, and D.F. Dinges, Sustained attention performance during sleep deprivation: evidence of state instability. Arch Ital Biol, 2001. 139(3): p. 253-67.
7. Pope, A.T., E.H. Bogart, and D.S. Bartolome, Biocybernetic system evaluates indices of operator engagement in automated task. Biol Psychol, 1995. 40(1-2): p. 187-95.
8. Kramer, A.F., Physiological metrics of mental workload: A review of recent progress, in Multiple task performance, D.L. Damos, Editor. 1991, Taylor & Francis: Washington, DC. p. 279-328.
9. Sterman, M.B. and et al., Concepts and applications of EEG analysis in aviation performance evaluation. Biol Psychol., 1995. 40(1-2): p. 115-30. Review.

10. Byrne, E.A. and R. Parasuraman, Psychophysiology and adaptive automation. Biol Psychol, 1996. 42(3): p. 249-68.
11. Prinzel, L.J., et al., A closed-loop system for examining psycho physiological measures for adaptive task allocation. Int J Aviat Psychol, 2000. 10(4): p. 393-410.

3B ANTHROTRONIX

AugCog TIE Report

Introduction

AnthroTronix, Inc. was contracted by Lockheed Martin, Advanced Technology Laboratories (ATL) to develop a suite of physiological sensors for monitoring stress/arousal using commercial off-the-shelf (COTS) sensors. The goal was to develop a gauge or set of gauges that combine data from multiple sensors into a single measure of stress/arousal at a time resolution of 2 seconds or less. Preliminary data was collected and analyzed during Phase I of the DARPA Augmented Cognition Program in order to quantify the results of the ATL interruption support system.

AnthroTronix' role in the March Technology Integration Experiment (TIE) was to address sensor integration issues and to identify functional requirements for Phase II of the Augmented Cognition program in order to further develop and validate a multiple sensor/gauge architecture for task manipulation.

We developed a gauge that would detect changes in arousal as indicated by changes in physiological parameters controlled by the Autonomic Nervous System (ANS). The ANS consists of two branches: the sympathetic and the parasympathetic. There is a known relationship between sympathetic nervous system activity and emotional arousal, although one cannot identify the specific emotion being elicited. The autonomic nervous system consists of sensory neurons and motor neurons that run between the central nervous system and various internal organs such as the heart, lungs, viscera, and glands. The contraction of both smooth muscle and cardiac muscle is controlled by motor neurons of the autonomic nervous system. Exocrine glands, glands whose secretions pass into a system of ducts that lead ultimately to the exterior of the body, such as the sweat glands, are also controlled by the ANS. The ANS is responsible for monitoring conditions in the internal environment and bringing about appropriate changes in those conditions.

Preliminary data collection was conducted using a variety of COTS physiological sensors. After assessing the various sensors we concluded that heart rate, respiratory rate, and skin conductance are the parameters most consistently and measurably affected by changes in task events and workload. (Also initially examined were skin temperature, blood volume, and respiratory volume. However, respiratory volume did not change at a sufficiently noticeable rate, blood volume provided information that was similar to, but less reliable than, heart rate data, and the response time for skin temperature was too slow to be effective in detecting arousal or driving a task.)

Cardiac activity is most commonly monitored using an electrocardiogram (EKG), which uses an array of three, or sometimes twelve, electrodes to detect the small electrical signal produces by the heart muscle each time it contracts. This signal produces a waveform, which indicates contracting and relaxing of the atria and ventricles. From this signal heart rate can be easily calculated.

Respiratory rate can be assessed in several ways. The method that is least obtrusive uses a simple strain gauge, secured around the subject's chest to detect expansion of the chest cavity. In order to account for abdominal, as well as thoracic expansion during inhalation, a second strain gauge is secured around the subject's abdomen.

Sweat gland activity is a relatively easy parameter to measure and can be assessed non-invasively using Galvanic Skin Response (GSR) sensors. GSR sensors typically consist of two electrodes placed at a slight distance from each other on the skin's surface, most often on the palm of the hand or on two fingers. A tiny electrical voltage is applied through the two electrodes, in order to establish an electric circuit in which the subject becomes a variable resistor. The real-time variation in conductance, which is the inverse of the resistance, is calculated. Standard measurement units for

skin conductance include micro-Siemens and micro-mhos, which are equivalent. For consistency, we will report all skin conductance data in micro-mhos.

Our combined sensor gauge integrated heart rate, respiratory rate, and skin conductance data, averaged over each wave, as well as averaged over each second, as presented at the TIE in San Diego. The premise behind this multi-sensor gauge was that changes in arousal would be evidenced by various changes in each of the physiological parameters being measured by the various sensors. Significant changes in cardiac, respiratory, and skin conductance data occurring simultaneously would be considered changes in arousal.

Figure 1, below, shows respiratory, cardiac, and skin conductance data, respectively from top to bottom, for a single 15 minute session. We hypothesized that by equally weighing each of these parameters and combining them into a single arousal gauge, it would be possible to identify changes in a subject's overall arousal in response to a given task.

These changes would be examined in relation to task performance so as to identify patterns of physiological activity occurring at times when task performance was high in comparison to patterns occurring at times of decreased task performance. It was hoped that consistent patterns of physiological activity would be identified in relation to optimal task performance, as well as situations in which the subject became cognitively overloaded, resulting in decreased task performance. By identifying these patterns, it would be possible to assess a subject's cognitive workload level, and to predict cognitive overload. Preliminary data collection and analysis conducted by AnthroTronix, using the ATL interruption support system, further validated this hypothesis.

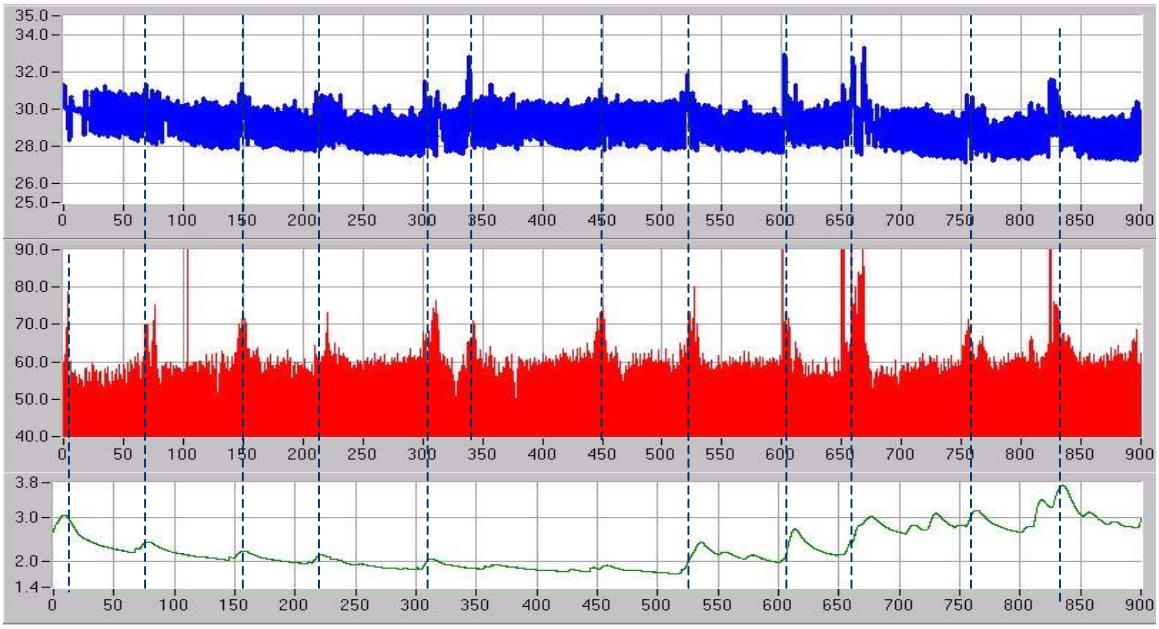


Figure 1: Respiratory, Cardiac, and Skin Conductance data (top to bottom respectively) for a Single 15 Minute Session

For the March TIE, we were asked to report findings correlating the average of our gauge data over 75 second intervals, to metrics of the Warship Commander Task (WCT) such as wave size, average RTiff, and number of errors. We did not observe a relevant correlation between the two; however, this was not surprising due to the fact that the natural rhythm of cardiac and respiratory activity is highly variable, and that the ANS works by constantly changing physiological parameters in response to various environmental stimuli.

We then examined the second by second data to look at event related effects, and discovered that for the WCT, heart rate and respiratory rate only acted as a scaling factor, and therefore changes in the gauge were being indicated primarily by changes in GSR. This does not invalidate our combined sensor gauge; it does, however, indicate that a simpler gauge may be able to correlate with events in the task. In order to more closely examine the relevant data, that which impacted upon the gauge in response to the task, we developed a secondary gauge, based entirely on changes in GSR.

While changes in overall GSR level are indicative of general arousal levels, we are particularly interested in sudden peaks in GSR because these abrupt changes can often be correlated to specific events. Fear, anger, startle response, orienting response and sexual feelings are all among the emotions that may produce similar GSR responses. Our hypothesis is that peaks in a subject's GSR will occur at times when the subject is experiencing a great deal of anxiety or stress. As anxiety increases, GSR will continue to increase at a significant rate until the subject becomes overloaded, at which point we expect to see a sudden drop in GSR (defined as a decrease of .01 micro-mhos or more within 2000 milliseconds or less), indicating that the subject has "given up". By identifying a pattern of physiological activity occurring at the time of cognitive overload, we hope to be able to predict, and therefore prevent, occurrences of cognitive overload.

In order to identify these peaks in arousal, we've developed an algorithm, which uses the first derivative of the raw GSR data to identify the rates at which GSR increases and decreases over time. We hypothesize that the onset of significant increases and decreases in GSR will correlate to specific task events driving these changes. Based on the data collected at the TIE, we have identified significant changes in GSR as having a delta of 0.01 micro-mhos or more over a time interval of up to 2000 milliseconds. While the GSR sensor samples at a resolution of 32 Hz, changes in GSR occurring as gradually as .01 micro-mhos per 2000 milliseconds are considered significant. For this reason, the temporal resolution of this gauge ranges from 31.25 to 2000 milliseconds.

In addition to identifying patterns of physiological activity related to cognitive overload, we also hope to identify trends among the task events triggering cognitive overload. We can then draw conclusions as to which aspects of the task are most taxing, and can use this information to drive the task, preventing stress and cognitive overload.

Approach/Method

At the March TIE in San Diego cardiac, respiratory and skin conductance data were collected from 6 subjects for a total of 19 sessions using the Warship Commander Task. Cardiac activity was monitored using a 225Hz, three-electrode electrocardiogram (EKG) sensor, applied to three sites on the subject's chest. Respiratory data was collected using two strain gauges wrapped around the chest and abdomen, respectively. Skin conductance data was collected using a dual electrode GSR sensor applied to the surface of the skin on the underside of the 2nd and 4th toes on the subject's left foot. Both the respiratory and GSR sensors sampled at a rate of 32Hz. All sensor data was saved using only the team number, participant number, and WCT scenario code.

The data collected from one subject during a single session was processed and analyzed using our combined sensor gauge. This gauge data, which combined heart rate, respiratory rate, and GSR, was averaged over each wave, as well as over each second, and was presented at the conclusion of the TIE on March 6th.

Following the TIE, further data processing and analysis was conducted at our lab in College Park, MD. A second gauge was developed, which uses the calculation of the first derivative of the raw GSR data to determine the rate of change of GSR over the course of each scenario, and to correlate peaks in the GSR data to specific task events.

Results

As predicted, there was a significant correlation between the wave size and task performance metrics, such as the average RTiff and number of errors per wave. As expected, we did not observe a correlation between wave size and our gauge averaged over each scenario (See Figure 2). It was clear to us, as it was to the other developers, that significant physiological changes would be found in the second by second data (See Figure 3). In particular, we expected to find a correlation between sudden changes in arousal and significant task events, such as the onset of a new wave, engaging of a track, and auditory task alerts.

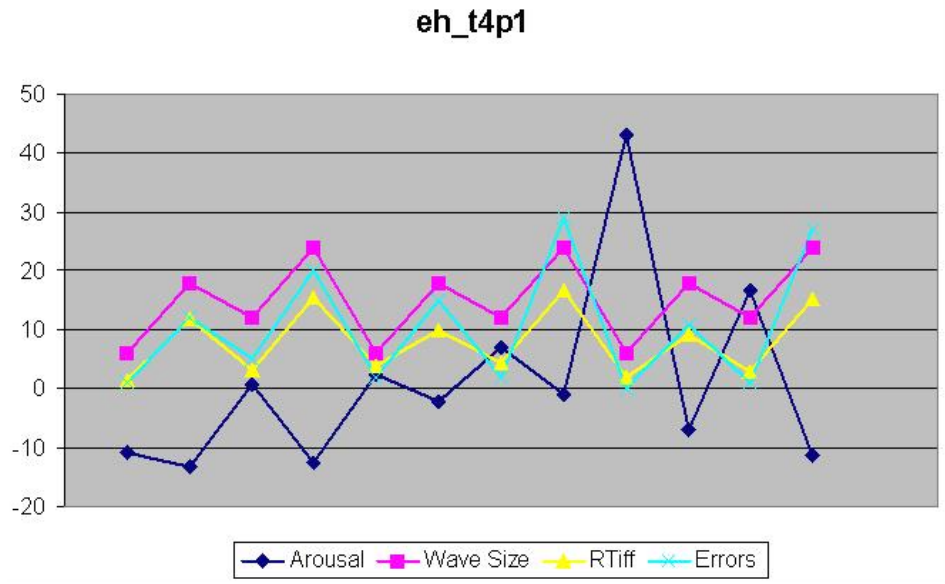


Figure 2: Average Arousal, Wave Size, RTiff, and Errors for Each Wave

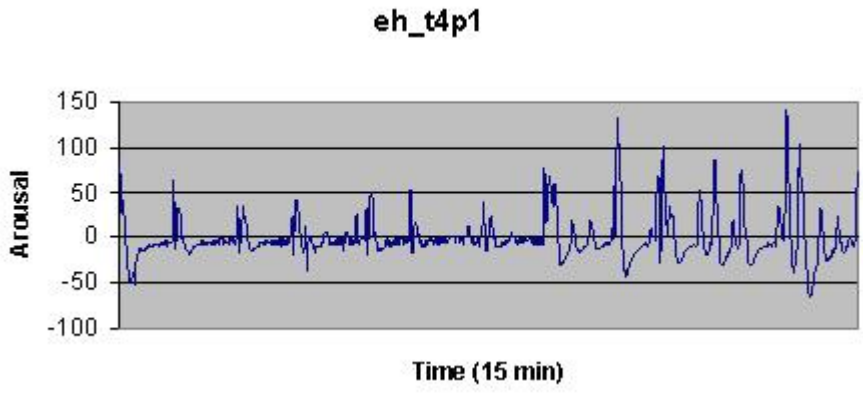


Figure 3: Second by Second Arousal Data for a Single Session Collected at the TIE

Discussion

As discussed in the introduction, our combined sensor gauge did not bear a significant correlation to event data (wave size, average RTiff, errors, etc.) when averaging the data over each wave. However, correlations were identified between peaks in the second by second gauge data and specific task events, as they occurred. When examining each of the gauge parameters separately, it became evident that the heart rate and respiratory data had little affect on the gauge, and that the peaks in the gauge data were being driven primarily by changes in GSR. This does not invalidate our combined sensor gauge; we feel that perhaps this particular task was not stressful enough as to induce significant changes in cardiac and respiratory activity, as detected by our sensors. It is likely that a more stressful task and higher resolution data acquisition would produce different results. For example, subjects operating the WCT were aware of the order of wave size, and therefore were able to prepare themselves for the larger waves and relax prior to the smaller waves. Subjects operating a task in which the level of difficulty was randomized might experience more anticipation and anxiety, as well as experiencing surprise, all of which impact arousal. We expect to further refine and employ our combined sensor gauge during Phase II research, and will continue testing this gauge on various task models.

For analysis of the TIE data we elected to develop a second gauge, based on the rate of change of GSR over the course of the session, and to correlate peaks in the gauge data to specific task events. We identified peaks in the subject's arousal as an increase of 0.01 micro-mhos or more within a 2 second time interval. Peaks were identified throughout the sessions and correlated to task events.

Common task events correlating to increased arousal included the onset of a new wave (AWAV), particularly waves of 18 or more tracks, engaging of a track (T##E), auditory messages (AM__), and auditory requests (AR__). Figure 4 identifies specific task events, which are immediately followed by peaks in arousal. Errors of commission were accompanied by auditory alerts, which conveyed to the user that a mistake had been made. Although, this elicited a slight response from the subjects, as shown in Figure 4 (AQCK), we feel that deduction of points and other such consequences would further increase this effect.

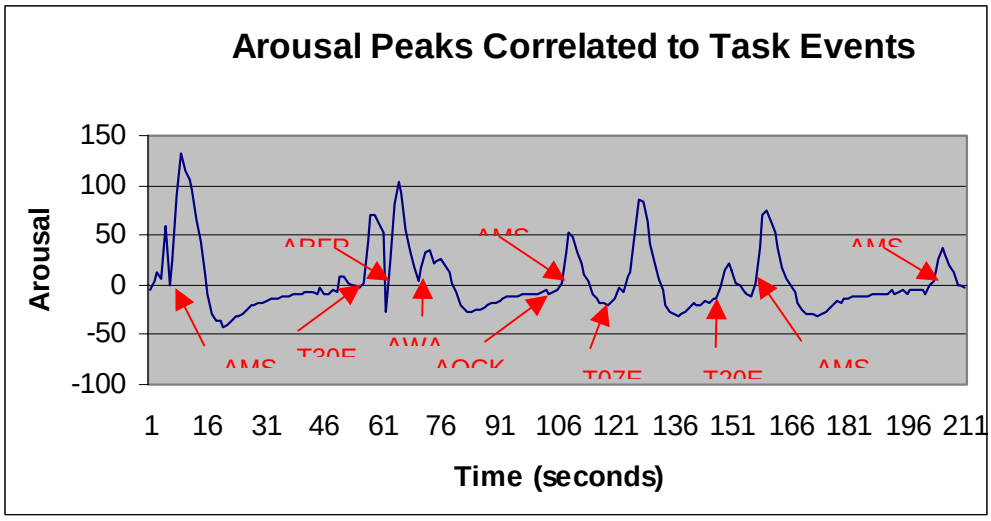


Figure 4: Arousal Peaks Correlated to Task Events

Having identified patterns of task events that elicit increased arousal allows us to predict cognitive overload. A common pattern which we observed displayed rapidly or progressively increasing arousal elicited by numerous task events, resulting in decreased task performance, followed by a sudden decrease in arousal despite continued decreasing task performance. This is characteristic of a subject "giving up". Figure 5 illustrates just such an example. In this case arousal begins to increase very quickly, corresponding to the onset of a new wave (AWAV). However, after a few seconds, arousal decreases continuously despite repeated errors and auditory alerts (AMSW, AMSR, ARFW).

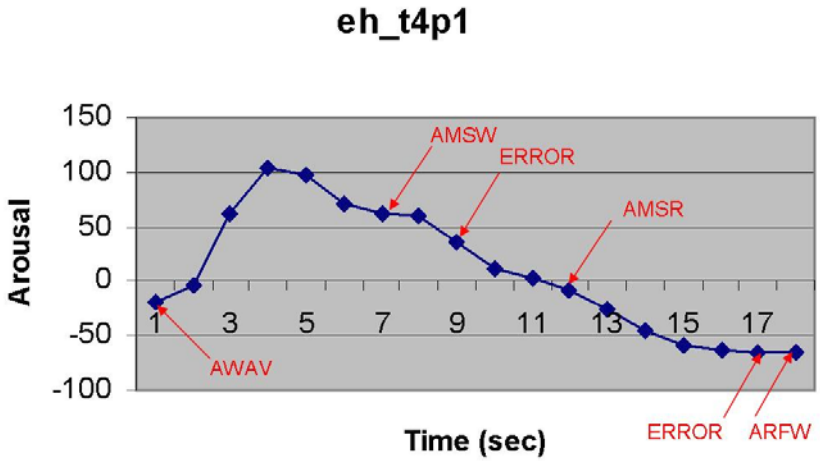


Figure 5: Example of a Subject "Giving Up"

References

[1] http://www.bio-medical.com/Gsr.html

[2] http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/P/PNS.html

[3] http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/E/ExocrineGlands.html

[4] http://www.bio-medical.com/Gsr.html

3C CLEMSON UNIVERSITY

A Summary of the Clemson TIE Data Collection Effort as Part of Team 1 (prepared by Dr. Eric R. Muth on 4/8/03)

On March 3-5, three Augmented Cognition research efforts at Clemson University, University of Pittsburgh and Geodesics were combined in a joint experiment at Pacific Science and Engineering (PSE), La Jolla, CA. Dr. Eric Muth brought his "Arousal Meter (AM)", Dr. Carey Ballaban brought his posture system and Dr. Don Tucker brought his EEG system and physiological data were collected simultaneously while participants completed the Warship Commander Task.

Method

Subjects

Nine participants were recruited by PSE. Two participants were run on Monday (T1P8 and T1P5), four on Tuesday (T1P1, T1P6, T1P2, T1P3) and three on Wednesday (T1P4, T1P7, T1P9). For the AM, data were usable for 7/9 participants. On Monday, the Clemson team was not at the TIE, but Roy Stripling was trained via the phone and used an on-line, desktop version of the AM to collect data for us, demonstrating the simplicity and ease of use of the AM. He successfully collected data from only 1/2 subjects, T1P5. Subject T1P8 was not useable. Both the EGI and Pittsburgh teams collected data on Monday. On Tuesday, the Clemson team was on site and successfully collected data from 3/4 subjects using an off-line, wearable version of the AM. Subject T1P3 was not useable. Both the EGI and Pittsburgh teams collected data on Tues. On Wednesday Clemson successfully collected data from 3/3 subjects. Only the Pittsburgh and Clemson teams collected data on Wednesday.

Procedure

When each subject arrived at the lab, he or she was first briefed about the nature of the study and signed multiple consent forms. The subject was then fitted with the necessary apparatus to detect physiological measures: heart rate, posture (head position and body position) and EEG. After a brief calibration procedure, the subject then completed the Warship Commander as specified in Table 1. One notable exception to Table 1 is participant T1P9. This was an additional participant that was run outside of the required 8. The participant was run through the same Warship Commander scenarios as participant T1P5. Immediately following each wave of the Warship Commander task, the participants also completed a brief subjective measure of workload. Those data are not presented here as they will be presented by another team member.

Table 1: Warship Commander Experimental Conditions

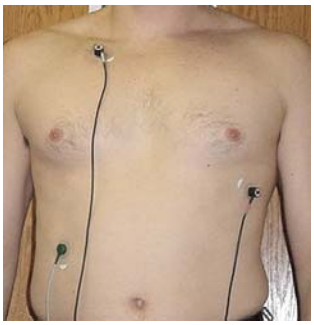
Participant	Team 1
1	K, K+, L+, L
2	K+, L, K, L+
3	F, F+, E+, E
4	F+, E, F, E+
5	G, G+, H+, H
6	G+, H, G, H+
7	J, J+, I+, I
8	J+, I, J, I+

Notes:

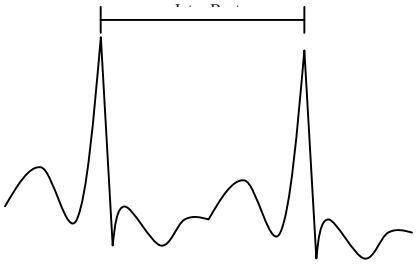
1. E, G, I, K – High Task Load (equivalent to C6)
2. F, H, J, L – Low Task Load (equivalent to D6)
3. + denotes the auditory task is run concurrently (on high)
4. The scenario column lists the scenario radio buttons to be selected for the trial.

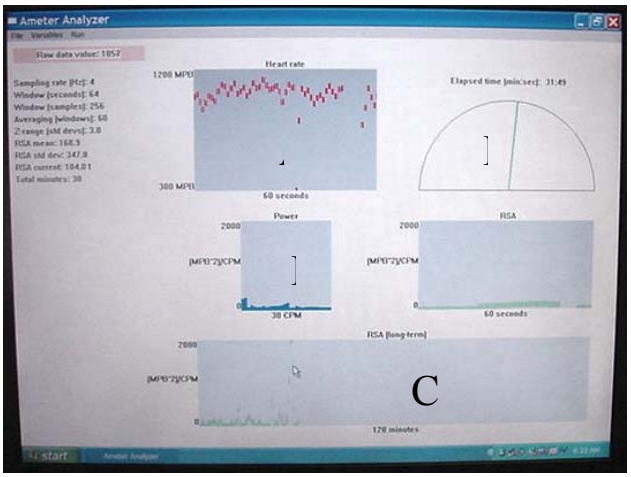
AM Technical Information

The Arousal Meter (AM) consists of two components, a heart rate sensor and analysis software. The heart rate sensor (shown at right) is manufactured by UFI (Morro Bay, CA) and marketed as the EZ-IBI unit. The person to be monitored is connected to the unit via 3 electrode leads as seen at left. Two active recording leads (black) are connected, one on the person's right side just below the collar bone and one on the left side just below the left breast. These two electrodes are connected to Fetrodes (field effect transistors) that serve as amplifiers and increase the signal to noise ratio. These leads are positioned to minimize electrode movement and be in line with the major vector of depolarization of the heart. The third lead (green with grey) serves as a reference for signal noise reduction.



The EZ-IBI unit is powered by a 9-volt battery and serves as an amplifier, filter and initial signal processor. The EZ-IBI unit derives inter-beat-intervals (IBIs) by monitoring the raw electrocardiogram (ECG) and recording the time interval between successive R-spikes of the cardiac QRS interval (as shown at right). IBIs typically range between 500 and 1200 msec. When IBIs are plotted against time, a waveform is evident in which three periodic fluctuations can be observed (low, mid and high frequency). When respiration is accounted for, the high frequency fluctuation has been related to the changing influence of parasympathetic nervous system (PNS) activity on the heart. PNS activity is inversely related to autonomic arousal.





The analysis software (shown at left), reads (via a serial port) and analyzes the IBI data from the EZ-IBI unit and is currently run on a laptop PC. IBIs are plotted over time (A) and processed using the Fast-Fourier-Transform (FFT—The mathematical equivalent of passing light through a prism and breaking it into its component colors/frequencies). FFT derived power is plotted across frequencies (B) to determine the high frequency (HF) peak associated with PNS activity (between 9 and 30 cycles per minute). This HF peak is then plotted over time in both long duration (C1) and short duration (C2) windows. The mean and standard deviation of the HF peak are continually re-calculated. A standardized "arousal" score is derived [-(x–μ/σ)] that drives the AM (D). In the case of the TIE data, a mean and standard deviation were calculated for the entire participant's recording. This mean and standard deviation were used to standardize all of the second by second arousal scores during the entire time data were recorded for a given participant. For each participant, average arousal scores were calculated for each wave of the Warship Commander task as well as the breaks that occurred between each of the Warship Commander scenarios. Increases in this score are associated with increased autonomic arousal and decreases with decreased autonomic arousal. A state shift has been operationally defined as a score that changes from negative to positive. The AM has approximately a 1 sec resolution but performs analyses 4 times a second for redundancy.

Results

Wave Size

The primary analysis was a by wave size analysis, the data from the 12 waves that were presented were reduced down to the 4 different wave sizes (6, 18, 12 and 24). Hence, an average arousal by wave size was calculated for each participant before the analyses were conducted.

Figure 1 shows the average arousal by wave size in the Warship Commander task with standard deviation plotted as the error bars. The x-axis shows wave size and the y-axis shows arousal score. Note that the x-axis is not ordered by size in the graph (small to large). Rather it is arranged by the order in which the wave sizes were repeatedly presented.

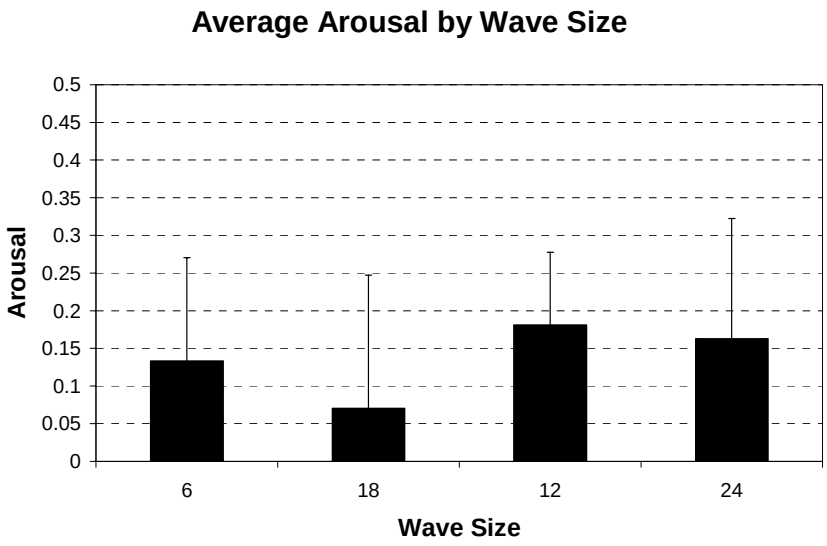


Figure 1. Average arousal by wave size.

The first thing that is evident from the graph is that during all of the waves, arousal was in the positive direction, indicating that the participants, on average were aroused during the Warship Commander task. The second thing that is evident is that the participants were not very aroused. The arousal score is in standard deviation units and the maximum average arousal score seen for a given wave size was less than 0.2. This is quite low. Further, given the standard deviations, it is probably not significant from 0. A repeated-measures analysis of variance (Anova) was run comparing arousal over the 4 wave sizes and no differences were found $F\ [1/6]=0.735$, $p > 0.05$.

Wave Number within a Scenario

A secondary analysis was conducted to examine changes over time within scenario. For this analysis, the 12 waves from each of the 4 scenarios were averaged together by wave number.

Figure 2 shows the change in arousal score by wave number. There were no differences in arousal by wave number $F\ [11/66]=1.21$, $p>0.05$. However, an interesting trend appears in Figure 2 showing that participants appeared to have the lowest arousal at the beginning and end of each scenario and the highest arousal in the middle of each scenario.

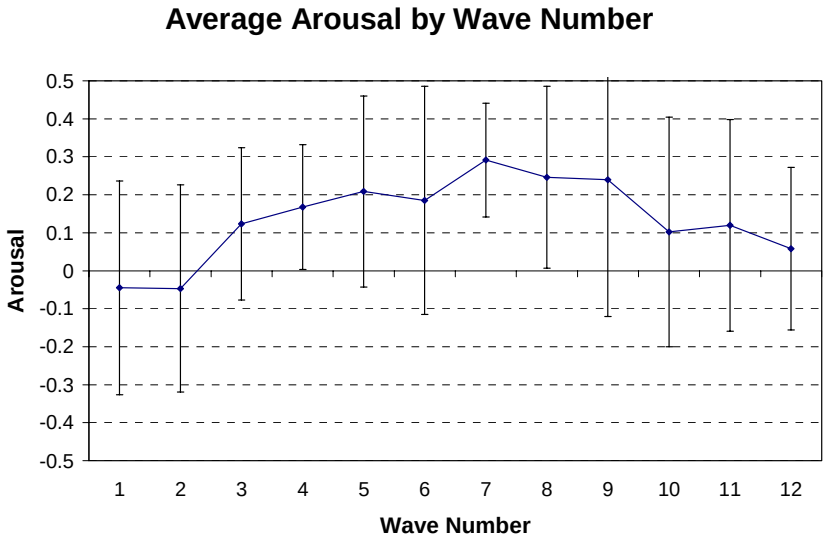


Figure 2. Average arousal by wave number.

Arousal over Time

A secondary analysis was also performed that examined the fluctuation in arousal score across the entire experimental session. For this analysis data were reduced to average arousal score for each break period between scenarios and average arousal score by wave. Arousal score was examined from the beginning of recording (B0) through the end of the last scenario (waves 37-48).

Figure 3 shows the arousal score over time. Means plus/minus 1 standard deviation error bars are shown. The x-axis is labeled with the 4 break periods (B0, 1, 2 and 3). The 12 points following each break period represent the 12 waves that occurred in each of the four scenarios. An analysis of these data reveled a significant change over time $F\ [51/255] = 1.52$, $p<0.05$. It is important to note that for the repeated measures ANOVA only 6 participants could be used because participant T1P5 had missing data for waves 45-48. Post-hoc follow-up tests revealed that the significant F was primarily due to three things: 1) a difference between the break periods and task periods; 2) a difference between the second scenario and the other three scenarios, with the second scenario generating higher arousal; and 3) a within scenario fluctuation in which arousal started off low, reached a peak mid-scenario and trailed off at the end of the scenario (as seen in Figure 2 above).

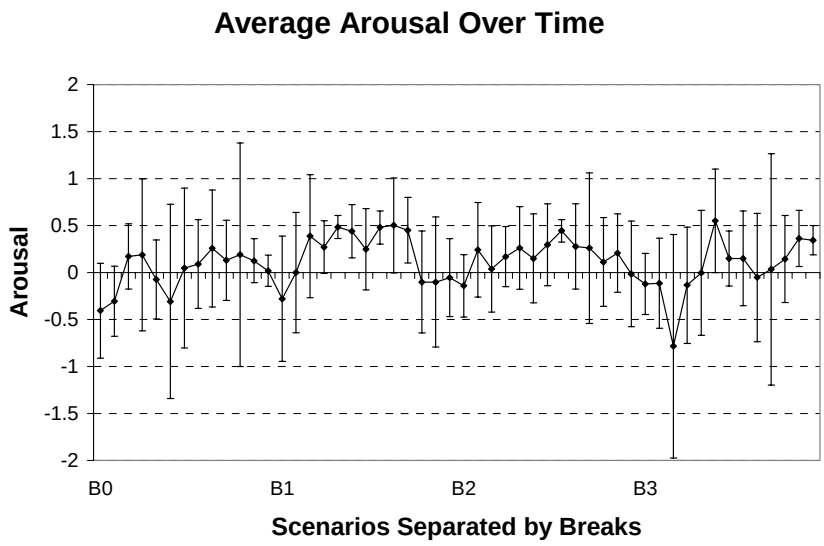


Figure 3. Average arousal over time.

Arousal by Break and Scenario

In order to illustrate the differences between break periods and task periods, revealed in the "arousal over time" analysis, a follow-up analysis was conducted looking at the average arousal score for each break vs. the average arousal score for each of the four scenarios. Figure 4 displays the arousal score over time averaged for each break and each scenario mean plus/minus 1 standard deviation. The x-axis is labeled with B0, B1, B2 and B3 for the 4 breaks and S1, S2, S3 and S4 for the four scenarios. A repeated-measures ANOVA revealed only a marginally significant difference $F\ [7/42] = 2.12$, $p < 0.10$ ($p=0.064$). Post-hoc follow-up tests were conducted with caution based on the significant results in the "arousal over time" analysis. The follow-up tests revealed that the only significant differences were between the break periods and the task periods.

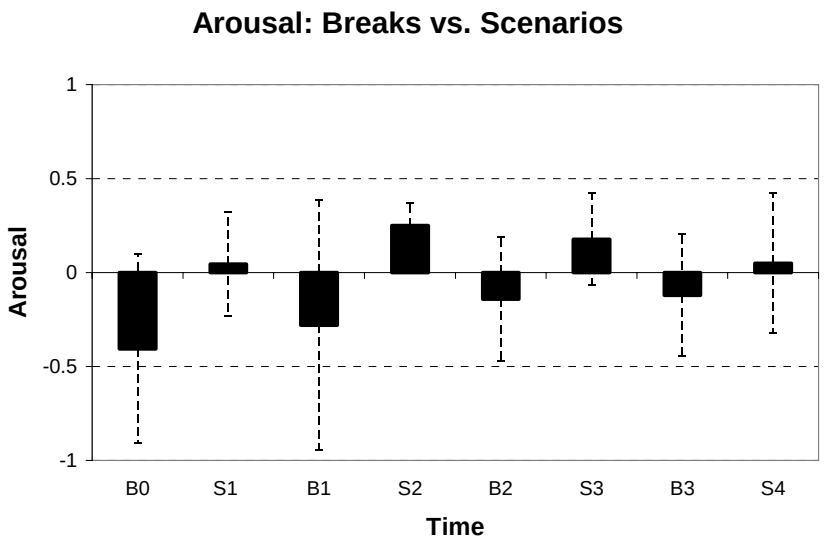


Figure 4. Arousal: Breaks vs. Scenarios.

Discussion

The results indicate that autonomic arousal was higher during task performance than during breaks before and after each task. Arousal did not change significantly by wave or within a given scenario. However, two interesting trends were noticed. First, a trend appeared indicating that during a scenario, arousal started low, peaked during the middle of the scenario and ended low. Second, for some reason, the second scenario seemed to result in higher arousal than the other three scenarios. These two results are trends and must be considered as such. The population run through the task consisted of experts at the task. It would be interesting to see if the within and between scenario trends are confirmed in a more novice population and to examine how these trends vary as an individual learns the task. The within scenario trend, if in the future proves to be significant, is easily interpreted as an effect of time where the participant (particularly a well trained participant who knows what to expect) starts out relaxed, peaks during the middle of the task when workload is high and he/she is concentrating, and then begins to relax toward the end of the scenario. Unfortunately, because each subject completed slightly different scenarios, it is difficult to say why scenario two might result in more arousal. However, this might also be an effect of time.

Conclusion

The arousal meter revealed significant changes over long time periods (break vs. task). However, the arousal meter failed to reveal short duration differences within tasks. This is likely due to the expert population and the minimum amount of arousal variability seen during task performance. The arousal meter currently has second by second accuracy. The current data shows that the arousal meter is effective at showing meaningful changes that occur over a matter of minutes. Whether or not second by second data are actually meaningful remains to be seen.

3D DREXEL UNIVERSITY

Functional Near Infrared (fNIR) Sensor Measurements in Warship Commander Task

AUGCOG - TIE

Drexel University - fNIR Group

School of Biomedical Engineering, Science and Health Systems, Philadelphia PA.

Summary

Functional near infrared (fNIR) sensor measures hemodynamic changes in the cortex using a portable, safe, affordable and negligibly intrusive optical imaging system. We hypothesize that there is a positive correlation between blood oxygenation in the relevant areas of dorsolateral prefrontal cortex and cognitive effort defined as attention and working memory. In this report, we present the deployment and real-time statistical analysis of fNIR in assessing the cognitive state of the user. This report is based on data collected during the DARPA Augmented Cognition - Technical Integration Experiment session. The experimental protocol for this session used a complex task called the Warship Commander Task (WCT). In the WCT, the primary task is air warfare management. While performing the air warfare task, task difficulty and task load are manipulated by changing the number of tracks per wave, the number of yellow tracks, and the presence or absence of an auditory memory task. The fNIR data analysis explored; 1) the correlations among cognitive workload, the participant's performance, and changes in blood oxygenation levels of the dorsolateral prefrontal cortex, and 2) the effect of divided attention as manipulated by the secondary component of the WCT (the auditory task).

The fNIR data is grouped into two measurement classes to observe and compare the spatial changes in the prefrontal cortex. Each group is dealt with as an output of the following sub-gauges: i) fNIR (Left): the measurements acquired from the left forehead, and ii) fNIR (Right): measurements from the right forehead. Eight participants performed a series of four WCT scenarios. Changes in blood oxygenation levels for left and right hemispheres were submitted to repeated-measures ANOVAs for each of the four scenarios. The results, acquired in the context of the Warship Commander Task, indicated that the fNIR gauge output was significantly sensitive to wave size changes. They also suggested that there is a positive correlation between a participant's performance and oxygenation responses in dorsolateral prefrontal cortex as a function of task load.

The main effects for complexity and for the divided attention task, despite increasing overall cognitive effort, were not associated with significant changes in blood oxygenation. There are a number of potential reasons for these findings. First, the current analyses focused on only two fNIR parameters, average change in oxygenation and rate of change of oxygenation. It is possible that other parameters could add predictive power in these complex cognitive tasks. We are currently working on developing a parametric model for the oxygenation pulse to extract additional features such as peak amplitude, pulse width, latency, etc. Hence, further development in the algorithm, fine-tuning and increasing the number of features, are expected to enhance consistency and efficacy of the gauge. Second, the current sensor was applied over a limited area of the frontal poll and dorsolateral prefrontal cortex. Some of these manipulations may have had effects in areas of the cortex that are accessible to fNIR, but were not measured with the current sensor. This question remains for future generations of sensor to determine. Finally, the WCT itself is complex, and numerous cognitive and emotional functions are occurring during the execution of the task. It is possible that these various

tasks have differential effects on the hemodynamic response. For example, recent research using PET indicates that various areas of cortex show increases in oxygenation during a divided attention task relative to a full attention task, whereas other areas demonstrate decreases in oxygenation during the same task [5]. Further work is needed to more fully explicate our understanding of brain function during what may be common everyday, and yet extremely complex tasks.

1. Introduction

Near infrared (NIR) spectroscopy (NIRS) enables the measurement of changes in the concentration of deoxygenated hemoglobin (deoxy-Hb) and oxygenated hemoglobin (oxy-Hb) noninvasively during functional brain activation in human [1-3]. The technology allows the design of portable, safe, affordable, non-invasive and negligibly intrusive monitoring systems. This makes NIRS suitable for the study of cognition related hemodynamic changes under many working conditions and in the field.

Biological tissues are relatively transparent to light in the near infrared range between 700 to 900 nm. In this "optical window" the two primary absorbers are oxy-Hb and deoxy-Hb, which are two biologically relevant markers for brain activity monitoring [6]. The wavelengths in this "optical window" pass easily through tissue and their absorption provides information about brain functions including motor and visual activation, auditory stimulation and performance of cognitive tasks [3,4].

Typically, an optical apparatus consists of a light source by which the tissue is radiated and a light detector that receives light after it has interacted with the tissue. Photons that enter tissue undergo two different types of interaction, namely absorption and scattering. According to the modified Beer-Lambert Law [6], the light intensity after absorption and scattering of the biological tissue is expressed by the equation:

$$I=GI_o e^{-(\alpha_{HB} C_{HB} + \alpha_{HBO2} C_{HBO2})*L} \qquad (1)$$

where G is a factor that accounts for the measurement geometry and is assumed constant when concentration changes. I_o is input light intensity, α_{HB} and α_{HBO2} are the molar extinction coefficients of deoxy-Hb and HbO_2, C_{HB} and C_{HBO2} are the concentrations of chromophores, deoxy-Hb and HbO_2, respectively and L is the photon path which is a function of absorption and scattering coefficients μ_a and μ_b.

By measuring optical density (OD) changes at two wavelengths, the relative change of oxy- and deoxy-hemoglobin versus time can be obtained. If the intensity measurement at an initial time is I_b (baseline), and at another time is I, the OD change due to variation in C_{HB} and C_{HBO_2} during that period is:

$$\Delta OD = \log_{10} \frac{I_b}{I} = \alpha_{HB} \Delta C_{HB} + \alpha_{HBO_2} \Delta C_{HBO_2} \qquad (2)$$

Measurements performed at two different wavelengths allow the calculation of ΔC_{HB} and ΔC_{HBO_2}. Oxygenation and blood volume can then be deduced:

$$Oxygenation = \Delta C_{HBO_2} - \Delta C_{HB} \qquad (3)$$

$$BloodVolume = \Delta C_{HBO_2} + \Delta C_{HB} \qquad (4)$$

2. System Implementation

2.1 Acquisition Hardware System

The fNIR system used in this study was originally described by Chance et al. (1998) [1]. The current flexible probe developed in our laboratory consists of 4 LED light sources and 10 detectors. Figure 1 below shows a block diagram of the continuous wave (CW) fNIR sensor system to monitor brain activity. The main components are the probe that covers the entire forehead of the participant, a control box for data acquisition, power supply for the control box and a computer for the data analysis software. The communication between the data analysis computer and the task presentation computer is established via serial port to time-lock the fNIR measures to the task events.

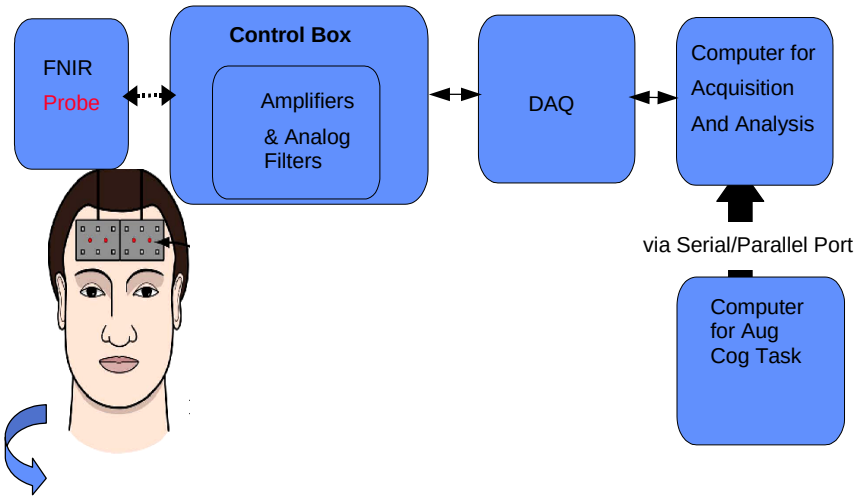


Figure 1. Block diagram of functional Near Infra Red (fNIR) sensor system: The control box hosts analog filters and amplifies; data acquisition board (DAQ) is used for switching the LED light sources and detectors, which collect the reflected light.

The flexible probe is a modular design consisting of two parts: a reusable, flexible circuit board that carries the necessary infrared sources and detectors, and a disposable, single-use cushioning material that serves to attach the probe to the participant (Figure 2). The flexible circuit provides a reliable integrated wiring solution, as well as consistent and reproducible component spacing and alignment. Because the circuit board and cushioning material are flexible, the components move and adapt to the various contours of the participant's head, thus allowing the sensor elements to maintain an orthogonal orientation to the skin surface which dramatically improves light coupling efficiency and signal strength. The source detector separation is 2.5 cm.

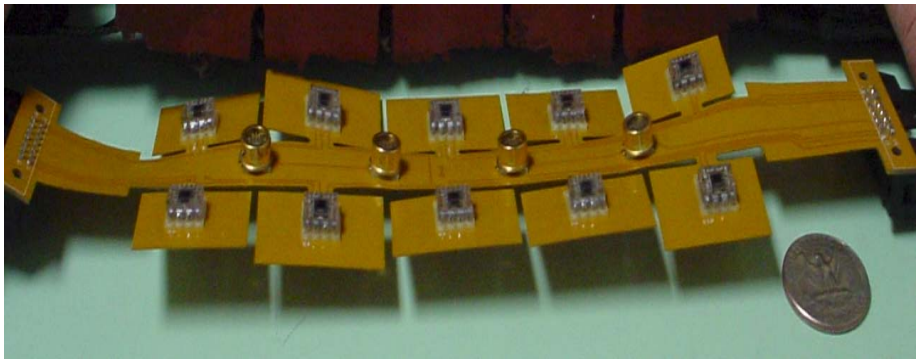


Figure 2.a. Flexible Probe

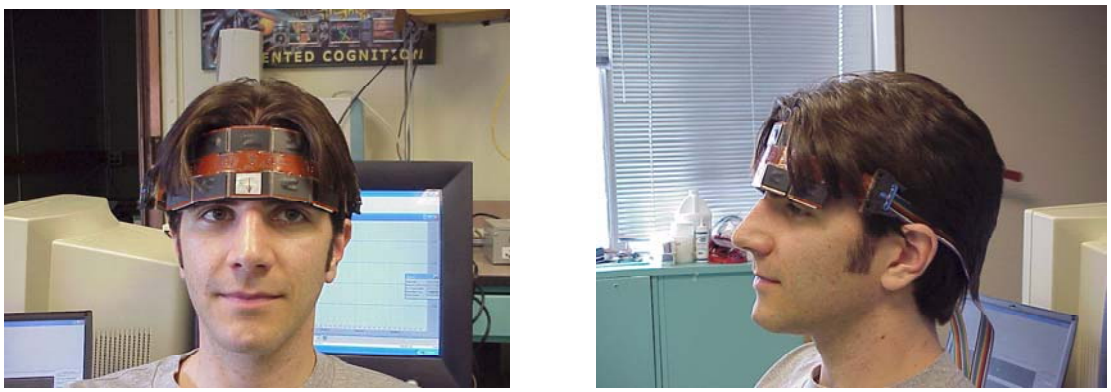


Figure 2.b. Participant wearing flexible probe.

2.2 Data Presentation System

Both online and off-line data presentation is available with fNIR system (Figure 2):

Online: Real time 16-channel gauges simultaneously monitor the hemodynamic changes during data collection (Figure 3). The behavioral events (e.g. onset of stimulus, type of stimulus, onset of response and type of response) are synched up with the fNIR measures for further event specific quantification and analysis. During the AugCog experiment sessions, each wave start and end events were recorded by the on-line fNIR data acquisition system.

Off-line: Testing and analysis platform is designed to process and present the averaged hemodynamic changes subsequent to data collection from a given participant. The WCT data is managed and processed using this off-line platform.

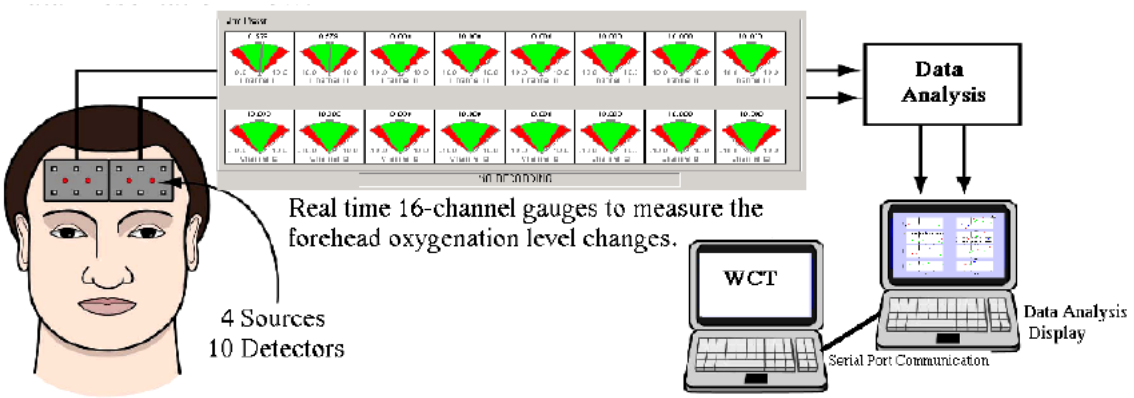


Figure 3. Data analysis and presentation system flow.

2.3 Signal Preprocessing and Conditioning

In order to increase the signal to noise ratio (SNR) of the fNIR device in extracting the hemodynamic response under cognitive tasks, signal processing algorithms have been implemented to identify, to eliminate and or to compensate for noise and other signal distortion such as electronic drift. In particular, optical data suffers from artifacts caused by head motion. Head movement may displace the sensors and cause them to capture ambient light or direct path light that has not passed through tissue. Adaptive filtering technique is implemented to detect and remove motion artifacts in the fNIR data collected during the AugCog experiment sessions.

2.4 Data Processing and Feature Extraction

As explained in section 2.2, the testing and analysis platform incorporates algorithms, which enumerate, sort and average the event segments marked during data collection. Each marker embedded into the fNIR data represents an event and thus the algorithm is able to extract these segments and to sort them in order to calculate averaged response to the specific stimuli. Using equations 2 and 3, the blood oxygenation is calculated to assess the response to each stimulus. The calculated blood oxygenation reveals relative changes to the baseline.

Feature extraction and pattern recognition algorithms provide the tools to test the hypothesis that underlies specific experimental protocols and to perform trend analysis. Feature extraction module is also part of the testing and analysis platform and automatically calculates averaged oxygenation and rate of oxygenation change over time. Rate of oxygenation change feature is extracted from the fNIR data by segmenting the data as explained above and fitting a line to the trend, i.e. increase or decrease, in each segment. The slope of line or rate of oxygenation change provides relevant information to brain activity versus baseline state.

3. Method

3.1 Subjects

A total of 8 healthy subjects were participated in the DARPA Augmented Cognition TIE study. The TIE was conducted at Pacific Science and Engineering in San Diego, CA.

3.2 Warship Commander Task

The Warship Commander Task description, the design and implementation of the task are detailed in the TIE Report document. Briefly, the primary task is air warfare management. While performing the air warfare task, task difficulty and task load is manipulated by changing the number of aircraft in a wave. Each subject completed multiple sessions with 3 waves of 6, 12, 18 and 24 tracks (presented in the order 6, 18, 12, 24).

3.3 Analyses

For each wave of 75 seconds, the rate of change in the oxygenation was calculated from the fNIR measurements. For the purpose of these analyses, blood oxygenation values were averaged across the channels or pixels covering left and right hemispheres. To test the primary hypothesis that increased workload would be associated with relative increases in blood oxygenation in the frontal poll and dorsolateral prefrontal cortex, we computed a 2 (High vs Low Complexity) x 2 (Full vs Divided Attention) x 4 (Wave Size) repeated measures analysis of variance (ANOVA) for each hemisphere. To test the hypothesis that blood oxygenation would predict performance, we computed within-subjects correlations between blood oxygenation scores and their RTIFF scores for each wave.

4. Results

4.1 Task Load and Performance Analysis

In support of our primary hypothesis, the results indicated a main effect for wave size for both the left (F_{left}= 14.87, $p<0.01$) and right (F_{right} = 11.73, $p<0.01$) hemispheres (see Figure 4, 5,6). This finding suggested that an increasing rate of change in blood oxygenation was sensitive to increasing workload as indexed by the number of planes that had to be managed. However, the rate of change parameter did not differentiate between levels of complexity (High vs Low number of unknown planes; F_{left}= 0.41, p=0.55; F_{right} = 0.33, p=0.58). The main effect for the Full vs Divided Attention also did not attain significance (F_{left}= 1.20, p=0.32; F_{right} = 1.10, p=0.33).

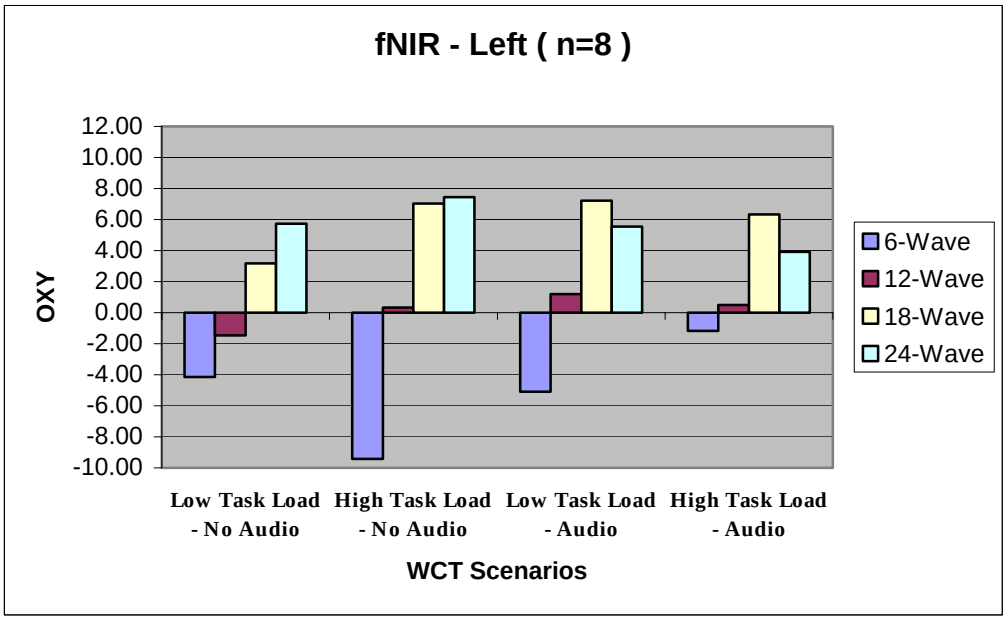


Figure 4. fNIR (Left) Averaged Oxygenation Data (n=8)

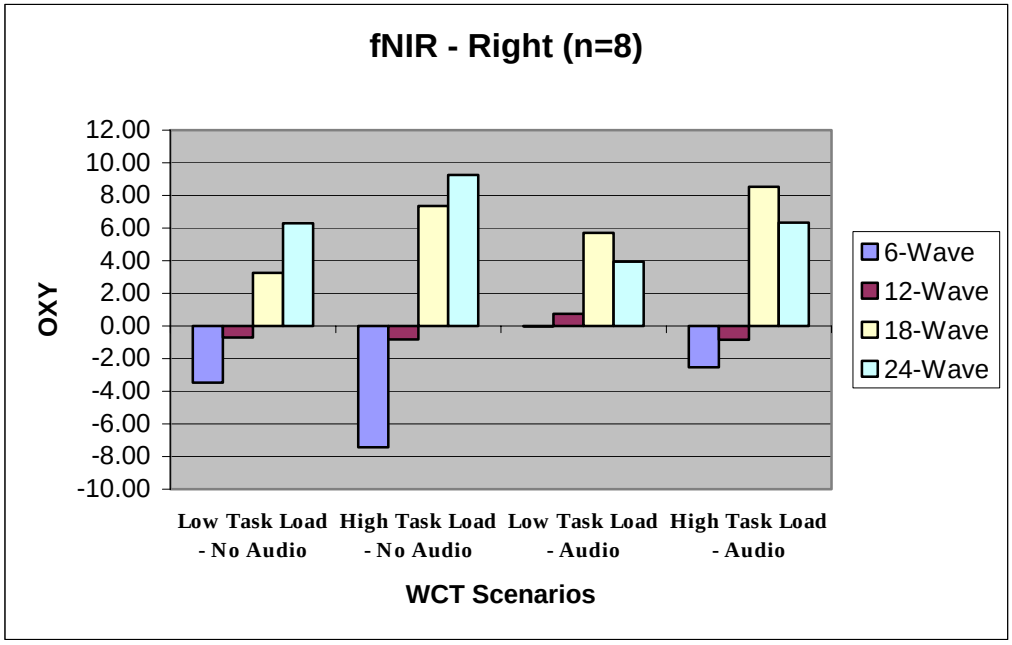
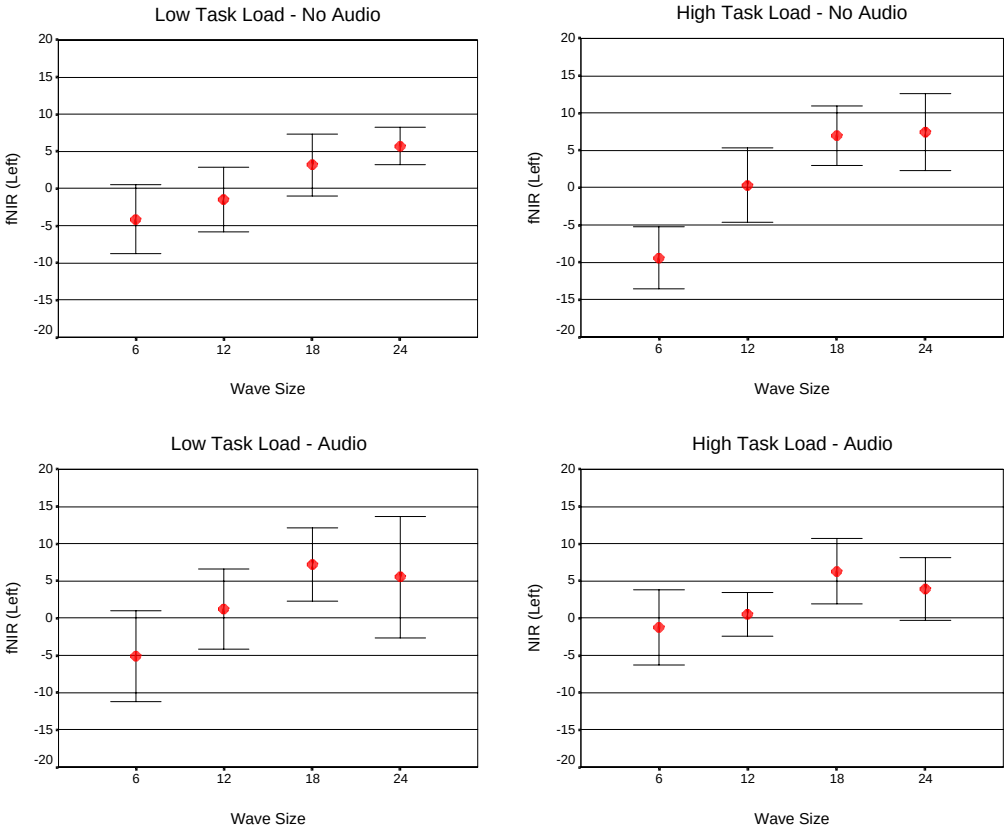


Figure 5. fNIR (Right) Averaged Oxygenation Data (n=8)



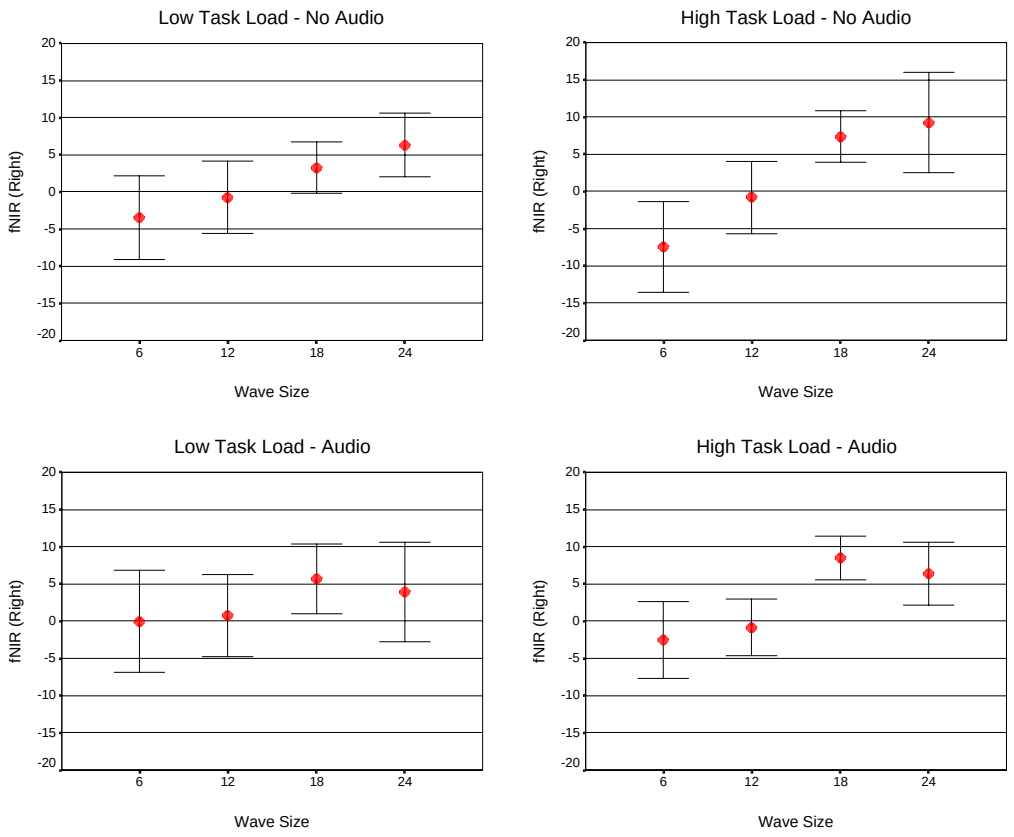


Figure 6. Detailed analysis of 4 different WCT scenarios.

A Pearson's correlation between fNIR gauge output and the RTIFF performance measure confirmed a positive relationship between prefrontal blood oxygenation and performance across all conditions (Left Hemisphere: Pearson's r = .31; Right Hemisphere: Pearson's r = .32). These data are graphically represented by wave size in Figure 7-8.

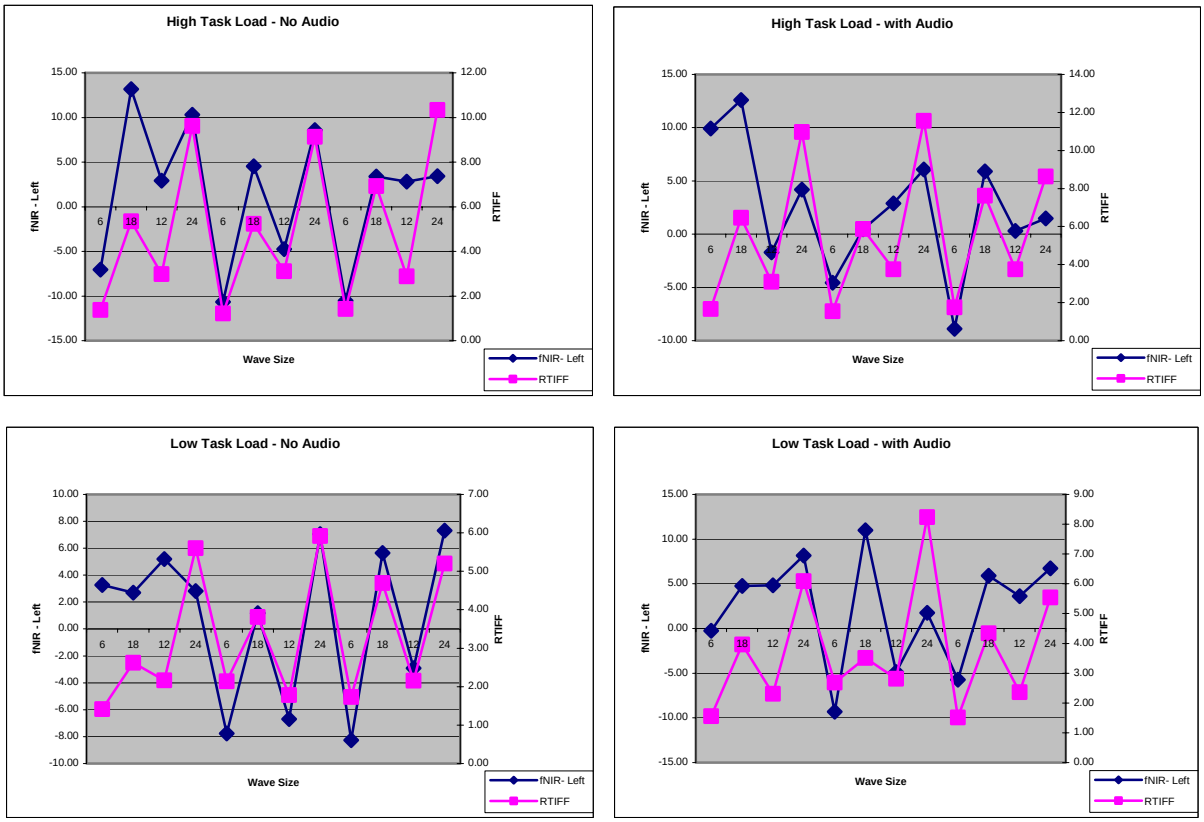


Figure 7. fNIR (Left) measurements vs. RTIFF (n=8). The plot reflects the correlation between rate of change in left prefrontal oxygenation and performance measure RTIFF for each of the 12 waves of the scenario.

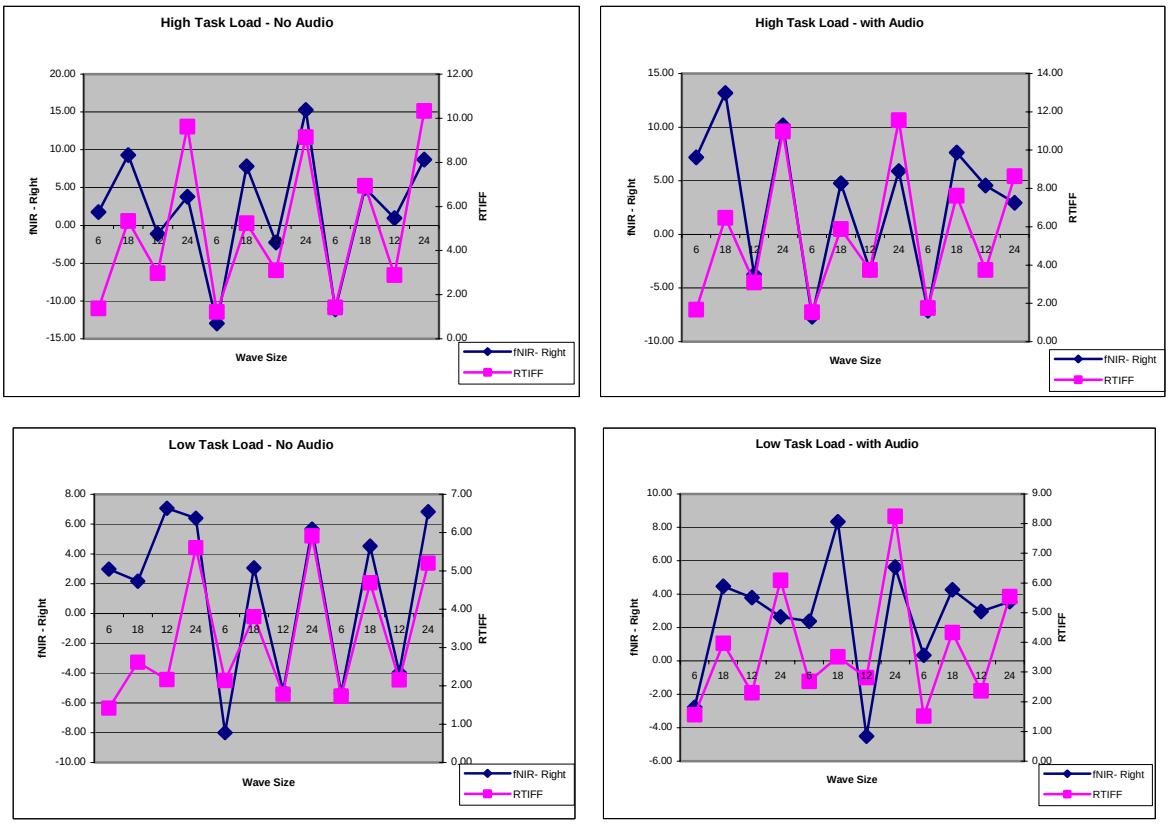


Figure 8. fNIR (Right) measurements vs. RTIFF (n=8). The plot reflects the correlation between rate of change in right prefrontal oxygenation and performance measure RTIFF for each of the 12 waves of the scenario.

4.2 Individual Participant Analysis

To measure correlation between number of tracks and the fNIR gauge values for each participant, we calculated 48 mean values of the blood oxygenation change rates for 3 waves of 6, 12, 18 and 24 tracks within each WCT session. These calculated values were analyzed and the results are presented in Table 1 and Table 2.

Table 1. Effects of track size on the fNIR (Left)

Subjects	F	p	R
P1	14.929	0.000	0.621
P2	19.980	0.000	0.672
P3	3.155	0.034	0.363
P4	1.931	0.138	0.306
P5	5.258	0.030	0.428
P6	0.312	0.817	-0.023
P7	2.769	0.053	0.395
P8	3.597	0.021	0.300

Table 2. Effects of track size on the fNIR (Right)

Subjects	F	p	R
P1	13.196	0.000	0.649
P2	8.331	0.000	0.537
P3	0.574	0.635	0.064
P4	3.006	0.040	0.342
P5	2.527	0.070	0.308
P6	1.151	0.339	0.169
P7	1.491	0.230	0.290
P8	2.152	0.107	0.281

* R: Mean correlations between number of tracks and the fNIR gauge.

The individual analysis in The Table 1 and 2 suggests that the fNIR gauge is significantly sensitive to number of tracks per wave for some participants. Interestingly, for subject P6 both left and right fNIR measurements failed. Unfortunately, we do not have the post interview with the subject. Hence, without further measurements and post interviews, we cannot explain this result.

4.3 Ship Status Task Results

To examine the effects of a divided attention task in which the secondary tasks utilized auditory memory resources, we compared scenarios with the Ship Status Task (SST) ON and with the SST OFF (Figure 9). To explore the differences between encoding and recall, the auditory memory task was first divided into epochs involving encoding and recall. Then, we computed baselines - two seconds prior to stimulus presentation – and responses - 4 seconds post stimulus presentation. As shown in Figure 10, there is significant difference between baseline and response to auditory stimuli, yet moderate difference between encoding and recall.

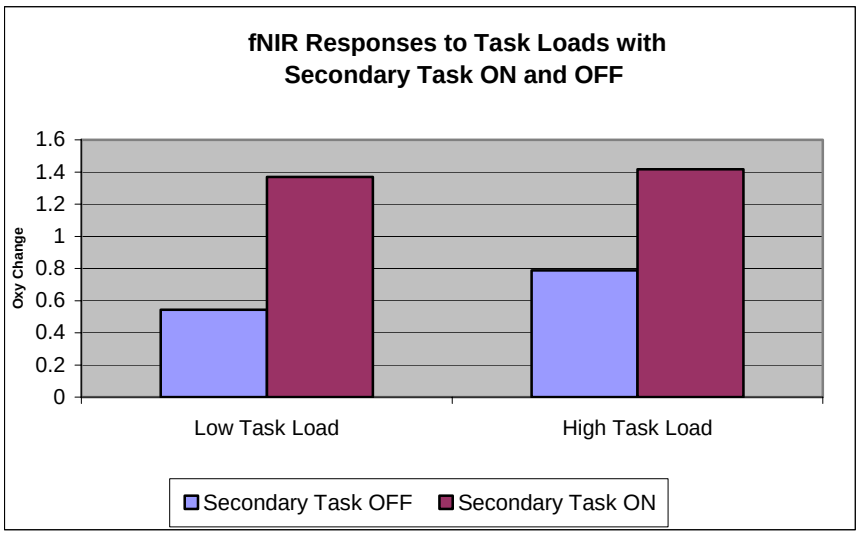


Figure 9. fNIR measurements for the Low and High Task Loads with the secondary task ON and OFF. (n=8)

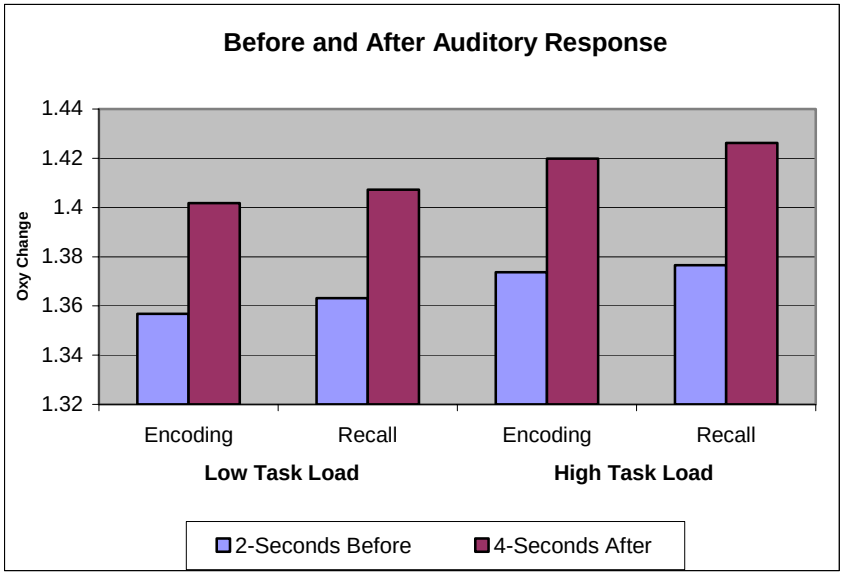


Figure 10. fNIR measurements: Before and After response to the encoding and recall stimuli. (n=8)

5. Discussion

Functional near infrared spectroscopy, a portable, safe, affordable and negligibly intrusive optical imaging system, can be used to measure hemodynamic changes in the cortex. In this study, our task was to use fNIR as a gauge of cognitive workload in a complex, "realistic" task. We expected to find that changes in blood oxygenation in relevant areas of the frontal poll and dorsolateral prefrontal cortex would be associated with increasing cognitive effort defined as attention and working memory. The results, acquired in the context of the Warship Commander Task, suggest a reliable association between cognitive workload and increases in the oxygenation responses under circumscribed conditions. They also suggested that there is a positive correlation between a participant's performance and oxygenation responses in dorsolateral prefrontal cortex as a function of task load. As an initial endeavor, these results are promising for the use of fNIR in the creation of a symbiotic relationship between the operator and the operational environment.

The main effects for complexity and for the divided attention task, despite increasing overall cognitive effort [7], were not associated with significant changes in blood oxygenation. There are a number of potential reasons for these findings. First, the sample size was small, which limits the power of these analyses. It is possible that a larger sample could result in more positive results. Second, the current analyses focused on only two fNIR parameters, average change in oxygenation and rate of change of oxygenation. It is possible that other parameters could add predictive power in these complex cognitive tasks. We are currently working on developing a parametric model for the HbO_2 pulse to extract additional features such as peak amplitude, pulse width, latency, etc. Hence, further development in the algorithm, fine-tuning and increasing the number of features, are expected to enhance consistency and efficacy of the gauge. Third, the current sensor was applied over a limited area of the frontal poll and dorsolateral prefrontal cortex. Some of these manipulations may have had effects in areas of the cortex that are accessible to fNIR, but were not measured with the current sensor. This question remains for

future generations of sensor to determine. Finally, the Warship Commander Task itself is complex, and numerous cognitive and emotional functions are occurring during the execution of the task. It is possible that these various tasks have differential effects on the hemodynamic response. For example, recent research using PET indicates that various areas of cortex show increases in oxygenation during a divided attention task relative to a full attention task, whereas other areas demonstrate decreases in oxygenation during the same task [5]. Further work is needed to more fully explicate our understanding of brain function during what may be common everyday, and yet extremely complex tasks.

Acknowledgments

We acknowledge with thanks Mark St. John, David Kobus, Jeff Morrison, Gary Kollmorgen and their colleagues at PSE, SPAWAR and BMH Inc. for organizing and hosting the pre-TIE and TIE sessions.

This work has been sponsored in part by funds from the Defense Advanced Research Projects Agency (DARPA) Augmented Cognition Program and the Office of Naval Research (ONR), under agreement numbers N00014-02-1-0524 and N00014-01-1-0986.

References

[1] Chance B, Anday E, Nioka S, Zhou S, Hong L, Worden K, Li C, Murray T, Ovetsky Y, Pidikiti D, Thomas R (1998). A novel method for fast imaging of brain function, non-invasively, with light. Optics Express, 2, 10.

[2] Chance, B. Zhuang Z, UnAh C, Alter C, Lipton L (1993). Cognition-activated low-frequency modulation of light absorption in human brain. Proceedings of the National Academy of Science of the United States of America 90, 3770–3774.

[3] Villringer A, Chance B (1997). Non-invasive optical spectroscopy and imaging of human brain function. Trends in Neuroscience, 20, 435-442.

[4] Izzetoglu K, Yurtsever G, Bozkurt A, Yazici B, Bunce S, Pourrezaei K, Onaral B (2003). NIR Spectroscopy Measurements of Cognitive Load Elicited by GKT and Target Categorization. Proceedings of the 36th Hawaii International Conference on System Sciences

[5] Iidaka T, Anderson ND, Kapur S, Cabeza R, Craik FIM (2000). The effect of divided attention on encoding and retrieval in episodic memory revealed by positron emission tomography. Journal of Cognitive Neuroscience, 12:2, 267 –280.

[6] Cope, M. (1991). The Development of a Near-Infrared Spectroscopy System and Its Application for Noninvasive Monitoring of Cerebral Blood and Tissue Oxygenation in the Newborn Infant. Univ.College London, London.

[7] St. John M, Kobus DA, et al. (2002). A Multi-Tasking Environment for Manipulating and Measuring Neural Correlates of Cognitive Workload," Proceedings of the IEEE 7th Conference on Human Factors and Power Plants; 7:10 –14.

3E ELECTRICAL GEODESICS, INC.

During the TIE we were able to collect data from subjects 1, 2, 3, 5, 6, and 8. Below is the summary analysis for subjects 1, 2, 5, 6 and 8. The analyses were conducted for all scenarios. The data from subject 3 was lost; we are unable to find the data files.

Gauge Description

We have two gauges: 1) Motor Effort and 2) Auditory Working Memory (Auditory Effort). These correspond to the theta responses to the KIFF and AHTH events, respectively.

Hypotheses

As workload increases, increased cognitive effort is required. We predict that the motor effort and auditory effort gauges will predict participants' subjective effort ratings. Moreover, both gauges will provide non-redundant information.

Device

The device used to assess participants' cognitive state is a dense-array (128-channel) EEG system. This system allows us to sample the entire potential field and allows us to track local cortical networks responsible for different cognitive operations.

Method

We assume that during a task, the state of the brain cannot be accurately assessed unless the assessment is done relative to some processing event. That is, there is no particular global brain state that we can assess that will give us information about the processing capacity of the brain. Therefore, we selected two events (KIFF and AHTH) from the WCT to focus our analysis.

Another reality of analyzing the neurophysiological signal of the brain is that the frequency component of the signal carries information about function. Therefore, we have to focus our initial analyses on those frequencies that may be most relevant to the question. Because we are interested in assessing workload, the candidate frequency is between 4-7 Hz (theta).

The reality of brain function is that information processing is not carried out across the entire cerebral network. Rather, cognitive functions appear to involve local networks. We must analyze the activity of these networks to assess particular cognitive capacities. Therefore, we must address another dimension of complexity in the analysis. Because we acquired dense-array EEG data, we can focus on particular regions for the analyses, which are guided by existing literature.

For the EEG analysis (gauges), we obtained a measure of theta averaged .5 sec before and after the AHTH and KIFF events. Because the KIFF and AHTH events represent processing capacity in different domains, motor and auditory, respectively, we focus the analysis on those sensor positions that overlie the somatosensory motor cortex for the KIFF event and over the medial prefrontal cortex for the AHTH. It should be noted that the data over the medial prefrontal cortex reflect activity from both the auditory cortex and the medial prefrontal cortex.

Data

AHTH Event

The first analysis we did was to obtain the correlation between subjects' effort rating with our EEG measure of effort (theta). For subject one the correlations were quite impressive. For the K+, L+, and L scenarios the correlation coefficient are .53, .63, and .76, respectively (see Figure 1). For scenario K the correlation coefficient is quite small (-.11).

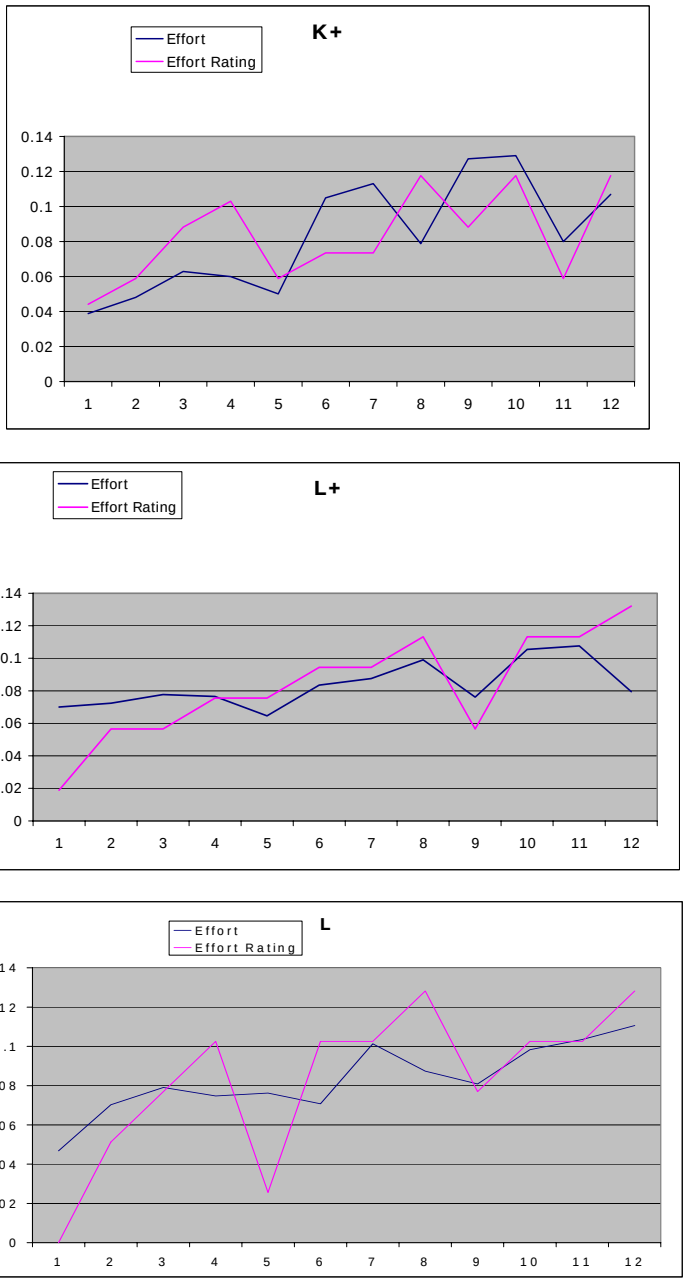


Figure 1. Effort X Effort Ratings: Values on X axis represent wave order. Values on Y axis are scaled units (percentage).

For subject 2, the results were less impressive. For the K+, L+, L, and K scenarios the correlation coefficients are .06, -.27, -.38, and -.39. For subject 5, the correlation coefficients for scenarios G, G+, H, and H+ are -.03, .42, -.34, and .27, respective. Figure 2 shows the relation between effort and effort ratings.

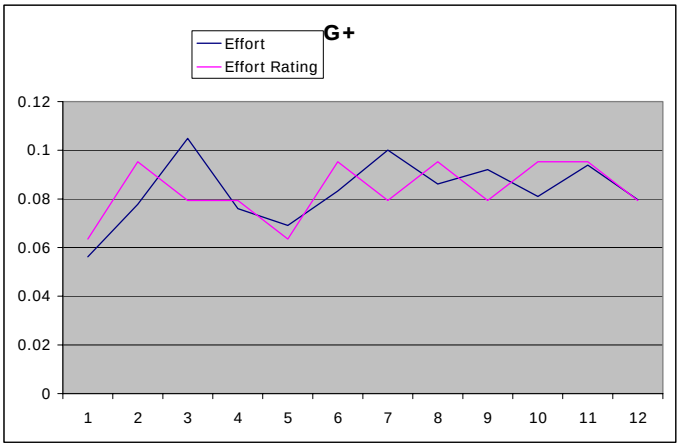
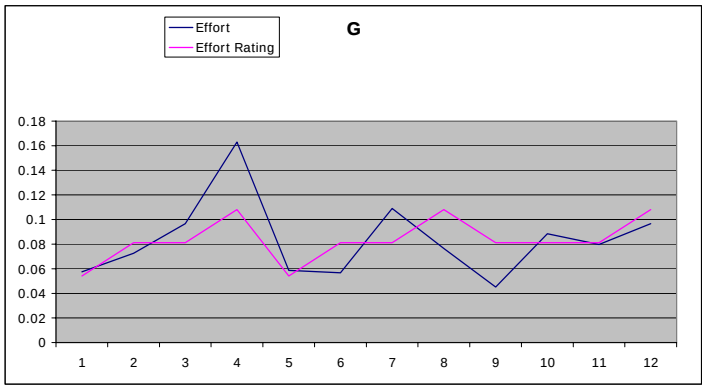


Figure 2. Effort X Effort Ratings: Values on X axis represent wave order. Values on Y axis are scaled units (percentage).

For subject 6 the correlation coefficients are for scenarios G, G+, H, and H+ are .59, .31, .22, and -.41, respectively. The data for scenarios G and H+ are illustrated in Figure 3.



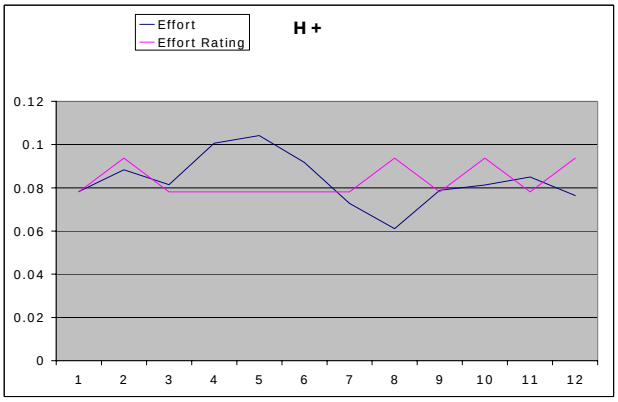


Figure 3. Effort X Effort Ratings: Values on X axis represent wave order. Values on Y axis are scaled units (percentage).

For subject 8 the correlation coefficient are -.09, -.40, -.52, and -.70 for scenarios J, J+, I, I+, respectively.

KIFF Event

For subject 1, the analysis of showed that the correlation coefficients for scenarios K, K+, L, L+ are -.54, .51, .18, .-06. For subject 2 the correlation coefficients are .18, .40, -.12, and .54, for the same scenarios. Figure 4 shows the effort X effort ratings for scenario L+.

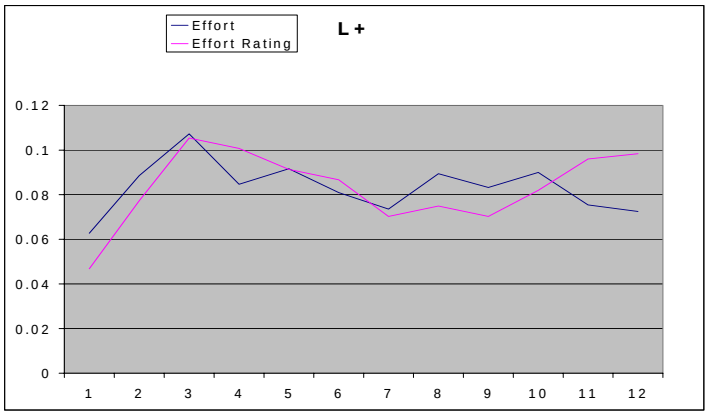


Figure 4. Effort X Effort Ratings: Values on X axis represent wave order. Values on Y axis are scaled units (percentage).

For subject 5 the correlation coefficient for the G, G+, H, and H+ scenarios are .44, .42, .20, and -.69. For subject 6 the correlation coefficients are .24, .57, .60, and .18 for the same scenarios. Figure 5 shows the data for subject 6 for the G+ and H+ scenarios.

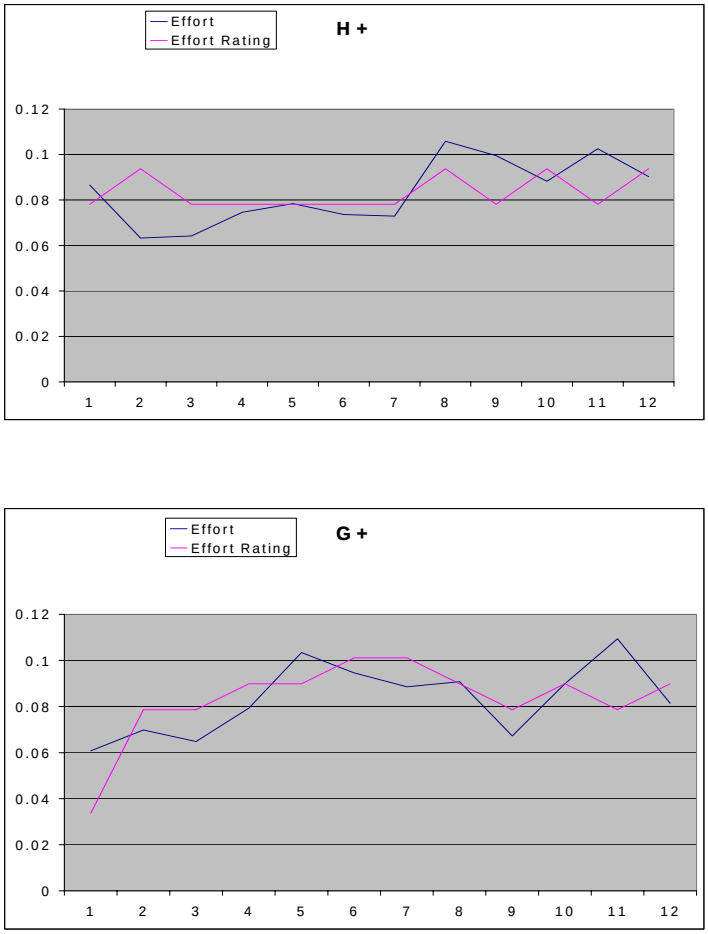


Figure 5. Effort X Effort Ratings: Values on X axis represent wave order. Values on Y axis are scaled units (percentage).

For subject 8, the relation between effort and rated effort are -.13, -.26, -.09, and .03.

Predicting Subjective Effort

Table 1 shows the R square (right column) for each subject for each scenario. The R square represents the proportion of variance in subjective effort ratings accounted for by motor and auditory effort gauges. The middle column presents the correlation between the motor and effort gauges. The median R square is .30. That is, 30 percent of the variance in participants' subjective effort ratings can be accounted for by the motor and effort gauges. The median correlation between the motor and auditory gauges is .18. These results show that a reasonable prediction of task workload, as measured by subjective effort, can be made by the motor and auditory effort gauges. Moreover, the gauges do not appear to provide redundant information.

These results confirm our prediction and suggest that additional gauges can be built around events that occur in the context of the task. This is promising with regards to aim of the AugCog program to build multiple gauges that can tap into different cognitive functions and resources.

Table 1. Left column lists the participant and scenario. Middle column lists the correlations between the auditory and motor effort gauges. Right column lists the R square (proportion of variance accounted for in the subjective effort ratings by the auditory and motor gauges).

	r (AHTH & KIFF)	Rsquare
P1K	0.17	0.29
P1K+	0.8	0.3
P1L	0.24	0.58
P1L+	0.15	0.49
P2K	0.07	0.19
P2K+	0.1	0.16
P2L	0.53	0.15
P2L+	0.63	0.3
P5G	2.6	0.22
P5G+	0.48	0.24
P5H	0.12	0.17
P5H+	0.53	0.48
P6G	0.31	0.35
P6G+	0.7	0.33
P6H	0.1	0.39
P6H+	-0.48	0.17
P8J	-0.47	0.04
P8J+	0.19	0.19
P8I	-0.15	0.3
P8I+	-1.6	0.5

3F QINETIQ

Introduction

The Cognition Monitor (CM) system was designed to provide real-time analysis of operator state and was initially developed as part of the DERA Cognitive Cockpit program (e.g. Taylor et al., 2001). The term 'operator state' refers to a constellation of behavioral, physiological and contextual factors. We prefer the term 'operator state' to 'workload' (Wickens, 1987) or 'cognitive load' (Pleydell-Pearce et al., 1995), as we take the view that it is not only load that affects the ability of an individual to perform a task effectively. These states may include anxiety, drowsiness and even changes in levels of situational awareness. The CM is a system that monitors behavioral, physiological, contextual and situational data either within a cockpit simulator or in conjunction with laboratory-based tasks. In the cockpit simulator, CM processes a host of low-level variables and then employs data fusion routines to provide estimates of higher aspects of pilot state, which are then passed to a Tasking Interface Manager system that is concerned with adaptive aiding of operators (see Bonner et al., 2000; Taylor et al., 2001). Lower-order variables include physiological and behavioral inputs. Examples of higher-order state estimates include alertness, stress, visual load, auditory load, verbal load, motor load, spatial load and executive load. Higher-order outputs also involve identification and tracking of ongoing tasks. This is important because such a facility enables real-time prediction of operator intent, and also provides a means by which specific tasks can be identified as overloading the operator. In such cases, the Cognitive Cockpit can instigate routines to aid the pilot (Bonner et al., 2000; Shadbolt et al., 2000). For example, aiding routines may recommend that a specific task be taken under automatic control.

Overview of CM Architecture

CM consists of a stand-alone physiological monitoring system (CogPhys) capable of analyzing EEG and a host of autonomic variables (e.g. cardiac, respiratory, and electrodermal activity). It also includes hardware and software that analyze ambient luminance, sound and vibration/movement of the participant. The vibration detectors have been used to control for physiological artifacts associated with participant movement and vibration.

Meaningful analysis of physiological data requires a good understanding of behavioral activity exhibited by participants, and of the cognitive demands of the task. For example, task demand and workload typically depend on the skill and motivation of participants. In some cases, tasks that are intended to impose a particular level of demand upon participants can change their character if participants make mistakes or fail to deal appropriately with events (e.g. McCallum & Pleydell-Pearce, 1993). A good understanding of tasks and their cognitive components also allows fine-grained comparisons between physiology and information processing concepts. Levels of verbal, spatial and memory load vary considerably across Warship Commander Task (WCT) scenarios. Furthermore, the ability to model the information processing demands of a task can lead to real-time task tracking and analysis. In some cases this can provide information of a quality that cannot be matched by any physiological methods available either now or in the foreseeable future.

The Cognition Monitor system includes task modeling software that provides an ability to track and analyze behavioral activity in real time. This system is known as the Task Taxonomy and was first described by Pleydell-Pearce et al., (2000). The Cognition Monitor Task Taxonomy (CMTT) was initially implemented within the DERA/QinetiQ Cockpit Program (Taylor et al., 2001). The CMTT was adapted to function in conjunction with the WCT and was demonstrated, operating in real time, at the Aug Cog Technical Integration Experiment (TIE).

CMTT consists of layers of detection algorithms. At the lowest level all possible interactions with the task platform are encoded at an ‘event level’ that is defined as a non-decomposable tangible interrupt (e.g. button press, touchscreen activity or voice-activated request). At higher levels in the taxonomy, more complex tasks are defined in terms of combinations of events that have a temporal ordering that uniquely defines a higher-order 'task' (e.g. pressing an area of a touch-screen when a particular digital map is displayed). At progressively higher levels of abstraction, combinations of tasks (which themselves are each conjunctions of events) provide information about more global goals and operator intention.

The CMTT can therefore be regarded as an emulation or model of the environment. Interactions with the environment instantiate particular states in the model, which mirror activities in the environment, in real time. In addition, CMTT contains task knowledge derived from knowledge elicitation (KE) sessions with subject matter experts (SMEs). This knowledge describes the cognitive demands of specific events and event conjunctions (i.e. higher-order tasks). This knowledge also includes logical inferences about situation awareness.

During the AugCogTIE, a partial version of the taxonomy was employed. It was populated with event level knowledge and KE data derived from SMEs who learned the WCT task.

In summary, CogPhys collects and analyses physiological data and CMTT collects and analyses behavioral data. Outputs from both these systems are passed to a third system, CogWeb, and it is here that behavioral and physiological data are fused to provide higher-order measures of cognitive demand and affective state. During the TIE the full CM system was demonstrated deriving a number of gauge estimates for different aspects of cognitive-affective status. Only estimates of Executive Load have been analyzed and reported here.

The Gauge Concept

The CM system does not rely upon any single dependent variable, but instead monitors a large number of inputs in parallel, and in real-time. These variables are fused to provide high level descriptors of cognitive-affective status. Behavioral and physiological data are typically fused within CM, but for the purposes of the TIE we have maintained a separation of the two classes of measure. CM makes inferences about operator state in two general modes: bespoke and generic. Bespoke predictions are based on coefficients or ‘weights’ that are derived for each participant, on the basis of calibration data and previous encounters with the individual participant (e.g. Pleydell-Pearce et al., 2003). Generic predictions are based upon weights that are derived from across-participant analyses. These generic weights are based upon previous research, and have been systematically improved over the years.

There are a number of core philosophical assumptions that underlie our approach to operator state; these are described in the following paragraphs.

Operator Variability: From the outset, we have maintained that there are important individual differences in the way that people approach workload environments. These differences can be termed 'Cognitive Style' and we have shown that variability in cognitive style influences brain activity observed during demanding task performance (e.g. Riding et al., 1997, cf. Pleydell-Pearce, 1994). Because CM can focus upon individual differences it can provide a more accurate assessment of operator state than can be produced by approaches that utilize across-participant statistics. Individual differences are also important given the well known differences in size and structural organization of individual brains; and, in terms of EEG, topographic differences in factors such as skull thickness are also important. Individual differences in autonomic reactivity are also very important and are well known in the literature concerned with arousal and emotion.

Context Specificity: CM is designed to learn about particular contexts, for example specific tasks. It was initially designed to work in a cockpit environment, but has recently been extended to the WCT. Different working environments impose different kinds of cognitive demand. We argue that the varieties of cognitive demand, across platforms, mean that it is virtually certain that no single measure will provide enough information to predict operator state.

Inclusive Measures: CM analyses a very large number of variables, in some cases in excess of 30,000 dependent variables every second. This approach is based upon our argument that the predictive power of data sources must be assessed within the context to which they are to be applied. When such analyses take account of operator variability the result is a powerful system that is tailored to a specific environment and a specific individual.

Generic Analysis: Although we have stressed the importance of individual differences, CM is capable of analyzing data on the basis of across-participant knowledge. In other words, it can analyze performance based upon prior encounters with many participants, and assume that all participants will react in the same way. This can be performed after a short calibration period of about 10 minutes.

The Warship Commander task (WCT)

The WCT has been developed to enable a demonstration of the utility of the various gauges developed under Phase 1 of the DARPA Augmented Cognition program. The task involves the participant controlling a fictional area of airspace in front of a maritime convoy. It requires the participant to locate and identify new contacts, track the progressions of each contact, and engage contacts if they cross a line of engagement. Further to this, participants may be prompted to remember verbal messages relating to ships' status for subsequent recall. The demands of the task may be varied by changing the number of contacts on the screen at any one time, or by inclusion of the ships' status task (SST). The classification of the contacts, either as friend, enemy or unidentified (yellow), enables a further manipulation of the requirements of the task. As yellow tracks require a further process of enquiry before classification, a degree of contact uncertainty can be introduced by manipulating the number of yellow tracks presented.

From a cognitive task analysis perspective, the WCT has visual, auditory and motor components. There is also some short- and longer-term memory demand, and the requirement for spatial and verbal processing.

Executive Load

We have defined Executive Load gauges that are used to describe activity during the performance of the WCT. Two such gauges are reported here: one using the physiological measure approach we have explained elsewhere (Pleydell-Pearce et al., 2003); the second based upon tracking of behavior using the CMTT.

All events passed from the WCT were analyzed for their intrinsic meaning, in terms of the cognitive processes they represent. These events — such as button presses or auditory stimuli — were subsequently coded into the taxonomy and rated for executive load. Examples of the type of event that contributed to this gauge are (1) the cognitive operations invoked (for example, if working memory processes were involved, or if task events indicated a dynamic switch of attention, executive demand would increase); (2) subjective ratings from experienced operators; and (3) tasks with motor responses that required selective choice among competing alternatives and hence raised executive demand. Verbal tasks also increased executive demand. The executive demand inferences were

maintained in real time and tried to match demand on a moment-to-moment basis. This involved taking account of the probable (or demonstrable) duration of specific events.

METHOD

Procedure

Firstly, participants read and signed the consent forms. Then, sensors were affixed and participants were prepared for running the calibration routines and the task. Data were recorded from 14 electrodes within an electrode cap, at sites around the scalp according to the 10–20 system of placement (Jasper, 1958), and are illustrated in Figure 1 below. Two electro-oculogram channels were recorded from electrodes placed above and below the right eye, and at the outer canthi. In addition, data were recorded from electrodes placed around the chest to measure electro-cardiogram (ECG), and on the second phalange of the third and fourth finger of the right hand to measure electro-dermal activity (EDA). Accelerometers were attached to the hand via gloves, and to the electrode cap on the participant's head to measure movement. Participants also wore a strain gauge around the chest to measure respiration, and a throat microphone to record digitized vocal output. Ambient sensors recorded sound and screen luminance.

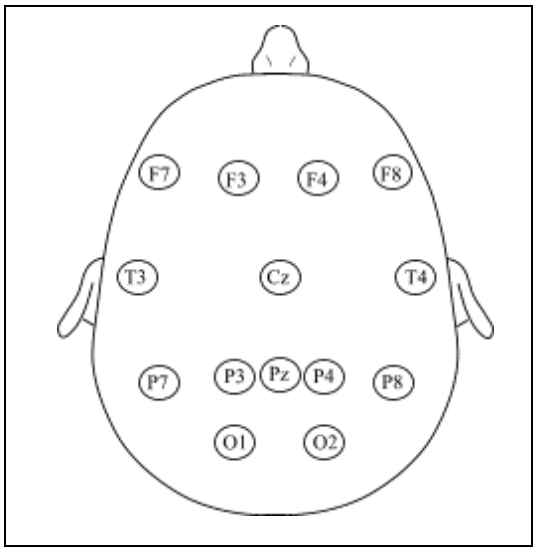


Figure 1 Electrode placements used in AugCog TIE

Participants initially performed two calibration tasks prior to the main block of trials. Both calibrations used scenario 'So' (three waves of six aircraft, high yellows, Ship Status task off). During the first calibration, subjects performed the task as they had been trained to do. When presented with this scenario on a second occasion, participants were asked to relax, and avoid eye movements by focusing on a point at the center of the VDU that presented the WCT. Owing to time constraints, we were unable to perform calibrations associated with the presence versus absence of the ship status task.

RESULTS

Behavioral Data

Analysis of behavioral data focused upon the task taxonomy that provides information about a host of distinct performance measures derived from the available outputs of the WCT. The executive load measure, derived from tracking behavior and scaled between 1 and 5, was analyzed in a three-way repeated measures analysis of variance (ANOVA) with two levels of ship status task, four levels of aircraft per wave (6, 12, 18 and 24) and two levels of number of yellow tracks (low versus high). Table 1 displays results for executive load. Inspection of the data, and results of statistical analysis, indicate a reliable discrimination between load levels.

Table 1: Measures of executive demand derived from the CM Task Taxonomy. Results of three-way repeated measures ANOVA show highly significant main effects

Executive Load

Condition Means	*Ship status off*		*Ship status on*	
Yellow	*low*	*high*	*low*	*high*
6 Aircraft	2.06	2.02	2.56	2.69
12 Aircraft	2.26	2.79	2.88	3.21
18 Aircraft	2.52	3.00	3.16	3.62
24 Aircraft	3.26	3.97	3.70	4.09

Source	*df*	*F*	*p*
Yellow (low/high)	1,5	61.66	<0.001
Ship Status task (off/on)	1,5	108.72	<0.001
Aircraft (6,12,18,24)	3,15	20.24	<0.001
Yellow x Ship	1,5	1.53	ns
Yellow x Aircraft	3,15	3.78	<0.05
Ship x Aircraft	3,15	2.71	ns
Yellow x Ship x Aircraft	3,15	2.37	ns

Physiological measures of Executive Load

For the purposes of the TIE, CM derived 5786 physiological variables. Initially, we wished to demonstrate how many of these variables were significantly different in a contrast between the relaxed calibration and the load calibration (three waves of six aircraft, scenario ‘S’). To achieve this aim, the means, across sequential analysis windows, were calculated within both the load and relaxed calibration blocks. These two values were then compared using a correlated t-test, for each of the 5786 variables. EEG data were corrected for contamination by eye movements using the procedure described by Gratton et al. (1983) but with modifications suggested by Conway et al. (2001). Of these contrasts, 1106 (19.11%) were significant ($p<0.05$). More importantly, 1000 (17.28%) of these contrasts involved EEG electrodes only. The exclusion of ECG, respiration, accelerometer and photocell data indicates that there are large differences between relaxed wakefulness and performance on the WCT. We wish to stress that the number of significant differences is far greater than would be expected on the basis of chance (see Pleydell-Pearce et al., 2003, for similar findings and discussion). These findings do not, of course, guarantee that these same variables will discriminate levels of task demand. For example, some of the variables may correlate with non-specific levels of alertness, or may reflect movement artifacts within EEG caused by interaction with

WCT. However, it is our experience that those variables that do show differences between relaxed wakefulness and moderate task performance are critical predictors, and it is from these variable that the Executive Load gauge was developed.

Bespoke Analysis: Executive Demand

For the purposes of the TIE, each participant was individually calibrated prior to data collection proper. The calibration involved analysis of both relaxed wakefulness and actual WCT performance (see above). CM calibration routines then selected variables that were highly predictive of the difference between the two states, for each participant. The criteria for selection include analysis of conjoint predictive power, and regression analyses can be used to achieve this (see Pleydell-Pearce et al., 2003). Attachment 1 provides information about some of the measures selected for Participant 1.

During the TIE we focused upon demonstrating our measures of Executive load, derived from physiological data. Executive Load is based upon analysis of inter-electrode coherence, spectral power and spectral power ratios. It should be noted that regression coefficients are much easier to employ when expressed in terms of standard scores and thus means cited here are in standard and not raw score form.

Figure 2 displays the mean inferred executive load averaged across all participants, scenarios and wave sizes derived from both physiological data and from the behavioral analysis (CMTT). Although the similarity between the curves is reassuring, the reason for showing these data is to indicate that, overall, each wave starts and ends with low levels of workload. For this reason, analyses that collapse values within each wave size are likely to be less reliable than those that focus upon the central portion, where there is greater activity. For this reason, we focus upon the period from 8.52s to 68.18s within each wave.

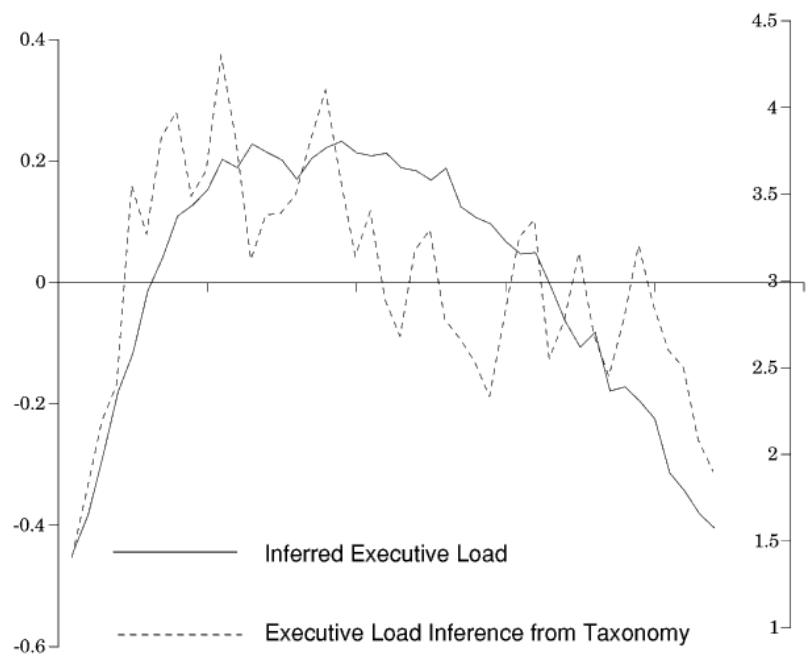


Figure 2: Comparison between measures of executive load derived from physiology (continuous trace) and from the CM Task Taxonomy (dotted trace). Data are averaged across all participants, conditions and wave sizes. The left vertical axis refers to the physiological measure, and the right vertical axis refers to the taxonomy measure

Figure 3 displays mean values for executive load derived from physiological data. The data represent an average taken from the most demanding phase of each wave (see earlier). The data indicate good separation between wave sizes, and good separation between low and high yellows for wave sizes 6 and 12. However, the data for the 24 aircraft wave size indicate lower levels of executive load when the ship status task (SST) was included.

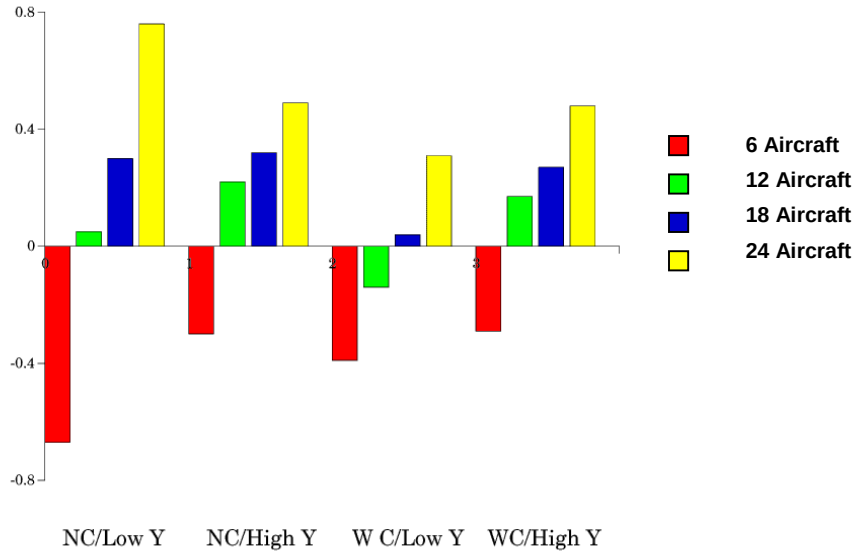


Figure 3: Estimates of executive demand derived from analysis of physiology. NC = SST off, WC = SST on, Low Y = low number of yellows, High Y = high number of yellows

Data representing Executive Load were subjected to a three-way ANOVA, with two levels of frequency of yellow (FOY), two levels of SST and four levels of aircraft. The main effect of aircraft was significant, $F(3,15) = 49.78$, $p < 0.0001$. Post hoc contrasts employed t-tests (this procedure was employed only when main effects of interactions were significant). These revealed that, with the exception of the contrast between 12 aircraft and 18 aircraft, all contrasts between wave sizes were significant, greater numbers of aircraft being associated with higher inferred levels of executive load.

The interaction between FOY and aircraft was significant, $F(3,15) = 11.69$, $p < 0.001$. Post hoc t-tests indicated that there were significant differences between high and low yellows for wave sizes of 6 and 12 aircraft, but not for 18 and 24 aircraft wave sizes. This finding may indicate that, because larger wave sizes are more demanding, the effects of FOY are less apparent since participants are active throughout. Contrasts between wave sizes, within both low and high number of yellows, revealed a virtually identical pattern to the main effect of wave size described above, with one exception: for low FOY, the contrast between 12 and 18 aircraft was significant.

The interaction between communication levels and aircraft was also significant, $F(3,15) = 13.63$, $p<0.01$. Post hoc analyses revealed that there was a significant difference between high and low communications but only for the 24 aircraft wave size: higher levels of inferred workload were associated with the *absence* of communications. We think that these data may indicate that the verbal task leads to an interruption of workload that is particularly marked for 24-aircraft scenarios. This conclusion is also supported by analysis of performance data. For example, time to Identification of Friend or Foe (IFF) was significantly slower during scenarios that involved communications.

Contrasts between wave sizes within both levels of communications revealed an identical pattern to the main effect of wave size described above.

Finally, the three-way interaction was significant, $F(3,15) = 19.27$, $p < 0.001$. When the lowest-order interaction is significant, then main effects and higher-order interactions may be confounded, and need to be interpreted with caution (Winer, 1971). The significant three-way interaction permits 166 unique pairwise comparisons; of these, 104 were significant. Examination of these comparisons indicates that the vast majority of significant contrasts are in accord with the prediction that executive load is increased as the number of tracks increases, with the inclusion of the SST, and as the number of yellow tracks increases. However, there are two puzzling findings. First, there is little evidence for any significant differences between high and low comms trials. The only significant findings were for 6 tracks with low yellows where SST-present is greater than SST-absent, and 24 tracks with low yellows where SST-absent is greater than SST-present. No SST differences are found for 12 or 18 tracks, with tracks and yellow level held constant. Second, the executive load gauge for the 24-aircraft condition seems different from other wave sizes. In terms of wave size, the 24 tracks condition is associated with greater executive load than all other wave sizes, at high and low yellow, and with or without SST. (The only exception is high yellows, SST absent, 24 versus 18 tracks.) However, the 24 tracks condition is associated with greater executive load for low versus high yellows, when SST is absent, and for SST absent versus present when yellows are low. In this sense, the trend in the data for 24 gauges is in the opposite direction to that for other wave sizes.

Verbal Load Reconsidered — Scenarios run with and without SST

We wish to make some further comments about the problems encountered in discriminating scenarios with and without communications. First, our analysis of responses to communication requests indicated that, for some participants, reaction time (RT) increased and accuracy decreased for scenarios with higher numbers of aircraft. However, RT for IFF did not reveal a significant main effect of SST, and therefore RT IFF did not discriminate between the presence or absence of the additional verbal task ($F(1,5) = 5.53$, $p>0.05$). However, one participant failed to respond to SST requests on all 24 track waves with either high or low yellows, and all 18 track waves with high yellows. When this participant is removed from the data set, the main effect of SST, for RT IFF, becomes highly significant, ($F(1,4) = 21.58$, $p<0.01$), with SST present associated with slower RT IFF.

Before moving on, it is instructive to consider RTs in the SST task. Data corresponding with SST trials were subjected to a two-way ANOVA, with two levels of yellow and four levels of wave size. We excluded the one participant who made no attempts to respond to SST at higher numbers of tracks, since the absence of any valid responses made the data unusable. We also calculated RT only on trials where the SST response was correct. In an analysis of the remaining five participants, we found a significant main effect of yellows ($F(1,4) = 10.67$, $p<0.05$). This reflected slower RTs during trials with a high number of yellow tracks (3165 ms) compared to the condition in which a lower number of yellow tracks were presented (2811 ms). In addition, the main effect of wave size was significant, ($F(3,12) = 16.19$, $p<0.05$). Overall, this reflected progressively longer RTs with higher numbers of tracks per wave (W6: 2366ms, W12: 2764ms, W18: 3401ms and W24: 3422ms). Post hoc contrasts revealed that, with the exception of 18 versus 24 tracks, all pairwise comparisons were significant. It is unfortunate that the design of the WCT does not enable us to determine whether these RT increments reflected longer processing time (e.g. memory search) or postponement of responses to a convenient period. The latter strategy is clearly likely, given that participants were not explicitly instructed to respond to SST interrupts with immediate effect. The strategy adopted by participants for dealing with the increased demand of the SST may have been effectively to

task-switch between the core WCT and the SST. This effect of task switching would not necessarily lead to increased executive load, except for the additional burden placed by the SST requirement for short-term memory. However we are unable to make estimates of the relative loads on these two tasks given the design of the study.

We also noted that scenarios run with the SST on produced very large differences in our measures of eye-movement activity, and this reflects the fact that responses to communication requests required a lateral eye movement to the left. Thus our eye movement systems were capable, albeit serendipitously, of discriminating verbal demand. For example, consider low frequency horizontal eye movement activity reflected in delta activity (derived from Fourier analysis). A three-way ANOVA revealed a highly significant main effect of ship status task, $F(1,5) = 37.18$, $p<0.001$, with the presence of communication requests associated with greater levels of horizontal eye movement. We observed a host of similar findings in other measures of eye-movement activity. A further finding of interest was that the near DC component for an aggregate of left posterior EEG channels was significantly associated with the presence vs absence of the ship status task: $F(1,5) = 125.91$, $p<0.0001$. This finding replicates increments in left posterior low frequency activity during verbally demanding tasks that we have reported elsewhere (e.g., Pleydell-Pearce, 1994).

Although we did not calibrate presence versus absence of the verbal ship status task, during previous studies we have found that activity within the theta bandwidth is particularly sensitive to verbal demand, and that the topography of this effect depends upon the modality used to convey verbal information.

GENERAL DISCUSSION

The executive load gauge appeared to perform well over the study, particularly with respect to the discrimination of number of tracks and percentage of yellow contacts. We were less able to distinguish between the manipulations of verbal load introduced by the SST. However, we feel that the gauge might have performed better if we had been able to include a calibration condition for the SST, and were not compromised by a participant who made no effort to respond to requests for information during SST performance.

If we are making claims for second-by-second accuracy in gauge readings, it is important to get a second-by-second view of actual task demand. These effects are all too easily masked by averaging data over extended periods of time. Indeed we take the view that the co-registration of behavioral and physiological measures will enable the fine-grained analysis of psycho-physiological variables, and may be one of the only viable methods for the examination of gauge performance.

The AugCog TIE provided a useful opportunity to demonstrate the capabilities of the CM system. In particular, the partial extension of the CM Task Taxonomy to the WCT was a very useful exercise. Here we have largely focused upon a single measure provided by CogPhys, the derivation of a measure of executive load based upon EEG activity alone. We have also presented results of our real-time Task Taxonomy system that can provide extensive information about cognitive activity and workload. Our decision to focus upon a single CM physiological measure has allowed a thorough methodological and statistical appreciation of the philosophy underlying our approach.

Our aim is to discover which variables predict load. This requires that measures be validated across test sessions, and we have shown that our bespoke variables are stable in repeat sessions separated by as much as two months (Pleydell-Pearce et al., 2003). Overall, then, our use of very large data sets is motivated by a neutral stance with respect to dependent variables. In our view, dependent variables should prove themselves, and we avoid *a priori* favoritism. This approach also means that we can embrace new variables, and indeed we are continually examining new mathematical approaches to data.

REFERENCES

Bonner, M. C., Taylor, R M., Miller, C. A. (2000). Tasking Interface Manager: affording pilot control of adaptive automation and aiding. In P. T. McCabe, M. A. Hanson, S. A. Robertson (Eds), Contemporary Ergonomics 2000. Taylor & Francis, London. pp70-74

Conway, M. A., Pleydell-Pearce, C. W., Whitecross, S. E. (2001). The neuroanatomy of autobiographical memory: A slow cortical potential study of autobiographical memory retrieval. Journal of Memory and Language **45**; 493-524.

Gratton, G., Coles, M. G. H., Donchin, E. (1983). A new method for off-line removal of ocular artefact. Electroencephalography and Clinical Neurophysiology **55**; 468-484.

Jasper, H.H. (1958). The ten twenty electrode system of the International Federation. Electroencephalography and Clinical Neurophysiology **20**; 371-375

McCallum, W. C., & Pleydell-Pearce, C. W. (1993). Brain slow potential changes associated with visual monitoring tasks. In W. C. McCallum and S. H. Curry (Eds.) Slow Potential Changes in the Human Brain. Plenum, New York. pp165-189

Pleydell-Pearce, C.W. (1994). DC potential correlates of attention and cognitive load. Cognitive Neuropsychology **11 (2),** 149-166.

Pleydell-Pearce, C. W., Dickson, B. T., Whitecross, S. E. (2003). Multivariate analysis of EEG: Predicting cognition on the basis of frequency decomposition, inter-electrode correlation, coherence, cross phase and cross power. Proceedings of the Thirty-Sixth Annual Hawaii International Conference on System Sciences. p 131 Abstr. IEE, Computer Society. The Printing House, USA.

Pleydell-Pearce, K., Dickson, B., Whitecross, S. (2000). Cognition Monitor: A system for real time pilot state assessment. In P. T. McCabe, M. A. Hanson and S. A. Robertson (Eds) Contemporary Ergonomics 2000. Taylor and Francis, London. pp65-69.

Pleydell-Pearce C. W., McCallum, W. C., Curry, S. H. (1995). DC shifts and cognitive load. In G. Karmos, M. Molnar, V. Csepe, I. Czigler, J. E. Desmedt (Eds). Perspectives of Event Related Potentials Research. Supplement 44 to Electroencephalography and Clinical Neurophysiology. Elsevier, Amsterdam. pp302-311.

Riding, R. J., Glass, A., Butler, S. R., Pleydell-Pearce, C. W. (1997). Cognitive style and individual differences in EEG alpha during information processing. Educational Psychology 17(1&2); 219-234.

Shadbolt, N. R., Tennson, J., Milton, N., Howells, H. (2000). Situation assessor support system: a knowledge based systems approach to pilot aiding. In P. T. McCabe, M. A. Hanson, S. A. Robertson (Eds), Contemporary Ergonomics 2000. Taylor & Francis, London. pp 60-64

Taylor, R. M., Bonner, M. C., Dickson, B., Howells, H., Miller, C., Milton, N., Pleydell-Pearce, K., Shadbolt, N., Tennison, J., Whitecross, S. (2001). Cognitive cockpit engineering: Coupling functional state assessment, task knowledge management and decision support for context sensitive aiding. Human Systems IAC Gateway. Vol XII: (1).

Wickens, C. D. (1987). Attention. In P. A. Hancock (Ed.) Human Factors Psychology. Elsevier Science Publications, Holland. pp 29-80.

Winer, B. J. (1971). Statistical Principles in Experimental Design. Second Edition. McGraw-Hill, USA.

Attachment 1: Examples of Bespoke Calibration Data Selected for Participant 1

Coherence: LF/F4: gammaH
Coherence: LF/F4: gammaA
Coherence: RF/F3: gammaH
Coherence: RF/F4: 84Hz
Coherence: LP/O1: beta1
Coherence: RP/O2: beta1
Coherence: F7/F3: gammaL
Coherence: F7/F3: 30-40Hz
Coherence: F7/F3: beta2
Coherence: F7/F4: beta2
Coherence: F8/F3: 84Hz
Coherence: F8/F3: gammaH
Coherence: F8/F3: beta2
Coherence: F8/F4: gammaL
Coherence: F8/F4: 30-40Hz
Coherence: F8/F4: beta2
Coherence: F3/F7: gammaL
Coherence: F3/F7: 30-40Hz
Coherence: F3/F7: beta2
Coherence: F3/F4: gammaH
Coherence: F3/F4: gammaL
Coherence: F3/F4: gammaA
Coherence: F3/F4: 30-40Hz
Coherence: F3/F4: beta2
Coherence: F3/P3: 84Hz
Coherence: F3/P3: gammaH
Coherence: F3/P4: 84Hz
Coherence: F3/P4: gammaH
Coherence: F4/F8: gammaL
Coherence: F4/F8: 30-40Hz
Coherence: F4/F8: beta2
Coherence: F4/F3: gammaH
Coherence: F4/F3: gammaL
Coherence: F4/F3: gammaA
Coherence: F4/F3: 30-40Hz
Coherence: F4/F3: beta2
Coherence: F4/T3: gammaA
Coherence: F4/T3: beta2
Coherence: F4/T4: 84Hz
Coherence: F4/P3: gammaH
Coherence: F4/P3: gammaA
Coherence: F4/P3: beta2
Coherence: F4/P4: 84Hz
Coherence: F4/P4: beta2
Coherence: T5/T6: 30-40Hz
Coherence: T5/O1: beta1
Coherence: T6/T5: 30-40Hz
Coherence: T6/O2: beta1
Coherence: O1/T5: beta1
Coherence: O2/T6: beta1
Spectral Power: LF: gammaH
Spectral Power: LF: gammaA
Spectral Power: P3: gammaA
Spectral Power: P4: gammaH
Spectral Power: P4: gammaL
Spectral Power: P4: gammaA
Spectral Power: P4: 30-40Hz
Spectral Power Ratio: Cz: beta1/beta2 (i.e. divisions of bandwidths within electrodes)
Note:
GammaL = 40=50Hz
GammaM = 50-70Hz
GammaH = 70-100Hz
GammaA = 30-100Hz
LF = Aggregate of left frontal electrodes
RF = Aggregate of right frontal electrodes
LP = Aggregate of left posterior electrodes
RP = Aggregate of right posterior electrodes
Residual electrode labels (e.g. T5, O1, F4 follow the convention of the International 10:20 system).

RESULTS: ICA ANALYSES

The PSE Report of the Technology Integration Experiment gives a comprehensive analysis of the overall results of the study. This section provides a further study of the Ship Status Component of the task and also focuses on cognitive workload analyses of individual participants as measured by the Index of Cognitive Activity. Because we tested with all four teams who participated in the TIE, we have labeled participants by team number and by participant number.

I. Data Problem And Resolution

The data analyzed were measures of pupil diameter acquired every 4 msec for each eye as the participant completed the scenario. Each scenario generated 225,000 data points for each eye.

During the TIE, we experienced significant interference in data recording that appeared in the eye data at a frequency of 10 Hz. Subsequent investigation by us and the manufacturers of our equipment detected a very small 10 Hz signal in the cameras themselves. The magnitude of this signal is usually extremely small and insignificant. For instance, we (and the manufacturers) have never before detected it in our evaluations, and we have tested over 500 individuals. Moreover, the manufacturers and their other customers have also never detected it.

However, at the TIE, the magnitude of the 10 Hz signal was approximately 30 times greater than normal. The cause of this large increase in amplitude is unknown. The worst instances occurred when we were testing with Teams 1 and 4, although there were isolated instances across all teams.

To evaluate the severity of the problem and to preserve as much data as possible, we initiated the following procedure. Blinks and blink artifacts were removed from the pupil signal using our standard algorithms. The data for each scenario were then divided into 900 1-second segments, each composed of 250 observations. Right and left eyes were analyzed separately. Each segment was examined for the presence of the abnormal 10 Hz signal using a fast Fourier transform. When the number of seconds with abnormal data exceeded 10% of the total scenario time, the entire scenario was removed from the analysis. For the remaining data, the seconds containing the abnormal signal were considered to be missing data and the pupil recordings made during these seconds were not analyzed. The data found to be within the acceptable boundaries of the 10 Hz signal were then analyzed to compute second-by-second Index of Cognitive Activity (ICA). Mean ICA values for each wave of each scenario were calculated based on the seconds with acceptable data.

All data for all participants were analyzed using this procedure, retaining those scenarios with fewer than 10% loss. We report here full data for seven participants: T1P4, T2P2, T2P6, T2P7, T3P4, T3P8, and T4P6.

II. Ship Status Task Analysis

Cognitive Processes

As reported earlier in this document, the Index of Cognitive Activity increased significantly for participants when the Warship Commander included the Ship Status Task. The Ship Status Task consists of two distinct types of audio communications:

- *Messages* provide information about a specific aspect of the ship (water supply, course change, radio frequency).
- *Queries* require the operator to recognize and convey specific information contained in a previous message.

Messages are given in three different voices (two male, one female), with each voice relaying information about a single feature such as water supply. Messages begin by naming the specific feature and then giving a value to it. For example, a typical message is "Fresh water level at six nine five." Messages are delivered in slow, deliberate tones.

Queries are given in a single male voice and begin with the statement: "This is the Captain" followed by a short question of the type "What is our fresh water level?" In contrast to Messages, Queries are delivered rapidly. Messages and queries last roughly 3 seconds each.

Messages and Queries differ in the amount and type of cognitive processing that they elicit. Both involve the cognition associated with detecting an auditory stimulus. Both require that the stimulus receive attention, and both necessitate translation of the stimulus into meaningful speech. However, the two types of communication differ with respect to their impact on working memory.

Messages can be considered to trigger the cognitive processes associated with storing information in working memory. When the Message arrives and is fully decoded, the individual either stores the pertinent information in working memory for a short time or fails to do so. The information will at best be retained only briefly, and other Messages may displace it or interfere with it. Furthermore, the information may be successfully stored in working memory but its trace may decay so rapidly that it is no longer available when needed. And, it is always possible that no attempt is made to store the information in working memory because the information is not attended to. An individual may simply ignore incoming Messages if workload from doing other tasks is too large. Individuals with a great deal of familiarity with the task might be expected to chunk information for storage in working memory according to the known categories that appear in Queries. Large numbers denote water levels, small numbers indicate ship's course, and call signs (Bravo, Echo) are communications channels. The experienced operator does not have to store all of the associated information—just the numbers and call signs.

On the other hand, Queries require additional aspects of cognitive processing by the participant. Once the Query is recognized and its content decoded, the participant must actively parse the Query to determine the desired target of information. Next, the participant must search working memory to extract any information that matches the target. And finally, the individual must match the working memory extraction to the correct information in the options list on the display and press a response button. Just as a Message may fail to receive attention, a Query may also be ignored. Some participants failed to respond to some of the Queries.

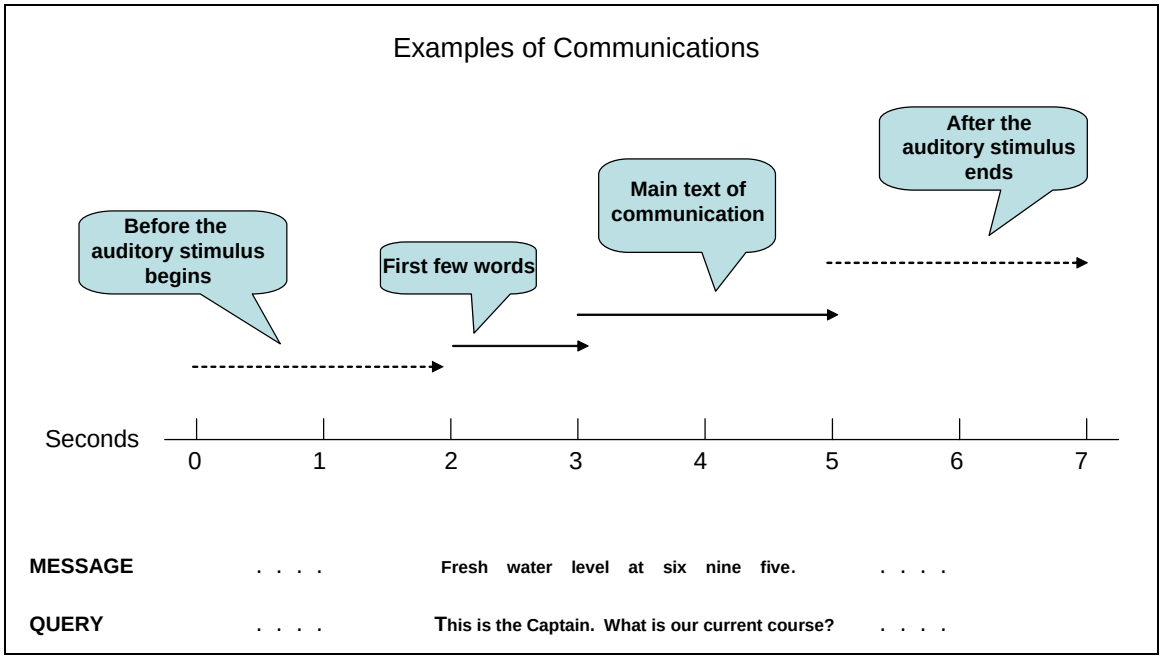
An experienced operator will learn to recognize quickly the difference between Queries and Messages from the opening words ("This is the Captain"). It is possible that this recognition triggers heightened attention to the auditory component of the Query and/or primes working memory to facilitate the search that is needed to answer the question.

There are a number of different questions we can ask of the data, including:

- Do Messages and Queries elicit the same levels of ICA? That is, is there a constant response to the audio stimulus?
- Is the ICA constant through the communication or does it vary?
- Does the ICA peak during the communication or after the communication has concluded?
- Does the ICA differ for incorrect or correct responses to Queries?

Method and Analyses

To study the relationship between the ICA and the hypothesized cognitive processes required by Messages and Queries, we looked at each communication over a 7-second time period, as shown in the following figure:



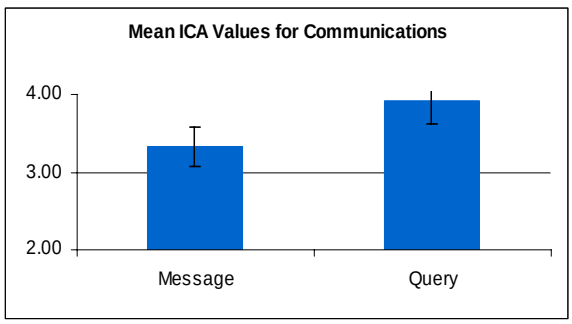
The 2 seconds prior to the initiation of the communication gives us information about the level of general workload occurring when the communication arrives. Once the communication begins, the first-second data lets us determine whether there is a rapid ICA response as the auditory stimulus is detected. The next two seconds require the decoding of the information contained in the communication. And the two seconds after the communication ends allows us to determine whether additional processing occurs after the conclusion of the communication.

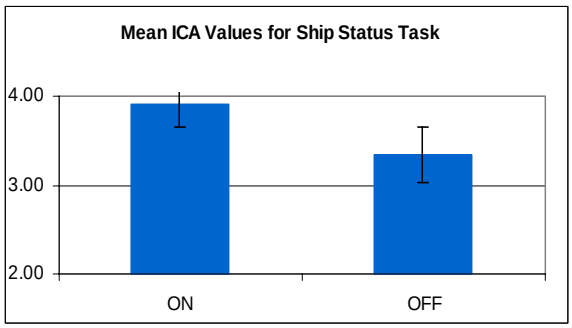
The operator is presented with 72 messages during each SST scenario, with 6 messages per wave. The operator responds to 36 queries in each scenario (3 per wave) based on the previous messages presented within the same wave. Each wave in the scenario spans 75 seconds. Messages and queries within each wave of each scenario occur at the following times:

Second	Communication
1	wave begins
6	1st message
12	2nd message
18	1st query
29	3rd message
35	4th message
41	2nd query
52	5th message
58	6th message
64	3rd query

For each participant on each scenario, mean values were computed for the 2 seconds prior to the target second associated with each communication, for the 2 seconds following the target second, and for the next 2 seconds after the communication has ended. One set of means was computed across the two scenarios that contained the Ship Status Task, and a second set of means was computed for the two scenarios that did not. Thus, we have ICA values for each participant's response to messages and queries across the four time periods of interest, and we have matching ICA values from exactly the same points in the same scenarios when the participant did not have to respond to the Ship Status Task.

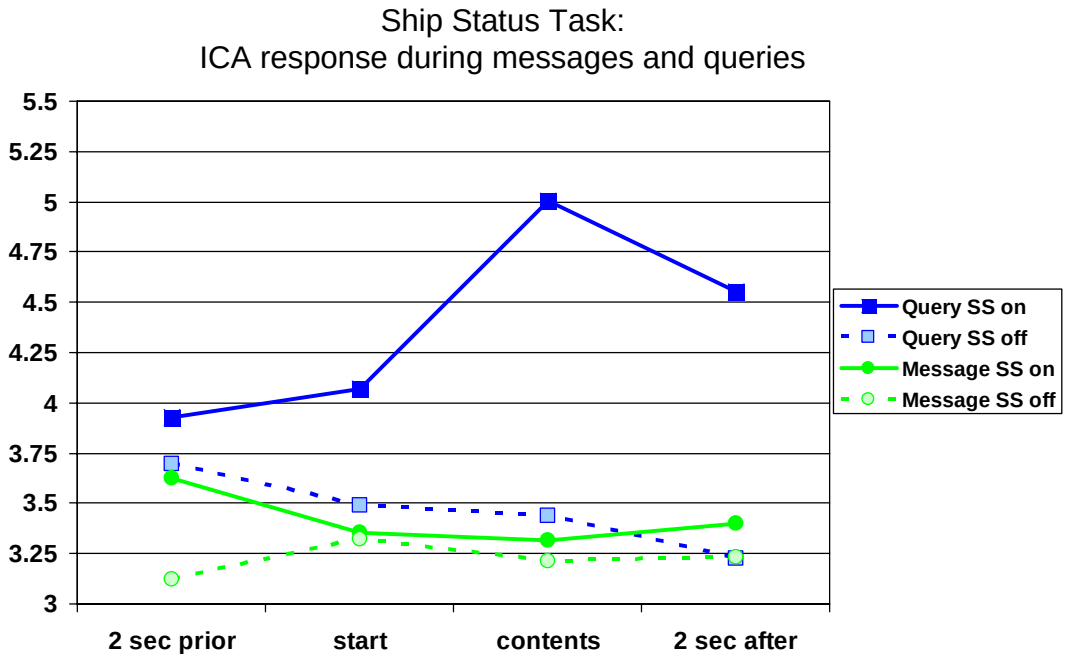
We analyzed these data using a 2x2x4 repeated measures analysis of variance. The factors were Ship Status Task (on or off), type of communication (Message or Query), and Timing (2 second pre-communication, initial second of communication, 2 second communication text, 2 second post-communication). Multivariate tests for main effects of Ship Status Task and type of communication were both significant, $F_{SS\ Task}(1,6)=9.21$ and $F_{Type}(1,6)=34.01$. The first finding confirms the earlier report that ICA is higher on scenarios in the presence of the Ship Status Task than in its absence. The second finding—the statistical significance between the two types of communications—is quite strong and confirms what we thought we saw during the TIE, namely, that the largest ICA responses appear in response to Queries, not to Messages. These results are shown in the figure below.





Of greater importance are the two significant interactions from the repeated measures analysis: the two-factor interaction of Communication by Ship Status Task (F=19.05, df=1,6) and the three-factor interaction (F=13.99, df=3,4). We focus here on the three-way interaction because it influences the interpretation of the two-factor interaction as well as the interpretation of the main effects.

The figure below graphically captures the important information in the interaction. The four timing intervals are shown on the X-axis, beginning with the 2 seconds prior to the time that a communication is given, the 1st second of the communication (labeled "start" in the figure), the 2 seconds containing the body of the communication (labeled "contents" in the figure), and the 2 seconds immediately following the communication.



The main result is obvious: the ICA responds quickly and strongly to Queries. The solid blue line is significantly higher than the other three lines for all timing intervals accept the 2 seconds prior to the message.

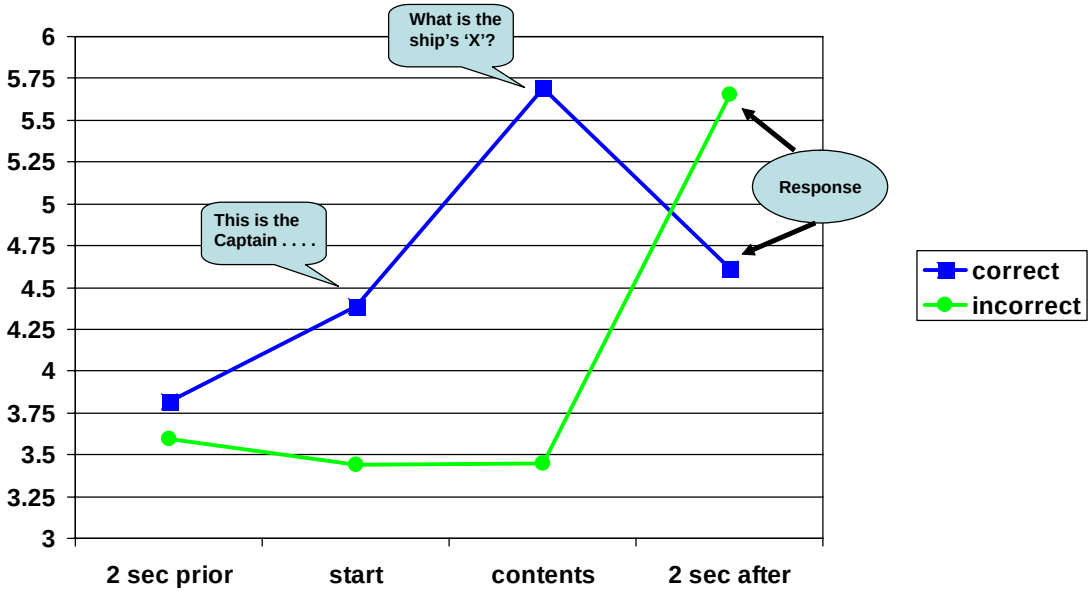
A question readily answered by this analysis is whether the relationship between the ICA and communications is constant across the two types of communication present in this task. The answer is clearly negative. The participants in this study showed significantly higher ICA in response to Queries than to Messages, as shown by the solid blue line versus the solid green line. This finding suggests that the ICA is responding to the specific cognitive demands of a Query and not just to the audio stimulus itself. If the ICA rose in response to the audio stimulus alone, we would observe high values for Messages as well as for Queries.

In terms of the cognitive processes involved, this analysis suggests that the ICA reflects the effort required to retrieve information from working memory rather than the effort to store information. The ICA for messages was no higher than the ICA computed for the same time periods when no audio stimulus was present.

What can we learn by looking at the ICA across the timing intervals for Queries that were answered correctly versus those that were not? To answer this question, we carried out a second analysis using

only the Query data. A 2 x 4 repeated measures analysis of variance included the factors of Correct (yes or no) and Timing (with the same 4 intervals as before). Results yielded statistically significant values for both main effects ($F_{Correct}$=6.3; df=1,6; p=.046 and F_{Timing}=12.46;df=3,4; p=.012). The interaction term fell just short of conventional significance at α=.05 ($F_{Interaction}$=6.00; df=3,4; p=.058). The figure below presents these results.

Ship Status Task:
ICA as a function of correct and incorrect responses to ship status



Participants responded correctly to most Queries, with an average of 26 correct responses. The mean ICA values were quite different for correct and incorrect answers, as shown above. When they answered correctly, participants had generally the same pattern shown in the overall analysis, namely, the ICA began to rise with the start of the Query and peaked during the text as the participant determined the focus of the query. In contrast, for the incorrect responses, the participants did not respond to the Query initiation or the text but had dramatically increased ICA during the response period that followed the Query.

These analyses were based on only seven participants across 4 scenarios each. The findings are strong in the data but should be interpreted cautiously given the small number of subjects and tasks.

III. Individual Participant Analyses

For each participant, we took the 48 mean ICA values for each wave as our data and examined the task structure, participant performance, and cognitive workload as measured by the Index of Cognitive Activity.

(1) HOW WELL DOES THE TASK STRUCTURE PREDICT THE INDEX OF COGNITIVE ACTIVITY FOR EACH INDIVIDUAL?

To answer this question, we regressed the ICA on four variables:

- Wave size (6,12,18, and 24)
- Task complexity (high and low, based on the number of unknown yellow tracks)
- Audio task (present or absent)
- Order (wave number within a scenario,1-48, included to examine fatigue effects)

Thus, we examined the common model: ICA = CONSTANT + β_1WAVE + β_2COMPLEXITY + β_3AUDIO + β_4ORDER for all participants. The regression in all cases was across the 4x12=48 waves of the test scenarios. For each participant, there were: 12 instances each for each wave size of 6, 12, 18, and 24; 24 instances each of high complexity and low complexity; 24 instances each of audio present and audio absent; and one instance each of possible wave order 1-48. The ICA was the average computed value for both eyes across data found to be free of the 10 Hz contaminations.

The model was statistically significant for all participants, and the proportion of variance accounted for by the model is shown in the table below for each participant.

participant	R^2
T1P4	.587
T2P2	.430
T2P6	.461
T2P7	.730
T3P4	.501
T3P8	.662
T4P6	.521
Overall average	.556

Thus, using only the structural information from the scenarios—wave size, wave complexity, presence or absence of audio task, and scenario order—we were able to account for a sizable amount of the variance observed in the ICA for every participant, ranging from a low of 43% to a high of 73%. These results confirm that the ICA reflects the underlying structure of the Warship Commander Task.

The model was successful for every participant. However, the importance of the independent variables was not necessarily constant. Each participant reacted to the WCT in his or her own way, and the coefficients of the independent variables are not necessarily the same across individuals. The table below shows how the individuals differed.

	wave	complex	audio	order
T1P4	***		***	***
T2P2	***			
T2P6			***	
T2P7	***		***	***
T3P4	***		***	***
T3P8	***		***	***
T4P6		***	***	***

We conclude that the best model in general for predicting the Index of Cognitive Activity for any individual will utilize all four task variables.

(2) HOW WELL DOES THE ICA PREDICT TASK DIFFICULTY?

We do not have a single measure of task difficulty, but the three variables of wave size, presence of the audio component, and complexity are all candidates. We examined each of them.

Three regressions were run with the data for each participant, with dependent variables of audio component, wave size, and complexity. (Note: Multiple linear regressions were run for all three analyses. Separate logistic regressions confirmed the results of the linear regressions for the two binary variables of audio and complexity. For comparability with wave size results, the results from the linear regressions for audio and complexity are shown here.)

The independent variables were ICA and RTIFF. Four previous studies by our research group have consistently shown that both variables have independent statistical significance for predicting task difficulty as measured by wave size.

The three structural variables of difficulty were related to ICA and RTIFF in varying degrees. The multiple correlation coefficients (R^2) values are shown below for the three analyses for all participants. All results reported are statistically significant at $p<.05$ unless otherwise stated.

	R^2 Audio	R^2 Wave	R^2 Complex
T1P4	0.28	0.54	ns
T2P2	ns	0.70	0.14
T2P6	0.33	0.62	0.17
T2P7	0.35	0.61	ns
T3P4	0.10	0.47	0.17
T3P8	ns	0.66	0.16
T4P6	0.44	0.46	0.14
Average	0.30	0.58	0.15

*ns indicates non-significance

The statistical significance of the independent variables for each of these tests is shown below:

	DV=Audio		DV=Wave		DV=Complex	
	ICA	RTIFF	ICA	RTIFF	ICA	RTIFF
T1P4	***			***		
T2P2			***	***		***
T2P6	***			***	***	***
T2P7	***		***	***		
T3P4	***		***	***		***
T3P8				***	***	
T4P6	***			***		***

Clearly, task difficulty as determined by the audio component is related only to the ICA; RTIFF did not reach significance for any participant in these analyses. In contrast, RTIFF was dominant in determining task difficulty as defined by wave size. The relationship between RTIFF and wave size is very strong for all participants. Similarly, task difficulty as defined by complexity is dependent upon both variables but as shown in the R^2 table above, the relationship with the independent variables of the model is much weaker than for audio or wave size.

How well does the ICA predict Performance?

In the main body of this report, six measures of performance were identified. Three are reaction time measures: time required to identify the track (RTIFF), time required to warn the track (RTWarn), and time required to engage the track (RTEng). In addition, the percentage of total possible score achieved for each wave plus two error terms were defined.

A simple examination of the correlations for each participant between the ICA and the six measures of performance is informative:

	T1P4	T2P2	T2P6	t2P7	T3P4	T3P8	T4P6
RTIFF	***	ns	ns	***	***	***	ns
RTWarn	***	***	ns	***	***	***	ns
RTEng	***	***	ns		***	***	ns
PctGS	ns	***	ns	***	***	***	ns
EC	ns	ns	ns	***			ns
EO	ns	ns	ns	***	***	***	ns

An inexplicable finding in this table is that for participant P6, ICA is not related to any measure of performance. This participant had some physical discomfort during his session with Team 4, but it would be unlikely to cause this result. It is interesting to note that the fNIR results as reported in the

preliminary report also show unusual results for P6. A second unusual finding is that the correlations for participant P8 are negative: as performance increases, his ICA decreases. Without further testing, we cannot explain the inverted results.

As can be seen in the table, for three participants, ICA is significantly related to 5 of the 6 performance measures. For two others, ICA is significantly related to 3 of the 6 measures. Thus, ICA appears to predict performance reasonably well for participants operating the WCT.

3H SARNOFF

Technical Accomplishment this Period:

Technical Interchange Experiment

The technical goal for the TIE in our case was to translate the single trial detection of evoked responses into a meaningful cognitive gauge as envisioned by the PM. To this end we assumed that for instance the ERN magnitude would be modulated by task difficulty. Therefore monitoring the magnitude of the ERN or related activity may give an indication of *perceived* task difficulty. Gevins proposed earlier a similar approach in the context of attention modulation using the well-known P300 activity. P300 is an EEG evoked response to an unexpected stimulus (typically auditory oddball), which is modulated by the available attentional resources of the user. We wanted to take a similar approach, but instead of using task unrelated stimuli we intended to use the response to the stimuli within the task. We explored a number of options and decided to focus on the response to the auditory warning signals as compared to other auditory feedback signals. The goal therefore was to detect a modulation of the response to warning signals with any of the task parameters. Our two primary findings from the WarCom experiments in San Diego are:

The differential activity elicited by warning signals as opposed to other auditory feedback had a frontal midline distribution similar to the ERN (see figure 1).
Its magnitude is correlated to task difficulty as measured by the number of errors the subject makes within a wave (see figure 2).

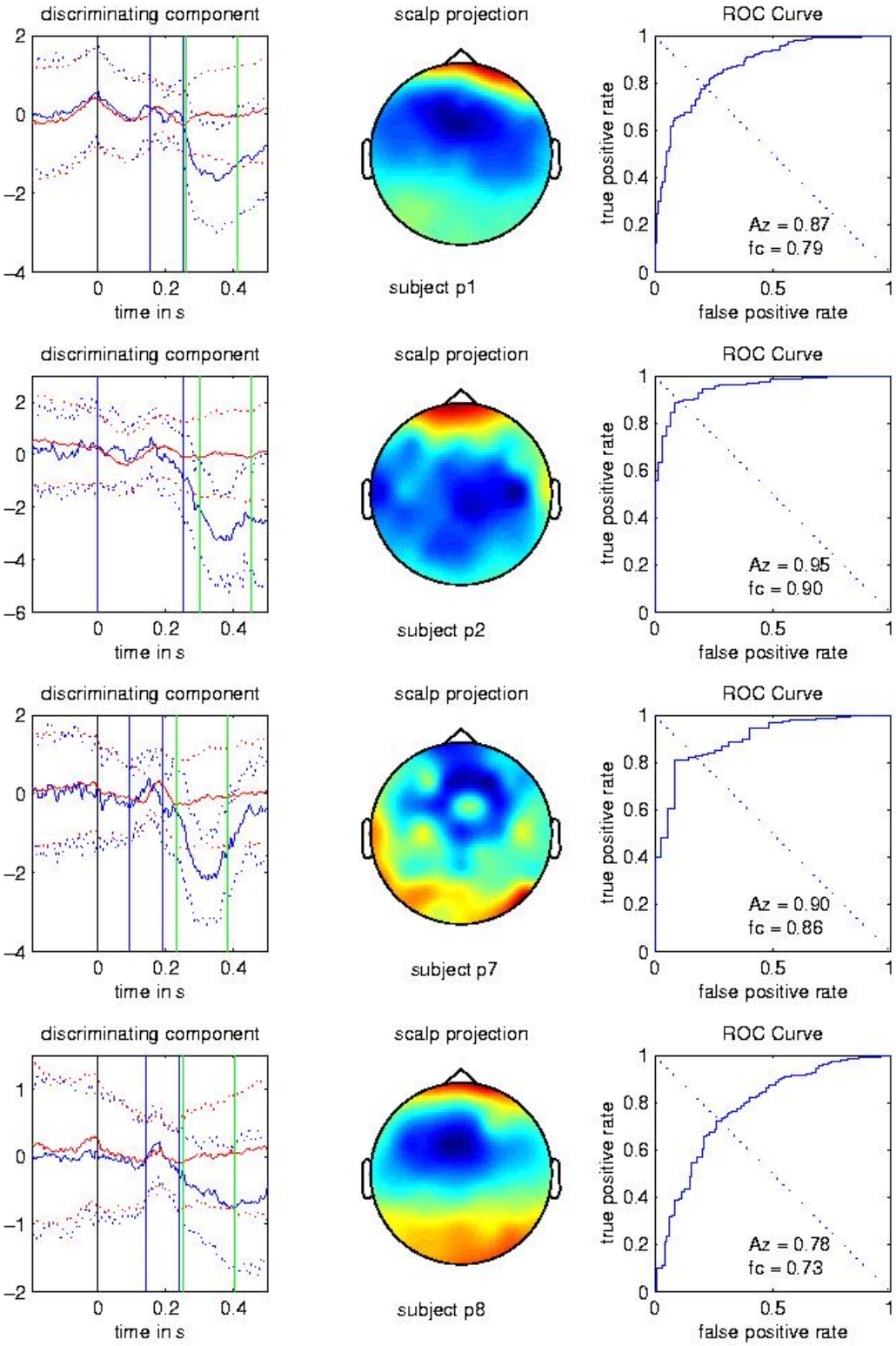


Figure 1. Time course and scalp distribution of the activity associated with warning signals as compared to other auditory feedback to user. Detection accuracy as indicated in the ROC curve give an indication of how reliable this activity is on a single trial basis. (Please ignore the scalp distribution for subject 7, which is in error due to technical difficulties during recording.)

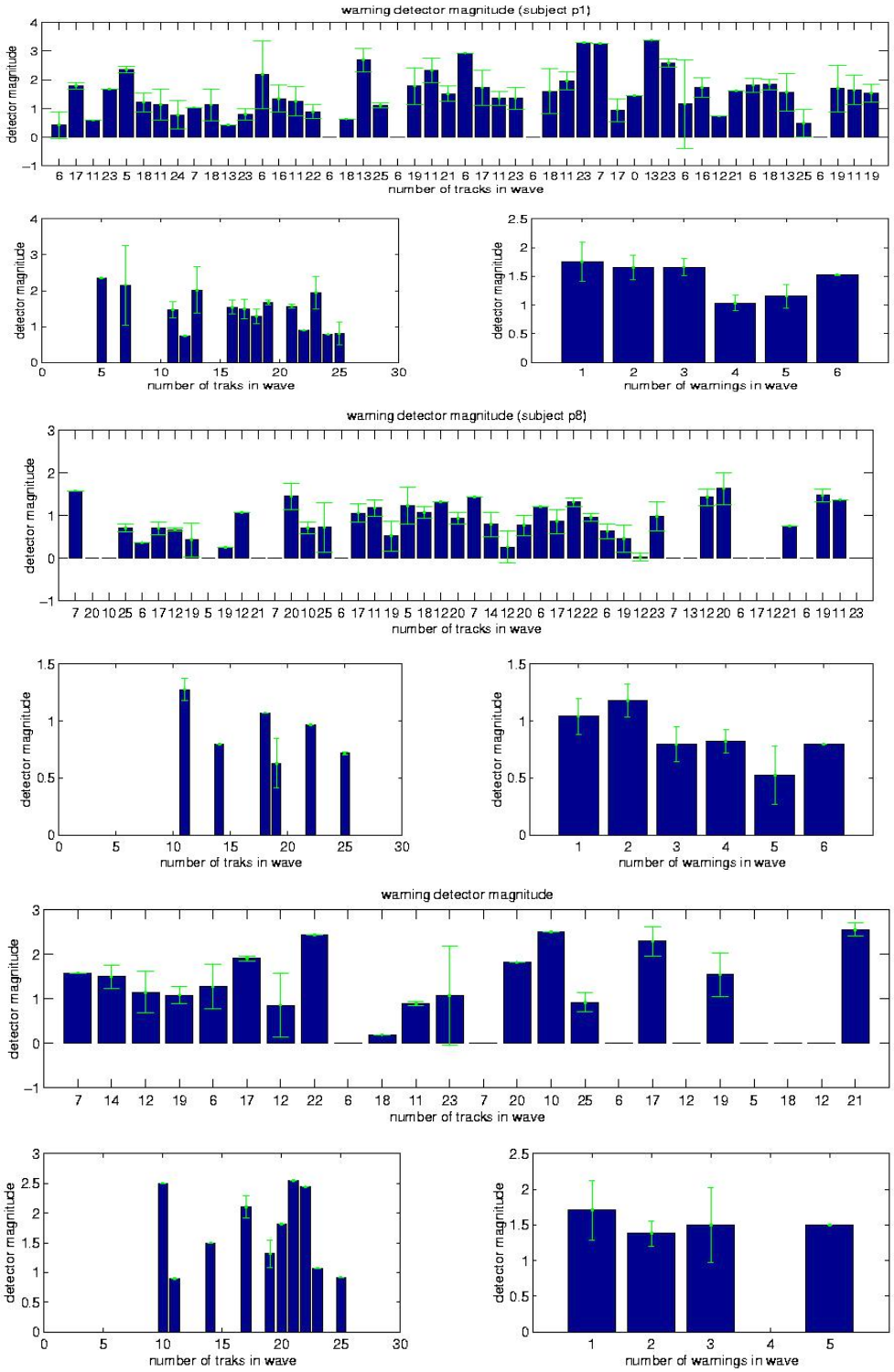


Figure 2. Intensity of the EEG response to warning signals averaged over all warnings within a wave of planes. The three charts for each subject show the average activity for each wave (across), the average for all waves with a given number of planes per wave (left), and average for all waves with a given number of warnings per wave (right).

Finally we would like to note that in the context of WarCom we had to develop a more robust eye-movement artifact removal algorithm. The WarCom task involves considerable eye motion, which makes most of the data unusable. Our new method uses a short calibration sequence of ca. 30 seconds where the subject is instructed to follow a cross on the screen. From this data one can extract projections for eye blinks, vertical, and horizontal eye motion. These projections are then used to subtract eye muscle artifacts as shown in figure 3.

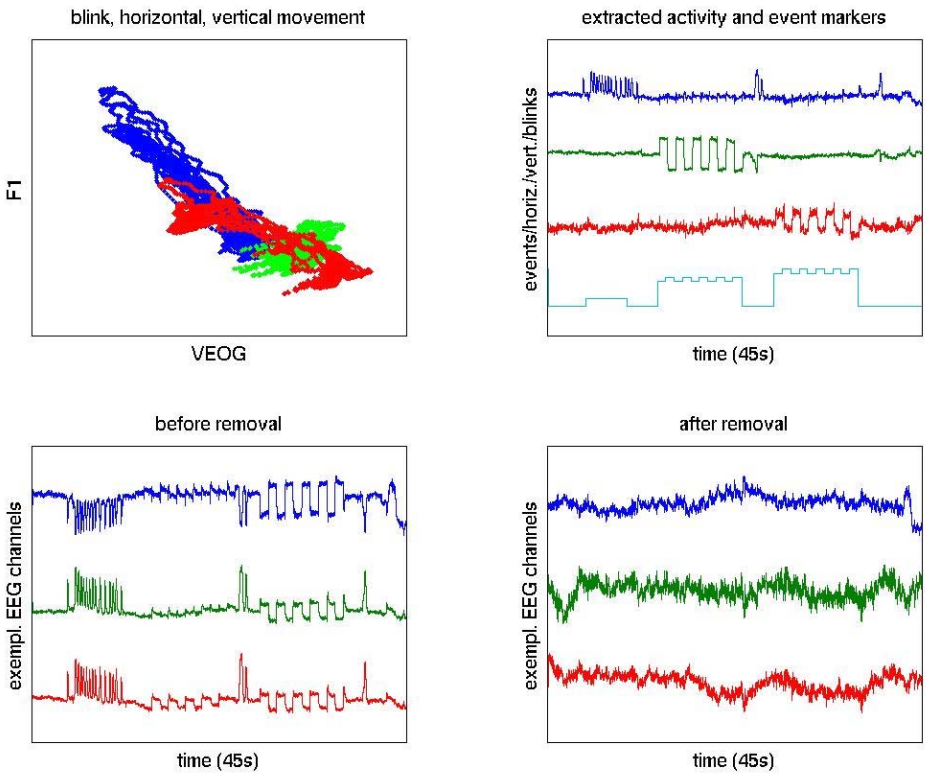


Figure 3: Eye muscle artifact subtraction. The subspace of the corresponding eye motion activity is estimated using a calibration sequence. Subsequently the activity is removed. *Top left*: Scatter plot of two of the 63 EEG coordinates color coded to represent the activity during eye blinks (blue), vertical eye motion (red), and horizontal eye motion (green). *Bottom, left*: Trace of EEG activity for three different EEG electrodes during the calibration sequence. *Top, right*: The projection of the 63 EEG coordinates onto the vectors that have been identified with the three different types of activity (same color code as before). The bottom trace reflects the instructions given to the subject during the recording of the calibration sequence. *Bottom, right*: Same as to the left but after subtracting eye-activity projections.

3I UNIVERSITY OF HAWAII

No Reports Submitted

3J UNIVERSITY OF NEW MEXICO

Introduction

We have devoted our effort to the understanding of brain activity measured by EEG while a human operator is performing a realistic and field relevant task (specifically the Warship Commander Task, WCT). The primary goal is to analyze data to determine the specific brain regions whose activity is particularly relevant to the performance of the operator. Typically, EEG data are collected under laboratory conditions, which depart drastically from any real world operational setting. Under such laboratory conditions, identical stimuli are presented numerous times under otherwise the same condition. Repetition of hundreds of trials and control of eye movements have been necessary for modeling the underlying neuronal sources. Furthermore, in a typical visual task, subjects are often asked to fixate on one particular stationary point on the screen to minimize associated variance.

In the real world, such tight control is not possible. Therefore, identifying underlying neuronal activation without these experimental constraints represents a major technical challenge. Previously, we applied second order blind identification (SOBI), an independent component analysis algorithm (Vigario et al 2000; Stone 2002), to continuous MEG data. We were able to recover neuronal sources of activation that could not be recovered using conventional methods of source modeling (Tang et al, 2002a) and we were able to measure single-trial response onset times from a relatively noisy data set (Tang et al, 2002b). Here, we extended this method of signal processing from MEG to EEG data, collected during the performance of WCT.

We aimed to recover neuronal sources of activation associated with aspects of WCT. We were able to overcome the difficulty in analyzing and interpreting EEG data obtained under poorly controlled non-laboratory conditions. Specifically, in contrast to most EEG/ERP studies that report sensor data, we derived the time course of activations from specific brain regions involved in the task and the spatial filters that allowed for spatial localization of these sources.

Methods

EEG Acquisition: Data was collected from 4 subjects performing the Warship Commander Task. 64 channel EEG data was collected at 500Hz acquired using NeuroCog Acquisition Software. EEG cap and amplifier were provided by Princeton University, Department of Psychology.

SOBI Analysis: Data was processed offline. First data was preprocessed using an independent component analysis (ICA) algorithm, second-order blind identification (SOBI). This ICA algorithm takes raw sensor data and recovers putative sources of neuronal activations. These sources were then epoched and averaged around the communications (comms, event 69) button press using the interval 1000ms before response to 3000ms after.

Spatial Localization: Components that showed response during this time period were localized using equivalent current dipoles. Components that localized to physiological and anatomically meaningful areas were then further processed to determine theta power.

Required Analysis: Theta (3.0625 – 7.8125Hz) power was calculated as the average power during the 2 seconds after the press of the comms button. Wave by wave means were obtained by averaging the theta powers retrieved during all comms event windows. Level by level means were obtained by averaging the theta powers over all waves with the same number of tracks, 6,12,18,24.

Results

We summarized main findings in the following files:

1. PowerPoint Presentation named tang_unm_Q1_2003.ppt

2. Excel Workbook named level means xls, containing:

- Event by event theta power, representing the theta power at the time of each comms button press in each scenario
- Wave by wave averages for the theta powers of the comms button presses
- Level by level averages for the theta power of the comms button press
- For control, we also included theta power averages for noise sources that showed no response during the time window around the comms button presses.

Conclusions

We were able to identify three types of neuronal sources of activation during WCT

1. Frontal-ocular (FO) activation that appears to be associated with eye movement

2. Occipital-parietal (V) activation indicative of typical early visual system processing

3. Synchronized anterior-posterior (SAP) activation that may indicate the activation of an attentional network

These three types of sources can be reliably identified across subjects and across task scenarios.

As predicted theta power of the frontal-ocular source (FO) under the high-yellow condition is greater than theta power under the low-yellow condition**.** Theta band activity may be particularly relevant to the generation of ocular motor output during the active monitoring and maintenance of the visual scene required by WCT

For the synchronized anterior-posterior (SAP) source, theta power under the high-yellow condition is less than theta power under the low-yellow condition. Communication between the anterior and posterior parts of the brain during the WCT may be supported by activity within frequency bands other than theta.

References

Stone JV (2000) Independent Component Analysis: an introduction. Trends in Cognitive Sciences 6(2):59-64

Tang AC, Pearlmutter BA, Malaszenko NA, Phung DB, & Reeb BC (2002a) Independent Components of Magneto encephalography: Localization. Neural Computation 14:1827-1858

Tang AC, Pearlmutter BA, Malszenko NA, & Phung DB (2002b) Independent Components of Magnetoencephalography:Single-Trial Response Onset Times. NeuroImage 17: 1773-1789

Vigario R, Sarela J, Jousmaki V, Hamalainen M, & Oja E (2000) Independent component approach to analysis of EEG and MEG recordings. IEEE Transactions on Biomedical Engineering 47(5):589-593

3K UNIVERSITY OF PITTSBURGH

Dynamic Postural Assessment Chair (DPAC)

Carey Balaban, Jarad Prinkey, Mark Redfern, Joseph Cohn, Roy Stripling

Introduction

This project is based upon the hypothesis that changes in automatic behaviors (e.g., postural adjustments and respiratory rate) are advantageous for assessing cognitive awareness in a warfighter. This approach takes advantage of the intricacy of these coordinated movements and benefits from the fact that measures of these behaviors are minimally intrusive. This hypothesis derives in part on recent observations demonstrating an impact of cognitive load on postural control. Dault et al (2001), for example, showed that varying levels of a modified Stroop task degraded postural adaptability, while Andersson et al (1998) demonstrated degraded postural control that corresponded to heightened mental activity on a mental rotation task. Importantly, these results highlight the reciprocal nature of this coupling. While postural stability was degraded with increasing mental task complexity, mental activity was impaired by increasingly difficult iterations of the dynamic posture tests. These observations suggest that gauges measuring postural behavior will be particularly important for warfighters aboard ships, planes, or moving ground vehicles.

For our purposes, the postural movements of our warfighters can be decomposed into those that are 'automatic', 'voluntary' and 'noise.' Analysis from data collected at the TIE indicated that gauges based upon the automatic movements are able to detect decrements in vigilance and cognitive capabilities. The first suite of gauge being developed based on this data set uses seated postural bracing and head stabilization as a first-line metric of changes in functional cognitive state. We also envision this gauge as an early level filter to cue specialized processing of information from other gauges. In the future, we plan to use postural sway responses in moving environments to derive other "higher level" gauges that parse predictive ("scheduled") versus reactive ("event-related") components of seated postural control.

Materials

The basic system consists of an Operator's chair from Lockheed-Martin's Sea Shadow ship. The chair has been reupholstered with slipcovers containing 16X16 pressure sensor arrays, both in the seat bottom and the seat back. A Flock of Birds (Ascension Technologies) tracker was placed on each subject's head and chest. Both sets of sensors provide insight into underlying postural behavior. The pressure pad sensors provide a direct indication of overall postural based 'bracing behaviors'. This information is, for the current data set, most meaningful for second by second time scales. The Flock of Birds system, gives a direct indication of head movement associated with monitor engagement. It provides spatiotemporal resolution of small dynamic postural changes, as well as larger time-scale information.

Methods

Head Movement: Flock of Birds Data. These data are provided as a series of worksheets within the 'Gauge summary' file. Two 'gauges' are provided at the wave-level ('Wave Ave.' worksheet), a 'Monitor Coupling Response' (Gauge 1), which is the wave-level magnitude (from a least squares fit of the position data: offset - magnitude * exp(-t/20)) divided by the root mean square residual (RMS), and a measure of wave-level 'Control Variability' (Gauge 2), which is 1/RMS. The second by second data for these two 'head gauges' are also provided in this excel file ('Sec by Sec' worksheet), as gauges 1 and 2

(the reciprocal of the second-by-second reciprocal standard deviation of a de-trended second of data, an analog of the algorithm used to calculate variability for the wave-by-wave measure).

Back Pressure Pad Data: The 'accuracy of back bracing' gauge in the 'Wave Ave.' worksheet is the probability of an low bracing (Back Bracing Gauge from the Second by Second Worksheet value greater than 0.2) coinciding with a change in Task Pending value between zero and –2 over the previous second. (The session values of this gauge are summarized as a supplement in the 'low back bracing & pend task Worksheet'). Two gauges are also provided from the seat back within the "Sec by Sec' worksheet. The pressure data from each sensor in the array were first differentiated in the time domain to yield a matrix of instantaneous changes in pressure at each sampled time point (rate: 4.58 Hz). The 'Back Bracing' gauge (Gauge 3) is the standard deviation of these pressure changes across the entire pad (256 sensors) at each time point, divided by the maximum pressure during the session (a form of coefficient of variation). The second-by-second gauge values for the file were obtained by resampling these standard deviation values (at 1 sec intervals) with a cubic spline algorithm (MATLAB spline.m). The 'Back Contact' gauge (Gauge 4) is simply a second by second count of the number of sensors with which the subject's back is making contact.

Results

Head Movement: As seen in figure 1 below, AP head position accurately marks the onset/offset of each wave during a round of Warship Commander. The time course of AP head movement during the onset of each wave can be fit with an exponential curve enabling the tightness of fit to this curve (estimated by the rms deviation from the curve value) to serve as a measure (or gauge) of engagement. Figure 2 illustrates this concept by depicting the time course of head position over two different waves in the Warship Commander task. Note the easier wave (left panel, with 6 targets) fits the curve poorly, while the more challenging wave (right panel, with 24 targets) conforms well to the fitted curve. Both the 'monitor coupling response' and the 'control variability' gauges yield larger values for the more challenging wave.

Back Pressure Pad Data. Empirical analysis of the data produced by the 'back bracing' gauge indicates that values less than 0.2 are indicative of strong bracing (low sd from one instant to the next), a value between 0.2 and 0.4 indicates weak bracing and a value greater than 0.4 indicates movement. In figure 3, second-by-second values for both 'targets on screen' (top) and 'targets pending' (bottom) are plotted vs time during the Warship Commander session. Back bracing gauge values greater than 0.2 (indicating weak bracing to movement) are indicated by the red '+' symbol, while values lower than 0.2 are shown as black dots. Note that low back bracing primarily occurs in between waves and as the task load is declining (measured either as targets on screed or as tasks pending). A value of greater than 0.2 on the back bracing gauge indicated with approximately 80% probability that there was a change in tasks pending between 0 and –2 (i.e., stable or steadily dropping workload) during the previous second. The data are summarized by condition in the Excel spreadsheet in the 'low back bracing & pend tasks' Worksheet.

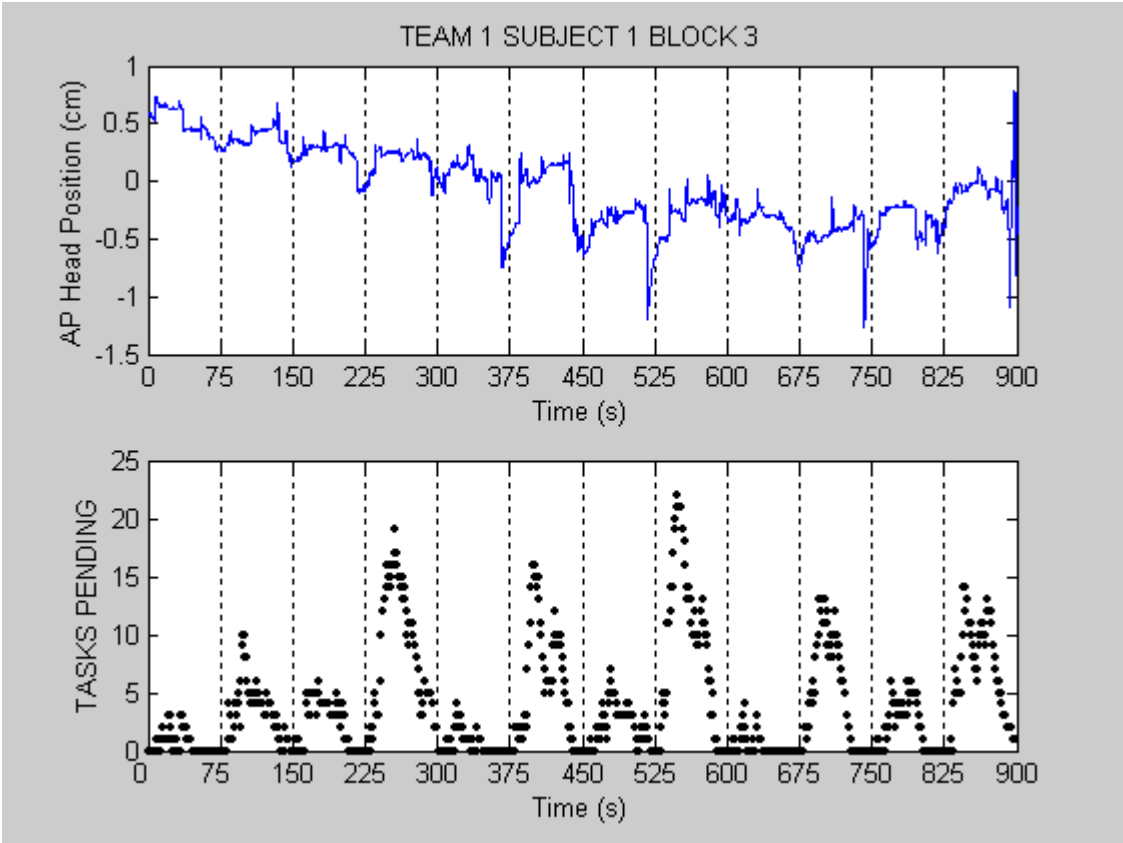


Figure 1.

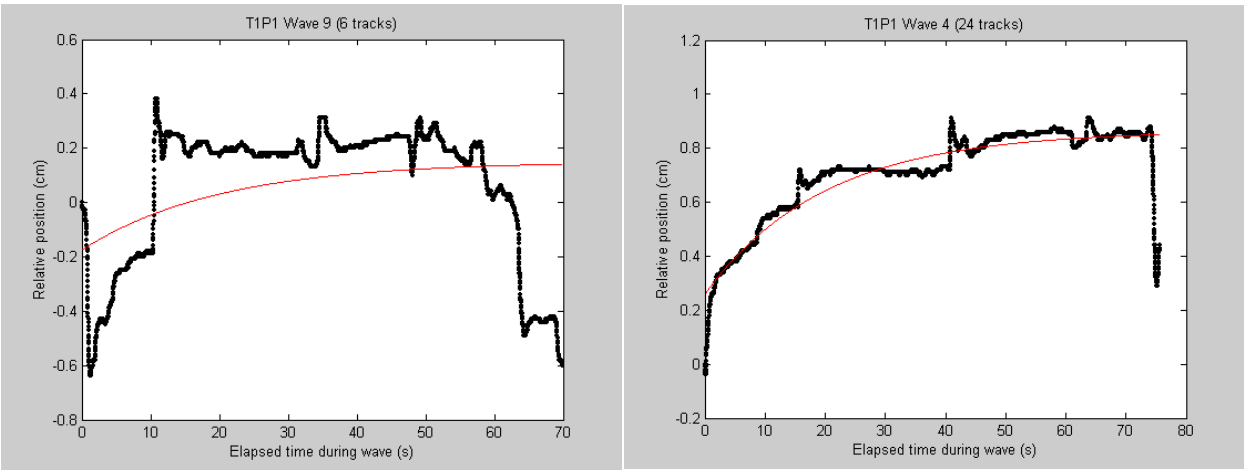


Figure 2.

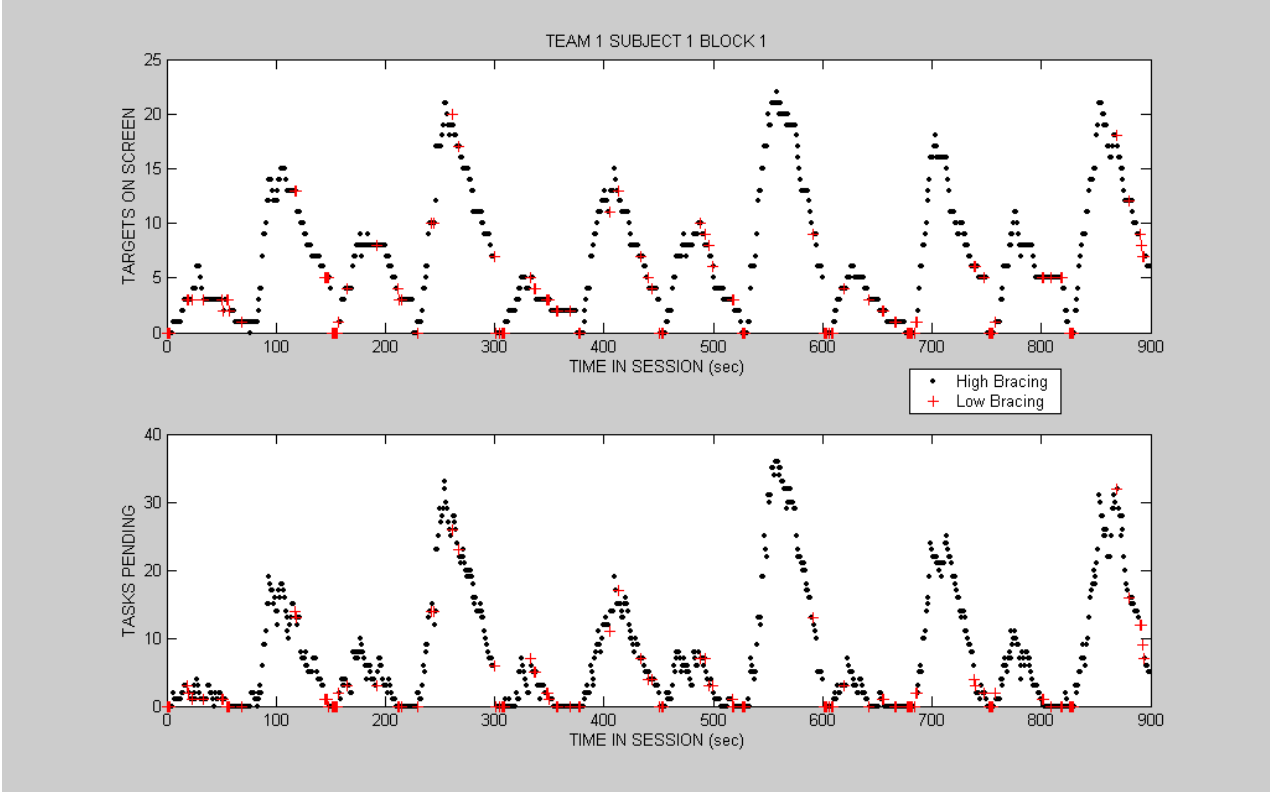


Figure 3.

Discussion

For Augmented Cognition systems, accurate and reliable assessment of cognitive state is an essential first step. However, no individual sensor (and possibly no suite of sensors) can unambiguously detect a discrete cognitive state. Each sensor (or suite of sensors) carries some level of uncertainty, but this uncertainty can be greatly limited by evaluating the data in the functional context under which they were collected. Figure 4 illustrates this point by demonstrating that high back bracing gauge values (indicating subject body movement) are observed both when tasks are increasing as well as when tasks are decreasing. Because of its much stronger correlation both with low and decreasing task loads (see excel worksheet, Wave-by-Wave Gauge 3: 'Accuracy of Back Bracing Gauge'), high gauge values appear to signal disengagement between the subject and the Warship Commander Task. Just as drivers traveling a winding road brace themselves until reaching stretches of straight roadway, operators of Warship Commander brace until they perceive an easing of the workload. This suggests that the rare occurrence of high gauge values during increasing task loads are indicative either of a misperception on the part of the subject (erroneously anticipating a decrease of workload, when an increase is occurring) or of an overwhelming urge on the part of the subject to reposition him/herself in the face of increasing challenge (which may in turn indicated the onset of overwhelming fatigue, frustration, irritability, or other form of distraction).

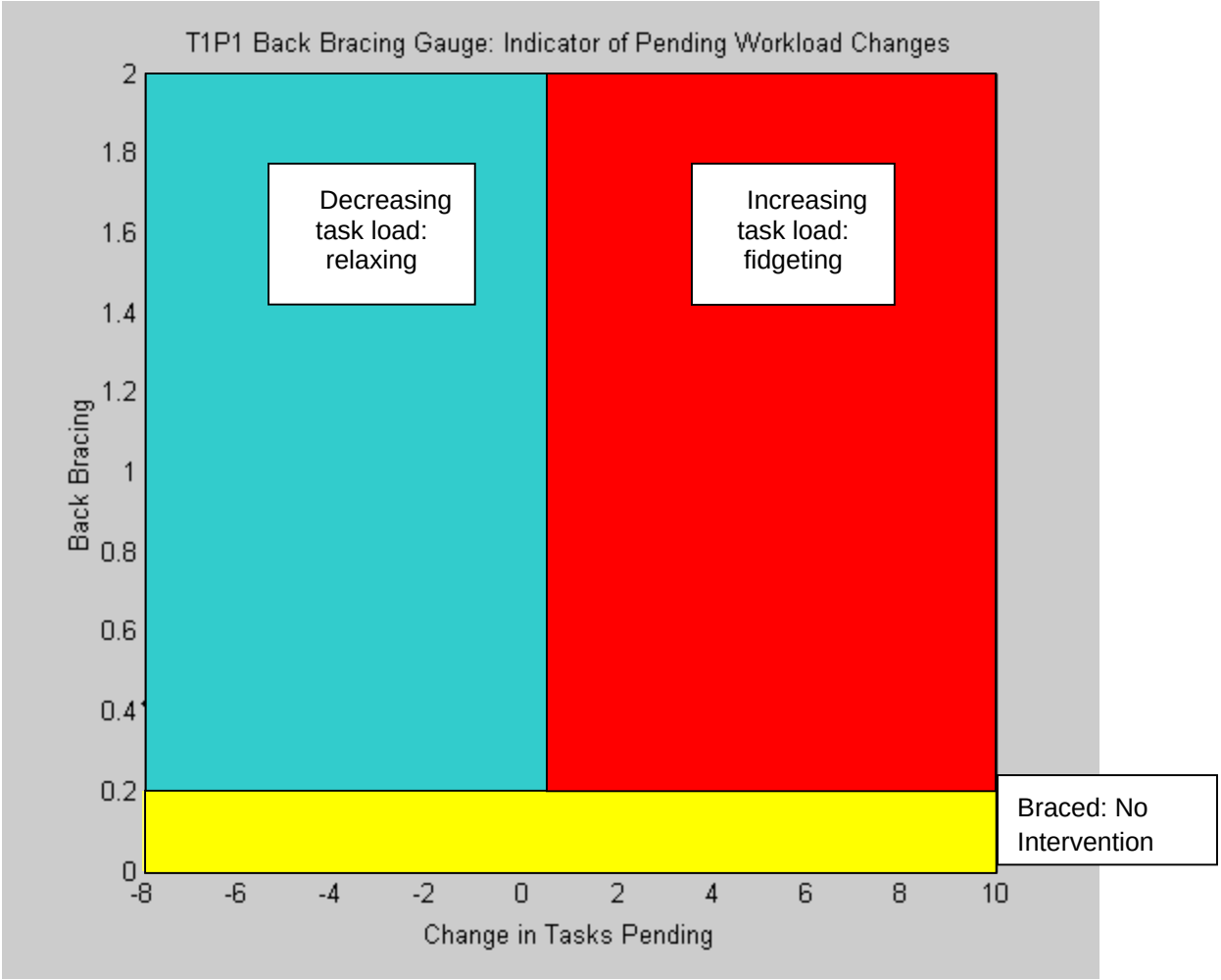


Figure 4

Clearly, additional research is needed to validate the cognitive correlates of this and the other posture-based gauges, but even with this simple characterization one can envision closed-loop applications: where back bracing gauge values are high AND task load is increasing, apply measures to reduce workload; where this gauge's values are high AND task load is low, increase workload to prevent boredom and to maximize productivity; finally when this gauge's values are low no intervention may be necessary, as the operator is well engaged in the task at hand.

References

Andersson G, Yardley L, Luxon L. 1998. A dual-task study of interference between mental activity and control of balance. Am J Otol 19:632-637.

Dault MC, Frank JS, Allard F. 2001. Influence of a visuo-spatial, verbal and central executive working memory task on postural control. Gait Posture 14:110-116.

APPENDIX 4: A COMMENTARY BY ALAN GEVINS AND MICHAEL SMITH

A Commentary on the DARPA AUGCOG Phase I TIE

Alan Gevins & Michael Smith

SAM Technology & San Francisco Brain Research Institute

The body of work represented by completion of the Technical Integration Experiment, and more generally by Phase I of the Augmented Cognition program as a whole, is usefully seen in its historical context. The problem of user overload in the context of modern, information-rich computer-based work environments has long been recognized by the human factors and ergonomics community, and as a result it has been the topic of extensive empirical investigation. In fact, over two decades ago DARPA's Biocybernetics program aimed to improve human-system performance using EEG brain function monitoring. Although that program did not lead to the desired thought-control of fighter planes, it did help substantially advance research on brain event-related potentials, and focused attention on the importance of trying to understand how cognitive brain activity is affected by performance in complex and stressful operational environments. Even so, despite many years of research effort and a great deal of theoretical discussion, no general consensus has yet emerged on the two most basic issues in the study of cognitive workload: the intertwined problems of how to define it and how to measure it. That this field of inquiry is lacking in clear central guiding principles and standardized methodologies is evidenced by the diverse array of measurement approaches included in the TIE. An important strength of the Augmented Cognition program to date has been the fact that it has not gotten bogged down in this amorphism and intellectual baggage, and instead has encouraged performers to pursue many independent creative directions. This has allowed performers to usefully revisit older problems and approaches, and to accelerate the development of new ones. Indeed the AugCog Phase I program has succeeded in re-energizing a critically important area of science and engineering in a very short amount of time. Given the lack of a pre-existing gold standard for either the "what" or the "how" of cognitive workload assessment, a particularly notable aspect of the TIE experiment was the fact that it brought together many people, some with little prior experience or preconceptions about how to best measure cognitive workload, arousal, or alertness, and succeeded in getting them to coordinate their diverse efforts and work hard on the problem of assessing these qualities of a human performer during a complex videogame that simulates a demanding operational setting. While this type of problem has been successfully addressed in previous studies using unimodal approaches, the TIE represents a unique accomplishment in terms of its multimodal, collaborative nature. Its eclectic approach provided a context for identifying cognitive assessment techniques, alone and in combination, which might be gainfully employed in an adaptive automation context to improve overall performance of complex human-controlled systems.

While completing an integrated physiological data collection effort under the highly demanding circumstances of the TIE represents an impressive technical achievement in and of itself, the scientific results produced by this exercise must be seen as ambiguous at this time. In this regard, some performers' self-assessments of their results may prove to be somewhat overly optimistic when subjected to critical peer-review. The behavioral results suggest that the various task manipulations

(number of tracks per wave, variation in track difficulty, and imposition of a secondary verbal task) all succeeded in modulating task difficulty. Despite this, there was a surprising amount of variation in the degree to which the outputs of the various gauges covaried with those manipulations. And, to the extent that outputs of some gauges did appear to respond to the task manipulations, it is difficult to know what exactly was driving those responses. In part this uncertainty arises from unknowns with respect to what the gauges are actually measuring-- in some cases the nature of the input signals is not very clear, and in all cases the potential exists for contamination of the intended signals by spurious artifactual sources of physiological and instrumental variance. More generally though, interpretation is clouded by the fact that in the context of the Warship Commander Task (and most real world tasks for that matter) changes in visual, motor, and auditory complexity are directly related to manipulations of task difficulty. For instance, as a "thought experiment" it is instructive to imagine what type of outputs might have been produced by a set of "cognitive workload" gauges consisting of the integrated output of a set of photodiodes attached to the video monitor, a software agent counting mouse clicks and trackball movements, and a microphone placed near the subject. How much useful additional information would remain in the output of the brain-, other-body-organ-, or behavior-based workload gauges after regressing out variation directly measurable in such physical parameters of the test environment? And, which approaches would be most reliably capable of providing this type of independent information? Answers to these questions seem critical for assessing the value-added by subject-centered cognitive workload assessment techniques such as those employed in the TIE, as well as for defining the most promising paths for future development. While the preliminary results described in this report are very promising, the critical answers will have to await further research. Despite such lack of definitiveness, the results reflect an ambitious, novel, and important undertaking that deserves broader dissemination in the peer-reviewed literature.

As a final observation, it is also important to recognize that although newer neuromonitoring technologies are clearly very exciting, they too are prone to significant potential confounds and interpretational difficulties…as are all the more established modalities. For example, to date Near Infrared Spectroscopy has received mixed reviews in clinical brain monitoring contexts. Some of this skepticism comes from the fact that at short distances between emitters and detectors the cortical NIR signal can be confounded by changes in oxygenation of scalp muscles (Germon et al, 1994). While such confounds are less of a concern with greater emitter-detector separation (Germon et al, 1998), increased separation decreases the overall signal-to-noise ratio and is influenced more by blood flow beneath the cortical surface. Even at some "optimal" emitter-detector separation NIR signal strength is affected by regional differences in the thicknesses of the skull and scalp tissues and in the geometry of the brain surface. This problem is compounded by (or perhaps contributes to) the finding of poor levels of consistency in between-subject measurements even under carefully controlled conditions (Henson et al., 1998), and periodic reports of higher levels of brain oxygenation in dead subjects than in some healthy control subjects (Kytta et al, 1999; Schwarz, 1996)! Such complexities do not detract from the intrinsic value of further developing advanced brain monitoring tools. They do though suggest that the timeframe for transitioning such measures into routine use in the context of augmenting human-system cognitive capabilities may be longer than is sometimes appreciated.

References

Germon, T.J., Kane, N.M., Manara, A.R., Nelson, R.J. (1994). Br. J. Anaesth., 73: 503-506.

Germon, T.J., Evans, P.D., Manara, A.R. et al. (1998). Clin. Monitor. Comput., 14:353-360.

Henson, L.C., Lalalang, C., Temp, J.A., Ward, D.S. (1998). Anesthesiology, 88:58-65.
Kytta, J., Ohman, Tanskanen, P., Randell, T. (1999). J. Neurosurg. Anesthesiol., 11:252-254.
Schwarz, G., Litscher, G., Kleinart, R., Jobstmann, R. (1996). J. Neurosurg. Anesthesiol., 8, 189-193.

APPENDIX 5: GLOSSARY

Alpha Level: Established levels of chance used for computing statistical tests. An alpha level of 0.05 is generally used as the cutoff for statistical significance. For exploratory studies, an alpha level of 0.1 is sometimes used to indicate "marginal" statistical significance.

Analysis of Variance (ANOVA): statistical test that computes the chance that the observed differences between conditions are either real or due to random fluctuations. The test computes an "*F ratio*" of the variance within conditions to the variance between conditions. The "*p*" value associated with the *F* is the probability that the observed differences between conditions are due to chance. *P* values less than 0.05 are considered "statistically significant."

AugCog: Augmented Cognition

Correlation (r): measures the strength of the linear relationship between *x* any *y*. The stronger the correlation, the better *x* predicts *y*.

EEG: Electrical Encephalography

ERP: Event-Related Potential (of an EEG signal)

Errors of Commission (EC): Number of errors committed during a wave of a scenario in the Warship Commander Task.

Errors of Omission (EO): Number of tasks neglected during a wave of a scenario in the Warship Commander Task.

Eta Squared (η^2): Effect size. The proportion of the variability in the dependent variable that can be accounted for by the variation in the independent variable. Thus, the larger the Eta Squared value the greater the degree to which the variation in the measure is attributed to the different levels of the independent variable. Eta Squared is computed by the Analysis of Variance using the ratio of the sum of squares effect / sum of squares total.

fNIR: functional Near Infra-Red imaging. Gauge that uses LEDs and photodiodes as sensors to collect data.

GSR: Galvanic Skin Response

Latin Square: method for assigning the order of conditions within a test session to eliminate order effects by varying the order of conditions across participants.

Line of Engagement (LOE): Horizontal red line across the screen in the Warship Commander Task that designates areas where warning and engaging tracks can take place.

Number of Tracks per Wave: task load factor manipulated in the experiment – the number of aircraft (tracks) appearing in a wave during a scenario of the Warship Commander Task.

Percent Game Score (PctGS): The percentage of total game points for a wave that a participant was able to accumulate.

***p*-value:** the probability that the observed differences between experiment conditions is due to chance. *P* values less than 0.05 are considered "statistically significant."

Repeated Measures Design: experiment and statistical test design in which all factors are presented to each participant.

RTEngage: Response Time to Engage. The mean time from when tracks became eligible for engagement until the participant selected each track and pressed the Engage button.

RTIff: Response Time to Identify Friend or Foe. The mean time from when tracks appeared on the screen until the participant selected each track and pressed the IFF button.

RTWarn: Response Time to Warn. The mean time from when tracks crossed the LOE and became eligible for warning until the participant selected each track and pressed the Warn button.

Secondary Verbal (Memory) Task: task load factor manipulated in the experiment. The presence or absence of the Ship Status Task which required verbal and memory processing.

Ship Status Task: The name of the secondary verbal memory task in the Warship Commander Task.

Tasks Pending (Pending): Sum of tasks pending across each second of a wave.

TIE: Technical Integration Experiment

TLX, NASA TLX (NASA Task Load Index questionnaire): Questionnaire for measuring subjective workload on a task. The NASA TLX is a multi-dimensional rating procedure that derives an overall workload score based on a weighted average of ratings on six subscales. These subscales are Mental Demands, Physical Demands, Temporal Demands, Own Performance, Effort, and Frustration.

Track: aircraft

Track difficulty: task load factor manipulated in the experiment – percentage of difficult yellow tracks appearing in each wave.

Variance (σ): a description of the distribution of data values around the mean of all values. It is calculated by summing the squared differences between each data point from the mean, divided by the sample size minus one. A small variance indicates a low amount of variability in the measurement (which suggests uniformity).

WCT: Warship Commander Task

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-01-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden to Department of Defense, Washington Headquarters Services Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*
12–2003

2. REPORT TYPE
Final

3. DATES COVERED *(From - To)*

4. TITLE AND SUBTITLE
DARPA Augmented Cognition
Technical Integration Experiment (TIE)

5a. CONTRACT NUMBER
N66001-99-D-0050

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
0602301E

6. AUTHORS

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

8. PERFORMING ORGANIZATION REPORT NUMBER
TR 1905

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Defense Advanced Research Projects Agency (DARPA)
3701 North Fairfax Drive
Arlington, VA 22203-1714

10. SPONSOR/MONITOR'S ACRONYM(S)
DARPA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
This is the work of the United States Government and therefore is not copyrighted. This work may be copied and disseminated without restriction. Many SSC San Diego public release documents are available in electronic format at http://www.spawar.navy.mil/sti/publications/pubs/index.html

14. ABSTRACT
The Defense Advanced Research Projects Agency (DARPA) Augmented Cognition program will result in demonstrable, quantifiable augmentations to human cognitive ability in realistic operational environments. Towards this goal, the first phase of the Augmented Cognition program was to empirically assess the utility and validity of various psychophysiological measures in dynamically identifying changes in human cognitive activity as decision-makers engaged in cognitive tasks. This report is the culmination of Phase I – *Measuring Cognitive State.* It describes the empirical results of a Technical Integration Experiment (TIE) involving the evaluation of 20 psychophysiologically derived measures (cognitive state gauges) that were developed under Phase I of the Augmented Cognition program.

15. SUBJECT TERMS
Mission Area: Command and Control
Augmented cognition

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Clem Urban
U	U	U	UU	234	19B. TELEPHONE NUMBER *(Include area code)*

SSC San Diego

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Approved for public release; distribution is unlimited.