2015

Enhancing the effectiveness of Human-Robot teaming with a closed-loop system

Grace Teo
University of Central Florida

Part of the Psychology Commons

Find similar works at: <https://stars.library.ucf.edu/etd>

University of Central Florida Libraries <http://library.ucf.edu>

This Doctoral Dissertation (Open Access) is brought to you for free and open access by STARS. It has been accepted for inclusion in Electronic Theses and Dissertations, 2004-2019 by an authorized administrator of STARS. For more information, please contact STARS@ucf.edu.

STARS Citation

Teo, Grace, "Enhancing the effectiveness of Human-Robot teaming with a closed-loop system" (2015).
Electronic Theses and Dissertations, 2004-2019. 5160.
<https://stars.library.ucf.edu/etd/5160>

ENHANCING THE EFFECTIVENESS OF HUMAN-ROBOT TEAMING
WITH A CLOSED-LOOP SYSTEM

by

GRACE W. TEO
M.A. University of Central Florida, 2014
B.S. National University of Singapore 1996

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Psychology
in the College of Sciences
at the University of Central Florida
Orlando, Florida

Fall Term
2015

Major Professor: James L. Szalma

© 2015 Grace Teo

ABSTRACT

With the development in robotics and the increasing deployment of robots, human-robot teams are set to be a mainstay in the future. However, our understanding of the effectiveness and impact of this new form of teaming is limited. Previous experience with technology and automation has shown that technological aids do not always result in the intended consequences of improved performance and alleviation of workload and stress. No doubt a large part of this is due to the fact that the relationships among taskload, workload and performance are complex as human operators interact dynamically with tasks and technology. Measures of workload are also varied and differentially sensitive. There is also the added challenge posed by multi-tasking environments which typify most real-world situations. Given all this, efforts in designing technological aids, such as an adaptive robot aid in the context of human-robot teaming, would require a workload model that reflects the intricate relationship between taskload and the individual operator's experience of workload. Such a model can then be used to drive a closed-loop system on which adaptive robot aiding can be based. The present research sought to investigate the effectiveness of a closed-loop system, based on a model of workload, in enhancing performance in a simulated military mission involving a human-robot team. Results showed that adaptive robot aid driven by workload needs as assessed by physiological measures resulted in greater improvements in performance compared to robot aid that was imposed by the system.

ACKNOWLEDGMENTS

This work was supported in part by the U.S. Army Research Laboratory (ARL). The views and conclusions contained in this document are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of ARL or the U.S. Government.

This research would not have been possible without the help, advice and support of many. I wish to thank the members of my committee: Dr James Szalma, for serving as my primary advisor through my graduate career, Dr Peter Hancock, Dr Florian Jentsch, Dr Lauren Reinerman-Jones, for providing the opportunity for me to work on this project, and Dr Gerald Matthews, for his ingenuity with the alternative modeling technique and algorithms. I am also grateful to Jonathan Harris and Dr Daniel Barber for building the closed-loop system, Brandon Sollins for help with the data-wrangling, Avonie Parchment, James Tyson and Ryan Wohleber for assisting with the data collection, Dr Julian Abich, Nelson Lerma, Ben Sawyer, Audrey Fok, and the other colleagues and friends who were my sounding board. I also wish to thank my friends, John Carpenter, Ray and Helen Reynolds, and my family for their love and encouragement throughout this process.

TABLE OF CONTENTS

LIST OF FIGURES	xii
LIST OF TABLES	xiv
LIST OF ACRONYMS	xvii
CHAPTER ONE: INTRODUCTION.....	1
Current and Future Use of Robots: From Tools to Teammates.....	2
CHAPTER TWO: LITERATURE REVIEW	4
Taskload, Workload and Operator Functional State, and Performance.....	4
Taskload of the human teammate in an ISR mission.....	4
Workload and Operator Functional State (OFS).	7
Central Nervous System measures.....	11
EEG.....	11
TCD.....	11
fNIR.	12
Peripheral Nervous System measures.....	12
ECG.....	12
Ocular measures.....	12
Performance of the HR Team.	13
Predicting performance from workload measures.	14

Impact of Technology and Robots on Workload and Performance.....	15
Unintended consequences of technology and mitigating strategies.	18
Implementing levels of adaptive autonomy through human-robot team roles.	21
Modeling and Classifying Workload/OFS.....	24
Basis for using previous data for workload modeling	26
Training, Testing and Cross-validation of the workload model	27
Research Plan.....	28
Study 1 Objectives.	29
Study 1 Hypotheses.....	29
Study 2 Objective.....	30
Study 2 Hypotheses.....	31
CHAPTER THREE: METHODOLOGY (STUDY 1)	32
Study Goal	32
Study Design.....	32
Study Scenarios.....	32
Power Analysis	33
Participants.....	33
Experimental Task/Context	34
Change Detection Task.	35

Peripheral Task.	38
Measures	39
Physiological measures.....	39
Subjective measures.....	43
Performance measures.	44
Procedure	45
CHAPTER FOUR: RESULTS (STUDY 1)	46
Taskload and Performance.....	46
Taskload and Workload	47
Physiological workload measures.....	47
Subjective workload measures.....	49
CHAPTER FIVE: MODELING (STUDY 1)	53
Workload modeling with logistic regression	53
Results of the logistic regression	55
Alternative approach to modeling.....	57
Matching of Difference Scores	59
Algorithms to combine physiological measures	60
Algorithm 1: Proportion of repeated markers.....	60
Algorithm 2: Distance between all difference scores.	62

Algorithm 3: Number of sign matches.....	62
Algorithm Evaluation.....	63
Mean workload indices obtained from similar task pairs and by chance.	64
Distributions of workload indices from similar and random task pairs.....	68
Simulation of the closed-loop system and setting cutoff levels.....	72
Sensitivity of the algorithms.	78
Study 1 hypotheses addressed.....	82
Summary	82
CHAPTER SIX: METHODOLOGY (STUDY 2).....	84
Study Goal	84
Study Design.....	84
Study Scenarios.....	84
Robot aid.....	86
Power Analysis	87
Participants.....	88
Experimental Task/Context	88
Change Detection Task.	88
Threat Detection Task.....	89
Procedure	91

Measures	92
Performance measures.	92
Subjective measures.....	93
Physiological measures.....	93
CHAPTER SEVEN: RESULTS (STUDY 2)	94
Time-on-task effects	95
Analyses approach	96
Change Detection Performance	96
Effects of Taskload, Block/Order and Aid.	97
Order effects.....	102
Aid and Taskload.	103
Effects of Aid and Trigger Type.....	104
Threat Detection Performance	111
Effect of Block and Aid.	111
Effect of Aid and Trigger type (Dual task).....	112
Perceived/Self-reported Workload and Stress States.....	114
ISA.	114
TLX.....	117
DSSQ	120

Physiological workload index.....	123
Effects of Task and Block.....	123
Differences between Trigger Type groups.	125
Workload index values across different studies.....	127
Study 2 hypotheses addressed.....	128
CHAPTER EIGHT: DISCUSSION.....	130
Taskload effects	130
Aid effects	131
Imposed aid vs. adaptive aid	132
Taskload-Performance relationship	132
Online physiological workload index	134
Other issues with performance and the aid	135
Theoretical implications.....	139
Practical applications	141
Future research.....	142
Conclusion	143
APPENDIX A: UCF APPROVAL OF HUMAN-SUBJECTS RESEARCH	145
APPENDIX B: INFORMED CONSENT.....	148
APPENDIX C: DEMOGRAPHICS SURVEY	155

LIST OF REFERENCES	158
--------------------------	-----

LIST OF FIGURES

Figure 1. Simple four-stage model of human information processing and types of automation..	17
Figure 2. Screen capture of the MIX testbed	35
Figure 3. Icons used to represent entities for the change detection task.....	37
Figure 4. Electroencephalogram (EEG) electrode placement.....	40
Figure 5. Electrocardiogram (ECG) electrode placement.....	41
Figure 6. Functional Near-Infra Red (fNIR) sensors.	41
Figure 7. Transcranial Doppler Ultrasonography (TCD).	42
Figure 8. Desktop Eye tracker	43
Figure 9. Distribution of the workload index under algorithm 1.....	69
Figure 10. Distribution of the workload index under algorithm 2.....	70
Figure 11. Distribution of the workload index under algorithm 3.....	71
Figure 12. Workload index under algorithm 1 over the entire condition	74
Figure 13. Workload index under algorithm 2 over the entire condition	75
Figure 14. Workload index under algorithm 2a over the entire condition.....	76
Figure 15. Workload index under algorithm 2b over the entire condition	77
Figure 16. Scenarios in Study 2	85
Figure 17. Video feed showing characters in the “ground level” view	89
Figure 18. Examples of characters in the TD task.	90
Figure 19. Performance measures for the Change Detection task	102
Figure 20. Percent correct detections by Taskload and Aid	104
Figure 21. Number of false alarms by Aid and Trigger Type in Single task blocks	108

Figure 22. Percent correct by Aid and Trigger type in Dual task blocks.....	110
Figure 23. Average response time by Aid and Trigger type in Dual task blocks	110
Figure 24. ISA ratings by Taskload and Block.....	116
Figure 25. Physiological workload index over all samples in scenarios SSS and SDD.....	125
Figure 26. Physiological workload index over scenario SSS.	126
Figure 28. Physiological workload index over scenario SDD.....	127

LIST OF TABLES

Table 1 Components of the Soldier's taskload in the ISR mission under various configurations..	5
Table 2 Responses of physiological workload measures to increases in workload.....	10
Table 3 Workload measures sensitive to taskload changes in the CD task (Abich, 2013).....	27
Table 4 Scenarios in Study 1	33
Table 5 Levels of the Peripheral Task in Study 1	39
Table 6 Percent correct detections on the Change Detection task for each scenario (N=40)	47
Table 7 Percent correctly-answered prompts on the Peripheral Task for each scenario (N=40) .	47
Table 8 Physiological measures sensitive to Scenario effects (N=40)	48
Table 9 Physiological measures sensitive to CD taskload changes (N=40)	48
Table 10 ISA workload ratings in each scenario (N=40)	49
Table 11 TLX workload ratings by Scenario.....	50
Table 12 Results of model validation	56
Table 13 Scenarios in Abich (2013) Study	57
Table 14 Scenarios in the Abich (2013) study.....	63
Table 15 Scenarios in Study 1	64
Table 16 Description of Task Pairs Used in Algorithm Evaluation	64
Table 17 Mean and SDs of Workload Indices from various Task Pairs under each algorithm....	66
Table 18 Task Pairs for the Closed-loop Mock-up (Abich, 2013)	73
Table 19 Possible Outcomes of the Aiding Decision	78
Table 20 Decision outcomes for different cutoffs under the various algorithms	79
Table 21 Sensitivity of the Various Cutoff Scores by Algorithm.....	81

Table 22 Definition of terms used in the modeling approach.....	83
Table 23 Status of Adaptive Aid under Different Classifications of Workload State	87
Table 24 Behaviors associated with the different Robot roles	90
Table 25 Performance measures and their descriptions.....	92
Table 26 CD performance in Abich (2013) and Study 2	97
Table 27 Analyses of the Effects of Taskload, Block/Order and Aid (CD Performance).....	97
Table 28 CD performance measures by Taskload (N=43)	98
Table 29 CD performance measures by Block (N=43)	99
Table 30 CD performance measures by Taskload by Block (N=43).....	100
Table 31 Analyses of the Effects of Aid and Trigger Type (CD Performance)	104
Table 32 No. of participants with Physio-triggered aid or System-triggered aid by Condition .	105
Table 33 Description of levels for the factors Aid and Aid trigger type	106
Table 34 CD performance by Aid (Single task) (N=62).....	107
Table 35 CD performance by Aid (Dual task) (N=62)	108
Table 36 Analyses of the Effects of Block and Aid (TD Performance)	111
Table 37 TD performance across Blocks (N=43)	112
Table 38 Analyses of Aid and Trigger Type (TD Performance)	112
Table 39 TD performance by Aid (N=62)	113
Table 40 Analyses of the Effects of Taskload, Block and Aid (ISA).....	114
Table 41 ISA workload ratings by Block (N=40).....	115
Table 42 ISA workload ratings by Taskload and Block (N=40)	115
Table 43 Analyses of the Effects of Aid and Trigger Type (ISA).....	116
Table 44 Analyses of the Effect of Scenario (TLX)	117

Table 45 TLX subscales by Scenario (System trigger participants only, N=43)	118
Table 46 Analyses of the Effect of Trigger Type (TLX).....	119
Table 47 Analyses of the Effect of Scenario (DSSQ).....	120
Table 48 DSSQ subscales by Scenario (System-trigger participants only, N=43).....	121
Table 49 Analyses of the Effect of Trigger Type (DSSQ)	121
Table 50 DSSQ Pre-Post Worry by Trigger Type group as grouped in scenario SDD (N=62) .	122
Table 51 Analyses of the Effect of Taskload and Block (Physio.Workload Index).....	123

LIST OF ACRONYMS

AA	Adaptive automation
ANN	Artificial neural networks
ANOVA	Analysis of variance
AOI	Area of Interest
BPM	Beats per minute
CD	Change detection
CR	Correct rejection(s)
DSSQ	Dundee stress state questionnaire
FA	False alarm(s)
fNIR	Functional near-infra red spectroscopy
HR	Heart rate
HRV	Heart rate variability
IBI	Interbeat interval
ICA	Index of cognitive activity
ISR	Intelligence, Surveillance and Reconnaissance
LIDAR	Light detection and ranging (sensor)
LOA	Levels of automation
MANOVA	Multivariate analysis of variance
MRQ	Multiple resources questionnaire
MRT	Multiple resources theory
OFS	Operator functional state

OOTL	Out-of-the-loop
PT	Peripheral task
SDT	Signal detection theory
SPD	Spectral power density
SWDA	Stepwise discriminant analysis
TCD	Transcranial Doppler ultrasonography
TD	Threat detection
TLX	NASA taskload index
UGV	Unmanned ground vehicle

CHAPTER ONE: INTRODUCTION

The rise of automation began in the Industrial Revolution where machines relieved humans of difficult or tedious mechanical tasks (Sheridan & Parasuraman, 2005). This was followed by the Digital Revolution where machines and automation began taking over tasks that were more cognitive in nature (e.g., performing calculations). Gradually, advances in the field of computers and robotics gave rise not just to the automation of tasks requiring higher-order cognitive functioning, but also to the development of autonomous agents and robots (Satell, 2014). Although they have been utilized in dangerous search and rescue, as well as recovery and exploratory missions in inhospitable environments, the limited capabilities of robots have prevented their use from being more extensive. However, in recent years, the military and various stakeholders have articulated a vision of robots that are able to work more closely with humans (Gaudin, 2013), and to transit from being mere tools to being teammates. The NSF, NIH, USDA and NASA have also announced new investments of about \$38 million for the research and development of robots that “cooperatively work with people to enhance individual human capabilities, performance, and safety” (NSF, 2013). A key component of such co-operative work is for robots to acquire the ability for multimodal communication, which includes implicit communication and sensing of operator fatigue, workload and other psychological states.

The present study investigates the use of a closed-loop system that monitors operator workload state so that adaptive aid by a robot teammate can be rendered when workload state is high. Such a system is expected to manage operator workload by maintaining taskload at a manageable level. This study is the culmination of a series of research studies on workload and per-

formance conducted to further the application of multimodal human-robot communication in human-robot teams. The research included the development of an experimental testbed (Reinerman-Jones, Barber, Lackey, & Nicholson, 2010) and design of experimental tasks. These tasks enabled the study of the effects of task characteristics, such as task congruency and task switching (Reinerman-Jones, Taylor, Spouse, Barber, & Hudson, 2011), and automation type (Taylor, 2012), on workload and performance. In addition, substantial work has also been done on the effects of the task manipulations on multiple workload measures (Abich, 2013).

Current and Future Use of Robots: From Tools to Teammates

Traditionally, robots have been assigned tasks that are “dirty, dangerous, and dull” (Takayama, Ju, & Nass, 2008, p.1). Thus far, robots have been deployed in military applications such as intelligence and reconnaissance, bomb disposal, search and rescue missions, as well as civilian applications such as in manufacturing (Guizzo & Ackerman, 2012), medicine (da Vinci Surgery, 2013), entertainment, as domestic help (Aki, 2012), or assistance for the elderly or disabled (“Service & Entertainment Robots”, n.d.). However, despite their capabilities, robots at present are still largely teleoperated (i.e., operated via remote control) and work within a relatively well-defined problem space. For robots to transit from being tools to teammates, robots should possess “task-level autonomy” (DRC, 2013) and have capabilities that support richer, more sophisticated human-robot interaction. This includes the capacity for natural language processing of speech and gestures, as well as the ability to sense the operator’s physical fitness for the task, or psychological states such as fatigue, stress and workload. Such enhancement of the robot’s sensing capabilities would allow formulation of robot responses and behaviors that are better adapted to operator needs and to the dynamic operational environment.

The development of a closed-loop system that drives robot adaptive aid requires the understanding of (i) the interrelationships of taskload, workload and performance, (ii) the impact of technology, robots, and human-robot team roles on operator workload and performance, and, (iii) issues related to trust and safety in human-robot teams (e.g., Jentsch et al., 2012; Bainbridge, Hart, Kim, & Scassellati, 2008; Bartneck, C., Kulic, D., Croft, E., & Zoghbi, S., 2009; Heerink, Krose, Evers, & Wielinga, 2010), although the latter two areas of trust and safety are not the focus of the present study.

CHAPTER TWO: LITERATURE REVIEW

Taskload, Workload and Operator Functional State, and Performance

An understanding of the task and how it contributes to workload and output of performance is needed before a system to manage operator workload can be developed. Hence, this review begins with a description of the task domain followed by a discussion of issues in the measurement of operator state.

Taskload of the human teammate in an ISR mission.

The context for the present study is that of a military intelligence, surveillance and reconnaissance (ISR) mission. An ISR mission is typically a dangerous military operation involving a high degree of uncertainty about the physical environment and potential enemy presence therein. As with most military operations, the mission may begin with a relatively low taskload, but task demands usually increase over time as the pace of events is accelerated. Such transitions in task demands are one of the challenges in military operations as they increase workload and impair performance (e.g., Cox-Fuenzalida & Angie, 2005; Cox-Fuenzalida, Beeler & Sohl, 2006).

The objective of an ISR mission is to gather information about the presence and activities of an enemy or potential enemy in an area of interest (AOI). The information may be displayed in a variety of formats, e.g., video feeds, maps with icons showing overview, audio and text messages that are interchanged etc. When the ISR mission is conducted by a team, the taskload of a Soldier comprises collecting intelligence and detecting changes in the environment, as well as coordinating the team's activities to accomplish the mission. If a robot is part of the team, the Soldier's taskload includes detecting changes in the environment as well as being aware of the robot teammate's whereabouts and functioning as it explores the AOI.

A simulation of the taskload of the Soldier on an ISR mission in a human-robot team include: (a) a change detection (CD) task requiring detection and identification of changes represented as icons in the AOI, (b) a threat detection (TD) task which entails identifying potential enemy presence from a video feed, and (c) a peripheral task (PT) of maintaining awareness of the robot teammate's whereabouts and surroundings as it scouts the AOI gathering reconnaissance information. This latter task would capture the component of the taskload that results from being in a team (see Table 1).

Table 1
Components of the Soldier's taskload in the ISR mission under various configurations

	Main task of the ISR mission	Peripheral task from being in a team
Soldier alone on ISR mission	<ul style="list-style-type: none"> • Attending to the display(s) continuously in order to detect changes or enemy presence as they occur. • Identifying the changes and enemy presence and responding appropriately. • Adapt to changes in task demands across task duration. 	n.a.
Soldier-Soldier team on ISR mission	<ul style="list-style-type: none"> • Attending to the display(s) continuously in order to detect changes or enemy presence as they occur. • Identifying the changes and enemy presence and responding appropriately. • Adapt to changes in task demands across task duration. 	<ul style="list-style-type: none"> • Being aware of human teammate's whereabouts within the AOI • Coordinate efforts with teammates.
Soldier-Robot team on ISR mission	<ul style="list-style-type: none"> • Attending to the display(s) continuously in order to detect changes or enemy presence as they occur. • Identifying the changes and enemy presence and responding appropriately. • Adapt to changes in task demands across task duration. 	<ul style="list-style-type: none"> • Being aware of robot teammate's whereabouts within the AOI. • Oversee robot teammate's functioning.

The change detection task (CD) (Abich, 2013; Taylor, 2012) is predominantly a visual task that requires detection of changes in an array, and the identification of the type of change that occurred. The array comprises a visual set of abstract icons. During the task, three types of changes may occur – icons may appear, disappear, or move. The detection of the changes requires visual scanning as the array is larger than the central visual field of 30 degrees (Spector,

1990). This type of detection task, i.e., saccade-contingent changes (Rensink, 2002), has been associated with change blindness (Simon, 1996; Simon & Levin, 1997; Simon & Ambinder, 2005; Rensink, 1997; 2002). Change blindness may occur because some items (e.g., those in the peripheral visual field) are less likely to be detected because these are less fully processed and attended to (Kahneman & Treisman, 1984; Kahneman, Treisman & Gibbs, 1992; Rensink, 2000), or be due to the suppression of perception by mechanisms that generate eye movements (i.e., saccadic suppression) (Simon & Ambinder, 2005). The taskload of the CD task arises from attending to the array of visual stimuli, which requires continual visual scanning at a pace that is contingent on the rate at which changes occurred, as well as the response demand from classifying the change type.

The threat detection task (TD; Abich, 2013; Taylor, 2012) involves a video feed of the AOI displaying the environment that the UGV is in, which includes human characters and features in the environment. The characters may be friendly Soldiers, friendly civilians, enemy soldiers or enemy insurgents. Enemy soldiers and insurgents are threats or critical signals. The Soldier is required to view the feed and identify the threats in the environment. The threat detection task has characteristics of a vigilance task in that critical signals must be detected in the midst of “noise” or distracters, and this offers the possibility of four outcomes (i.e., hits, false alarms, misses and correct rejections). However, unlike traditional vigilance tasks which utilize static stimuli presented in trials (e.g., Becker, Warm & Dember, 1994, Helton et al., 2010; Funke et al., 2010, Szalma et al., 2006, Szalma & Teo, 2012, Warm et al., 2009), the threat detection task is a dynamic vigilance task, where the flow of stimuli (i.e., critical signals, distracters) is continuous

rather than discrete (Szalma et al., 2014). Such a dynamic task is more representative of the vigilance task in an ISR mission where threats in the environment are detected as the Soldier or robot moves through the AOI.

The peripheral task involves maintaining awareness of the robot's whereabouts while it is actively patrolling and gathering information in the AOI. In an actual mission, the Soldier would need to issue certain commands should the robot asset be captured, or to call for support via radio if the robot teammate encounters any mechanical problems. In the present study, the Soldier would need to be aware of the robot teammate's direction of travel as well as some features (e.g., the people and vehicles) within the robot's vicinity. (see Methods section for more details on the specific tasks used in this research).

Workload and Operator Functional State (OFS).

The impact of the ISR taskload on the Soldier can be conceptualized in terms of mental resources and considered within the broader framework of the operator functional state (OFS). Resources refer to the overall capacity of available mental, physical, physiological "energetic resources" or reserves, while OFS is broadly defined as the "background" state of the operator as s/he takes on and performs a task (Hockey et al., 2003). OFS is the result of the dynamic interaction of other cognitive tasks (e.g., primary, secondary tasks and subtasks), as well as biological and emotional goals already active within the individual's (cognitive-energetic) motivational system (Hockey et al., 2003). The OFS can be used to predict the ability of the operator to undertake and cope with the demands of the task at hand, and is typically described and evaluated in terms of the amount of available resources (Hockey et al., 2003). Within a more specific context of a task, OFS is assessed as the level of workload experienced. Workload relates to the operator's

response to a task, as well as the resources available for a certain taskload. Workload, a multidimensional latent construct, or an “intervening variable” (Gopher & Donchin, 1986, p. 41-44), is the means of explaining the complex relationship between taskload and performance. One definition of workload is that it is the “mental construct that reflects the mental strain resulting from performing a task under specific environmental and operational conditions, coupled with the capability of the operator to respond to those demands” (Cain, 2007, p.2). This capability to meet the taskload may be diminished by fatigue, illness and demands from other competing processes (activated to achieve task-irrelevant emotional and biological goals), as all these contribute to the OFS and workload (Hockey et al., 2003). Support for the construct of workload comes from the convergence of physiology, performance, and subjective measures (Tsang & Vidulich, 2006), as well as from task analysis and computational modeling (Wickens, 2002).

The measurement of workload is necessary to quantify the mental cost of performing tasks and to predict operator performance (Cain, 2007). O’Donnell and Eggemeier (1986) proposed that workload measures should: (i) be sensitive to changes in resource demand and reflected in changes in taskload, (ii) be diagnostic and indicative of the sources of workload variation, (iii) be non-intrusive and should not significantly add to the workload, and not impact performance on the primary task, (iv) be reliable in that its relationship with performance should be consistent over time, and (v) meet general implementation requirements and be relatively easy to measure.

There are performance-based measures, subjective (self-report) measures, and physiological measures of workload (Meshkati, Hancock, & Rahimi, 1989; Tsang & Vidulich, 2006). Performance-based workload measures assume that any increases in taskload or demands are typically followed by an increase in workload, which usually precedes performance declines (Rubio,

Diaz, Martin, & Puente, 2004). However incongruences in the form of workload dissociations and insensitivities have been observed that challenge the limits of this relationship. Performance-workload dissociations occur when the expected relationship of performance and workload (i.e., performance decreases with increases in workload: both measures indicating a change in resource capacity in the same direction) is not observed (Yeh & Wickens, 1988). Such dissociations may be diagnostic of how operators handle tasks (Hancock, 1996). For instance, when an individual allocates more effort into a task, both performance and subjective ratings of workload increase (e.g., Vidulich & Wickens, 1986); or when the operator is already maximally allocating his resources to a task, increasing task demands degrades performance but may not increase subjective workload ratings any further (Tsang & Vidulich, 2006). Both of these are examples of performance-workload dissociations because they run counter to the expected inverse relationship between workload and performance. Another reason that performance-based measures may not be suitable is that they are usually available only at the end of a task or after substantial blocks of performance trials, and so are unsuitable as workload measure for the present study which seeks to manage workload to prevent performance declines.

On the other hand, subjective measures assume that the individual can accurately assess and report his/her experience of the taskload and demands. Although the advantages of subjective measures include their ease of administration and low cost, they can only be obtained when the operator is not engaged too deeply in the task as they require overt responses and may be disruptive if administered during the task. This inability of subjective measures to capture workload continuously renders them unsuitable for the purposes of the present study.

The basis of physiological measures (e.g., heart-rate, ocular activity, brain activity, hemodynamics) lies in the notion that, with the activation of various cognitive processes required to

process task demands and execute the required responses, there are corresponding physiological responses in both the central (e.g., brain activity, cerebral perfusion as indicated by level of oxygen saturation, rSO₂) and peripheral nervous system (e.g., ocular activity as captured by the Index of Cognitive Activity (ICA), Marshall (2002)) that reflect this cognitive activity. Physiological measures have shown adequate test-retest reliability in studies on pilots, indicating their potential suitability as measures in OFS assessment (Wilson, 2003; 2003a; 2003b), and multiple studies have demonstrated the sensitivity of these measures to changes in OFS (Gevins et al., 1997; Wilson & Eggemeier, 1991; Taylor et al., 2010). Physiological measures allow continuous assessment of workload, provide high temporal resolution, and do not require any overt response from the operator that may interfere with task performance. These advantages render them particularly suitable as indicators of operator workload in adaptive systems. Commonly-utilized physiological measures in workload studies include heart rate, heart rate variability, respiration rate, brain activity, pupil size (diameter), electrodermal activity among others (Meister, 1986). (see Table 2 for a more extensive, but not exhaustive list):

Table 2
Responses of physiological workload measures to increases in workload

Physiological measure	Response of measure to INCREASES in workload
Heart rate	Increases (Wilson. & Eggemeier, 1991)
Heart rate variability	Decreases (Mulder, De Waard, & Brookhuis, 2005)
Pupil diameter	Increases (Backs, 1992; Casali, 1983; May, 1990)
Eye fixation duration	Increases (Callan, 1998); Decreases (Schulz et al., 2010)
No. of eye fixations	Increases (Van Orden, Limbert, Makeig, & Jung, 2001)
Theta waves (from EEG)	Increases (Hankins & Wilson, 1998)
Alpha waves (from EEG)	Decreases (Hankins & Wilson, 1998)
Beta waves (from EEG)	Increases (Kurimori & Kakizaki, 1995)
Cerebral bloodflow velocity (CBFV)	Increases (Warm & Parasuraman, 2007)
Oxygen saturation (rSO ₂)	Increases (Sassaroli et al., 2008)
Electrodermal activity	Increases (De Waard, 1996)

Adapted from Meister (1986)

Physiological workload measures can be subdivided into measures that tap central nervous system activity, and those that assess peripheral nervous system activity. These measures can thus be used in combination to provide a more comprehensive assessment of workload state.

Central Nervous System measures.

EEG.

The electroencephalogram (EEG) detects neuronal activity that results from changes in the amount and type of mental workload, as well as changes in overall state of arousal (Gevins & Smith, 2007). When mental workload increases, alpha waves are substituted by beta waves (Sabatini, 1997), and increases in theta waves with workload have also been reported (Hankins & Wilson, 1998). Specifically, increases in spectral power in frontal midline theta signal has been observed in attention-demanding tasks (Gevins et al., 1997; Miyata, Tanaka, & Hono, 1990). EEG has also been used in the detection of workload from more realistic tasks such as in the MATB which simulates aviation tasks (Comstock & Arnegard, 1992; Smith et al., 2001) as well as workload changes in real time (Heger, Putze & Schultz, 2010; Baldwin & Penaranda, 2012).

TCD.

Cerebral bloodflow velocity (CBFV), assessed by the Transcranial Doppler sonography (TCD) reflects brain metabolic activity that is expected to vary with task load. Increases in mental processing increases brain metabolic activity and its byproducts, such as CO₂. This requires increased bloodflow to remove the byproducts and waste resulting from the metabolic activity. Changes in CBFV are thus indicative of the level of mental activity (Aaslid, 1986). CBFV has been shown to index workload in a variety of motor, cognitive, and perceptual and tasks (Duschek & Schandry, 2003; Stroobant & Vingerhoets, 2000), including vigilance tasks (Shaw et al., 2013; Tripp & Warm, 2007).

fNIR.

The functional near-infrared spectroscopy (fNIR) provides an indication of regional oxygen saturation (rSO₂), a hemodynamic response, in the prefrontal cortex when placed on the forehead. As with CBFV, the level of rSO₂ is indicative of the metabolic activity that results from mental processing. Increases in rSO₂ have been reported with increased mental and cognitive load (Izzetoglu et al., 2004; Son et al., 2005; Sassaroli et al., 2008). The fNIR has also been utilized to monitor changes in workload in complex tasks such as a UAV piloting task, showing correlations with performance and other measures (Ayaz et al., 2010; 2012).

Peripheral Nervous System measures.

ECG.

The electrocardiogram (ECG) detects cardiac responses which are enervated by the peripheral nervous system. ECG measures such as heart rate (HR) and heart rate variability (HRV) have been shown to relate to workload changes (Wilson & Eggemeier, 1991; Mulder et al., 2005). However, as HR is easily influenced by physical activity and other extraneous variables, HRV is considered to be more indicative of workload. Increases in workload are associated with declining HRV (Byrne & Parasuraman, 1996; Wilson, 1992).

Ocular measures.

Eye fixation duration, number of fixations, pupil diameter have been reported to correspond to workload. Longer fixation durations correspond to increases in cognitive load (Callan, 1998), although there are studies that show the opposite trend (e.g., Schulz et al., 2011). Higher fixation frequencies reflect an increase in mental workload and a more effortful search (Van Orden et al., 2001). The nearest neighbor index (NNI) of workload, derived from eye fixation patterns, has been found to vary with differing task loads from a UGV navigation task (Fidopiastis

et al., 2009). Pupil diameter increases as a function of higher mental workload (Backs & Walrath, 1992; May et al., 1990; Beatty, 1982), as does the Index of Cognitive Activity (ICA), which is derived from pupil diameter but filters out the pupil's reflex response to light, thereby isolating the response from cognitive stimuli (Marshall, 2002).

Performance of the HR Team.

In many applied settings, the discussion of taskload and workload is relevant primarily in the context of understanding how these relate to performance. This is because, from a practitioner's perspective, the eventual goal for understanding taskload and workload is to predict and improve performance, the criteria of interest. In general, performance degradations tend to (i) occur small decrements, (ii) be selective, affecting less critical aspects of task first, (iii) be more likely to occur in laboratory studies than in operational environments where the negative consequences of performance failures are more severe, (iv) observed when the work is prolonged, (v) less pronounced when activation and effort are high (Hockey et al., 2003). However taskload and performance do not always covary because operators seem to accommodate increases in task demand (to a point) through certain compensatory mechanisms. Hockey's Compensatory Control model of performance regulation (Hockey, 1997) explains this relative stability in performance by positing a compensation mechanism that uses a "reserve effort budget for meeting additional demands, unpredictable changes in the demands-resource balance, or the additional burden associated with stressful environments" (Hockey, 1997, p. 78). The model also predicts that maintenance of task performance under extreme high task demands would have an "effect on mood states reflecting the affective response to emergency and sustained coping effort" (Hockey, 1997, p. 86).

While Hockey's model describes a compensatory mechanism that accounts for how task performance may be maintained at the expense of cognitive states etc., the Maximal Adaptability Model (Hancock & Warm, 1989) extends the concept of compensation to a more general notion of adaptability. The model postulates that there are zones of adaption that account for situations where stress from the environment and taskload may increase without necessarily being accompanied by a reciprocal drop in performance, and recognizes that stress occurs with both overload and underload in task demands. Furthermore, the zones of adaptation are nested within each other, suggesting that different forms of adaptation occur in a particular order. For instance, increases in taskload would be met with a level of psychological adaptability such that performance is maintained, but at a cost of subjective comfort (the experience of stress and workload). If task demands continue to increase or fall to a level that is beyond the individual's ability to adapt psychologically (i.e., to maintain acceptable levels of performance), then the individual enters the zone of physiological adaptability, where the task demands begin to affect the individual psychologically, and performance may begin to decline (Hancock & Warm, 1989; Szalma & Teo, 2012).

Predicting performance from workload measures.

If there are multiple zones of adaptation that occur in a sequence, then early indicators that forewarn of impending decrements in operator performance in the human-robot team (manifested as increased workload and stress) may be used as inputs for a regulatory mechanism such as a closed-loop system, in which these indicators will trigger interventions such as an aid, so that performance declines can be averted. This concept has been applied in other domains. For instance, the percentage of eyelid closures (PERCLOS) is indicative of driver drowsiness (Wierwille et al., 1994) and aids such as the *Copilot* (Grace & Stewart., 2001), can help avert serious

accidents by monitoring the driver's PERCLOS and providing an auditory advisory tone with a visual indicator when a driver is in danger of dozing off at the wheel. Other physiological workload measures have also been found to predict performance changes. For example, the amplitude of the P300 have been found to be proportional to the amount of cognitive resources allocated to the target, enabling subsequent prediction of performance deficits in target discrimination from diminishing P300 amplitudes (Israel, Wickens, Chesney, & Donchin, 1980). Enhanced frontal theta activity, coupled with reduced alpha power, has also been associated with increases in resource allocation to the task (Gevins & Smith, 2003), and can be used to predict performance on the task.

To minimize performance declines, the robot teammate can be designed to provide aid to the Soldier when workload levels indicate impending performance failures. However, research on automation has shown that technological aids do not always lead to the intended outcome of improved performance (Parasuraman & Riley, 1997). The operator's workload experience may also be affected by the robot's behavior (including its aiding behavior) in unexpected ways as technology can actually increase workload and impact performance adversely (e.g., Parasuraman & Riley, 1997; Endsley & Kiris, 1995). Such findings underscore the importance of using technology appropriately to enhance performance.

Impact of Technology and Robots on Workload and Performance

For many decades, researchers and practitioners have been trying to determine how best to use technology and automation to improve performance and aid in work. Fitts' (1951) Men-Are-Better-At/Machines-Are-Better-At (MABA-MABA) list described the tasks and functions that man and machines are better at than the other, and has been a starting point for efforts at

function allocation. However the list is simplistic, static and outdated, and unlikely to be applicable to the systems and technologies of today (Hancock & Scallen, 1996; Hoffman et al., 2002); although the list was never intended by Fitts to be prescriptive (Hancock, 2009). More importantly, the list does not acknowledge the fact that humans adapt to technology just as they adapt to changing task demands and their own workload (i.e., Hancock & Warm, 1989). Operators often change their work practices when they work with automation (Dekker & Woods, 2002), and do so both in positive and negative ways. In their seminal work, Parasuraman and Riley (1997) discussed how technology and automation can be misused, disused and abused.

The effects of automation can be understood in terms of the stages of human processing. Parasuraman, Sheridan, and Wickens' (2000) model of human processing and types of automation provides a framework for understanding how types of automation can be used to alleviate workload and improve performance. It posits that there are four main stages of processing: (i) sensory processing, corresponding to automation involving information acquisition, (ii) perception/working memory, supported by automation that performs information analysis, (iii) decision making, corresponding to automation that aids with decision selection, and (iv) response selection, supported by automation that deals with action implementation (Parasuraman, Sheridan, & Wickens, 2000) (see Figure 1). Examples of technology aiding in the information acquisition stage include sensors such as surveillance cameras and seismometers, while various computational software and algorithms are instances of technology assisting with information analysis. Expert systems and decision algorithms aid in decision selection, and technologies such as the autopilot or cruise control in automobiles help in action implementation. While some automation affect individual stages, other types of automation impact multiple stages.

Figure 1. Simple four-stage model of human information processing and types of automation
Adapted from Parasuraman, Sheridan, and Wickens (2000).

An autonomous robot can be thought of as being the outcome of automating the entire information-processing chain, and the resultant system installed into a physically mobile vehicle equipped with actuators and that allows interaction with the physical environment. As defined by Robot Industries Association (RIA), a robot is “a reprogrammable, multifunctional manipulator designed to move material, parts, tools, or specialized devices through variable programmed motions for the performance of a variety of tasks” (Hamilton & Hancock, 1986, p. 70). With respect to the four-stage information processing model, an autonomous robot would acquire information from its environment, analyze the information, make decisions about the information and possible actions to take, and be able to interact with the physical world and entities within it when it implements the selected action. It would be self-directing and self-governing at least for a period of time. Viewed this way, a robot can be considered as an extension of automation. This relationship and the interrelatedness of both the fields of robotics and automation provide justification for drawing upon research on automation to guide predictions about human-robot interaction and

teaming. Note, however, that further research is needed to confirm that models of automation are valid for extension to human-robot interaction and teaming.

Unintended consequences of technology and mitigating strategies.

A major finding in automation research is that although automation has often improved productivity and reduced the burden of work on the human, it is not the case that having more automation is always better. Under some circumstances, automation may increase mental workload (Parasuraman & Riley, 1997; Wiener, 1988). High levels of automation can leave the human out-of-the-loop (OOTL), resulting in various adverse consequences such as the loss of situational or system awareness, reduced trust, greater complacency, over-reliance issues, decreased user acceptance, skill atrophy, performance degradations, and unbalanced mental workload (Carmony & Gluckman, 1993; Endsley & Kiris, 1995; Parasuraman et al., 1993). Loss of situational/system awareness occurs when the human operator's awareness of system and situational dynamics is diminished as a result of the automation taking over certain functions, especially the decision making functions (Endsley & Kiris, 1995; Kaber, Omal, & Endsley, 1999). This can result in errors and delays when the automation fails and human operator needs to resume control (Billings, 1997; Wickens & Holland, 2000; Wiener & Curry, 1980). In addition, even when operators are allowed to choose when to use automation (i.e., "adaptable automation"; Opperman, 1994), findings suggest that this cognitive evaluation of whether or not to use the automation contributes to operator workload (Kirlik, 1993), and operators are not always able to accurately assess when they would benefit most from automation (Kirlik, 1993; Neubauer, Matthews, Langheim, & Saxby, 2012). The two main strategies commonly employed to address these problems with automation use are: (i) to employ different levels of automation (LOA), and, (ii) to use adaptive automation (AA) (Kaber & Endsley, 2004).

The LOA strategy adjusts the degree of automation so that the operator is aided but is not left entirely out-of-the-loop. Higher LOA represents increased machine autonomy while lower LOA denotes greater human involvement and diminished automation (Parasuraman et al., 2000). With respect to the 4-stage model of information processing (refer back to Figure 1), an automation operating at a high LOA in the information acquisition stage would provide the operator “decluttered” and filtered information that has already been categorized according to certain criteria such that the “raw” data is unavailable to the operator (Yeh & Wickens, 2001). On the other hand, medium LOA in the same stage only tentatively classifies incoming data, allowing the operator to see the “raw” data (Parasuraman et al., 2000). A system can have a higher LOA for one stage and a lower LOA for another. The Theater High Altitude Area Defense system, which intercepts missiles (U.S. Army, 2007) has high LOAs for the information acquisition, information analysis, and decision selection stages, but low LOA for the action automation so that the human ultimately controls the firing (Parasuraman & Wickens, 2008).

The other strategy to keep the human in-the-loop while aiding with technology uses automation adaptively. While the LOA approach focuses on the extent to which various subtasks and processes are automated, the adaptive automation (AA) approach emphasizes when automated is invoked (Hancock & Chignell, 1987, Rouse, 1988, Scerbo, 1996). Rouse (1977) identified some advantages of employing adaptive automation that include the maintenance of reasonable operator workload, reduction of stress (Szalma, Hancock & Hancock, 2012), the support of human performance, and dynamic definition of task roles for operators (Kaber & Endsley, 2004). Using the AA scheme during high traffic conditions have resulted in the smallest increase in mental workload among Air Traffic Controllers (Hilburn, Jorna, Byrne, & Parasuraman, 1997),

and improved detection of system failures in a multitask flight simulation (Parasuraman, Mouloua, Molloy, & Hilburn, 1993; Parasuraman, Mouloua, & Molloy, 1996).

Strategies for implementing AA are based on (i) system functionality, where specific functions are identified to be automated, (ii) certain goals or criteria, such as when performance reaches a particular level or when critical events occur, or on (iii) models of performance and workload that drive adaptive logic (Morrison & Gluckman, 1994; Rouse & Rouse, 1983). These are similar to the major invocation strategies by Parasuraman et al. (1992) that specify that the schedule of AA can be based on factors such as (i) when critical environmental events happen, (ii) when operator performance reaches particular levels, (iii) when the operator experiences a critical level of workload as determined by modelling various workload measures and performance, (iv) when the operator is deemed to be in certain states as indicated by physiological measures, or (v) a hybrid of these strategies (Parasuraman et al., 1992).

For the context of an ISR mission in the present study, it is difficult to specify when critical events are expected to occur when performing the ISR mission tasks because there are no specific task events that are expected to invoke high workload. This is unlike the aviation domain, where take-off and landing are typically the most demanding events in flight. Using performance levels as criteria is *ex post facto* as the system would invoke automation when performance has already declined. It would be more appropriate to use an adaptive strategy that is based on a workload model as the driver for the adaptive aid.

Workload models can differ on the algorithms, the input/output parameters, and the assumptions about the relationships among taskload, workload, and performance. Models have been based on discrete-event simulations (e.g., Donmez et al., 2010) and task analysis which assume a direct relationship between taskload and workload (e.g., Pomranky & Wojciechowski

(2007) utilizing the Improved Performance Research Integration Tool (IMPRINT)). However, workload is not synonymous to taskload, but is a response of the operator to the taskload. For this reason, a model that utilizes physiological measures of workload, which allow continuous monitoring of workload responses to the task, would be more appropriate. Using physiological measures in such a way is consistent with a neuroergonomics approach (Parasuraman, 2003).

Implementing levels of adaptive autonomy through human-robot team roles.

In the context of a human-robot team, the LOA and AA strategies can be implemented by having different human and robot roles associated with different levels of autonomy. Scholtz's (2003) model classifies human-robot roles into (i) Supervisor, (ii) Peer, (iii) Subordinate, and, additional roles of (iv) Mechanic, and (v) Operator for the robot. Role assignment should be dynamic as a means to implement adaptive robot autonomy. The level of robot autonomy can be adapted to the operator's OFS or workload state as assessed by physiological workload measures. This is akin to the idea of "sliding robot autonomy" (Dias et al., 2008).

The closed-loop system, which uses the operator's workload level as the input, drives the adaptive aid by the robot, which in turn, relieves taskload and consequently, workload. Presumably, the system reaches a point of equilibrium when the closed-loop system results in a balanced workload where the operator does not experience the extremes of over- or underload. This can be expected to minimize OOTL performance issues such as having unbalanced workload and automation complacency (Parasuraman & Hancock, 2008).

The closed-loop system.

At the start of the ISR mission, the Soldier assumes the role of Supervisor, and the robot teammate adopts the Subordinate role with a low level of autonomy. The robot essentially functions as a ‘tool’, awaiting commands and executing them as they are issued by the Soldier. As the mission progresses and as the pace of events rises, taskload may increase, just as they might in typical military operations. Increases in taskload are expected to result in a rise in workload, and the closed-loop system will then classify the Soldier as being in a high workload state. This will trigger a shift in the robot’s team role from a Subordinate to a Peer. The Peer role, having a higher level of autonomy, is associated with behaviors that resemble that of a ‘teammate’. This can include (i) taking action without the need for explicit step-by-step commands, (ii) showing more initiative by aiding with tasks without being instructed to, (iii) anticipating needs by “pushing up” information, and (iv) initiating action like radioing for help instead of requiring the Soldier to do it himself. Such behaviors are likely to relieve the Soldier’s taskload and alleviate his workload. In the present study, when the robot is in Peer mode, it would aid the Soldier with the Change Detection task by providing an audible ‘BEEP’ that would occur with every change event. In doing so, the aid increases the saliency of the change events, and, because the aid is auditory, it is unlikely to overtax resources in the visual channel (Wickens, 2002). Such an action initiation by a robot in Peer mode is expected to relieve workload. A previous study with the ‘BEEP’ aid for the Change Detection task has shown that such an aid improves performance in terms of detection rates (Taylor, 2012; 2013). The Soldier would still need to classify the change type as soon as the event is detected.

For the present study, the robot teammate would continue to be in Peer mode for the remaining duration of the task condition following the occurrence of the first mode switch. However, in practice, the reduction of workload should be detected by the closed-loop system, which should then revert the robot team role to that of a Subordinate when Soldier workload is low, thereby returning full control to the Soldier until such time when the Soldier workload is high once again. These contingencies follow that of a negative feedback loop which has been linked to improved performance and system stability (Freeman et al., 1999). Under the negative feedback contingency, when automation is triggered in response to increases in workload, workload decreases, which in turn, reduces the level of automation (Freeman et al., 1999). Such a closed-loop system can result in an equilibrium and regulation of workload.

The anticipated sequence of events in the closed-loop system for the present study was as follows:

- i. At the start of the mission, Soldier (i.e., the participant) assumes role of “supervisor” and robot teammate adopts the “subordinate” role in the HR team.
- ii. Participant workload is assessed continuously (via physiological measures) and robot teammate obtains regular updates on participant workload state as the workload index is computed and updated frequently (e.g., every 30 seconds).
- iii. As taskload increases, participant workload is expected to increase. This should be reflected in the physiological workload index.
- iv. When workload index shows a degree of stability (e.g., reaches the cutoff level for a number of consecutive data samples; the debouncing factor), the closed-loop system would classify the participant as being in a state of HIGH workload. Robot teammate will switch to assume “Peer” mode. To inform the participant of this role switch, a

- notification message (e.g., auditory message) will be provided. In “Peer” mode, robot aids by providing an audible ‘BEEP’ with each change event. Aid is not revoked for the remaining duration of the condition.
- v. Participant taskload is expected to be alleviated, which should reduce workload, and performance declines should be averted.

Modeling and Classifying Workload/OFS

Although some researchers have determined workload state with only a few measures (e.g., EEG: Prinzel et al., 2000, 2003; HRV: Hoover et al., 2012), it is more judicious to base classification of workload state on multiple physiological measures. First, physiological measures assess workload differently and do not respond to changes in the operator’s workload state at the same rate (Sciarini & Nicholson, 2009). For instance, measures such as CBFV and rSO₂ that tap metabolic responses reflecting a more global state respond more slowly to changes in workload state compared to the duration of eye fixations, which respond more immediately. In addition, due to the multidimensional nature of workload, multiple physiological measures are better able to provide a more comprehensive picture of the workload experienced from demands of a task (Hankins & Wilson, 1998). This is especially true for the workload experienced in dynamic, multi-tasking environments, because while one measure may respond to certain workload manipulations, others may not (Wilson & O’Donnell, 1988; Matthews et al., 2015a). For example, Matthews et al. (2015a) reported that eye fixation duration and an EEG-based workload index were sensitive to the single/dual task workload manipulation, but HRV and rSO₂ discriminated between different levels of certain single tasks (Matthews et al., 2015a).

Different physiological workload measures have been used to classify OFS/workload. For instance, Taylor et al. (2010) compared the ability of various physiological measures in discriminating high and low task demand conditions (assumed to be indicative of workload) in a multi-tasking study and reported that EEG measures had the greatest potential for classifying OFS. Venables and Fairclough (2004) found that for their tracking, system monitoring and resource management tasks, EEG (specifically at sites Cz, Pz, P3, P4), respiration rate, blink duration and galvanic-skin response (GSR) predicted OFS stress states. Craven et al. (2006) successfully classified cognitive workload in real time from EEG, ECG, HRV, GSR measures of workload during a battery of memory tasks.

Other approaches combine various physiological measures forming a workload index that was used to classify OFS/workload and as the basis for providing adaptive aid. The Engagement Index (EI), computed with various ratios of alpha, beta and theta EEG bands, relates to arousal and task engagement, and was used as the input for adaptive aiding systems (Freeman et al., 1999; Pope et al., 1995; Prinzel, Freeman, Scerbo, Mikulka, & Pope, 2000). Another workload index is the Task Load Index (TLI; Gevins & Smith, 2003), a ratio of theta EEG band from frontal sites to alpha EEG band from parietal sites. The TLI was able to distinguish performance on the Multi-Attribute Task Battery (MATB) on various levels of taskload (Smith et al., 2001).

In addition to using selected physiological measures or derived indices, researchers have employed the use of classifiers such as stepwise discriminant analysis (SWDA) (e.g., Wilson & Russell, 2003a) or artificial neural networks (ANN) (Wilson & Russell 2003b; 2007; Baldwin & Penaranda, 2012). Although the latter method with ANNs show much promise (i.e., classifica-

tion accuracy often exceed 80%), the level of diagnosticity with the ANN approach is low as algorithms derived from the inputs (physiological measures) are not easily accessible for further design of adaptive aid (Zhang et al., 2013).

Basis for using previous data for workload modeling

To implement a closed-loop system that responds to operator workload, two things are required: (i) a workload index derived from a workload model based on a combination of physiological measures, and (ii) a threshold level of the index that signifies impending performance declines. The workload model for the present work was developed from data collected from a previous study (i.e., Abich, 2013) that utilized the same task in which low and high levels of taskload did not result in any workload dissociations or insensitivities (i.e., taskload increases were met with corresponding increases in subjective workload and performance degradations). Results from the Abich (2013) study showed that the taskload changes in the CD task were detected by a variety of performance-based, self-report, and subjective workload measures, and these showed that, for the CD task and the levels of the task investigated, the relationship between taskload and workload was positive (see Table 3). This relationship was observed both when the CD task was administered alone as well as when the CD task was part of a dual task. Furthermore, additional analyses of the Abich (2013) data (as reported in Matthews et al., 2015a) showed that some of the psychophysiological measures were sensitive to changes in task demand and workload, indicating that they could be used to index workload/OFS in the present study.

Table 3

Workload measures sensitive to taskload changes in the CD task (Abich, 2013)

	CD in a single task condition	CD as part of a dual task condition
Performance-based work-load measures	<ul style="list-style-type: none"> • Percent correct • Percent appeared correct • Percent disappeared correct • Percent movement correct 	<ul style="list-style-type: none"> • Percent correct • Percent appeared correct • Percent disappeared correct • Percent movement correct
Self-report workload measures	<ul style="list-style-type: none"> • ISA • All TLX subscales • TLX Global workload • MRQ-Manual • MRQ-Short Term memory • MRQ-Spatial attentive • MRQ-Spatial concentrative • MRQ-Spatial emergent • MRQ-Spatial positional • MRQ-Visual lexical • MRQ-Visual temporal 	<ul style="list-style-type: none"> • ISA • All TLX subscales • MRQ-Spatial concentrative • MRQ-Spatial emergent • MRQ-Visual temporal
Physiological workload measures	<ul style="list-style-type: none"> • Alpha band SPD at F3, F4, C3, POz • Beta band SPD at F3, Fz, C3, Cz • Beats per minute (BPM) • Inter-beat interval (IBI) • Heart rate variability (HRV) • Index of cognitive active (ICA) 	<ul style="list-style-type: none"> • Alpha band SPD at F4, C3 • Beta band SPD at Fz • Theta band SPD at C3 • HRV • Fixation duration • Convex-Hull NNI

*Workload measures showed a positive, monotonic relationship with CD taskload across the 3 levels of CD taskload examined.

Training, Testing and Cross-validation of the workload model

The procedure for developing and cross validating models typically involve (i) a training dataset which is used to obtain or “train” the model, (ii) a testing dataset against which the model is tested, and (iii) a validation dataset which is used to cross-validate the model (Jensen, 2013; Hastie et al., 2009). For predictive modeling, the cross-validation set is necessary to minimize model overfitting (Starkweather, 2011; Jensen, 2013; Hastie et al., 2009). Cross-validation enables the amount of variance explained by the model, as well as the predictive accuracy of the

model to be determined (Jensen, 2013). Indications of model fit include the root mean square error (RMSE) and R^2 indices.

For cross-validation of the workload model, data from a separate sample was collected. The sample was administered the same change detection task, but instead of the threat detection task as in the Abich (2013) study, the peripheral task, developed for the current research, was used as the second task instead. The purpose was to determine the robustness of the model when applied to a study in which the change detection task was paired with a task that captured another component of the ISR mission undertaken by a human-robot team. For successful cross-validation, when the predictors in the model are forcibly entered and applied to the testing set, the resultant R^2 should be similar to that in the training set. In addition, there should not be a large discrepancy between the predicted (i.e., from model derived from training set) and actual scores in the validation set (Tabachnick & Fidell, 2006). The added cross-validation provided additional support that the workload model was able to classify the workload resulting from the multi-task-load of an ISR mission. The study scenarios were designed to allow the relationships among taskload, workload and performance to be replicated, and included having the change detection task administered alone, as well as paired with various levels of the peripheral task. Workload was operationalized by the same physiological measures collected in the Abich (2013) study.

Research Plan

The purpose for the present work aimed to investigate the feasibility of using a closed-loop system that was based on physiological workload indices to drive adaptive robot aiding in a human-robot team during a simulated ISR mission. This endeavor entailed two successive phases corresponding to two studies with the following objectives and hypotheses:

Study 1 Objectives.

Objective 1.1. To verify the relationship between taskload and workload as a function of the demands posed by the change detection and peripheral tasks.

Objective 1.2. To derive a workload model from the Abich (2013) study and to test the model with the testing dataset.

Objective 1.3. To cross-validate the workload model with a new sample of participants who will be administered the change detection and peripheral tasks.

Study 1 Hypotheses.

Hypothesis 1.1: Taskload would inversely relate to performance. Given the findings in the Abich (2013) study, a negative relationship between taskload and performance was expected. For Study 1, when the Change Detection task was performed alone (Lower taskload), performance on the Change Detection task was hypothesized to be superior to that in the dual task conditions when the Change Detection task was performed with a Peripheral task (Higher taskload). When the two tasks were administered together, performance on each task was expected to be superior when the Change Detection task was paired with lower levels of the Peripheral task (Lower taskload) compared to a condition in which it was paired with higher levels of the Peripheral task (Higher taskload).

Hypothesis 1.2: Taskload and workload would be positively associated. The Abich (2013) study found that workload increased when taskload increased. The workload model derived from that study presupposed this relationship as taskload was be regarded as a proxy for workload. For Study 1, workload was expected to be lower when the Change Detection task was performed alone (Lower taskload) compared to when the Change Detection task was performed

concurrently with a Peripheral task (Higher taskload). Workload was hypothesized to be lower when the Change Detection task was paired with a lower level of the Peripheral task (Lower taskload), than when it was paired with a higher level of the Peripheral task (Higher taskload).

Hypothesis 1.3: The physiological workload model would generalize to the new sample.

Because the new sample was administered the same Change Detection task in the same ISR context as in the Abich (2013) study, the model derived from the Abich (2013) study was expected to yield a good fit with the data from the new sample.

Study 2 Objective.

Objective 2.1: To determine the effects of adaptive robot aid, based on physiological measures of workload, on performance, workload and stress.

In order to evaluate the effects of the closed-loop that adapts robot aid to workload as assessed by physiological measures, it was necessary to have a “control” condition/block when the robot aid was not adapted to the participant’s workload needs. This control was needed to address competing hypotheses that performance improvements were due merely to the presence of the aid should it be observed that performance in the presence of robot aid was indeed superior to that when aid was absent. In fact, a previous study with the same change detection task and aid manipulation showed that the aid was effective in improving CD performance (Taylor, 2012; 2013). Hence Study 2 design included conditions/blocks in which participants were provided adaptive aid based on their physiological measures of workload (a physio.-based trigger), as well as a condition/block where aid was imposed, i.e., aid was not adapted/ matched to participants’ workload needs (a system-based trigger). Study 2 design entailed the same CD and TD tasks used in Abich (2013).

Study 2 Hypotheses.

Hypothesis 2.1: Taskload would inversely relate to Performance. The Abich (2013) study showed a negative relationship between taskload and performance. Since Study 2 utilized the same CD task in the same ISR context, the same relationship was expected to be observed.

Hypothesis 2.2: Taskload and Workload would relate positively. Study 2 utilized the model derived from Abich (2013) which found a positive relationship between taskload and workload. This relationship was expected therefore to hold in Study 2 which employed the same CD task and ISR context.

Hypothesis 2.3: Performance would improve with aid. The aid in the form of an auditory BEEP alert was expected to increase the saliency of the change event without taxing the same visual channel as the tasks (Wickens, 1980). A previous study that examined the same CD task and aid showed that the auditory aid led to improved performance both under static and adaptive automation (Taylor, 2012; 2013). It was expected that performance will improve with the provision of aid in the present study.

Hypothesis 2.4: Performance improvements from the aid would be greater for those who were provided aid based on physiologically-based workload measures compared to those who had aid imposed. As the physiologically-based workload index should reflect the individual's workload needs more accurately, the closed-loop system should allow the schedule of the aid to more closely follow the individual's changes in workload needs, leading to better performance.

CHAPTER THREE: METHODOLOGY (STUDY 1)

Study Goal

To recapitulate, the goals for Study 1 were to (i) to determine the relationship among taskload, workload and performance for the change detection and peripheral tasks, (ii) to derive a workload model from the Abich (2013) study and test the model with the testing dataset, and (iii) to cross-validate the workload model with a new sample of participants who were administered the change detection and peripheral tasks.

Study Design

A one-way repeated measures ANOVA with 4 levels (peripheral task absent, peripheral task level 1, peripheral task level 2, peripheral task at level 3) repeated measures ANOVA design was adopted for Study 1.

Study Scenarios

The study scenarios are summarized in Table 4. All scenarios began with the CD task at low level for the first 5 minutes, followed by a high level of the task for the next 7 minutes. CD task level was manipulated by varying event rate. This was to simulate the “ramping up” of taskload commonly encountered in military operations. Scenario 1, with only the CD task, was a close replication of a scenario in the Abich (2013) study. Scenarios 2, 3 and 4 consisted of the CD task paired with various levels of the Peripheral task.

Table 4
Scenarios in Study 1

Time from start of block	Level of CD taskload	Level of peripheral taskload
Scenario 1: Change Detection task only (Peripheral task absent)		
0-5 minutes (5 min)	Low (6 changes/min)	N.A.
5-12 minutes (7 min)	High (24 changes/min)	
Scenario 2: Change Detection task & Peripheral task at Level 1		
0-5 minutes (5 min)	Low (6 changes/min)	LOW (5 prompts/3min*)
5-12 minutes (7 min)	High (24 changes/min)	
Scenario 3: Change Detection task & Peripheral task at Level 2		
0-5 minutes (5 min)	Low (6 changes/min)	MEDIUM (7 prompts/3 min*)
5-12 minutes (7 min)	High (24 changes/min)	
Scenario 4: Change Detection task & Peripheral task at Level 3		
0-5 minutes (5 min)	Low (6 changes/min)	HIGH (9 prompts/3 min*)
5-12 minutes (7 min)	High (24 changes/min)	

*after Reinerman-Jones et al., 2010.

Power Analysis

From previous studies (i.e., Abich, 2013) that utilized the same Change Detection task and workload measures, the effect to be detected was expected to be medium in size. For a medium effect size (i.e. $f=0.25$), and at alpha 0.05, the required sample size for 4 groups (one-way within-subjects ANOVA) is 36 based on computations using G*Power (Faul, Erdfeld, Buhner & Lang, 2007).

Participants

Forty students (22 males, 18 females) aged between 18 to 40 years ($M=19.95$) from the University of Central Florida psychology research participation pool served as participants. They were awarded course credit for participation. Participants were all self-reported to be right-

handed, had normal color vision as verified by the Ishihara color vision test, and normal or corrected-to-normal vision (self-report). They were required to refrain from consumption of any alcohol, caffeine, sedative medication 24 prior to the study.

Experimental Task/Context

The context of the study was that of a military operation in an entirely PC-based, virtual and simulated environment. Each participant was told that s/he was a Soldier on an Intelligence, Surveillance and Reconnaissance (ISR) mission with an Unmanned Ground Vehicle (UGV) robot teammate that has been equipped with a range of sensors (e.g. high-definition cameras, light detection and ranging (LIDAR) sensors, radars, radiant heat-sensing units), enabling it to sense entities and activities over a wide radius. During the mission, the UGV patrolled the area collecting reconnaissance information, which would be sent back to the Soldier, who was located at a base. The feed from the UGV would be presented to the Soldier in his display as an aerial map with icons representing the enemy activities/entities sensed. For the mission, the Soldier was to detect and identify changes (reconnaissance information) presented on his display, and monitor the whereabouts of the UGV teammate patrolling within the reconnaissance area. The ISR mission was represented as two tasks – (i) the change detection task, and (ii) a peripheral task of monitoring the robot as it scouted the AOI for reconnaissance information.

The Soldier's display comprised 3 windows: (a) a main change detection map, (b) a smaller aerial map of the same geographical terrain showing the route taken by the UGV, and (c) a video feed that showed the UGV's "ground level" view of what it is viewing through its front-facing cameras. The Soldier performed the change detection task within the main change detec-

tion map window, and monitored the UGV's whereabouts and maintained awareness of its surroundings for the peripheral task through the smaller aerial map and the video feed ("ground level" view). The blue outlined area refers to the "ground level" view of the UGV, the yellow outline denotes the "route map", and the red outline indicates the Change Detection map and response buttons.

Figure 2. Screen capture of the MIX testbed (outlines overlaid to differentiate each window but were not present in the actual task)

Change Detection Task.

The change detection task was developed and administered on the Mixed Initiative eXperimental testbed (MIX testbed; see (Figure 2); Reinerman-Jones, Barber, Lackey, & Nicholson, 2010), which simulated an operator control unit (OCU) for the Unmanned Ground Vehicles (UGVs). The main window where the change detection task was administered consisted of a map

depicting an aerial view of a geographical terrain that is strewn with icons representing “points of interest”. Immediately above the map, there were three response buttons, labeled “Appeared,” “Disappeared,” and “Movement,” corresponding to the three types of changes. Participants were to detect and identify any changes to the icons as soon as that they occur. Three types of changes/events were possible: (i) an appeared event, where new icon(s) were added to the map, (ii) a disappeared event, where existing icon(s) were deleted from the map, and (iii) a movement event, where the change is in the location of icon(s). The participant was instructed to click on the appropriate button indicating the type of change that occurred to the icons as soon as s/he detected a change. In all movement events, the change in location amounted to moving the icon(s) a small distance away (approximately 100 pixels) from its previous location, in any direction. This was so that the movement event would not be misconstrued as a disappeared event that is immediately followed by an appeared event at a farther location. Participants were told that only one type of change would occur at any time and that they were not required to click on the location of the change on the map itself.

The icons used in the task were adapted from the Department of Defense (2005), and symbolized certain types of military assets and statuses (see Figure 3). However, participants were not informed of their associated meaning and were not required to distinguish among them. They were merely instructed to detect and identify any change that occurred to the icons. To ensure that the icons were visible on the map and could be easily viewed as distinct and separate entities, the size of the icons was set to 28 x 28 pixels and they were separated by a minimum of 100 pixels. Throughout the task, there were approximately equal numbers of icons of each color (i.e., red, blue, yellow) and shape.

Figure 3. Icons used to represent entities for the change detection task.

As the task parameters and manipulations in previous studies (i.e., Abich, 2013; Taylor, 2012) have been shown to successfully elicit low and high workload as indicated by a multiple measures, the same task parameters and manipulations were used in the present study. However, for the present ISR mission, the change detection task comprised a 5-minute “Low” taskload segment that was followed by a 7-minute “High” taskload segment. This was to allow for a ‘ramping up’ into high taskload. Taskload levels were manipulated by changing the event/change rate. In the “Low” taskload condition, 6 changes occurred per minute, while in the “High” taskload condition, there were 24 changes per minute. As the event rate was to be the only manipulation of taskload, the set size (i.e., the average number of icons present on the map) was held relatively constant at 24 icons to avoid the potential influence of set size on performance (Tollner-Burngasser, Riley, & Nelson, 2010). In addition, because the number of icons involved in a change event would affect the level of saliency of the change (i.e., 1 icon appearing vs. 5 icons appearing at once), every change involved 2 icons appearing/disappearing/moving at once (i.e., “medium level of saliency” in Abich, 2013). Participants were not explicitly informed of these task parameters.

Peripheral Task.

Being part of a team, the Soldier would need to be aware of the UGV robot teammate's whereabouts and surroundings as it patrols the AOI. The peripheral task was designed to simulate this aspect of the mission. From the route map and video feed, participants were able to monitor the whereabouts of the robot teammate. Auditory prompts administered at predetermined time-points assessed the extent to which the participant was aware of the robot teammate's whereabouts during the ISR mission. The prompts required a verbal response, and included questions such as, "In which direction was the robot heading before the last turn?" and "Did the robot pass any men since the last turn?" It was especially crucial that adding the peripheral task minimally modified the visual display and visual scan pattern from that in the Abich (2013) study. Hence, the display layout and visual stimuli in the Abich (2013) study were retained, and the prompts were administered auditorily and required verbal responses instead of a visuo-motor one such as a mouse click or key stroke. This was also to prevent response competition with the response mode in the change detection task (i.e., mouse clicks). The intention of having the two tasks in different stimulus and response modes was to reduce the interference of the peripheral task on workload experienced from the change detection task as predicted by the Multiple Resource Theory (MRT; Wickens, 1980; 1984). Furthermore, as the study adopted a repeated measures design, the routes traveled by the robot teammate in needed to be different across scenarios, but yet similar enough so that difficulty level of the peripheral task was equated across scenarios. A route with multiple turns and intersections was set in the virtual environment and this was rotated and/or reflected across scenarios. This ensured that the route for all scenarios had comparable number of cross junctions and T-intersections, as well as North, South, East and West facings.

For Study 1, three levels of the peripheral task were tested to determine the level of peripheral task that, when paired with the change detection level task, would result in a substantial level of taskload that warrants robot aiding. In a previous study that involved the same change detection task paired with an auditory task, the frequency of the auditory prompt was set at 7 prompts per 3 minute block (Reinerman-Jones et al., 2010). Based on this, the three levels of the peripheral task, differing on event rate, were determined, and these are summarized in Table 5:

Table 5
Levels of the Peripheral Task in Study 1

	Peripheral task Level 1	Peripheral task Level 2	Peripheral task Level 3
Base rate for prompts	5 prompts/3 min	7 prompts/3 min*	9 prompts/3 min
No. of prompts in Low level segment of CD task	5	7	9
No. of prompts in High level segment of CD task	12	16	21
Total no. of prompts for entire scenario	17	23	30

*after Reinerman-Jones et al., 2010.

Measures

Physiological measures.

At the start of the mission, the participant was fitted with various physiological sensors that captured different measures of workload. These included oxygenation levels and velocity of cerebral blood flow, heart rate and brain activity among others.

Electroencephalogram (EEG).

Brain activity was monitored and recorded with the B-Alert X10 electroencephalogram system from Advanced Brain Monitoring (see Figure 4). Using the international 10-20 system,

brain activity at nine sites Fz, F3, F4, C3, Cz, C4, P3, Pz, P4 were recorded. Additional electrodes placed on either mastoid served as the reference and ground. At each site, three bandwidths, alpha (8-13 Hz), beta (14-30 Hz) and theta (4-7 Hz) were sampled at a rate of 256 samples per second and subject to power spectral density (PSD) analysis. By selectively combining the bandwidths from the nine individual sites, brain activity at the left and right hemispheres, as well as the frontal, temporal and parietal lobes were computed.

Figure 4. Electroencephalogram (EEG) electrode placement.

Electrocardiogram (ECG).

Measures of heart rate (HR), heart rate variability (HRV), and interbeat interval (IBI) were obtained with the Advanced Brain Monitoring System's B-Alert X10 system (see Figure 5). Sampling at 256 samples per second, single-lead electrodes were placed on the center of the right clavicle and on the lowest left rib to maximize the amplitude of the R-wave (Henelius et al., 2009). HR was measured in beats per minute, IBI was computed from the intervals between successive R-wave peaks, while HRV comprised the statistical variance of the time elapsed between heartbeats across a time interval.

Figure 5. Electrocardiogram (ECG) electrode placement.

Functional Near-Infra Red (fNIR).

Cerebral oxygen saturation (rSO_2) levels were recorded in real time with Somantics' In-vos Cerebral/Somantic oximeter (model 5100C; see Figure 6). The sensors, held in place with a headband, were placed on the left and right sides of the forehead and thus measured oxygenated hemoglobin (oxy-Hb) and deoxygenated hemoglobin (deoxy-Hb) of the left and right prefrontal cortex (Chance, Zhuange, UnAh, Alter, & Lipton, 1993; Ayaz et al., 2010, Ayaz et al., 2011).

Figure 6. Functional Near-Infra Red (fNIR) sensors.

Transcranial Doppler ultrasonography (TCD).

Recordings of cerebral blood flow velocity (CBFV) from the medial cerebral artery (MCA) were obtained by Spencer Technologies' ST³ Digital Transcranial Doppler system, model PMD150 (see Figure 7). A Marc 600 headframe held two 2-MHz ultrasound transducers that were secured proximally to the zygomatic arch on either side of the skull, along the temporal

bone. Ultrasound gel was applied between the transducers and the skin to enhance the signals. CBFV measures were derived from high pulse repetition frequency (PRF).

Figure 7. Transcranial Doppler Ultrasonography (TCD).

Eye tracker.

The faceLAB 5 desk-mounted eye tracking system by Seeing Machines was used to capture ocular data such as the number and duration of fixations, and saccades, taking into consideration head position and eye gaze direction (see Figure 8). The device comprised stereo cameras and two infrared light sources which accommodated a variety of head positions, eye gaze directions, as well as most types of eyewear. In addition to measures such as number of fixations, fixation duration, the data collected enabled computation of an Index of Cognitive Activity (ICA), based on pupil diameter.

Figure 8. Desktop Eye tracker

Subjective measures.

Perceived workload.

Perceived workload was assessed in three ways: (i) general workload perceived during different time points in the task (Instantaneous Self-Assessment; ISA), (ii) workload perceived in terms of the appraisals of task demands and responses to these demands (NASA-Task Load Index; TLX).

The Instantaneous Self-Assessment (ISA; Hulbert 1989, Jordan 1992) developed by the UK Civil Aviation Authority assesses immediate subjective workload during the performance of a task (Tattersall & Foord, 1996). It consists of an auditory prompt, “Please rate your workload”, which was administered in the middle of both the “Low” and “High” taskload segments of the Change Detection task in all the scenarios. Participants verbally responded with their ratings of perceived workload at that moment. Ratings were made on a 5-point Likert scale, ranging from a rating of “1” denoting “Under-utilized”, to a rating of “5”, indicating “Excessive” workload.

Another measure of workload was the NASA Task Load Index (NASA-TLX; Hart & Staveland, 1988). Apart from a global workload index, the measure identifies six sources of workload, three of which reflect appraisals of the task (Mental Demand, Physical Demand, and Temporal Demand), and three of which reflect appraisals of the individual's response to the task (Performance, Effort, and Frustration). The NASA-TLX was administered after each of the 4 scenarios.

Stress states.

The Dundee Stress State Questionnaire (DSSQ; Matthews et al., 2002; 2013) was used to assess perceived stress, which was conceptualized as comprising three broad dimensions: (i) Task Engagement, which refers to the appraisal theme of commitment to effort towards the task and the energetic and motivational aspects of stress, (ii) Distress, which involves evaluations of the task that pertain to overload of information processing capacity and includes the affective component of the stress response to the task, and (iii) Worry, which entails self-evaluative appraisals and the self-reflective aspects of stress (Matthews et al., 2002). The pre-task version of the DSSQ was administered to acquire a baseline level of stress, while post-task version of the DSSQ was administered after each scenario.

Performance measures.

As the ISR mission comprised the main task of change detection and the peripheral task of being aware of the UGV's whereabouts and its orientation in the virtual environment, there were two performance measures. The first pertained to the accuracy of change detection assessed by the percentage of correct detections. The second related to the extent to which the Soldier was in-the-loop and aware of the robot teammate's whereabouts. The Soldier responded verbally to

prompts to directional and situational information pertaining to the robot's environment. Performance on the peripheral task was assessed by the percentage of prompts that were answered correctly.

Procedure

After giving their informed consent and being screened for any deficiencies in color vision with the Ishihara color blindness test, and for any exclusion criteria with the restrictions checklist, participants completed a demographics questionnaire (see Appendix C). They were then fitted with the various physiological sensors and provided a brief explanation of each measure. After the setup, a 5-minute physiological resting baseline was recorded in which participants were instructed to relax, keep their eyes open, and not focus on anything in particular. Following that, participants completed a series of pre-task questionnaires. Subsequently, they were briefed on the ISR mission, instructed on the tasks and completed a practice mission. They then proceeded with the four study scenarios, the order of which were counterbalanced across participants. In-between each scenario, participants completed post-scenario questionnaires. When all scenarios have been administered, the physiological equipment and sensors were removed and participants were debriefed, thanked, and dismissed.

CHAPTER FOUR: RESULTS (STUDY 1)

Study 1 data were analyzed to determine the relationship among taskload, workload and performance for a sample that completed a different second task in the same ISR context.

Taskload and Performance

A 2(CD level) x 4(Scenario) repeated measures ANOVA was performed for CD performance. However CD levels were not counterbalanced as the block with low CD taskload always preceded the block with high CD taskload (to preserve operational fidelity where taskload is “ramped up”), the duration of CD Low was 5 min compared to CD High (7 min), and effects of the taskload transition are thus likely to be found in the CD Highblock. These confounds obscure the effects of CD level, so any findings involving the effects of CD level should be interpreted with the appropriate caution.

Results of the ANOVA indicated a significant main effect of CD level for CD performance, $F(1,39)=252.670$, $p<0.001$, $\eta_p^2=0.866$, where poorer CD performance was observed for the High level of CD ($M=35.054$, $SE=1.372$) compared to the Low level of CD ($M=54.597$, $SE=1.982$; $d=1.812$). There was a significant main effect of Scenario, $F(3,117)=9.242$, $p<0.001$, $\eta_p^2=0.192$, in which the mean correct detection was highest for the scenario with the lowest taskload (i.e., Scenario 1), and worst in the scenario with the highest taskload (i.e., Scenario 4). However post-hoc analyses showed that only CD performance in Scenario 1 (Single task condition) was significantly different from that in the other scenarios (Dual task conditions; see Table 6). The effect of PT taskload on CD performance was not significant as there were no statistically significant differences in CD performance among Scenarios 2, 3 and 4, which differed on PT

level. The CD level by Scenario interaction effect was also not significant ($p>0.727$ in each case).

Table 6

Percent correct detections on the Change Detection task for each scenario (N=40)

Scenario	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1 (CD only)	49.181	1.611	45.922	52.441
2 (CD + PT Low)	44.141	1.947	40.204	48.079
3 (CD + PT Med)	43.301	1.729	39.804	46.797
4 (CD + PT High)	42.679	1.895	38.846	46.511

Note: CD=Change detection task; PT=Peripheral task

For PT performance, a one-way repeated measures ANOVA for the three scenarios with the PT task, was performed. Results indicated a significant main effect of PT taskload, $F(2,74)=15.939$, $p<0.001$, $\eta_p^2=0.301$, where PT performance in the scenario with the lowest PT level (i.e., Scenario 2) was significantly worse than that of the other two scenarios, which did not differ significantly between each other (see Table 7). The effects of CD level on PT performance could not be determined because CD taskload varied within scenarios.

Table 7

Percent correctly-answered prompts on the Peripheral Task for each scenario (N=40)

Scenario	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
2 (CD + PT Low)	58.514	2.771	52.900	64.128
3 (CD + PT Med)	68.993	2.759	63.402	74.584
4 (CD + PT High)	69.386	2.610	64.099	74.673

Note: CD=Change detection task; PT=Peripheral task

Taskload and Workload

Physiological workload measures.

A series of 2(CD level) x 4(Scenario) repeated measures ANOVAs were performed for the various physiological workload measures. Results indicated a significant main effect of Scenario for Alpha Frontal, $F(2.520,98.289)=4.154$, $p=0.012$, $\eta_p^2=0.096$, Beta Parietal,

$F(3,117)=2.741, p=0.046, \eta_p^2=0.066$, and ICA mean, $F(3,102)=3.224, p=0.026, \eta_p^2=0.087$ (see Table 8).

Table 8
Physiological measures sensitive to Scenario effects (N=40)

Workload measure	Scenario 1 CD only	Scenario 2 CD+PT Low	Scenario 3 CD+PT Med	Scenario 4 CD+PT High
*Alpha Frontal ^a	$M= -0.299$ $SE=0.040$	$M= -0.366$ $SE=0.037$	$M= -0.307$ $SE=0.046$	$M= -0.339$ $SE=0.042$
*Beta Parietal ^a	$M=0.014$ $SE=0.037$	$M=0.066$ $SE=0.050$	$M=0.126$ $SE=0.059$	$M= 0.083$ $SE=0.049$
ICA mean ^b	$M=0.490$ $SE=0.019$	$M= 0.465$ $SE=0.018$	$M=0.456$ $SE=0.016$	$M=0.469$ $SE=0.016$

*: EEG measures were analyzed as the mean percent change of spectral power density (SPD) from baseline

^a: None of the pairwise differences were significant ($p>0.053$ in each case)

^b: Difference between Scenarios 1 and 3 are significant ($p=0.41$)

Significant main effects of CD level was observed for Beta Frontal, $F(1,39)=7.842, p=0.008, \eta_p^2=0.167$, Beta Parietal, $F(1,39)=5.387, p=0.026, \eta_p^2=0.121$, Theta Frontal, $F(1,39)=12.210, p=0.001, \eta_p^2=0.238$, HRV, $F(1,39)=22.792, p<0.001, \eta_p^2=0.369$, IBI, $F(1,39)=25.497, p<0.001, \eta_p^2=0.395$, Number of Fixations, $F(1,37)=317.141, p<0.001, \eta_p^2=0.896$, Fixation Duration, $F(1,37)=8.753, p=0.005, \eta_p^2=0.191$, ICA mean, $F(1,34)=4.573, p=0.040, \eta_p^2=0.119$, ICA std. dev., $F(1,34)=5.647, p=0.023, \eta_p^2=0.142$ (see Table 9).

Table 9
Physiological measures sensitive to CD taskload changes (N=40)

Workload measure	CD Low level	CD High level
*Beta Frontal ^c	$M=0.008$ $SE=0.038$	$M=0.051$ $SE=0.040$
*Beta Parietal ^c	$M=0.049$ $SE=0.043$	$M=0.096$ $SE=0.046$
*Theta Frontal SPD ^c	$M=0.017$ $SE=0.040$	$M=0.062$ $SE=0.041$
HRV ^c	$M=0.114$ $SE=0.033$	$M= -0.024$ $SE=0.034$
IBI mean ^c	$M=0.060$ $SE=0.011$	$M=0.041$ $SE=0.010$
Number of Fixations ^c	$M=1456.082$ $SE=46.560$	$M=1930.563$ $SE=59.262$

Workload measure	CD Low level	CD High level
Mean Fixation Duration ^c	<i>M</i> =241.496 <i>SE</i> =4.913	<i>M</i> =234.926 <i>SE</i> =4.010
ICA mean	<i>M</i> =0.465 <i>SE</i> =0.016	<i>M</i> =0.475 <i>SE</i> =0.016
ICA std.dev.	<i>M</i> =0.131 <i>SE</i> =0.003	<i>M</i> =0.128 <i>SE</i> =0.003

*: EEG measures were analyzed as the mean percent change of spectral power density (SPD) from baseline

^c: Pairwise difference significant ($p < 0.05$)

Although the CD level x Scenario interaction effect was significant for Beta Occipital, $F(3,117)=3.298$, $p=0.023$, $\eta_p^2=0.078$, and Theta Occipital, $F(2.128,82.981)=3.908$, $p=0.02$, $\eta_p^2=0.091$, the subsequent simple effects analyses that yielded no statistically significant results.

Subjective workload measures.

ISA.

A 2(CD level) x 4(Scenario) repeated measures ANOVA was conducted for ISA ratings. There was a significant main effect of CD level, $F(1,36)=53.348$, $p < 0.001$, $\eta_p^2=0.597$, and Scenario, $F(3,108)=10.693$, $p < 0.001$, $\eta_p^2=0.229$ on ISA ratings. Subsequent analyses revealed that ISA workload ratings were significantly higher when CD taskload was high ($M=3.736$, $SE=0.108$), relative to when CD taskload level was low ($M= 2.932$, $SE=0.123$; $d=1.098$). Across scenarios, differences in ISA ratings were mostly between Scenario 1 (Single task condition) and the other scenarios (Dual task condition) (see Table 10).

Table 10
ISA workload ratings in each scenario (N=40)

Scenario	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1 (CD only)	2.973	.124	2.722	3.224
2 (CD + PT Low)	3.541	.132	3.273	3.808
3 (CD + PT Med)	3.297	.119	3.057	3.538
4 (CD + PT High)	3.527	.119	3.285	3.769

Note: CD=Change detection task; PT=Peripheral task

TLX.

A one-way repeated measures MANOVA was performed for the perceived workload ratings on the TLX subscales as these were moderately correlated. Wilk's lambda revealed a significant effect of Scenario, $\Lambda = 0.283$, $F(18,22)=3.096$, $p=0.007$, $\eta_p^2=0.717$, on perceived workload. Univariate analyses of the Scenario effect indicated significant results for Mental Demand, $F(2,308,90.031)=18.143$, $p<0.001$, $\eta_p^2=0.317$, Temporal Demand, $F(3,117)=13.908$, $p<0.001$, $\eta_p^2=0.263$, Effort, $F(2,527,98.567)=7.736$, $p<0.001$, $\eta_p^2=0.166$, and Frustration, $F(3,117)=5.602$, $p=0.001$, $\eta_p^2=0.126$. With the exception of the Effort subscale, workload ratings were significantly lower for Scenario 1 (Single task condition) compared to the Dual task conditions (see Table 11). For Effort, significantly lower ratings were obtained for Scenario 1 compared to Scenarios 3 and 4. The pairwise difference between Scenarios 1 and 2 were not statistically significant ($p=0.709$). No analyses were conducted on the Global workload measure of the TLX as the MANOVA of the TLX subscales already indicate the results of global workload.

Table 11
TLX workload ratings by Scenario

Measure	Scenario	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Mental Demand*	1 (CD only)	69.875	3.250	63.302	76.448
	2 (CD + PT Low)	82.125	2.569	76.929	87.321
	3 (CD + PT Med)	84.000	2.238	79.474	88.526
	4 (CD + PT High)	84.750	2.221	80.257	89.243
Physical Demand	1 (CD only)	29.375	4.655	19.959	38.791
	2 (CD + PT Low)	33.375	4.488	24.298	42.452
	3 (CD + PT Med)	35.625	4.576	26.370	44.880
	4 (CD + PT High)	35.625	4.741	26.036	45.214
Temporal Demand*	1 (CD only)	60.875	3.580	53.634	68.116
	2 (CD + PT Low)	73.750	3.266	67.144	80.356
	3 (CD + PT Med)	75.875	2.691	70.433	81.317
	4 (CD + PT High)	77.750	2.802	72.083	83.417

Measure	Scenario	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Effort†	1 (CD only)	68.250	3.170	61.838	74.662
	2 (CD + PT Low)	74.375	3.545	67.204	81.546
	3 (CD + PT Med)	78.750	2.708	73.273	84.227
	4 (CD + PT High)	82.000	2.270	77.408	86.592
Frustration*	1 (CD only)	51.875	4.216	43.346	60.404
	2 (CD + PT Low)	61.250	4.186	52.783	69.717
	3 (CD + PT Med)	62.875	4.236	54.307	71.443
	4 (CD + PT High)	64.000	4.279	55.345	72.655
Performance	1 (CD only)	47.500	3.629	40.159	54.841
	2 (CD + PT Low)	53.000	4.026	44.856	61.144
	3 (CD + PT Med)	52.375	4.019	44.246	60.504
	4 (CD + PT High)	54.125	4.274	45.480	62.770

*Significant univariate result, Scenario 1 significantly lower than all other scenarios. ($p < 0.037$ in each case)

†Significant univariate result, Scenario 1 significantly lower than Scenarios 3 and 4 ($p < 0.028$ in each case)

Note: CD=Change detection task; PT=Peripheral task

DSSQ.

Analyses on the DSSQ were performed on the change scores from Pre-Task DSSQ. A one-way repeated measures MANOVA was performed for the ratings on the DSSQ stress state subscales as these were moderately correlated. Wilk's lambda revealed a significant effect of Scenario, $\Lambda = 0.567$, $F(9,31) = 2.633$, $p = 0.022$, $\eta_p^2 = 0.433$, on stress states. Univariate analyses of the Scenario effect indicated significant results for Distress, $F(2.510, 97.880) = 6.541$, $p = 0.001$, $\eta_p^2 = 0.144$, and Worry, $F(3, 117) = 3.250$, $p = 0.024$, $\eta_p^2 = 0.077$, but not for Task Engagement, $p = 0.757$. The increase in the level of Distress from Baseline for Scenario 1 was significantly smaller ($M = 8.200$, $SE = 0.736$) compared to that in Scenario 3 ($M = 11.325$, $SE = 1.152$) and Scenario 4 ($M = 11.900$, $SE = 1.124$). Subsequent analyses for Worry did not yield any significant results, $p > 0.112$ in each case.

Results from Study 1 data confirmed the stability of the Taskload-Performance and Taskload-Workload relationships for the CD task observed previously in the Abich (2013) study. The

data from the current Study 1 were then used for cross-validation in the subsequent workload modeling effort.

CHAPTER FIVE: MODELING (STUDY 1)

Workload modeling with logistic regression

Logistic regression was conducted to identify significant indices/predictors of taskload condition of the Abich (2013) data. Due to the stable positive relationship between taskload and workload, and the absence of workload dissociations/insensitivities for the levels of CD taskload examined, taskload condition was used as a proxy for workload, i.e., low/high workload would have been elicited when taskload is low/high. Using taskload instead of workload as the criteria also avoided the circularity of utilizing workload measures to predict workload.

Data preparation included Winsorization to reduce the influence of outliers (Tabachnik & Fidell, 2006) and imputation of missing values by the series mean for the respective conditions. Variable selection was conducted by (i) reviewing literature for physiological measures that indexed workload, (ii) performing paired-samples t-tests to identify variables that discriminated between Low and High taskload levels, (iii) conducting correlational analyses and Principle Component Analysis (PCA) to reveal underlying components and relationships among the variables. This was to ensure that the number of variables was appropriate for the sample size (Green, 1991) and that variables were not highly-correlated, minimizing the likelihood of multi-collinearity. The following 13 variables were selected and entered in the subsequent logistic regression procedure using the stepwise backward method:

1. Mean fixation duration (Fixation duration)
2. Number of fixations (No. of fixations)
3. Mean pupil diameter (Pupil diameter)
4. Mean Index of Cognitive Activity (ICA)
5. Heart rate variability (HRV)
6. Mean heart inter-beat interval (IBI)
7. Mean SPD of Theta at frontal midline (Theta Fz)

8. Mean SPD of Alpha at parietal midline (Alpha POz)
9. Taskload index (TLI) (i.e., Frontal Theta/Parietal Alpha (Gevins & Smith, 2003))
10. Mean rSO₂left (rSO₂ left)
11. Mean rSO₂right (rSO₂right)
12. Mean CBFV left (CBFV left)
13. Mean CBFV right (CBFV right)

After identifying the potential predictors, the Abich (2013) data were then divided into a training set and a testing set. Each participant had an equal chance of being assigned to either the training or the testing set. The training set was be used to develop/train the model using backward stepwise logistic regression, which minimizes suppressor effects (Fields, 2013), while the testing set, comprising participants not in the training set, was be used to test the model. To determine stability and model robustness, the data were divided into the training and testing sets five times so that there were five sets of training and testing sets, which resulted in five regression models. The model from each training set was tested against its corresponding testing set. Model test results were considered favorable when the R^2 that resulted from forcibly entering the model's predictors to the testing set was similar to the R^2 of the training set. Good model fit was obtained if the scores from applying the model to the testing set, and the actual testing set scores, showed minimal discrepancy (Tabachnick & Fidell, 2006). To incorporate the five models into the closed-loop system, the following architecture was proposed:

Step 1: For each of the 5 models, determine the significant predictors of workload and the corresponding regression coefficients.

Step 2: Determining the value of p (i.e., probability that the physiological measures had originated from an individual experiencing high workload) from the regression coefficients of the significant predictors for that model and value of the physio measure at each timepoint (a timepoint comprised the data of the physiological measures over last 2

minutes, with a rolling window of 30 seconds). For each model, if $p > 0.5$, then workload at that timepoint would be classified as being HIGH.

Logistic regression equation:

$$\text{Logit}(p) = \text{Constant} + \beta_1(X_1) + \beta_2(X_2) + \dots \beta_n(X_n)$$

$$p = \frac{1}{1 + e^{-\text{logit}(p)}}$$

where p = probability that the physiological measures originated from a High workload state (if p is > 0.5 , classify as HIGH).

β = standardized logistic regression coefficient

X = value of the physiological measure at a specific timepoint.

Step 3: The number of models that has to indicate HIGH workload for the classification of HIGH at a specific timepoint.

Step 4: The number of consecutive timepoints with the classification of HIGH that must be obtained before adaptive aid is invoked.

Results of the logistic regression

The results of the logistic regression modeling with the training sets and the attempts to test the models with the corresponding testing sets are summarized in Table 12. Successful model validation occurred when the R^2 of the training set and that of the testing set after applying the model, are comparable. A small discrepancy (i.e., as indicated by a statistically significant correlation) between the predicted (i.e., from model derived from training set) and actual scores in the testing set would indicate that the model is generalizable (Tabachnick & Fidell, 2006).

Table 12

Results of model validation

	<i>Predictors in model from Training set</i>	<i>Training set R^2</i>	<i>Testing set R^2</i>	<i>Correl. of pred. & actual scores in Testing set*</i>
Training set 1 & Testing set 1	1. No. of fixations 2. Pupil diameter mean 3. HRV 4. IBI 5. Theta Fz 6. Alpha POz 7. TLI 8. CBFV left	$R^2 = 0.184$	$R^2 = 0.067$	$r = 0.194$; $r^2 = 0.0376$
Training set 2 & Testing set 2	1. HRV 2. Theta Fz	$R^2 = 0.145$	$R^2 = 0.038$	$r = 0.074$; $r^2 = 0.0055$
Training set 3 & Testing set 3	1. rSO ₂ Right 2. Alpha POz 3. TLI	$R^2 = 0.119$	$R^2 = 0.066$	$r = 0.021$; $r^2 = 0.0004$
Training set 4 & Testing set 4	1. HRV 2. CBFV left 3. ICA 4. Theta Fz	$R^2 = 0.199$	$R^2 = 0.072$	$r = -0.033$; $r^2 = 0.0011$
Training set 5 & Testing set 5	1. HRV 2. Theta Fz	$R^2 = 0.104$	$R^2 = 0.009$	$r = 0.066$; $r^2 = 0.0004$

*All correlations were not statistically significant

The modeling approach using logistic regression did not result in any stable workload model. The models from each of the five training sets yielded different predictors and coefficients that could not be validated against the respective testing set, suggesting that perhaps individual differences in physiological responses to changes in workload were too varied to enable an adequately stable model to be obtained. This notion that there are large individual differences in workload measures has been raised by other researchers (e.g. Moray, 1984; Meshkati & Loewenthal, 1988; Roscoe, 1993; Johannes & Gaillard, 2014). For a stable model to be achieved, the changes in the physiological workload measures from the different workload levels needed to be larger in magnitude than the interindividual differences in workload responses. That this was not observed in the present study suggests that either (i) a more robust workload manipulation was

needed, (ii) more sensitive (O'Donnell & Eggemeier, 1986) workload measures needed to be identified, or (iii) a more idiographic approach to modeling workload was needed.

An alternative workload manipulation in the Abich (2013) study used single and dual tasks as manipulations of low and high workload respectively (Matthews et al., 2015a). The two tasks, which were set in the same context of an ISR mission, were the (a) Change Detection task, and (b) Threat Detection task. The taskload was expected to be lower when the participant was only performing the Change Detection task (single task), and to be higher in dual tasking conditions, when the participant performed both tasks. The following section discusses an alternative approach to modeling using this workload manipulation.

Alternative approach to modeling

Reexamination of the Abich (2013) data showed that the workload responses, specifically the NASA-TLX ratings, were more distinct between low and high workload levels when taskload was manipulated using single task vs. dual tasks, compared to when taskload was manipulated with event rate. (Matthews, personal communication). This led to further investigation of the single/dual task manipulation of taskload. The experimental conditions, administered in a repeated measures design, in the Abich (2013) study are described in Table 13:

Table 13
Scenarios in Abich (2013) Study

	<i>Scenario 1 (Single task)</i>	<i>Scenario 2 (Dual task)</i>	<i>Scenario 3 (Single task)</i>	<i>Scenario 4 (Dual task)</i>
Change Detection Task (CD)	Task at Low, Medium, High levels	Task at Low, Medium, High levels	-	Task at a constant Medium level
Threat Detection Task (TD)	-	Task at a constant Medium level	Task at Low, Medium, High levels	Task at Low, Medium, High levels

The physiological measures for each condition were computed as a percentage change from resting baseline values. All measures were standardized by converting values to z-scores to ensure comparability. Scenarios 1 and 3 were single task conditions, while Scenarios 2 and 4 were dual task conditions. Correlations of the workload response difference scores (i.e. increases or decreases in workload levels) on the physiological measures between Scenarios 1 (single CD task) & 2 (dual task with the same CD task paired with the TD task), and that between Scenarios 1 (single CD task) & 4 (dual task with CD + TD) were found to be positive and significant. Similar findings were observed for the workload response difference scores from other single-dual task pairs - i.e. Scenarios 3 (single TD task) & (dual task with CD +TD) and between Scenarios 3 (single TD task) & 4 (dual task with the same TD paired with the CD task). All correlations were positive and significant ($p < 0.05$), and ranged from 0.299 (Theta at F3) to 0.820 (mean fixation duration). On the contrary, for event rate manipulation of workload, the correlations of workload response difference scores (i.e. increases or decreases in workload levels as reflected in the physiological measures) between the Low and High event rate segments of the single task in Scenario 1 and that between Low and High event segments of the single task in Scenario 3) were largely non-significant. The same was observed with the correlation of workload response difference scores obtained from the Low and High event rate segments of Scenario 2 and that of Scenario 4, which were both dual tasks (Matthews, personal communication). The differences in workload between single and dual tasks were consistently larger than those between low and high event rate workload manipulations, i.e., the single/dual task manipulation of workload was much stronger than the low/high event rate workload manipulation. Hence, the single/dual task workload manipulation was selected for use in the workload modeling.

Matching of Difference Scores

The consistently large differences in the physiological measures between single and dual task conditions, and the persistent differences in the NASA-TLX ratings for the single and dual task conditions indicated that the single task condition reliably resulted in lower workload than the dual task condition. Thus, the single task condition was considered as the condition that evoked low workload, and the dual task condition as the condition that elicited high workload. Difference scores obtained from a task pair comprising a baseline single (i.e., low workload) and a baseline dual (i.e., high workload) task condition would then be compared to the difference scores obtained from another task pair formed by pairing the same baseline single task (i.e., low workload task) with the experimental task. If the difference scores from the two task pairs matched, then the new task would be considered to have evoked a similarly high workload as the baseline dual task which was known to elicit high workload. This provided an approach to identifying when high workload was evoked in the experimental condition.

Hence, in this approach, there is a *Single Task Baseline* which is a single task condition that is known to elicit low workload, and a *Dual Task Baseline* which is a dual task condition that has been shown to evoke high workload. A *Task Pair* would comprise the pairing of any two conditions (e.g., a Task Pair comprising a single and a dual task condition would be a Single-Dual Task Pair, while a Task Pair consisting of a single task condition and another single task condition would be a Single-Single Task Pair). When a Task Pair consists of a Single Task Baseline condition and a Dual Task Baseline condition, the difference score between them is the *Benchmark Difference Score*. When a Task Pair is formed by a Single Task Baseline condition

and an experimental or unknown condition, the difference score between them is the *Test Difference Score*. If the Benchmark Difference Score and Test Difference Score matches, then the unknown or experimental condition can be considered to have evoked the same level of high workload as the Dual Task Baseline. Because the Benchmark Task Pair is Single-Dual Task Pair, matches are more likely if the experimental condition were another dual task (i.e., resulting in a Single-Dual Test Task Pair), than if the experimental condition was a Single task (i.e., resulting in a Single-Single Test Task Pair).

Algorithms to combine physiological measures

While difference score matching provided a means of classifying workload state, an algorithm was still required to combine the difference scores from the various physiological measures for each individual participant and to quantify the extent of similarity or match between the Benchmark Difference Scores (i.e., computed from the physiological measures from the Single Task Baseline condition and that from the Dual Task Baseline condition that formed a Single-Dual Task Pair), and the Test Difference Scores, computed with the physiological measures from the Single Task Baseline condition and that from the experimental condition. The workload index would indicate the extent of match or similarity between the set of Benchmark Difference Scores and the set of Test Difference Scores. The algorithm would need to accommodate individual variability in physiological responses to workload. Several algorithms were explored and evaluated in turn (Matthews, personal communication).

Algorithm 1: Proportion of repeated markers.

For this algorithm, the workload index reflected the similarity in physiological responses between an experimental task condition to that in a Dual task condition (previously determined

to reliably elicit HIGH workload). Physiological markers of workload for the individual were first identified from a Single Task Baseline and a Dual Task Baseline condition which result in the set of Benchmark Difference Scores. The physiological markers were the measures that showed a sizeable workload response difference score (i.e., a threshold defining a sizeable response was set as ≥ 0.5 SD or ≤ -0.5 SD after the z-score conversion) between the Single Task Baseline and the Dual Task Baseline conditions. Next, the second set of workload response difference scores (i.e., the Test Difference Scores) were obtained by pairing the earlier Single Task Baseline condition with the condition that elicits the unknown level of workload (i.e., the experimental condition). Comparison of markers among the Benchmark Difference Scores to the number of markers among the Test Difference Scores reflected the extent to which the physiological response observed in the experimental condition resembled the response observed during the Dual Task Baseline condition. In this algorithm, similarity of physiological responses was quantified as the proportion of physiological markers found among the Benchmark Difference Scores that re-emerged as markers among the Test Difference Scores:

$$\begin{array}{l} \text{Workload index} \\ \text{(or Proportion of repeated markers)} \end{array} = \frac{\text{Markers observed in both Benchmark and Test Diff. Scores}}{\text{Markers observed in the Benchmark Diff. Scores}}$$

For example:

$$\begin{array}{l} \text{Workload index} \\ \text{(or Proportion of repeated markers)} \end{array} = \frac{\text{HRV, IBI, Fix.duration}}{\text{HRV, IBI, Fix.duration, Theta frontal SPD}} = \frac{3}{4} = 0.75$$

If the experimental condition elicited a physiological workload response that is similar to that seen in the Dual Task Baseline condition, then the ratio would approach 1, indicating that the workload experienced in the experimental condition was similarly high to that elicited by the

Dual Task Baseline condition. A cutoff score, based on inspection of the graph of workload differences over time, would define the level of similarity in workload response between the Dual Task Baseline and the experimental conditions.

Algorithm 2: Distance between all difference scores.

In this algorithm, the workload index was the Euclidean distance between the set of Benchmark Difference Scores and the set of Test Difference Scores. Smaller Euclidean distances indicated greater similarity between the Benchmark and Test Difference Scores . Although there was no set range for Euclidean distance, the cutoff score for similarity would still be needed. The Euclidean distance was computed as follows:

$$d(x, y) = \sqrt{\sum_i^n (x_i - y_i)^2}$$

Where:
x = a physio. measure in the Benchmark Task Pair
y = the corresponding physio. measure in the Test Task Pair
i = a physiological metric (e.g., i=1 denotes HRV, i=2 denotes IBI etc.)

Algorithm 3: Number of sign matches.

This algorithm quantified the similarity of the Benchmark Difference Scores and Test Difference Scores as the number of times the sign (i.e., positive or negative) for corresponding measures matched, and this also constituted the workload index. For example, if there was an decrease in HRV from the Single Task Baseline to the Dual Task Baseline (i.e., positive sign resulting from subtracting HRV in the Dual Task Baseline condition from the HRV in the Single Task Baseline condition), and if the experimental condition elicited a similar level of workload as the Dual Task Baseline, then the Test Difference Score, which was computed from the Single Task Baseline and experimental condition should likewise be positive. The cutoff for the algorithm should be between half and the total number of measures because the number of signs that would be expected to match by chance would be half the number of measures.

Algorithm Evaluation

The Benchmark Difference Score would always be computed from the Single Task Baseline and a Dual Task Baseline Task Pair, while the Test Difference Score would be computed from the Single Task Baseline and an experimental condition. If the experimental condition were another dual task condition (i.e., forming another Single-Dual Task Pair), then the Benchmark Difference Scores and Test Difference Scores should be similar, but if the experimental condition were a Single task condition (i.e., forming a Single-Single Task Pair), then the Benchmark Difference Scores and Test Difference Scores should be dissimilar.

Study scenarios from both the Abich (2013) dataset and Study 1 dataset were utilized to form the Task Pairs. Study 1 dataset served as the cross-validation set. In addition, random data (i.e., z-scores randomly drawn from a normal distribution) were generated to provide a gauge of the level of difference score similarity that would be obtained by chance. To recapitulate, the scenarios in the Abich (2013) study and the scenarios from Study 1 are summarized in Table 14 and Table 15 respectively.

Table 14
Scenarios in the Abich (2013) study

	<i>Scenario 1</i> <i>(Single task)</i>	<i>Scenario 2</i> <i>(Dual task)</i>	<i>Scenario 3</i> <i>(Single task)</i>	<i>Scenario 4</i> <i>(Dual task)</i>
Change Detection Task (CD)	Task at Low, Medium, High levels	Task at Low, Medium, High levels	-	Task at a constant Medium level
Threat Detection Task (TD)	-	Task at a constant Medium level	Task at Low, Medium, High levels	Task at Low, Medium, High levels

Table 15
Scenarios in Study 1

	<i>Scenario 1</i> (<i>Single task</i>)	<i>Scenario 2</i> (<i>Dual task Low</i>)	<i>Scenario 3</i> (<i>Dual task Med</i>)	<i>Scenario 4</i> (<i>Dual task High</i>)
Change Detection Task (CD)	Task at Medium level	Task at Medium level	Task at Medium level	Task at Medium level
Peripheral Task (PT)	-	Task at Low level	Task at Medium level	Task at High level

Algorithms were evaluated on (i) how distinguishable the mean workload index values from similar Task Pairs were from the mean workload index values that would be obtained by chance, (ii) how distinct the distributions of workload indices were from similar Task Pairs and from random data, (iii) how discriminating the workload indices were over time, which would enable the setting of a cutoff threshold (indicating HIGH workload) in a mock-up of the closed-loop system based on the algorithm, and (iv) the level of sensitivity (d') of the algorithm.

Mean workload indices obtained from similar task pairs and by chance.

Various task pairs were formed. Each task pair yielded a set of difference scores. The difference scores reflected the changes in physiological responses between (i) Single and Dual task conditions from the Abich (2013) study (i.e., rows 1, 2, 3 & 4), (ii) Single task condition and random data which comprised z-scores randomly-drawn from a normal distribution (i.e., rows 5 & 6), and (iii) Single and Dual task conditions from Study 1 data (i.e., rows 7 & 8; see Table 16).

Table 16
Description of Task Pairs Used in Algorithm Evaluation

Task Pair	Description*
<i>Abich (2013) data</i>	
1. Scen 1-Scen 2 (Single-Dual)	Diff. score=Physio. measure in Scenario1 – Corresp. physio. measure in Scenario2. (e.g., HRV in Scenario 1 – HRV in Scenario 2)
2. Scen 1-Scen 4 (Single-Dual)	Diff. score=Physio. measure in Scenario1 – Corresp. physio. measure in Scenario4. (e.g., HRV in Scenario 1 – HRV in Scenario 4)

Task Pair	Description*
3. Scen 3-Scen 2 (Single-Dual)	Diff. score=Physio. measure in Scenario3 – Corresp. physio. measure in Scenario2. (e.g., HRV in Scenario 3 – HRV in Scenario 2)
4. Scen 3-Scen 4 (Single-Dual)	Diff. score=Physio. measure in Scenario3 – Corresp. physio. measure in Scenario4. (e.g., HRV in Scenario 3 – HRV in Scenario 4)
5. Scen 1-Random (Single-Random)	Difference score=Physio. metric in Scen.1 – Randomly-drawn z-score (e.g., HRV in Scenario1 – z-score randomly-drawn from normal dist.)
6. Scen 3-Random (Single-Random)	Difference score =Physio. metric in Scen.1 – Randomly-drawn z-score (e.g., HRV in Scenario3 – z-score randomly-drawn from normal dist.)
<i>Study 1 data</i>	
7. Scen 1- Scen 2 (Single-DualLow)	Diff. score=Physio. measure in Scenario1 – Corresp. physio. measure in Scenario2. (e.g., HRV in Scenario1 – HRV in Scenario2)
8. Scen 1- Scen 4 (Single-DualHigh)	Diff. score=Physio. measure in Scenario1 – Corresp. physio. measure in Scenario4. (e.g., HRV in Scenario1 – HRV in Scenario4)

*All physio. measures were computed as percentage change from baseline and converted to standardized z-scores.

These difference scores were then paired (see Table 17) and the Task Pairs were compared. Only difference scores from the same dataset were paired as the intent was that the first set of difference scores (i.e., Benchmark Difference Scores) showed the difference in the individual's workload response between Low and High workload, while the second set of difference scores (i.e., Test Difference Scores) revealed the difference in the individual's workload response between Low workload and the level of workload elicited by the experimental condition. The Benchmark Difference Scores and Test Difference Scores were expected to match (i.e., be similar) when both sets of difference scores originated from single and dual task conditions (i.e., rows 1, 2, & 5). Benchmark and Test Difference Scores would not match as well when the Test Difference Scores involved random data which are not expected to show similarity to scores obtained from the Dual Task Baseline condition (i.e., rows 3 & 4).

Table 17

Mean and SDs of Workload Indices from various Task Pairs under each algorithm

Benchmark Diff. Scores	Test Diff. Scores	Expected similarity	Wkld index* (Algrthm. 1)	Wkld index† (Algrthm. 2)	Wkld index‡ (Algrthm. 3)
<i>Abich (2013) data</i>					
1. Scen 1-Scen 2 (Single-Dual)	Scen 1-Scen 4 (Single-Dual)	High	0.552 (0.24)	4.718 (2.04)	20.114 (6.48)
2. Scen 3-Scen 2 (Single-Dual)	Scen 3-Scen 4 (Single-Dual)	High	0.537 (0.23)	4.978 (2.18)	19.483 (5.92)
3. Scen 1-Scen 2 (Single-Dual)	Scen 1-Random (Single-Random)	Low	0.287 (0.13)	7.624 (1.77)	14.356 (3.91)
4. Scen 3-Scen 2 (Single-Dual)	Scen 3-Random (Single-Random)	Low	0.323 (0.14)	7.518 (1.55)	14.631 (3.98)
<i>Study 1 data</i>					
5. Scen 1- Scen 2 (Single-Dual)	Scen 1- Scen 4 (Single-Dual)	High	0.509 (0.22)	4.677 (1.90)	20.68 (4.70)

*Difference score match for algorithm 1: Mean proportion of markers/measures among the Benchmark Difference Scores that were observed as markers among the Test Difference Scores (SD in parenthesis)

†Difference score match for algorithm 2: Mean Euclidean distance of the Benchmark Difference Scores and Test Difference Scores (SD in parenthesis)

‡Difference score match for algorithm 3: Mean number of matched signs (SD in parenthesis)

For algorithm 1, when the Test Task Pair was similar to the Benchmark Task Pair in that they each comprised single and dual tasks (i.e., rows 1, 2, and 5 in Table 17), the mean workload index obtained were comparable and ranged from 0.509 to 0.552. This indicated that the proportion of physiological measures that showed a substantive change between the single and dual conditions in the Benchmark Task Pair, which also showed a sizeable difference between the single and dual conditions in the Test Task Pair (i.e., proportion of repeated markers), ranged from 50.9% to 55.2%. Hence, 50.9% to 55.2% of the physiological markers observed among the Benchmark Difference Scores emerged as markers again (i.e., repeated markers) among the Test Difference Scores. In contrast, when the Test Task Pair included random data (i.e., rows 3 and 4 in Table 17), the proportion of repeated markers declined to 23.87% to 32.3%. Using random

data for the test task pair gave an indication of the proportion of repeated markers that may be obtained by chance.

Similar pattern of results were obtained for algorithm 2. When the Benchmark and Test Task Pairs similar (i.e., rows 1, 2, and 5 in Table 17), the mean workload index were comparable and the Euclidean distances were smaller (denoting greater similarity), ranging from 4.677 to 4.978. When the Test Task Pair included random data which indicated the Euclidean distances that would occur by chance (i.e., rows 3 and 4 in Table 17), the mean workload indices obtained were larger, at 7.518 and 7.624, indicating greater dissimilarity.

Algorithm 3 resulted in a similar pattern of results. When the Benchmark and Test Task Pairs were alike (i.e., rows 1, 2, and 5 in Table 17), the mean number of matched signs was higher, ranging from 19.483 to 20.68. However, using random data for the Test Task Pair, the number of matched signs that may have been obtained by chance were in the 14.356 and 14.631 range, indicating that the physiological workload responses from similar Task Pairs were more matched than the workload responses from dissimilar Task Pairs involving random data.

Additional analyses were conducted on Task Pairs that were similar or dissimilar in various ways in order to ensure that difference score matches were due to similarity in the change in workload response between single (low workload) and dual (high workload) tasks. Comparisons were made with difference scores from (i) Task Pairs with event rate as the taskload manipulation, (ii) Task Pairs with mixed taskload manipulations (i.e., Benchmark Task Pair used Single-Dual task manipulation while the Test Task Pair used event-rate manipulation), and (iii) Task Pairs with the same Single-Dual manipulation but from different studies (i.e., Benchmark Task Pair used a single and dual task conditions from Abich (2013) study, and the Test Task Pair used

single and dual task conditions from Study 1). Results confirmed that the best matches were obtained when both Benchmark and Test Task Pairs were both Single-Dual Task Pairs from the same individual. Matches were poor when Task Pairs used conditions that differed on event rate, suggesting that the event rate taskload manipulation was not as effective as the Single-Dual taskload manipulation. Matches were extremely poor when the Benchmark and Test Task Pairs were from different individuals even when both Task Pairs utilized Single-Dual task taskload manipulation.

That for each of the algorithms, the data from the new sample (Study 1 dataset) yielded comparable workload index values (i.e., means and standard deviations) to those from the Abich (2013) dataset suggests that these algorithms are applicable to a separate dataset which utilized the CD task and a different second task. All these indicated that the workload index values would be comparable so long as the Benchmark and Test Task Pairs are similar (i.e., both are Single and Dual Task Pairs).

Distributions of workload indices from similar and random task pairs.

In addition to examining the means of the workload indices, the distributions of the workload index from various algorithms were also compared. The graphs in Figure 9, Figure 10, and Figure 11 show the distribution of the workload index obtained from the various task pairs under the different algorithms. For each graph, there appeared to be two distinct distributions corresponding to the similar (filled-in circles) and dissimilar (open circles) Benchmark Task Pair and Test Task Pair.

*Larger workload index values on the x-axis indicate greater similarity between the Benchmark Difference Scores and the Test Difference Scores.

Figure 9. Distribution of the workload index under algorithm 1.

In Figure 9, for the S1-S2 & S1-S4 Task Pairs, which were both Single-Dual Task Pairs, 50% of the workload index values (i.e., proportion of repeated markers) were approximately 0.57 and below (see solid arrow). This was close to the mean of 0.552 obtained in Table 17. However, when the Task Pairs were dissimilar, such as those in the S1-S2 & S1-Random Task Pairs, 90%, or almost all of the workload index values were around 0.48 and below (see dotted arrow). Half of the lowest workload index values for similar Task Pairs exceeded almost all the workload index values that were obtained from dissimilar Task Pairs, indicating that similar Task Pairs are very likely to produce workload index values that are much higher than the values that would have been obtained by chance.

*Smaller workload index values on the x-axis indicate smaller distances (more similarity) between the Benchmark Difference Scores and the Test Difference Scores.

Figure 10. Distribution of the workload index under algorithm 2.

Similarly, in Figure 10, 90% or almost all the workload index values from S1-S2 & S1-S4 (similar Task Pairs) fell below 7.7 (solid arrow), demonstrating that most of the values showed small distances or more similarity between the difference scores. On the other hand, when the Task Pairs were dissimilar, such as S1-S2 & S1-Random, only 50% of the smallest distances (most similar) fell below 7.2 (dotted arrow), with the other 50% of values reaching distances as high as 10.

*Larger workload index values on the x-axis indicate greater similarity between the Benchmark Difference Scores and the Test Difference Scores.

Figure 11. Distribution of the workload index under algorithm 3.

In Figure 11, most (90%) of the workload index values from dissimilar Task Pairs S1-S2 & S1-Random fell below 18.1 (dotted arrow). The lowest 50% of values from similar Task Pairs C1-C2 & C1-C4 were below 21.1 (solid arrow), which shows that values for similar Task Pairs were likely to be higher than most of the values obtained by chance. However, values from similar Task Pairs can reach as low as 12 (i.e., 10% of the workload index values from similar Task Pairs S1-S2 & S1-S4 were below 12.1), indicating a fairly high likelihood that a particular observed value may be miscategorized.

The range of workload index values within which the graphs for similar and dissimilar Task Pairs overlapped seemed largest for algorithm 3 compared to algorithms 1 and 2. This extent of overlap represented a type of “effect size” (Nolan & Heinzen, 2010). The greater the overlap, the smaller the effect size. From the means and standard deviations in Table 17, effects sizes (Cohen’s *d*) obtained for algorithm 1 ranged from 1.009 to 1.373, that for algorithm 2

ranged from 1.343 to 1.605, while d for algorithm 3 ranged between 0.962 to 1.463. For algorithm 3, the workload index values that would have been obtained by chance and the workload index values from similar Task Pairs were only about 1 standard deviation apart (e.g., Similar Task Pairs: $M=20.114$, $SD=6.48$; Chance/Random level: $M=14.356$), and the overlap of the two distributions seemed greater compared to that for the other algorithms (see Figure 11). From these observations, the decision was made to exclude algorithm 3 from further consideration, as the discriminability of workload levels were closer to chance level relative to algorithms 1 and 2.

Simulation of the closed-loop system and setting cutoff levels.

Mock-up of the closed-loop system.

Algorithms 1 and 2 were selected to be “mocked up” in a closed-loop system using the Abich (2013) data. The results of the simulation would enable a cutoff or threshold level of the workload index to be determined. In addition to algorithms 1 and 2, two derivatives of algorithm 2 that computed the Euclidean distance using five and ten difference scores (identified from the Single-Dual Task Baseline Pair) of the largest absolute magnitude were also evaluated. This was a further attempt to individualize the algorithm such that measures that were not relevant to the individual (i.e., not markers of workload for the individual) were not included in the distance computation.

In the simulation, the data were “streamed” into the closed-loop application as samples of data comprising 2-minute blocks of data every 30 seconds (i.e., a 2-minute moving or “rolling” window) to simulate the updating of the workload index when the system goes ‘live’. Instead of the Test Difference Score being the difference score between the Single Task Baseline condition

and an entire experimental condition, the Test Difference Score in the simulation was the difference between the Single Task Baseline condition and the data from a sample of the 2-minute “rolling” window, i.e., a *Rolling Test Difference Score*, that was computed from the *Rolling Test Task Pair*. Furthermore, although there were study scenarios in the Abich (2013) study which included varying event rates, only the data from the medium event rate segment of each scenario were streamed in the mock-up (refer to Table 14 for study conditions). To verify the robustness of the algorithms, additional Task Pairs were formed from the study scenarios in the Abich (2013) data (see Table 18). The workload indices that result from the algorithms should reflect the level of similarity expected between the Benchmark Task Pair and Rolling Test Task Pairs relative to each other. Scenario 1 (Single task) and Scenario 2 (Dual task) were always used as the Benchmark Task Pair.

Table 18
Task Pairs for the Closed-loop Mock-up (Abich, 2013)

Benchmark Task Pair	Rolling Test Task Pair	Expected similarity between the Benchmark Difference Scores and Test Difference Scores
Scen 1-Scen 2 (Single-Dual)	Scen 1-Scen 1 (Single-Single)	<i>Very dissimilar</i> : The Benchmark Task Pair is a Single-Dual task pair, while the Rolling Test Task Pair is a Single-Single task pair, using the same condition.
Scen 1-Scen 2 (Single-Dual)	Scen 1-Scen 2 (Single-Dual)	<i>Very similar</i> : Both the Benchmark Task Pair and Rolling Test Task Pair are Single-Dual Task Pairs, comprising the same single task and dual task Scenarios.
Scen 1-Scen 2 (Single-Dual)	Scen 1-Scen 3 (Single-Single)	<i>Dissimilar</i> : The Benchmark Task Pair is a Single-Dual Task Pair, while the Rolling Test Task Pair is a Single-Single Task Pair using different single task conditions.
Scen 1-Scen 2 (Single-Dual)	Scen 1-Scen 4 (Single-Dual)	<i>Similar</i> : Both the Benchmark Task Pair and Rolling Test Task Pair are Single-Dual Task Pairs, but with different dual task scenarios.

In all the following graphs, difference scores from Task Pairs S12_S12 (i.e., Scenario1-Scenario2 as the Benchmark Task Pair, and Scenario1-Scenario2 as the Rolling Test Task Pair)

showed greatest similarity¹ as expected. This was reflected by the highest mean proportion of repeated markers for algorithm 1 and the smaller Euclidean distances for algorithm 2 and its derivatives. In contrast, difference scores from Task Pairs S12_S11 were expected to be most dissimilar since in S11, the second condition is the same as the first. Difference scores from S12_S13 and that from S12_S14 were most likely to be confounded. The algorithms were evaluated on the extent to which they could distinguish them across all data samples.

Figure 12. Workload index under algorithm 1 over the entire condition
(Sample number ranged from 8 to 14 because only the data from the Medium event rate segment in the conditions were used)

For algorithm 1, the expected order of task pairs from most similar (S12_S12) to most dissimilar (either S12_S11 or S12_S13) was obtained. Closer inspection of the workload index

¹ S12_S12 vectors were not completely similar because the Test Difference Score data were streamed as 2-min blocks, but the Benchmark Difference Score data were not.

values for S12_S13 and S12_S14 (i.e., the middle two sets of task pairs that were the most easily confounded) showed a substantive distance between the values for all of the samples, except sample number 14. In addition, workload index values for the different task pairs were relative stable across the samples, reflecting the constant taskload level. Setting the cutoff at 0.62 would adequately distinguish the workload index values as reflecting LOW (S12_S13) and HIGH (S12_S14) workload for all but one sample.

Figure 13. Workload index under algorithm 2 over the entire condition
 (Sample number ranged from 8 to 14 because only the data from the Medium event rate segment in the conditions were used)

In algorithm 2, the expected order of task pairs from most similar (S12_S12) to most dissimilar (either S12_S11 or S12_S13) was not observed. The values for S12_S13 (dissimilar task pairs) were smaller (denoting greater similarity) than that for S12_S14 (similar task pair).

Under this algorithm, workload index values observed did not reflect the extent of similarity of task pairs. Furthermore, the workload index values were more varied across the samples despite the constant taskload level. A possible cutoff score for this algorithm is 7.2, although at this cutoff, errors in categorizing samples 8, 10 and 14 would be highly probable.

Figure 14. Workload index under algorithm 2a over the entire condition
(Sample number ranged from 8 to 14 because only the data from the Medium event rate segment
in the conditions were used)

Algorithm 2a also used the Euclidean distance as a measure of response similarity. However, only the 5 markers that showed the greatest magnitude change were included in the workload index computation. By only including 5 markers, there was a lower possibility that markers irrelevant to the individual would be included in the computations of the index. The graphs showed that the differences in index values between similar and dissimilar Task Pairs

were not particularly large across all samples, and that values from similar and dissimilar Task Pairs could also be confounded (i.e., samples 9, 12, 13) should a cutoff of 3.4 be used.

Figure 15. Workload index under algorithm 2b over the entire condition (Sample number ranged from 8 to 14 because only the data from the Medium event rate segment in the conditions were used)

Under this algorithm, only 10 markers with the greatest magnitude of disparity in workload response between the Single and Dual Baseline Task conditions were used to compute the workload index. This was so that markers irrelevant to the individual would be left out of the index computations. The mock-up showed very close workload index values between the Task Pairs most likely to be confounded. For samples 9, 10, 12, 13, 14, the cutoff score of 4.4 would probably result in miscategorization.

In all the graphs, difference scores from Task Pairs S12_S12 (i.e., Scenario1-Scenario2 as the Benchmark Task Pair, and the same Scenario1-Scenario2 as the Test Task Pair) showed

greatest similarity² as expected. This was reflected by the highest mean proportion of repeated markers for algorithm 1 and the smaller Euclidean distances for algorithm 2 and its derivatives. In contrast, Test Difference Scores from Task Pairs S12_S11 were expected to be most dissimilar. Test Difference Scores from S12_S13 and that from S12_S14 were most likely to be confounded and the algorithms were evaluated on the extent to which they could distinguish them across all data samples. In this regard, algorithm 2, where the workload index comprised the Euclidean distance of all measures, did not show the expected pattern where difference scores for S12_S14 (both Single-Dual Task Pairs) are more similar than difference scores for S12_S13. As the workload index values from algorithm 2 failed to reflect the similarity between difference scores for S12_S14, and the dissimilarity between difference scores for S12_S13, algorithm 2 was excluded from further evaluation and consideration. On the other hand, algorithm 1 seemed to show the clearest distinction between the indices from difference scores that should be similar and that from difference scores that should be dissimilar.

Sensitivity of the algorithms.

From examining the graphs of the different Task Pairs, cutoff scores for each algorithm were proposed. With the cutoff score serving as a decision threshold for the adaptive aid, four outcomes were possible and these are shown in Table 19:

Table 19
Possible Outcomes of the Aiding Decision

	<i>Aid should be evoked</i>	<i>Aid should not be evoked</i>
	<i>S12 (Single-Dual) & S12 (Single-Dual)</i>	<i>S12 (Single-Dual) & S11 (Single-Single)</i>
	<i>S12 (Single-Dual) & S14 (Single-Dual)</i>	<i>S12 (Single-Dual) & S13 (Single-Single)</i>
<i>Aid was evoked</i>	HIT	FALSE ALARM (FA)
<i>Aid was not evoked</i>	MISS	CORRECT DECISION (CR)

² S12_S12 difference scores were not completely similar because the data for the Test Task Pair were streamed as 2-min blocks, but data for the Benchmark Task Pair were not.

As with any classification endeavor, there are false positives (i.e., experiencing low workload but classified as high workload state) and false negatives (i.e., experiencing high workload but classified as low workload state), and as appropriate decision criteria must be set to minimize one or both kinds of error. Similar Task Pairs (i.e., S12_S12 and S12_S14) resulted in a high rate of the aid being triggered (HIT), and low incidences of the aid not being provided (MISS). By contrast, dissimilar Task Pairs (i.e., S12_S11 and S12_S13) resulted in higher rate of the aid not being provided (CR, CORRECT REJECTION) and a low occurrences of the aid being triggered (FA, FALSE ALARM; see Table 20).

Table 20
Decision outcomes for different cutoffs under the various algorithms

Task pairs	Algorithm 1		Algorithm 2a		Algorithm 2b	
	Proportion of repeated markers		Euclidean distance of top 5 markers		Euclidean distance of top 10 markers	
Cutoff score evaluated	0.55	0.62	3.4	1.5	4.4	3.5
S12_S12*						
a) % given aid (HIT)	89.9%	88.6%	57.82%	36.73%	91.95%	79.87%
b) % not given aid (MISS)	10.1%	11.4%	42.18%	63.27%	8.05%	20.13%
S12_S14†						
c) % given aid (HIT)	68.5%	63.1%	68.71%	19.72%	75.17%	50.34%
d) % not given aid (MISS)	31.5%	36.9%	31.29%	80.27%	24.83%	49.66%
S12_S13‡						
e) % given aid (FA)	65.8%	61.7%	67.81%	14.38%	69.13%	44.30%
f) % not given aid (CR)	34.2%	38.3%	32.19%	85.62%	30.87%	55.70%
S12_S11§						
g) % given aid (FA)	36.2%	29.5%	9.46%	9.46%	68.76%	40.94%
h) % not given aid (CR)	63.8%	70.5%	90.54%	90.54%	36.24%	59.06%

*: S12_S12 task pair diff. scores should be very similar as the Benchmark and Test Task Pairs are identical.

†: S12_S14 task pair diff. scores should be similar as the Benchmark and Test Task Pairs are both Single-Dual Task Pairs.

‡: S12_S13 task pair diff. scores should be dissimilar as the Test Task Pair is Single-Single Task Pair.

§: S12_S11 task pair diff. scores should be dissimilar as the Test Task Pair is Single-Single Task Pair.

For each algorithm, two possible cutoff scores were evaluated. For Algorithm 1, shifting the cutoff from a less conservative 0.55 to more conservative 0.62 resulted in the expected decrease in Hit. The percentage in the Abich (2013) sample who would have been given aid in S12_S12 decreased from 89.9% to 88.6%, and decreased from 68.5% to 63.1% for S12_S14. False Alarms rates declined from 65.8% to 61.7% for S12_S3, and from 36.2% to 29.5% for S12_S11. For Algorithm 2a, shifting the cutoff from a less conservative 3.4 to more conservative 1.5 (lower values for Euclidean distance algorithms denote greater similarity), led to the expected decreases in Hit and False Alarm rate, but these decreases were much sharper: Hit rates decreased from 57.82% to 36.73% for S12_S12 and 68.71% to 19.72% for S12_S14, and False Alarm rates decreased from 67.81% to 14.38% for S12_S13 and there was no change for S12_S11. Similarly, for Algorithm 2b, moving the cutoff from a less conservative 4.4 to a more conservative 3.5 to 4.4 (lower values for Euclidean distance algorithms denotes greater similarity) resulted also in sharper decreases in Hit (91.95% to 79.873% for S12_S12 and 75.17% to 50.34% for S12_S14) and False Alarm rates (69.13% to 44.30% for S12_S13 and 68.76% to 40.94% for S12_S11). These findings suggest that the effects of cutoff scores were not as consistent under Algorithms 2a and 2b compared to that for algorithm 1.

To combine these outcome proportions from the different algorithms and the associated cutoffs, the signal detection measure of sensitivity, d-Prime (d') was used. The total number of participants for “Hits” was obtained from the S12_S12 and S12_S14 Task Pairs, while the number of participants for “False Alarms” was acquired from the S12_S11 and S12_S13 Task Pairs. Total Hit rate was obtained by averaging the Hit rates from S12_S12 and S12_S14, while total False Alarm rate was the average of the False Alarm rates from S12_S11 and S12_S13. These rates were then used to compute sensitivity by the following equation:

Sensitivity or $d' = Z(\text{Proportion of HITS}) - Z(\text{Proportion of FALSE ALARMS})$

The sensitivity of algorithm 1 was highest among the algorithms, and the cutoff of 0.62 was more sensitive than the 0.55 cutoff (see Table 21). Hence, algorithm 1, where the workload index was computed as the proportion of repeated markers, with the cutoff score of 0.62 (i.e., 62% of markers identified from the Benchmark Difference Scores were repeated among the Test Difference Scores), was selected for the closed-loop system.

Table 21
Sensitivity of the Various Cutoff Scores by Algorithm

	Average d' for all Task Pairs*
<i>Algorithm 1: Proportion of repeated markers</i>	
• 0.55 cutoff	0.788
• 0.62 cutoff	0.811
<i>Algorithm 2a: Euclidean distance of the top 5 markers</i>	
• 1.5 cutoff	0.604
• 3.4 cutoff	0.633
<i>Algorithm 2b: Euclidean distance of the top 10 markers</i>	
• 3.5 cutoff	0.574
• 4.4 cutoff	0.552

*Hit rates computed from S12_S12 and S12_S14, and False Alarm rates were from S12_S11 and S12_S13.

Study 1 hypotheses addressed

Hypothesis 1.1: Taskload would inversely relate to performance.

Results from the analyses involving CD performance supported the hypothesis as performance was significantly better during low CD taskload level relative to the high CD taskload level, and CD performance during the Single task condition (Scenario 1) was more superior than during the other Dual task conditions. That no significant differences were found across Scenarios 2, 3, and 4 suggests that the taskload manipulation for the PT task was a weak one. Results from the analyses with PT performance did not show the inverse relationship as PT performance was worse when PT taskload level was low compared to when it was higher. This hypothesis is partially supported.

Hypothesis 1.2: Taskload and workload would be positively associated.

Although not all workload measures were sensitive to the taskload manipulations, results indicate that for the workload measures that were sensitive, the level of workload experienced was positively associated with taskload levels. This hypothesis is supported.

Hypothesis 1.3: The physiological workload model would generalize to the new sample.

Results of the second modeling approach showed that the means and standard deviations of the workload index under algorithm 1 for the different task pairs obtained from the Abich (2013) data were comparable to that from the new sample with the CD and PT tasks (see Table 17). This suggests that algorithm 1 was generalizable to the new sample.

Summary

Main findings from Study 1 include the observation that workload measures differ their sensitivity to taskload, and this depended on how taskload was manipulated (i.e., by CD event

rate or by Single-Dual tasks). The type of taskload manipulation (i.e., Single-Dual tasks vs. event rate) seemed to have moderated sensitivity of workload measures. This concurred with previous findings involving the same CD task (Matthews et al., 2015a). More physiological workload measures were sensitive to changes in CD level taskload, than to changes across Scenarios that differed only on PT event rate. Nevertheless, the pattern of results suggested a positive relationship between taskload and workload as assessed by physiological measures, i.e., the higher the taskload, the higher the workload. By confirming that for the CD task, there were no dissociations that would adversely impact the model, and that the relationships among taskload, workload and performance found in the previous Abich (2013) study were replicated, Study 1 provided support for the use of the workload model with the same CD task in Study 2.

In addition, the workload model featuring a workload index based on physiological measures was developed. The model was cross-validated with a new sample (i.e., Study 1 data). The next study (Study 2) investigated the performance outcomes from applying the workload model to a closed-loop system that adapts robot aid based on workload needs. To recapitulate, the definition of terms used in this chapter are in Table 22:

Table 22
Definition of terms used in the modeling approach

Term	Definition
Single Task Baseline	Single task condition known to elicit low workload
Dual Task Baseline	Dual task condition known to elicit high workload
Task Pair	Pairing of any 2 task conditions for computing a difference score
Benchmark Task Pair	Pairing of the Single Task Baseline and Dual Task Baseline
Test Task Pair	Pairing of the Single Task Baseline and an experimental condition
Rolling Test Task Pair	Pairing of the Single Task Baseline and a sample of data from the 2-minute rolling window in an experimental condition
Benchmark Difference Score	Difference score computed from the Benchmark Task Pair
Test Difference Score	Difference score computed from the Test Task Pair
Rolling Test Difference Score	Difference score computed from the Rolling Test Task Pair

CHAPTER SIX: METHODOLOGY (STUDY 2)

Study Goal

The goal for Study 2 was to ascertain if adaptive robot aiding driven by a (individualized) workload index derived from physiological measures would result in performance improvements through managing workload.

Study Design

Study 2 adopted a one-way repeated measures ANOVA design with four levels:

- (i) a low taskload baseline scenario (i.e., Single Task Baseline condition)
- (ii) a high taskload baseline scenario (i.e., Dual Task Baseline condition)
- (iii) a low taskload experimental scenario with adaptive aiding
- (iv) a high taskload experimental scenario with adaptive aiding

Study Scenarios

Study 2 sought to examine (i) if low and high workload states could be distinguished through a physiologically-based workload index such that adaptive robot aiding can be given during the high workload state, and (ii) the effects of providing adaptive aid based on a workload index derived from physiological measures. This requires the index to show a sufficiently large deviation (above or below) from the low workload state “Single Task Baseline”, signaling a substantive change in workload state. To test this, the four scenarios were defined as illustrated in Figure 16:

Scenario S: Single Task Baseline Condition

Baseline block (BL)
|-----*Single task for 5 min*-----|

Scenario D: Dual Task Baseline Condition

Baseline block (BL)
|-----*Dual task for 5 min*-----|

Scenario SSS: Single Task Experimental Condition

(Blk1) (Blk2) (Blk3)
|-----*Single task for 5 min*-----|-----*Single task for 5 min*-----|-----*Single task for 5 min*-----|
Physio. triggered aid provided if workload index indicates high workload ↑ System or Physio.triggered aid
System will trigger aid if not already triggered

Scenario SDD: Dual Task Experimental Condition

Blk1 Blk2 Blk3
|-----*Single task for 5 min*-----|-----*Dual task for 5 min*-----|-----*Dual task for 5 min*-----|
Physio. triggered aid provided if workload index indicates high workload ↑ System or Physio.triggered aid
System will trigger aid if not already triggered

[Aid = auditory BEEP on every change event.]

Figure 16. Scenarios in Study 2

Substantial individual differences in physiological response to workload required a workload index that would accommodate the individual's *own* set of physiological markers of workload. For this, it was necessary to observe the individual in both low (Single task) and high (Dual

task) task/workload conditions, i.e., Single Task Baseline condition and the Dual Task Baseline condition. This first Single-Dual Baseline Task Pair and the Benchmark Difference Scores that result identified the physiological measures that comprised the workload response difference scores or markers for the individual. Then, together with the Single task baseline condition, a subsequent experimental condition would form the test task pairs. If the difference scores from the baseline and a test task pairs were similar, then the experimental condition would be considered to have elicited the same high level of workload as the Dual Task Baseline condition.

Each of the two baseline conditions (Scenario S and Scenario D) was five minutes in duration to avoid over-exposure to the task(s). Scenario S comprised a Single Change Detection task at the medium event rate (i.e., 12 change events per minute), while Scenario D (Dual task) consisted of the same Change Detection task at the medium event rate and a Threat Detection task also at a medium event rate (i.e., 30 events per minute). No intervention or aid was provided in these baseline conditions. Each of the two experimental conditions (Scenario SSS and Scenario SDD) are each 15 minutes in duration. Scenario SSS required participants to perform a Single task throughout, while Scenario SDD started off as a Single task for the first 5 minutes before taskload was ‘ramped’ up to a Dual task for the next 10 minutes.

Robot aid

For the experimental conditions, adaptive robot aid was provided if participants’ workload index met the cutoff value and was thus classified as experiencing HIGH workload. However, if there was no evocation of aid after 10 minutes, the aid was triggered even if the workload index did not indicate any significant change from the low workload baseline (i.e., level workload experienced was not different from that in the low workload baseline condition). The effect

of this unwanted/unneeded aid on workload was observed in the last 5 minutes of the experimental conditions. Nevertheless, to investigate the effects of an unneeded aid, the aid would have to be one that would alleviate workload if provided appropriately, but would adversely affect workload if not given at the right time. To simulate this cost-benefit equilibrium of aiding, the aid consisted of an auditory BEEP each time there is a change to be detected. From the results of previous studies with the same change detection task with the auditory BEEP aid (i.e., Taylor, 2012; 2013), having the auditory aid when workload is high may reduce workload, but during low workload, the aid can result in an overreliance on the BEEP which may increase detection rate, but may also negatively impact the accurate classification of the change type by the participant.

Table 23
Status of Adaptive Aid under Different Classifications of Workload State

	Low workload state	High workload state
Auditory BEEP aid given	Aid may increase detection rate but decrease classification of change type.	Aid is expected to alleviate workload and improve detection rate.
Auditory BEEP aid not given	Workload is manageable.	Workload is difficult to manage because no aid was given.

Power Analysis

Previous studies (i.e., Abich, 2013) utilizing the same Change Detection task and workload measures indicated that the size of the Single-Dual effect was medium. For a medium effect size (i.e. $f=0.25$ or $d=0.50$; Cohen, 1988), and at α level of 0.05 and a power level of 0.8, the required sample size for the repeated measures ANOVA is 36 as computed by G*Power (Faul, Erdfreld, Buhner & Lang, 2007).

Participants

Sixty-two students (34 males, 27 females) aged between 18 to 52 years ($M=20.31$) from the University of Central Florida psychology research participation pool served as participants. They were awarded course credit for participation. All participants were self-reported to be right-handed, had normal color vision as assessed by the Ishihara color vision test, and had self-reported normal or corrected-to-normal vision. They were required to refrain from consumption of any alcohol, caffeine, sedative medication 24 prior to the study.

Experimental Task/Context

Study 2 utilized the same context (i.e., ISR mission) and same two tasks as the Abich (2013) study from which the parameters for the individualized workload model was derived. The ISR mission was typified by a (i) Change Detection task (CD), and (ii) a Threat Detection task (TD). The low workload baseline task will comprise the single Change Detection task (CD), while the high workload dual task would consist of the CD task paired with a Threat Detection task (TD) task.

As with the Abich (2013) study and Study 1, the Soldier's display comprised 3 windows (see Figure 2). The Soldier would perform the change detection task within the main change detection map window, and view the people in the environment through the video feed showing the robot's "ground level" view.

Change Detection Task.

As in Study 1, the task required participants to detect changes in icons on a map. Icons may appear, disappear or move. Participants were instructed to click the appropriate change detection button as soon as they saw a change to indicate the change type detected. Since workload

was not manipulated through event rate, the change detection task was set at the medium event rate (i.e., 12 events per minute) through the duration of all scenarios. Each change event involved 2 icons at once, and the set size was maintained relatively consistently at 24 icons throughout (after Abich, 2013).

Threat Detection Task.

In keeping with the task in the Abich (2013) study, the threat detection (TD) task required participants to monitor the video feed in the UGV window (Figure 2, outlined in blue, also see Figure 17 below) for threats and to report them by first clicking on a “Threat Detect” button, and then clicking on the threat character in the environment. There were 4 categories of people: (i) friendly Soldiers, (ii) friendly civilians, (iii) enemy soldiers, and (iv) armed civilians/insurgents. Threats comprise enemy soldiers and armed civilians/insurgents (see Figure 18 for examples from each category). The event rate was the presence of a character in the UGV window, while threat probability was the ratio of threats to events. The event rate was set at 30 events per minute (medium rate, as designated in Abich, 2013), while the threat probability was 4 threats per 30 events/characters (medium threat probability, as defined in Abich, 2013).

Figure 17. Video feed showing characters in the “ground level” view

(From left to right: (i) friendly Soldiers, (ii) friendly civilians, (iii) enemy soldiers, (iv) armed civilians/insurgents)
Figure 18. Examples of characters in the TD task.

Each participant was told that s/he was a Soldier on an Intelligence, Surveillance and Reconnaissance (ISR) mission with an Unmanned Ground Vehicle (UGV) robot teammate. The participant would assume the role of “Supervisor” with the UGV robot as the Subordinate”. The participant was informed that the UGV robot may switch to the “Peer mode”, and would be more autonomous and will assist with the CD task. During the training, the participant would be informed of the differences between the Subordinate and Peer modes and will experience an instance of a mode switch. The details of the different robot modes as shown in Table 24:

Table 24
Behaviors associated with the different Robot roles

Robot mode	How Soldier is notified of mode switch	Change detection taskload of Soldier
Subordinate mode	Robot starts off in subordinate mode.	Soldier has to detect changes and classify change type.
Peer mode	Auditory alert: “Robot teammate now in peer mode and will be assisting with Change Detection task”.	Robot will aid with auditory BEEP on every change event, but Soldier will still need to classify change type.

Procedure

After giving their informed consent, participants were administered the Ishihara color-blindness test and completed a restrictions checklist. Those who were screened proceeded to complete the demographics questionnaire. As the participants were fitted with the physiological sensors, they were provided with a brief explanation of each sensor by the researcher.

When the setup of the physiological sensors was completed, a 5-minute physiological resting baseline was taken where participants were told to relax, keep their eyes open, and not focus on anything in particular. After that, participants completed a series of pre-task questionnaires. Then, they were briefed on the mission, trained on the change detection (CD) task, the threat detection task (TD), and were familiarized on how the robot teammate would switch to Peer mode. They underwent three short training missions where they practiced on the CD task alone, the CD and the TD tasks together, and both tasks together with the robot role change notification and the adaptive aiding with the CD task. This was followed by the Single and Dual Task Baseline conditions, the order of which were counterbalanced. In the midst of each condition, the ISA prompt was administered to capture perceived workload during the task. The NASA-TLX and DSSQ, being longer, were administered after each scenario.

After completion of the post-condition questionnaire for the second scenario (i.e., after the two Baseline conditions), the physiological data were processed and the workload model was applied to the individual's data, yielding the physiological markers for the individual. For the subsequent experimental conditions, the workload index (i.e., defined as the proportion of repeated markers) was computed with 2-minute data block every 30 seconds, after the first 2 minutes. Whenever the index was equal or greater than the cutoff set (i.e., 0.62), the workload

state for that data sample would be HIGH. When 3 consecutive data samples were HIGH, i.e., “debouncing factor” of 3, the closed-loop system would classify the participant as being in a stable HIGH workload state, and the aid would be evoked. Participants were not informed of the criterion used to trigger the aid – they were only told that aid may be provided during the experimental condition. The participant would be notified of the change in the robot role to Peer and there would be a BEEP for every change event in the Change Detection task. Once adaptive aid was provided, it was never revoked despite the value of the workload index, and would continue for the remainder of the experimental condition. If the aid was not triggered in the first 10 minutes of the experimental condition, then aid would be “imposed” in the last 5 minutes. In all scenarios, there was an ISA prompt in the middle of each of the three 5-minute block of task. Post-scenario questionnaires (i.e., NASA-TLX and DSSQ) were administered after every scenario. For the last scenario, the Human-Robot trust scale was also included. When the post-scenario questionnaires for the last scenario were filled out, the physiological equipment and sensors were removed and the participant was debriefed, thanked, and dismissed.

Measures

Performance measures.

Performance on the tasks was evaluated with the measures summarized in Table 25:

Table 25

Performance measures and their descriptions

Measure	Description
<i>Change detection performance</i>	
1. Percent correct	Percentage of change events that were correct detected and the change type correctly identified.
2. Percent misclassified	Mean of the following: <ul style="list-style-type: none"> • Percent of Appeared events misclassified as Disappeared

Measure	Description
	<ul style="list-style-type: none"> • Percent of Appeared events misclassified as Disappeared • Percent of Disappeared events misclassified as Appeared • Percent of Disappeared events misclassified as Moved • Percent of Moved events misclassified as Appeared • Percent of Moved events misclassified as Disappeared
3. Average response time	Average response time to correct detection (i.e. change events detected and correctly classified).
4. Number of false alarms	Number of the false alarms committed.
<i>Threat detection performance</i>	
1. Total correct	Number of threats correctly detected
2. Total incorrect	Number of incorrect detections
3. Total duplicate threats	Number of threats repeatedly identified

Subjective measures.

The self-report measures were the same as those used in Study 1.

Physiological measures.

The physiological measures were the same as those used in Study 1.

CHAPTER SEVEN: RESULTS (STUDY 2)

After excluding data from pilot participants, participants who were subjected to a different index threshold, and a participant for whom the closed loop application did not run, the sample comprised 62 participants. Data were checked for outliers and Winsorized by setting outliers to the 5th or 95th percentile values (Tabachnik & Fidell, 2006). This minimized the influence of outliers on the analyses. For all analyses, the degrees of freedom were adjusted for violations of sphericity using Box's epsilon (Maxwell & Delaney, 2004). Inflation of Type I error from multiple pairwise comparisons were minimized with the Bonferroni correction. All physiological measures were computed as percentage change from baseline.

There were four scenarios in Study 2: (i) Scenario S (Single Task Baseline condition), (ii) Scenario D (Dual Task Baseline condition), (iii) Scenario SSS (Single task experimental condition), and (iv) Scenario SDD (Dual task experimental condition) (refer to Figure 16 for Study scenarios). The baseline conditions were 5 minutes each in duration. For the following analyses, each of the two 15-minute experimental conditions were divided into three 5-minute blocks (i.e., Blk1, Blk2, and Blk3). Robot aid was always present in the last block in the experimental conditions (Blk3) either because the closed-loop system had detected high workload based on the physiological measures and triggered the aid earlier which was never revoked (i.e., "Physio. trigger"), or because aid was imposed (i.e., "System trigger") in the last block if no aid had been triggered earlier. While Scenario SSS consisted of only the Change Detection (CD) task throughout the entire condition, Scenario SDD began with only the CD task (Single task) for 5 minutes during the Blk1 block, but taskload was ramped up to a Dual task for the remaining 10 minutes for Blk2 and Blk3. Due to this setup of the experimental conditions, especially in Scenario SDD,

any significant effect of Block would reflect the influence of (i) being on task for a time period, or a Time-on-task effect, (ii) changes in taskload (i.e., from Blk1 to Blk2 in the Scenario SDD), and if the Blk3 block were included, then a significant effect of Block would also reflect (iii) the impact of the robot aid. However, for participants whose aid was “Physio. triggered”, the effects of the aid would be experienced earlier. Analyses of the subsequent pairwise comparisons were particularly critical in understanding the nature of the Block effect.

Time-on-task effects

Time-on-task effects on the measures of CD performance were estimated from previous study data. As with the experimental scenarios in the present data, the scenarios in the Abich (2013) were 15 minutes in duration and the data were split into the 1st, 2nd, and 3rd 5-min blocks. Only the scenario in which the level of the Change Detection task matched that used in the present study was analyzed, i.e., Scenario 4, where CD at a constant medium event rate was paired with TD (Dual task) (see Table 14). Results of the one-way repeated measures ANOVAs indicated that except for Mean Percent misclassified, there was no significant effect of Block (i.e., Time-on-task) on CD performance even after 15 min of the task. The majority of the effect sizes were small. The size of the Time-on-task effect were $\eta_p^2 = 0.014$ for Percent correct, $\eta_p^2 = 0.072$ for Percent misclassified, $\eta_p^2 = 0.015$ for Average response time, and $\eta_p^2 = 0.006$, for Number of false alarms. Overall, this result indicates minimal Time-on-task effect, although effects observed for Percent misclassified should be interpreted with more caution.

Analyses approach

The goals for the analyses were to investigate the effects of (i) taskload, (ii) the robot aid, and (iii) adapting aid based on the individual's workload needs (i.e., Trigger type: Physio. trigger or System trigger), on performance workload and stress states. To determine these effects, selected blocks from the baseline and experimental conditions were compared. However, because participants whose aid was triggered by their workload level (i.e., physio. triggered) would have had the aid for varying durations, there was a possible confound of length of exposure to the aid. Hence, analyses that did not involve any Trigger type effects were conducted only with those who had imposed aid, or the System-trigger group. This ensured that all in these analyses had the same stimuli and experience so that any differences observed could be attributed to the effects of interest. In each of the following analyses sections, a table will summarize (i) the type of analyses performed, (ii) the intent of the analyses, and (iii) the participants involved in the analyses in that section.

Change Detection Performance

Performance on the Change detection task when it was administered without aid, either as a Single task or part of a Dual task, was comparable to that in the Abich (2013) study (see Table 26). This suggested that the Single-Dual manipulation of high and low workload in Study 2 is likely to yield similar effects to that found in the Abich (2013) study, lending further support for the application of the model derived from the Abich (2013) study to Study 2.

Table 26

CD performance in Abich (2013) and Study 2

Performance measures	Abich (2013) study	Study 2
<i>Single task with CD task at medium level (i.e., 12 events per minute)</i>		
Percent correct	$M=61.67, SD=14.23$	$M=65.29, SD=11.34$
Percent appeared detected	$M=63.30, SD=16.26$	$M=67.18, SD=13.84$
Percent disappeared detected	$M=67.91, SD=16.14$	$M=73.06, SD=13.38$
Percent movement detected	$M=51.85, SD=20.26$	$M=53.23, SD=17.70$
<i>Dual task with CD task at medium level</i>		
Percent correct	$M=41.27, SD=14.49$	$M=42.51, SD=10.15$
Percent appeared detected	$M=43.59, SD=17.05$	$M=42.74, SD=12.37$
Percent disappeared detected	$M=44.86, SD=18.07$	$M=52.58, SD=14.79$
Percent movement detected	$M=33.88, SD=19.25$	$M=29.64, SD=14.93$

Effects of Taskload, Block/Order and Aid.

Table 27

Analyses of the Effects of Taskload, Block/Order and Aid (CD Performance)

Analyses	Analyses intent	Participants in analyses
2(Taskload: Single vs. Dual) x 4(Block: Baseline vs. Blk1 vs. Blk2 vs. Blk3) ANOVA	<ul style="list-style-type: none"> • Determine Order effect • Determine effects of Taskload • Determine effect of the Aid : Without aid= Baseline block : With aid= Blk3 block. • Determine effect of Aid by Task- load (i.e., if Aid effect is different for Single vs. Dual task). 	Only participants with system-triggered aid. All had the same stimuli throughout the study (N=43).

A series of 2(Taskload: Single vs. Dual task) x 4(Block: Baseline condition vs. Blk1 vs. Blk2 vs. Blk3) repeated measures ANOVA were computed to determine (i) the effects of taskload and block on the change detection performance measures, (ii) order effects, and (iii) the effects of the aid (see Table 25). Aid was provided in the Blk3 block. To minimize confounding results with the various aid trigger times in the physio-trigger group, the analyses was restricted only to participants who had the system-triggered aid (N=43) and who experienced the same pattern of stimuli throughout the experiment.

Analyses indicated significant main effects of Taskload for Percent correct, $F(1,42)=228.880$, $p<0.001$, $\eta_p^2=0.845$, for Percent Misclassified, $F(1,42)=11.307$, $p=0.002$, $\eta_p^2=0.212$, for Average response time, $F(1,42)=21.797$, $p<0.001$, $\eta_p^2=0.342$, but not for Number of false alarms, $F(1,42)=2.031$, $p=0.161$, $\eta_p^2=0.046$. (see Table 28). There were higher detection rates, lower percent misclassified, and faster response times during the Single task conditions. This suggested the notion that the Single-Dual task paradigm is an effective manipulation of taskload.

Table 28
CD performance measures by Taskload (N=43)

Measure	Taskload	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Percent correct	Single	65.069	1.690	61.657	68.480
	Dual	51.055	1.605	47.817	54.293
Percent misclassified	Single	6.836	.392	6.045	7.628
	Dual	8.152	.467	7.210	9.093
Average response time	Single	1.481	.027	1.427	1.535
	Dual	1.551	.024	1.503	1.599
Number of false alarms	Single	.721	.122	.475	.967
	Dual	.593	.094	.404	.782

The main effect of Block was significant for Percent correct, $F(1.701,106.860)=28.878$, $p<0.001$, $\eta_p^2=0.407$, for Percent misclassified, $F(1.701,71.439)=26.277$, $p<0.001$, $\eta_p^2=0.385$, for Average response time, $F(3,126)=8.091$, $p<0.001$, $\eta_p^2=0.162$, but not for Number of false alarms, $F(3,126)=2.576$, $p=0.061$, $\eta_p^2=0.058$ (see Table 29). Across both taskload levels, there were no significant differences between the Baseline and Blk2 block in the experimental conditions on all performance measures. Compared to the other blocks, Blk1 across Scenarios SSS and SDD had the highest Percent correct, and this was likely to be because there was a Single task taskload for that block in Scenario SDD. This was followed by Percent correct in Blk3, the block when aid

was provided. The Baseline and Blk2 were the blocks with the lowest Percent correct. Of all blocks, Blk3 had the highest Percent misclassified, a likely result of aid being provided in that block. Across Scenarios SSS and SDD, average response time were quickest during Blk1 and Blk3. These findings were likely due to the Single task taskload in Blk1 in Scenario SDD, and the aid provided in Blk3 in both experimental conditions.

Table 29
CD performance measures by Block (N=43)

Measure	Block	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Percent correct	Baseline block	54.340	1.543	51.225	57.454
	Blk1 block	64.274	1.988	60.262	68.285
	Blk2 block	53.547	1.916	49.679	57.414
	Blk3 block	60.087	1.632	56.793	63.382
Percent mis-classified	Baseline block	7.057	.381	6.287	7.826
	Blk1 block	5.921	.397	5.121	6.721
	Blk2 block	6.120	.363	5.387	6.853
	Blk3 block	10.878	.874	9.114	12.643
Average re-sponse time	Baseline block	1.548	.027	1.493	1.603
	Blk1 block	1.484	.029	1.426	1.542
	Blk2 block	1.548	.028	1.491	1.605
	Blk3 block	1.485	.022	1.440	1.529
Number of false alarms	Baseline block	.674	.123	.426	.923
	Blk1 block	.814	.141	.529	1.099
	Blk2 block	.686	.136	.411	.961
	Blk3 block	.453	.108	.236	.671

There was a significant Taskload x Block interaction effect for Percent correct, $F(3,126)=35.805$, $p<0.001$, $\eta_p^2=0.460$, for Percent misclassified, $F(3,126)=11.334$, $p<0.001$, $\eta_p^2=0.213$, for Average response time, $F(3,126)=5.064$, $p=0.002$, $\eta_p^2=0.108$, but not for Number of false alarms, $F(3,126)=1.447$, $p=0.235$, $\eta_p^2=0.033$ (see Table 30).

Table 30

CD performance measures by Taskload by Block (N=43)

Measure	Taskload	Block	Mean	Std. Error	95% Confidence Interval	
					Lower Bound	Upper Bound
Percent correct	Single	Baseline block	64.618	1.798	60.990	68.246
		Blk1 block	65.221	2.100	60.984	69.458
		Blk2 block	62.643	2.037	58.531	66.754
		Blk3 block	67.794	1.882	63.996	71.591
	Dual	Baseline block	44.061	1.643	40.745	47.378
		Blk1 block	63.326	2.105	59.078	67.574
		Blk2 block	44.451	2.125	40.164	48.739
		Blk3 block	52.381	1.732	48.886	55.877
Percent misclassified	Single	Baseline block	6.172	.458	5.247	7.098
		Blk1 block	6.368	.487	5.384	7.352
		Blk2 block	5.700	.418	4.856	6.544
		Blk3 block	9.105	.836	7.417	10.793
	Dual	Baseline block	7.941	.572	6.787	9.095
		Blk1 block	5.474	.465	4.535	6.413
		Blk2 block	6.541	.472	5.588	7.493
		Blk3 block	12.651	1.034	10.565	14.738
Average response time	Single	Baseline block	1.514	.030	1.453	1.575
		Blk1 block	1.485	.033	1.418	1.552
		Blk2 block	1.493	.030	1.433	1.554
		Blk3 block	1.432	.025	1.382	1.482
	Dual	Baseline block	1.582	.030	1.522	1.643
		Blk1 block	1.483	.029	1.424	1.542
		Blk2 block	1.603	.033	1.536	1.669
		Blk3 block	1.537	.024	1.489	1.586
Number false alarms	Single	Baseline block	.884	.189	.503	1.265
		Blk1 block	.907	.144	.616	1.198
		Blk2 block	.605	.160	.282	.928
		Blk3 block	.488	.177	.131	.846
	Dual	Baseline block	.465	.117	.229	.701
		Blk1 block	.721	.195	.327	1.114
		Blk2 block	.767	.166	.433	1.102
		Blk3 block	.419	.121	.174	.663

Subsequent analyses of the simple effects by taskload revealed that for Single task, there was a significant main effect of Block for Percent correct, $F(2.566, 107.751) = 3.470$, $p = 0.024$,

$\eta_p^2=0.076$, Percent misclassified, $F(2.104,88.358)=10.036$, $p<0.001$, $\eta_p^2=0.193$, Average response time, $F(3,126)=5.789$, $p=0.001$, $\eta_p^2=0.121$, but not Number of false alarms, $F(3,126)=2.373$, $p=0.075$, $\eta_p^2=0.053$. The differences in performance were largely driven by the aid provided in Blk3. The Blk3 block had significantly higher Percent correct than the Blk2 block, significantly higher Percent misclassified than all other blocks, and significantly faster Average response times compared to the Baseline and Blk2 blocks.

Similarly, for the Dual task, there was a significant main effect of Block for Percent correct, $F(2.932,123.137)=56.344$, $p<0.001$, $\eta_p^2=0.573$, Percent misclassified, $F(3,126)=31.149$, $p<0.001$, $\eta_p^2=0.426$, Average response time, $F(3,126)=7.276$, $p<0.001$, $\eta_p^2=0.148$, but not for Number of false alarms, $F(1.956,86.380)=1.594$, $p=0.208$, $\eta_p^2=0.037$. Percent correct was highest in Blk1, followed by Blk3, and both of these were significantly higher than that for Baseline and Blk2. Percent misclassified was greatest in Blk3 when aid was provided and lowest in Blk1 with the Single task. Average response time was fastest during the Single task in Blk1, followed by Blk3. These effect were driven largely by the Single task in Blk1 and the provision of aid in Blk3. The finding that performance was superior in the Single task blocks compared to the Dual task blocks, indicated that taskload and performance were negatively related (see Figure 19).

Figure 19. Performance measures for the Change Detection task (error bars denote SEs)

Order effects.

The results of the same omnibus ANOVA also showed that for all performance measures, as long as the taskload of the blocks in the comparison were of the same level (i.e., all single task blocks or all dual task blocks), performance in the baseline conditions were not significantly different from that in the experimental conditions, i.e., performance in Scenario S (Single Task Baseline) was not significantly different from the Blk2 block in Scenario SSS, and performance in Scenario D (Dual Task Baseline) did not significantly differ from that in Blk2 of Scenario

SDD ($p>0.05$). That this was observed even though the baseline conditions were always administered before the experimental conditions implies that order effects were minimal and not significant.

Aid and Taskload.

The same omnibus ANOVA also showed a robust effect of the aid. Examining the pairwise comparisons between the Baseline block (i.e., Without aid) and Blk3 (i.e., With aid), results indicated that Percent correct and Percent misclassified increased in Blk3, and Average response time was faster in Blk3 when aid was provided. Although the blocks Without Aid (i.e., Baseline) were always administered before the block With Aid (i.e., Blk3), earlier analyses revealed minimal order effects, suggesting that the differences are likely to be due to the presence of the aid. This aid effect was observed in both the Single and Dual task conditions.

When Aid was not present, Percent correct for Dual task was much lower than that for Single task (Single task: $M=64.618$, $SE=1.798$, Dual task: $M=44.061$, $SE=1.643$; $d=1.820$), but when aid was provided, the disparity in Percent correct between Dual task and Single task was not as stark (Single task: $M=67.794$, $SE=1.882$, Dual task: $M=52.381$, $SE=1.732$; $d=1.300$). For Single task, Percent correct rose by 3.176% on average with the provision of aid ($SD=12.899$), whereas for the Dual task, Percent correct increased by 8.320% on average for Dual task ($SD=12.155$) when aid was provided (see Figure 20). Such a Taskload by Aid interaction effect was not observed for the other performance measures.

Figure 20. Percent correct detections by Taskload and Aid (error bars denote SEs)

Effects of Aid and Trigger Type.

Table 31

Analyses of the Effects of Aid and Trigger Type (CD Performance)

Analyses	Analyses intent	Participants in analyses
2(Aid: Without aid vs. With aid) x 2(Trigger type: System trigger vs. Physio trigger) ANOVA	<ul style="list-style-type: none"> • Determine if the effect of Aid : Without aid = Baseline block : With aid= First 5min after aid onset (for System trigger group, this is also Blk3). • Determine effect of Trigger type • Determine effect of Aid by Trigger type, (i.e., if Aid effect is different for those with Physio. trigger vs. System trigger). 	Groupings into System and Physio trigger were different for scenarios SSS and SDD. Separate analyses were conducted for Single task (N=10 & 52) and Dual task (N=15 & 47).

Although the findings showed a strong effect of the aid on CD performance, the main focus of the study concerns the effect of having a closed-loop system, based on physiological workload measures, driving the adaptive robot aid. In both experimental conditions, the aid was

either triggered by the physiological measures earlier in the first two blocks and not revoked, or imposed and triggered by the system in Blk3. Analyses by these trigger types was critical in addressing the research question because aid rendered by the physiological trigger reflected aid that was provided during the individual's experience of HIGH workload (not just high taskload), while aid imposed by the system trigger denoted aid that was provided without consideration of workload state.

For each of the two experimental conditions, participants were divided into two groups: (i) those who had the physiological trigger, and (ii) those who had the system trigger. As the groupings were different for scenarios SSS and SDD (see Table 32), separate analyses were conducted for the Single task and Dual task experimental conditions.

Table 32

No. of participants with Physio-triggered aid or System-triggered aid by Condition

	Scenario SSS (Single Task Exptal Condition)	Scenario SDD (Dual Task Exptal Condition)
Participants with Physiological trigger*	10	15
Participants with System trigger	52	47
Total	62	62

*There were 6 participants for whom the aid was Physio-triggered in both conditions

First, equivalence of the two groups on CD performance during Single and Dual task was determined by one-way between-subjects ANOVAs. For Single task, no significant differences were observed in CD performance during the baseline block between those who received Physio. trigger and those with the System trigger, $p > 0.242$ for all measures. Similarly, there were no significant differences in CD performance between those who had the Physio trigger and those who had the System trigger in the Dual task baseline block, $p > 0.069$ for all measures. Hence, the

Physio trigger and System trigger groups showed comparable levels of CD performance and were equivalent at the start of the study.

To evaluate the effects of Aid and Trigger type, a 2(Aid: Without vs. With Aid) x 2(Trigger type: System trigger vs. Physio. trigger) mixed ANOVA with the second factor as a between-subjects factor was conducted for the Single task and Dual task blocks in turn. Earlier analyses yielding a non-significant order effect permitted the assignment of the Baseline condition as the “Without Aid” block (earlier analyses had indicated minimal order effects), and the “With Aid” block comprised the first 5 minutes after the onset of the aid. For those with the System trigger, this coincided with Blk3, whereas for those with the Physio. trigger, this was determined individually. Doing so ensured that the length of time participants were exposed to the aid was held constant for all participants (see Table 33).

Table 33
Description of levels for the factors Aid and Aid trigger type

		Trigger type	
		Physio. trigger	System trigger
Aid	Without Aid	Baseline condition of those whose physio. had triggered the aid	Baseline condition of those who had system-triggered aid.
	With Aid	First 5 min from start of Aid (variable by participant)	First 5 min from start of Aid, i.e., the Blk3 of experimental condition

Single task.

The ANOVA revealed significant effects of Aid for Percent misclassified, $F(1,60)=7.504$, $p=0.008$, $\eta_p^2=0.111$, and Average response time, $F(1,60)=6.125$, $p=0.016$, $\eta_p^2=0.093$, but not for Percent correct, $F(1,60)=2.850$, $p=0.097$, $\eta_p^2=0.045$, nor for Number of false alarms, $F(1,60)=0.343$, $p=0.560$, $\eta_p^2=0.006$. For both groups of participants, Percent misclassified was higher and Average response times were faster “With Aid”(see Table 34).

Table 34

CD performance by Aid (Single task) (N=62)

Measure	Aid	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Percent correct	Without aid	64.670	1.971	60.728	68.613
	With aid	68.136	1.940	64.255	72.016
Percent mis-classified	Without aid	6.640	.534	5.573	7.707
	With aid	9.218	.865	7.488	10.947
Average response time	Without aid	1.492	.034	1.423	1.560
	With aid	1.432	.029	1.373	1.490
Number of false alarms	Without aid	.921	.233	.454	1.388
	With aid	1.069	.208	.653	1.486

There were no significant effects of Trigger type for any of the performance measures, indicating that the groups were generally equivalent on CD performance in the Single task blocks ($p > 0.166$ in each case).

There was a significant Aid by Trigger type interaction effect for Number of false alarms, $F(1,60)=4.767$, $p=0.033$, $\eta_p^2=0.074$. Analyses of simple effects by Aid showed that the Number of false alarms for the groups were comparable when there was no aid (System-trigger group: $M=0.942$, $SE=0.188$, Physio-trigger group: $M=0.900$, $SE=0.428$; $d=0.031$), but was significantly higher for the group with Physio trigger after the provision of aid (System-trigger group: $M=0.538$, $SE=0.167$, Physio-trigger group: $M=1.600$, $SE=0.381$; $d=-0.882$). For the System-trigger group, the Number of false alarms declined with the provision of aid by an average of 0.404 ($SD=1.459$), but for the Physio-trigger group, the Number of false alarms increased instead by an average of 0.700 ($SD=1.494$) when aid was provided (see Figure 21).

Figure 21. Number of false alarms by Aid and Trigger Type in Single task blocks (error bars denote SEs)

Dual task.

The ANOVA revealed that that Aid had a significant effect on Percent correct, $F(1,60)=58.425$, $p<0.001$, $\eta_p^2=0.493$, Percent misclassified, $F(1,60)=26.568$, $p<0.001$, $\eta_p^2=0.307$, Average response time, $F(1,60)=14.070$, $p<0.001$, $\eta_p^2=0.190$, but not on Number of false alarms, $F(1,60)=0.121$, $p=0.729$, $\eta_p^2=0.002$. For both groups, Percent correct and Percent misclassified were higher, and Average response time was faster with Aid (see Table 35).

Table 35
CD performance by Aid (Dual task) (N=62)

Measure	Aid	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Percent correct	Without aid	41.251	1.485	38.281	44.221
	With aid	54.042	1.684	50.674	57.410
Percent misclassified	Without aid	7.604	.514	6.575	8.633
	With aid	12.252	.896	10.459	14.045
Average response time	Without aid	1.627	.030	1.568	1.687
	With aid	1.526	.025	1.476	1.576
Number of false alarms	Without aid	.567	.120	.328	.807
	With aid	.623	.133	.358	.889

As with the Single task blocks, there was no significant main effect of Trigger type, indicating that the two groups were generally equivalent on CD performance in the Dual task blocks as well ($p > 0.162$ in each case).

However, there was a significant Aid by Trigger type interaction effect for Percent correct, $F(1,60)=6.698$, $p=0.012$, $\eta_p^2=0.100$, and for Average response time, $F(1,60)=4.476$, $p=0.039$, $\eta_p^2=0.069$. When Aid was not present, the Physio trigger group had lower Percent correct (System-trigger group: $M=43.693$, $SE=1.461$, Physio-trigger group: $M=38.810$, $SE=2.585$; $d=0.488$), and slower Average response times (System-trigger group: $M=1.572$, $SE=0.029$, Physio-trigger group: $M=1.682$, $SE=0.052$; $d= -0.550$) compared to the System trigger group. When Aid was provided, the detection rate for the Physio trigger group surpassed that of the System trigger group (System-trigger group: $M=52.153$, $SE=1.656$, Physio-trigger group: $M=55.931$, $SE=2.932$; $d= -0.333$), and their Average response times (System-trigger group: $M=1.528$, $SE=0.025$, Physio-trigger group: $M=1.524$, $SE=0.044$; $d=0.023$) were comparable to those of the System trigger group. The two groups differed on their change in performance without and with aid. For the System-trigger group, Percent correct increased by 8.46% on average ($SD=11.720$), but for the Physio-trigger group, Percent correct improved by 17.12% on average ($SD=9.721$) with the provision of aid. The Average response time for the System-trigger group decreased by 0.0442 ms on average ($SD=0.175$), while the Average response time for the Physio-trigger group was reduced by 0.159 ms on average ($SD=0.203$) when aid was provided (see Figure 22 and Figure 23).

Figure 22. Percent correct by Aid and Trigger type in Dual task blocks (error bars denote SEs)

Figure 23. Average response time by Aid and Trigger type in Dual task blocks (error bars denote SEs)

All these findings demonstrate that while the aid improved performance, this effect of the aid was stronger during higher taskload (i.e., Dual task), and even stronger when it is provided during a state of HIGH workload as assessed by physiological measures. Nevertheless, the aid also resulted in greater number of misclassifications overall, and false alarms under certain conditions (i.e., Single task).

Threat Detection Performance

For Threat Detection (TD) performance, only data from the Dual taskload conditions (baseline and experimental) were analyzed because the TD task was only administered in Dual task conditions (i.e.. Scenarios D and SDD).

Effect of Block and Aid.

Table 36
Analyses of the Effects of Block and Aid (TD Performance)

Analyses	Analyses intent	Participants in analyses
One-way (Baseline vs. Blk2 vs. Blk3) ANOVA	<ul style="list-style-type: none"> • Determine Order effect • Determine effects of Aid. <ul style="list-style-type: none"> : Without aid block= Baseline block • : With aid block= Blk3 block. 	Only participants with system-triggered aid. All had the same stimuli throughout the study (N=43).

A one-way repeated measures ANOVA was conducted involving the Dual task Baseline, the Blk2, and Blk3 blocks of scenario SDD. Analysis was restricted to participants who had the System trigger to so that the length of exposure to the aid (for the CD task) was held constant across participants. Despite the presence of the aid in Blk3, there were no statistically significant effects of Block on all measures of TD performance, $p > 0.333$ in each case. TD performance did not differ between the Baseline and Blk2 block, $p > 0.99$ in each case, indicating no order effects

for TD performance, but also no effect of the CD aid on TD performance as the difference between the Baseline and Blk3 block (when aid was provided), was also not significant ($p>0.687$ in each case, see Table 37).

Table 37
TD performance across Blocks (N=43)

Measure	Block	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Total correct	Baseline (Scenario D)	15.837	.406	15.018	16.657
	Blk2 (Scenario SDD)	15.791	.419	14.945	16.636
	Blk3 (Scenario SDD)	15.395	.391	14.606	16.185
Total incorrect	Baseline (Scenario D)	.256	.100	.053	.458
	Blk2 (Scenario SDD)	.326	.119	.086	.565
	Blk3 (Scenario SDD)	.186	.083	.018	.354
Total duplicate threats	Baseline (Scenario D)	1.140	.259	.617	1.662
	Blk2 (Scenario SDD)	1.326	.276	.769	1.882
	Blk3 (Scenario SDD)	1.326	.319	.682	1.969

Effect of Aid and Trigger type (Dual task).

Table 38
Analyses of Aid and Trigger Type (TD Performance)

Analyses	Analyses intent	Participants in analyses
2(Aid: Without aid vs. With aid) x 2(Trigger type: System trigger vs. Physio trigger) repeated measures ANOVA	<ul style="list-style-type: none"> Determine if the effect of Aid : Without aid = Baseline block : With aid= First 5min after aid onset (for System trigger group, this is also Blk3). Determine effect of Trigger type Determine effect of Aid by Trigger type (i.e., if Aid effect is different for those with Physio. trigger vs. System trigger). 	Analyses were conducted by Physio-trigger and System-trigger grouping in scenario SDD (N=15 & 47).

Comparison of TD performance between the System trigger and Physio trigger groups with a one-way between subjects ANOVA on the Baseline condition revealed no significant differences between the groups, $p>0.240$ for all measures, indicating that the groups were equivalent on TD performance.

Designating the Baseline as the “Without aid” block, and the first 5-minutes after aid onset as the “With aid” block, a 2(Aid: Without aid vs. With aid) x 2(Triple type: System trigger vs. Physio trigger) mixed ANOVA with the second factor as a between-subjects factor was conducted. Results indicated a significant Aid effect for Total correct, $F(1,60)=7.041$, $p=0.010$, $\eta_p^2=0.105$, and Total duplicates, $F(1,60)=5.738$, $p=0.020$, $\eta_p^2=0.087$. When those with Physio-triggered aid were included in the analysis, the Total correct decreased and Total duplicates (i.e., number of threat characters who were repeatedly identified as threats) increased when aid was provided. This suggests that perhaps the aid had diverted resources away from the TD task and that this effect was more prominent when the Physio-trigger group, who had a longer exposure to the aid, was included. No other effects were significant ($p>0.067$ in each case).

Table 39
TD performance by Aid (N=62)

Measure	Aid	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Total correct	Without aid	16.003	.351	15.300	16.706
	With aid	14.913	.402	14.109	15.717
Total incorrect	Without aid	.350	.098	.155	.546
	With aid	.152	.077	-.003	.307
Total duplicate threats	Without aid	1.473	.336	.800	2.146
	With aid	2.148	.427	1.293	3.002

Perceived/Self-reported Workload and Stress States

ISA.

Effects of Taskload, Block and Aid.

Table 40

Analyses of the Effects of Taskload, Block and Aid (ISA)

Analyses	Analyses intent	Participants in analyses
2(Taskload: Single vs. Dual) x 4(Block: Baseline vs. Blk1 vs. Blk2 vs. Blk3) ANOVA	<ul style="list-style-type: none"> • Determine Order effect • Determine effects of Taskload • Determine effect of the Aid <ul style="list-style-type: none"> : Without aid= Baseline block : With aid= Blk3 block. • Determine if effect of Taskload by Aid (i.e., if Aid effect is different for Single vs. Dual task). 	Only participants who had system-triggered aid. All had the same stimuli throughout the study (N=43).

The Instantaneous Self-Assessment (ISA) measure of workload was administered at the same time point (i.e., 2.5 minutes into the block) within each block in all scenarios. Analyses were only performed on participants the System trigger so that the effects of taskload and block could be evaluated without being confounded by the potential presence of the aid in the Blk1 or Blk2 blocks. The 2(Taskload) x 4(Block) repeated measures ANOVA conducted revealed a statistically significant Taskload effect, $F(1,42)=51.020$, $p<0.001$, $\eta_p^2=0.548$, Block effect, $F(2.406,101.045)=38.780$, $p<0.001$, $\eta_p^2=0.480$, and Taskload X Block effect, $F(3,126)=16.457$, $p<0.001$, $\eta_p^2=0.282$. Subsequent analyses revealed that workload ratings were higher for Dual task blocks ($M=3.238$, $SE=0.094$) compared to Single task blocks ($M=2.657$, $SE=0.106$; $d=0.884$), indicating that perceived workload generally corresponded to taskload. The block main effect and interaction effect was likely driven by the Single block in Blk1 in scenario SDD.

Aid and Taskload.

Following the significant main effect of Block from the omnibus ANOVA, pairwise comparisons were examined (see Table 41). The effects of the aid was observed in specific pairwise comparisons of the omnibus ANOVA. For the Single task blocks, comparing the ISA ratings in the Single Baseline block, Blk1, Blk2, Blk3 blocks in scenario SSS revealed that the ratings in the Blk3 block (when aid was provided) was significantly higher than that in the other blocks, which were not different from each other. This indicates that in the Single task blocks, provision of aid increased workload.

For the Dual task blocks, ISA workload ratings were the lowest during the Blk1 block with only a Single Task, and highest in the Blk3 block when aid was provided. As with the Single task blocks, the provision of aid increased workload (see Table 42 and Figure 24).

Table 41
ISA workload ratings by Block (N=40)

Block	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Baseline	2.779	.102	2.574	2.984
Blk1	2.547	.105	2.335	2.758
Blk2	3.035	.103	2.827	3.243
Blk3	3.430	.112	3.204	3.657

Table 42
ISA workload ratings by Taskload and Block (N=40)

Taskload	Block	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Single	Baseline block	2.419	.126	2.165	2.672
	Blk1 block	2.558	.107	2.343	2.774
	Blk2 block	2.651	.128	2.392	2.910
	Blk3 block	3.000	.149	2.700	3.300
Dual	Baseline block	3.140	.103	2.932	3.347
	Blk1 block	2.535	.117	2.299	2.771
	Blk2 block	3.419	.116	3.184	3.653
	Blk3 block	3.860	.118	3.622	4.099

Figure 24. ISA ratings by Taskload and Block
(error bars denote SEs)

Effect of Aid and Trigger Type

Table 43
Analyses of the Effects of Aid and Trigger Type (ISA)

Analyses	Analyses intent	Participants in analyses
2(Aid: Without aid vs. With aid) x 2(Trigger type: System trigger vs. Physio trigger) ANOVA	<ul style="list-style-type: none"> • Determine if the effect of Aid : Without aid = Baseline block : With aid= First 5min after aid on-set (for System trigger group, this is also Blk3). • Determine effect of Trigger type • Determine effect of Aid by Trigger type, (i.e., if Aid effect is different for those with Physio. trigger vs. System trigger). 	Groupings into System and Physio trigger were different for scenarios SSS and SDD. Separate analyses were conducted for Single task (N=10 & 52) and Dual task (N=15 & 47).

To evaluate whether there were differences in self-reported workload between those with System trigger and Physio trigger, separate 2 (Aid: Without Aid vs. With Aid) x 2(Trigger type: System vs. Physio trigger) mixed ANOVAs with the second factor as a between-subjects factor

were conducted for Single and Dual tasks. However, because the ISA prompt was always administered mid-way through each 5-minute block, some in the Physio-trigger group may have had the aid when the prompt was administered in Blk1 and Blk2 blocks. Thus, the duration of exposure to the aid could not be held constant for the two groups when evaluating the effect of Aid.

Single Task.

The Trigger type group effect was not statistically significant ($p>0.05$), indicating that the Physio-trigger and System-trigger groups were equivalent. The ANOVA revealed only a significant effect of Aid, $F(1,60)=9.300$, $p=0.003$, $\eta_p^2=0.134$. Workload ratings were higher when aid was provided compared to when there was no aid (Without aid: $M=2.392$, $SE=0.131$; With aid: $M=2.940$, $SE=0.161$; $d= -0.474$). No other effects were significant, $p>0.790$ in each case.

Dual Task.

The non-significant Trigger type group effect ($p>0.05$) indicated that the Physio-trigger and System-trigger groups were equivalent. Results of the ANOVA showed a significant Aid effect, $F(1,60)=44.249$, $p<0.001$, $\eta_p^2=0.424$. Workload was significantly higher when aid was provided compared to when there was no aid (Without aid: $M=3.152$, $SE=0.105$; With aid: $M=3.813$, $SE=0.117$; $d= -0.755$). No other effects were significant, $p>0.537$ in each case.

TLX

Effects of Scenario

Table 44
Analyses of the Effect of Scenario (TLX)

Analyses	Analyses intent	Participants in analyses
One-way repeated measures (4 scenarios) MANOVA	<ul style="list-style-type: none"> Determine effects of Scenario. 	Only participants who had system-triggered aid. All had the same stimuli throughout the study (N=43).

The TLX was administered after each of the four scenarios. A one-way repeated measures MANOVA was performed for the perceived workload ratings on the TLX subscales as these were moderately correlated. Analyses were conducted only for data from participants who experienced the System trigger so that the effects of Taskload and Duration could be evaluated without being confounded by the potential presence of aid at different times in the Blk1 or Blk2 blocks. Wilk's lambda revealed a significant effect of Scenario on perceived workload, $\Lambda=0.205$, $F(18,254)=5.383$, $p<0.001$, $\eta_p^2=0.795$. Univariate analyses indicated that ratings on all subscales were significantly higher in Dual task conditions (i.e., Scenarios D and SDD) compared to the Single task conditions (i.e., Scenarios S and SSS), $F(2.486,104.422)=22.439$, $p<0.001$, $\eta_p^2=0.348$ for Mental Demand, $F(2.522,105.925)=12.840$, $p<0.001$, $\eta_p^2=0.234$ for Physical Demand, $F(2.686,112.831)=32.009$, $p<0.001$, $\eta_p^2=0.433$ for Temporal Demand, $F(2.412,101.287)=15.060$, $p<0.001$, $\eta_p^2=0.264$ for Effort, $F(2.138,89.814)=15.656$, $p<0.001$, $\eta_p^2=0.272$ for Frustration, and $F(3,126)=6.418$, $p<0.001$, $\eta_p^2=0.133$ for Performance workload (see Table 45).

Table 45
TLX subscales by Scenario (System trigger participants only, N=43)

Measure	Scenario	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Mental Demand	S ^{ac}	53.256	4.294	44.591	61.921
	D ^a	74.535	3.454	67.565	81.505
	SSS ^f	63.488	3.957	55.503	71.474
	SDD ^{cf}	79.302	2.975	73.300	85.305
Physical Demand	S ^{bc}	16.860	2.901	11.006	22.715
	D ^e	23.488	3.468	16.489	30.488
	SSS ^{bf}	28.140	3.765	20.541	35.738
	SDD ^{cef}	38.721	4.602	29.433	48.009
Temporal Demand	S ^{abc}	42.674	4.026	34.550	50.798
	D ^{ae}	62.326	3.275	55.717	68.934
	SSS ^{bf}	56.512	4.595	47.239	65.784
	SDD ^{cef}	75.581	2.497	70.542	80.621

Measure	Scenario	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Effort	S ^{ac}	50.465	3.560	43.280	57.650
	D ^{ae}	66.860	3.231	60.340	73.380
	SSS ^f	62.093	4.090	53.839	70.347
	SDD ^{cef}	75.349	2.726	69.848	80.850
Frustration	S ^{bc}	31.628	3.637	24.289	38.967
	D ^e	39.419	4.314	30.713	48.125
	SSS ^{bf}	45.233	4.700	35.747	54.718
	SDD ^{cef}	55.698	4.628	46.358	65.037
Performance	S ^{abc}	38.605	3.893	30.749	46.460
	D ^a	52.326	3.605	45.051	59.600
	SSS ^b	51.860	4.387	43.008	60.713
	SDD ^c	55.349	3.609	48.066	62.632

a,b,c,d,e,f: Pairwise differences are significant

No analyses were conducted on the Global workload measure of the TLX because global workload was already reported in the results of the MANOVA of the TLX subscales.

Effect of Trigger Type

Table 46
Analyses of the Effect of Trigger Type (TLX)

Analyses	Analyses intent	Participants in analyses
One-way between-subjects (System trigger vs. Physio trigger) MANOVA	<ul style="list-style-type: none"> Determine if effects across the Scenarios differed by groups. 	Groupings into System and Physio trigger were different for scenarios SSS and SDD. Separate analyses were conducted for Single task (N=10 & 52) and Dual task (N=15 & 47).

Scenario SSS

The one-way between-subjects MANOVA conducted on the System-trigger and Physio-trigger groups as grouped in scenario SSS yielded no significant differences in perceived workload ratings by the two groups across all four scenarios, $p > 0.05$ in each case.

Scenario SDD

Results of the one-way between subjects MANOVA showed no significant differences in perceived workload ratings between the System-trigger and Physio-trigger groups as grouped in scenario SDD for all four scenarios, $p>0.05$ in each case.

DSSQ

Effects of Scenario

Table 47

Analyses of the Effect of Scenario (DSSQ)

Analyses	Analyses intent	Participants in analyses
One-way repeated measures (4 scenarios) MANOVA	• Determine effects of Scenario.	Only participants who had system-triggered aid. All had the same stimuli throughout the study (N=43).

Apart from the Pre-task DSSQ, the Post-task DSSQ, like the TLX, was administered at the end of each of the four study scenarios. A one-way repeated measures MANOVA was performed as the DSSQ subscales were moderately correlated. Analyses were conducted only for participants who experienced the System trigger so that the effects of Taskload and Duration could be evaluated without being confounded by the presence of aid in the Blk1 or Blk2 blocks. Wilk's lambda revealed a significant effect of Scenario on stress states, $\Lambda = 0.280$, $F(9,34)=9.702$, $p<0.001$, $\eta_p^2=0.720$. The univariate analyses of the Scenario effect revealed that, except for scenario D, the other scenarios showed a decline in Task Engagement from Pre-Task levels, and the decline were greatest for scenario SSS. Distress levels increased from Pre-Task levels for all scenarios, and the rise was greatest in scenario SDD, $F(2.436,102.301)=21.675$,

$p < 0.001$, $\eta_p^2 = 0.340$ for Task Engagement, $F(2.608, 109.537) = 17.550$, $p < 0.001$, $\eta_p^2 = 0.295$ for Distress, $F(3, 126) = 1.152$, $p = 0.331$, $\eta_p^2 = 0.027$ for Worry (see Table 48).

Table 48
DSSQ subscales by Scenario (System-trigger participants only, N=43)

Measure	Scenario*	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Task En- gagement	S ^{ab}	-.860	.590	-2.052	.331
	D ^{ade}	1.116	.628	-.151	2.384
	SSS ^{bdf}	-4.372	.946	-6.281	-2.463
	SDD ^{ef}	-1.512	.802	-3.131	.108
Distress	S ^{ac}	3.488	.953	1.566	5.411
	D ^{ae}	5.419	1.026	3.348	7.489
	SSS ^f	5.651	.877	3.881	7.421
	SDD ^{cef}	9.442	1.067	7.288	11.595
Worry	S	-5.674	.734	-7.156	-4.193
	D	-6.372	.942	-8.274	-4.470
	SSS	-5.349	1.044	-7.456	-3.241
	SDD	-6.023	.869	-7.776	-4.270

*All scenario scores were computed as difference from Pre-Task level
a,b,c,d,e,f: Pairwise differences are significant

Effects of Trigger Type

Table 49
Analyses of the Effect of Trigger Type (DSSQ)

Analyses	Analyses intent	Participants in analyses
One-way between-subjects (System trigger vs. Physio trigger) MANOVA	<ul style="list-style-type: none"> Determine if effects across the Scenarios differed by groups. 	Groupings into System and Physio trigger were different for scenarios SSS and SDD. Separate analyses were conducted for Single task (N=10 & 52) and Dual task (N=15 & 47).

Scenario SSS.

Results of a one-way between-subjects MANOVA conducted on the System-trigger and Physio-trigger groups as grouped in scenario SSS revealed no significant differences in stress

states experienced by the two groups in Pre-Task, as well as after each of the four scenarios, $p>0.05$ in each case.

Scenario SDD.

Although Wilk's lambda did not reveal a statistically significant effect of the groups on stress states, $\Lambda = 0.770$, $F(15,46)=0.915$, $p=0.554$, $\eta_p^2=0.230$, the univariate analyses indicated that despite having similar levels of Pre-Task Worry, the decline in Worry for the Physio-Trigger group in scenario SDD was significantly smaller compared to the decline in the System-trigger group for three out of four scenarios. For stress states, $F(1,60)=5.999$, $p=0.017$, $\eta_p^2=0.091$ for Scenario S, $F(1,60)=4.505$, $p=0.038$, $\eta_p^2=0.070$ for Scenario D, $F(1,60)=2.190$, $p=0.144$, $\eta_p^2=0.035$ for Scenario SSS, and $F(1,60)=4.909$, $p=0.031$, $\eta_p^2=0.076$ for Scenario SDD) (see Table 50).

Table 50
DSSQ Pre-Post Worry by Trigger Type group as grouped in scenario SDD (N=62)

Scenario*	Trigger type	Mean	Std. Error	95% Confidence Interval	
				Lower Bound	Upper Bound
Pre Task Worry	System-trigger	13.596	.911	11.773	15.419
	Physio-trigger	13.067	1.613	9.840	16.294
†Scenario S Worry	System-trigger	-5.638	.664	-6.966	-4.311
	Physio-trigger	-2.333	1.175	-4.683	.017
†Scenario D Worry	System-trigger	-6.362	.825	-8.013	-4.711
	Physio-trigger	-2.800	1.461	-5.723	.123
Scenario SSS Worry	System-trigger	-5.234	.898	-7.030	-3.439
	Physio-trigger	-2.533	1.589	-5.712	.645
†Scenario SDD Worry	System-trigger	-6.021	.774	-7.570	-4.472
	Physio-trigger	-2.533	1.371	-5.275	.208

*All scenario scores were computed as difference from Pre-Task level

†Significant differences between System-trigger and Physio-trigger groups.

Physiological workload index

Effects of Task and Block.

Table 51
Analyses of the Effect of Taskload and Block (Physio.Workload Index)

Analyses	Analyses intent	Participants in analyses
2(Taskload: Single vs. Dual) x 3(Block: Blk1 vs. Blk2 vs. Blk3) ANOVA	<ul style="list-style-type: none"> • Determine effects of Taskload • Determine effect of Block • Determine Taskload X Block effect (i.e., if effects of Taskload differed by Block) 	Only participants who had system-triggered aid. All had the same stimuli throughout the study (N=43).

To evaluate the relationship between taskload, block and workload as measured by the physiological workload index, a 2(Taskload: Single vs. Dual) x 3(Block: Blk1 vs. Blk2 vs. Blk3) repeated measures ANOVA was computed. Analysis was confined only to participants who experienced the system-triggered aid and thus were exposed to the same stimuli throughout the experiment. Results revealed a significant effect of Taskload, $F(1,41)=15.651$, $p<0.001$, $\eta_p^2=0.276$, Block, $F(2,82)=21.694$, $p<0.001$, $\eta_p^2=0.346$, and Taskload x Block interaction, $F(2,82)=10.695$, $p<0.001$, $\eta_p^2=0.207$. The physiological workload index was higher for Dual task ($M=0.348$, $SE=0.017$) compared to Single task ($M=0.288$, $SE=0.020$; $d=0.493$), and highest during Blk3 when aid was provided ($M=0.376$, $SE=0.023$), followed by Blk2 ($M=0.310$, $SE=0.018$), and lowest in Blk1 ($M=0.268$, $SE=0.017$) with Single task for both scenarios SSS and SDD.

Additional analyses of the interaction by taskload level showed that for the blocks in scenario SSS, the physiological workload index was highest in Blk3 ($M=0.326$, $SE=0.029$), followed by Blk2 ($M=0.260$, $SE=0.022$), and then Blk1 ($M=0.278$, $SE=0.020$) with the difference between Blk2 and Blk3 being significant. Without the confound of varying taskload across the blocks in scenario SSS, it appears that the physiological workload index tended to be elevated at

the beginning of the scenario and increased further when aid was provided. For scenario SDD, the physiological workload index was again highest in Blk3 when aid was provided, but lowest in Blk1, which was probably the result of the Single task in that block (Blk1: $M=0.257$, $SE=0.020$, Blk2: $M=0.360$, $SE=0.019$, Blk3: $M=0.426$, $SE=0.024$).

Changes in the physiological workload index over all samples over time concurred with the results analyzed by Blocks. As expected, the workload index was higher during the Dual task blocks as the Rolling Test Difference Scores were more similar to the Benchmark Difference Scores. Rolling Test Difference Scores computed from Single-Single Task Pairs (i.e., using data from the Single Task Baseline condition and Single task blocks in the experimental condition only) were more dissimilar as evidenced by lower workload index values. Index values in the 2nd (Blk2) and 3rd (Blk3) blocks in scenario SDD, i.e., pairing the Single Task Baseline with Dual task blocks in the experimental condition, were higher, indicating that the Rolling Test Difference Scores that used Dual task blocks show greater similarity to the Benchmark Difference Scores, which were computed from the Single and Dual Task Baseline conditions. Aid was imposed in Blk3 in both conditions and resulted in a rise in workload index values (see Figure 25).

Figure 25. Physiological workload index over all samples in scenarios SSS and SDD. (Shaded areas correspond to the data used in the Block analyses. Error bars show 95% C.I.).

Differences between Trigger Type groups.

Scenario SSS.

No analyses on the physiological workload index were conducted by the Trigger type grouping variable because the physiological workload index value was the basis for the trigger type grouping. However, the workload index levels are more varied for the Physio-triggered aid group (participants received aid at different times in Blk1 and Blk2 blocks) compared to the System-triggered aid group. Even when aid was introduced to the System-triggered aid group in the Blk3 block, the increase in workload index levels was modest. Although the differences in variability between the groups could be, to an extent, explained by the differences in sample sizes

(N=52 for System-trigger group, N=10 for Physio-trigger group), the physiological responsivity of the two groups warrants further investigation (see Figure 26).

Figure 26. Physiological workload index over scenario SSS.
(Shaded areas correspond to the data used in the Block analyses. Error bars show 95% C.I.).

Scenario SDD.

No between-groups analyses were conducted since the groups (i.e., System-triggered vs. Physio-triggered groups) were defined by the workload index. As with scenario SSS, the Physio-triggered aid group exhibited more dramatic variability in workload levels. This is especially true between blocks Blk1 and Blk2 when the scenario transitioned from Single task to Dual task. In contrast, the workload index values in the System-trigger group showed only a gradual rise as taskload “ramped up” from Single task (Blk1 block) to Dual task (Blk2 block), and when the aid was provided in the Blk3 block. Even though the sample size difference of the groups (N=47 in the System-trigger group, N=15 in the Physio-trigger group) may account for these differences in

variability in this study, further research is still needed to understand the differences in psycho-physiological responsivity to workload/taskload (see Figure 27).

Figure 27. Physiological workload index over scenario SDD.
(Shaded areas correspond to the data used in the Block analyses. Error bars show 95% C.I.).

Workload index values across different studies.

Comparing the workload index values across studies, it appears that the values for Study 2 were lower than that obtained with the Abich (2013) and Study 1 data. The mean proportion of repeated markers in the Abich (2013) data was ($M=0.552$, $SD=0.24$) for the S12_S14 Single-Dual task pair, and ($M=0.537$, $SD=0.23$) for the S32_S34 Single-Dual task pair, and in Study 1, it was ($M=0.509$, $SD=0.22$) for the Single-Dual task pair (i.e., Scenario1-Scenario2 and Scenario1-Scenario4) (see Table 17). For the dual task blocks in Study 2, the mean proportion of repeated markers for the Single-Dual task pairs was ($M=0.395$, $SD=0.187$) for the System-trigger group, and ($M=0.457$, $SD=0.218$) for all participants (see Figure 25). Although the values are still larger than that with the random data or dissimilar task pairs, the finding is surprising given that

the same tasks and task parameters were used in both the Abich (2013) and Study 2. It is possible that the difference is due to the varying task durations of the Baseline conditions in the studies. In the Abich (2013), the scenarios that were used to compute the Benchmark Difference Scores were 15-minutes in duration. In Study 1, the scenarios that were used to compute the Benchmark Difference Scores were 12 minutes in duration. For Study 2, the Baseline conditions which would yield the Benchmark Difference Scores were only 5 minutes in duration to avoid overexposure to the task(s). It is possible that the duration of 5 minutes may not be sufficiently long for the physiological markers to be determined reliably.

Study 2 hypotheses addressed

Hypothesis 2.1: Taskload would inversely relate to Performance.

There were higher detection rates, lower percent misclassified, and faster response times during the Single task blocks compared to Dual task blocks, supporting the hypothesis that taskload and performance were inversely related and confirming that the single vs. dual task paradigm is an effective manipulation of taskload for the MIX testbed.

Hypothesis 2.2: Taskload and Workload would relate positively.

For the ISA, workload ratings during the Dual task blocks were higher than that in Single task blocks. For the TLX, highest ratings were obtained on all subscales for scenario SDD (highest taskload), and lowest workload was experienced in scenario S (lowest taskload). This supports the hypothesis that workload and taskload were positively associated (a performance-workload dissociation). Similar pattern of results were obtained for the Distress stress state subscale of the DSSQ. The greatest Distress level was experienced in scenario SDD, while the lowest

level was reported in scenario S. However, for Task Engagement the highest level was observed in scenario D, while the lowest level was reported in scenario SSS.

Hypothesis 2.3. Performance would improve with aid.

For the CD task, providing aid resulted in a higher detection rates, quicker response times but also more misclassified change events. It appears that the aid improved some aspects of performance in the CD task. The improvements in detection rate were also greater for Dual task blocks compared to Single task blocks. TD performance did not improve when aid was provided. This hypothesis was therefore partially supported.

Hypothesis 2.4. Performance improvements from the aid would be greater for those who were provided aid based on physiologically-based workload measures compared to those who had aid imposed.

For the CD task, providing the aid during Single task resulted in rise in false alarm frequency for those whose aid was Physio-triggered, but a decline in false alarm frequency for those whose aid was System-triggered. However, providing the aid during Dual task led to improved detection rates and faster response times for both groups, but the improvements were greater for the group with Physio-triggered aid. This hypothesis was largely supported, but the relatively small number of participants in the Physio-trigger group indicates that these results should be replicated in experiments that further explore the triggering criteria.

CHAPTER EIGHT: DISCUSSION

Taskload effects

Results indicated strong effects of taskload as performance was more superior, and self-reported workload, in the form of ISA ratings, in the Single task blocks were lower than that in the Dual task blocks. This supported the hypothesis and confirmed the effectiveness of the taskload manipulation using Single and Dual tasks. The physiological workload index, which provided a more objective measure of workload, was also sensitive to changes in taskload levels, with the Single task blocks eliciting lower workload than the Dual task blocks. On all TLX measures, perceived workload ratings were lowest in briefer Single Task Baseline condition (Scenario S) and highest in the lengthier Dual Task Experimental conditions (Scenario SDD). Workload differences between the shorter, but more difficult Dual Task Baseline condition (Scenario D) and lengthier, but easier Single Task Experimental condition (Scenario SSS) were mostly small and not statistically significant.

Taskload also influenced pre-post task changes in stress states. Task Engagement was slightly elevated from Pre-Task levels in the 5-minute Dual Task Baseline condition, but decreased from Pre-Task levels for the other three conditions, with the decline being the greatest in the lengthier Single Task Experimental condition (Scenario SSS). This may reflect an increased investment of mental resources to deal with the increased taskload, and a disengagement when performing a less challenging task for an extended period of time. On the other hand, Distress levels increased from pre-task levels across all conditions, with the lengthier and heavier taskload Dual Task Experimental condition (Scenario SDD) eliciting the greatest Distress. Dual task

blocks seem to encourage higher Task Engagement, but also evoked greater distress when performed for a longer period of time.

Aid effects

The effects of the CD task aid on performance were robust. When aid was provided, detection rates were higher, and responses times were faster, but percentage of classification errors also increased. The improvement in detection rates due to the aid, was larger when the taskload was heavier (i.e., in Dual task). Supplemental evidence from informal interviews after the study sessions revealed that, when the aid was provided, about half of the participants admitted to guessing the change type when they did not actually see the change whenever there was a BEEP from the aid. This may account for the increase in the classification errors when aid was provided. In addition, the CD aid did not seem to affect performance on the TD task.

The aid had the unexpected effect of increasing workload as assessed by the ISA and the physiological workload index. This was a performance-workload dissociation that occurred when additional resources are recruited to improve performance (Yeh & Wickens, 1988). It is likely that there were other processes at work, but unfortunately, limited inferences can be drawn as the effects of aid on perceived workload as measured by the TLX, and the effects of the aid on stress states cannot be determined because the TLX and DSSQ were administered after each scenario, whereas the aid was triggered within scenarios. Nevertheless, the pattern of data indicate that the aid did not improve performance by reducing workload.

Imposed aid vs. adaptive aid

The core of the research question for the present work pertains to the effects of the aid when adapted to an individual's workload needs. To address this, comparisons were made between the effects of the aid between the System-trigger group, who had imposed aid, and the Physio-trigger group, whose level of workload, as measured by the physiological workload index, indicated a need for the aid. The findings revealed differential effects of the aid for both groups. Under conditions of lower taskload (Single task), the aid increased frequency of false alarms when it was provided at the time it was needed (i.e., Physio-trigger group), but reduced false alarm frequency when it was imposed (i.e., System-trigger group). Under conditions of higher taskload (Dual task), although the aid resulted in an overall improvement in detection rates and responses times, it improved detection rates by a larger extent in the Physio-trigger group such that their detection rates surpassed that of the System-trigger group, and their slower response times levelled up to that of the System-trigger group. Hence, in high taskload situations, when the aid matched workload needs, detection rates and response times improved more without any accompanying increase in classification errors. This underscores the importance of providing aid at the right time, i.e., when the operator functional state indicates that it is needed.

Taskload-Performance relationship

For Study 1, CD performance was poorer (i) when CD event rate was high compared to when CD event rate was low, as well as (ii) when CD was paired with PT in a dual tasking condition compared to when CD was performed alone as a single task. However, interpretation of the results of Study 1 involving CD levels is caveated by the fact that the levels were not counterbalanced, the duration of the CD Low block was slightly shorter (5 min) than that of the CD

High block (7 min), and effects due to the taskload transition are likely to be observed in the CD High block. Nevertheless, the Single-Dual task manipulation indicated an inverse relationship between taskload and performance, as expected. For Study 2, the same Single-Dual taskload manipulation yielded similar results. CD performance was poorer in Dual task blocks compared to the Single task blocks. These results suggest that the taskload-performance relationship, for the levels of CD task used in these studies, is reliable for the Single-Dual taskload manipulations.

However, this inverse relationship was not observed with the PT task in Study 1. PT performance was poorest in the scenario with lowest PT taskload (lowest event rate for PT prompts). A similarly counter-intuitive finding was reported in Abich (2013) which found better TD performance with increasing TD taskload. Both the Peripheral and the TD tasks were developed on the same MIX platform and involved the same simulated environment that contained characters and features that resemble the stimuli found in a real-world environment. Both tasks required “moving” through, and monitoring the environment. In contrast to the CD task and other similar tasks used in workload studies that utilize symbols or more static stimuli, (e.g., Warm et al., 2009; Helton et al., 2008, 2010; Funke et al., 2010), performance on the PT and TD tasks may have improved with taskload because the stimuli in these tasks were more engaging, like that in video games. Another study that also used a video-game format reported an absence of the vigilance decrement in the dynamic vigilance task, and proposed that such formats reduce the boredom and monotony that would deplete resources in more traditional studies that use abstract or static stimuli (Szalma et al., 2014). Hence, the task structure for the TD task may have moderated the taskload-performance relationship observed.

Online physiological workload index

Study 1 explored various ways to derive a workload index based on physiological workload measures. As the models obtained from the logistic regression analyses were too disparate and could not be validated against the test sets, an alternative approach was adopted that accommodated individual differences in physiological responses. After evaluating several algorithms, the algorithm that was selected first identified the physiological markers of workload for the individual from the Benchmark Difference Scores, and classified the participant as experiencing High workload when a higher enough proportion of these markers later emerged in the Test Difference Scores, indicating that the current level of workload elicited a physiological response that was comparable to that elicited during the High workload level in the earlier Dual Task Baseline condition.

The physiological workload index values observed in Study 2 were lower than those in the Abich (2013) study and Study 1, despite the use of the same tasks and parameters. A plausible explanation is that the duration of the Baseline tasks in Study 2 may have been too brief for the physiological markers to be identified accurately and reliably. Permitting more time for the baseline conditions might allow a more reliable evaluation of the individual's physiological responses to workload to be obtained. Nonetheless, there is also the larger issue of the stability of the physiological workload response. Even with high workload Dual tasks, workload levels can change over time with learning, mastery of the task, adapting strategies which more often result in a reduction of workload (Moon et al., 2013), etc. Further investigation is needed on the diagnosticity of the physiological responses and how workload indices and triggering criteria should be determined.

Other issues with performance and the aid

Performance is a multi-faceted construct. In the present study, performance on the CD task was operationalized as Percent correct, Percent misclassified, Average response time to correct detections, and Number of false alarms. Performance on the TD task was assessed by Total number of threats correctly detected, Total incorrect detections, and Total duplicate detections. Although the aid improved detection rates and reduced response time, it also increased classification errors, suggesting a possible criterion shift. In certain situations, this outcome may not be favorable. For instance, for the human-robot team in the ISR mission, small improvements in response time is not worth the cost of higher misclassification rates, especially if it means diverting costly resources to the wrong location. It would be necessary to define the constituents of desirable performance in various phases of the task or mission.

The effects of adaptive aid warrants further investigation. In scenario SSS, the aid resulted in a slight decline in false alarm frequency in the System-trigger group, but an increase in false alarm frequency in the Physio-trigger group. On the other hand, in scenario SDD, the aid led to decrease in TD performance (i.e., fewer correct detections but greater number of duplicate threat identifications) for both groups, indicating that resources may have been diverted away from the TD task. These observations suggest that the aid may have resulted in a shift in response criterion. For scenario SDD, this criterion shift led to more duplicates in the TD, task, without affecting false alarm rates in the CD task. Conversely, for scenario SSS, where there was no second task, the shift in response criterion led to an increase in false alarms.

In the present study, aid was in the form of “BEEPs” that occurred with each change event, and was intended to make the changes more salient (Taylor, 2012; 2013). The use of an

auditory alert instead of a visual one was, according to the Multiple Resource Theory (Wickens, 2008), to avoid over-taxing resources of the same modality since the task is already a predominantly visual one. Given this, the aid should lead to improved performance with minimal increase in workload. Instead findings were the exact opposite as there was a performance-workload dissociation. This may have been a result of several factors. First, the aid could have elicited a type of orienting response when it was triggered. This may account for the initial rise in the index values immediately following the onset of the aid. Although the participants were familiarized with the aid and the “BEEPs” during training, and were notified of the impending aid immediately before it was triggered (i.e., auditory notification “Robot teammate now in Peer mode and will assist with change detection task”), it is possible that at aid onset, the “BEEPs” were newly-added auditory stimuli that led to a type of orienting response that consumed attentional resources. This may be akin to the rise in workload observed in driving studies when external events are encountered (e.g., Clarion et al., 2009).

Secondly, the aid may have altered the structure of the task. The aid may have modified the CD task such that instead of being a solely visual task involving change blindness, it became a visual and auditory task with the “BEEPs” increasing the saliency of the change events and cadence of the changes. This concurs with accounts from some participants who reported changing their strategy when the aid was triggered. Changes in strategy have been associated with increases in workload as strategy selection or adoption of a new strategy is resource consuming (Moon et al., 2013). A third explanation for the increased workload when aid was provided is that, although unintended, the aid inadvertently provided feedback to the participant on their performance on the CD task. From the informal interviews conducted post task, many participants reported that, in addition to directing the attention to the CD task, the aid informed them

of the change events they failed to detect. This feedback could have (i) motivated participants to do better on the task, or (ii) demotivated them as they were notified of their misses. According to the Feedback Intervention Theory (FIT), feedback or knowledge of results, influences the extent to which the participant engages in resource-consuming processes that are either task-relevant or task-irrelevant (Kluger & DeNisi, 1996, 1998). Feedback that contains information about the task is more likely to encourage resources to be directed towards task-relevant processes which lead to performance improvements, whereas feedback that directs resources towards task-irrelevant processes such as self-appraisal on the task, does not usually result in improvements. That detection rates improved with the BEEP aid suggests that the feedback from the aid was informative of the participant's performance on the task and thus resulted in more resources directed towards task-relevant processes. If this was the case, then performance improvements did not result from a reduction in workload, but resulted indirectly *because* of it. That is, there were processes (e.g., self-regulatory, task-relevant processes) that were activated as a result of the aid, which contributed to the higher workload experienced, but also enhanced performance. This particular instance of performance-workload dissociation where both workload and performance increase, is in direct contrast to the performance-workload dissociation that arise when the participant "gives up", leading to declines in both workload and performance (Hancock, 1996). Instead, the pattern of results observed in this study indicates that participants were able to improve their performance with the aid, but only with the energetic cost of higher workload. It may be that providing robot aid can increase workload in a manner similar to that observed in automation research (Parasuraman & Riley, 1997).

Hence, the extent to which the closed-loop system enhances performance depends on the type of aid and how the participant interprets and uses it. For instance, it appears that even in the

absence of any formal feedback, there are self-evaluative processes that participants engage in that encourage them to interpret information and stimuli as being indicative of their performance. Such examples support the transactional perspective of workload and task stress (Matthews et al., 2013).

Other processes that occur when participants perform tasks pertain to concentration, motivation, affect, self-regulation, as well as self-referent, cognitive interfering thought processes associated with stress from task(s) (Matthews et al., 2006; Matthews et al., 2013). Results from the DSSQ revealed that Task Engagement, which reflects cognitive appraisals related to commitment to effort, motivation and energy invested in a task, was lowest when performing a less challenging task with lower taskload for a longer period of time (i.e., Single Task Experimental condition), and highest when performing a more challenging task that was not as long (i.e., Dual Task Baseline condition). Increases in Distress levels from Pre-Task from all scenarios reflect the workload imposed by the tasks (Matthews et al., 2013), with scenario SDD (high taskload for a longer period of time) inducing the greatest rise in Distress.

Worry assesses the extent of self-referencing thoughts and cognitive interfering processes experienced (Matthews et al., 2013). Levels of Worry likely declined for all scenarios because Pre-Task levels of Worry often reflect anticipatory anxiety and appraisals about the task (Matthews & Funke, 2006). Among the scenarios, although the differences failed to reach significance, the trend of reductions in Worry seem greater for the more complex, multi-tasking Dual task conditions (i.e., $M = -6.372$ for D and $M = -6.023$ for scenario SDD) compared to the Single task conditions (i.e., $M = -5.674$ for S and $M = -5.349$ for scenario SSS). It is possible, that being the more complex conditions, the Dual task conditions were more effective in directing attention away from the self and towards the task (Matthews et al., 2013).

When the participants were grouped by whether they received System-triggered aid or Physio-triggered aid in scenario SDD, it was observed that the decrease in Worry was smaller for the Physio-trigger group (decline ranged from -2.333 to -2.800) compared to the System-trigger group (decline ranged from -5.234 to -6.362) for the Single Baseline, Dual Baseline, and Dual Experimental conditions, even though the two groups did not differ on Pre-Task Worry levels. Given that the group with Physio-triggered aid experienced greater workload (because their level of workload was higher enough to trigger the aid), their increased workload may be attributed to greater activation of self-referencing and cognitive interference that manifests as greater Worry, but also task-relevant processes induced by the feedback provided by the aid since this group also exhibited performance improvements in the form of higher detection rates and faster response times in scenario SDD.

Theoretical implications

The results of the present study appear to support the premise that robot aid should be adaptive and based on workload needs instead of being imposed or based merely on taskload. This is because workload, by also accounting for resources that are consumed by other processes activated as a result of the operator's interaction with the task, and his response to task demands provides a more accurate picture of the amount of resources available for allocation to the task. Such a perspective concurs with findings of the multidimensional nature of workload (Matthews, 2015a; 2015b). Results such as the performance-workload dissociation and the increases in workload from the aid, highlight the importance of these other processes that contribute to the complexity of the taskload, workload and performance relationship that the various theories of workload and stress attempt to describe. In particular, the study provides support for the notion

that reducing workload per se may not be the goal of adaptive systems. As proposed by the Maximal Adaptability Model, the zone of maximal adaptation does not occur at the lowest levels of workload (Hancock & Warm, 1989). Perhaps, it is not only *how much* workload is experienced, but also *what* the workload constitutes. It is plausible that some level of workload may be attributed to self-regulatory processes that are necessary to keep the operator within the zone of maximal adaptability in the same way that some behaviors in ADHD children (e.g., fidgeting) may be necessary for them to maintain alertness, and are workings of a compensatory mechanism (Sarver et al., 2015). This compensatory mechanism may not be unlike that described in Hockey's Compensatory Control model of performance regulation (Hockey, 1997), which is itself conceptualized as a closed-loop system. The workload that results from resources diverted to self-regulatory processes that helps the operator adapt and maintain performance is qualitatively distinct from workload that results from negative self-appraisals or cognitive interferences that are task-irrelevant and should be treated as such.

In the present work robot adaptive aiding was intended to be determined by the individual's workload needs because of the importance of the individual's workload experience and Operator Functional State (OFS), which are influenced by the various accompanying appraisal, coping, regulatory, compensatory processes etc., during performance on tasks. As Moray asserts: "Individual differences in workload research is far more important than has hitherto been acknowledged. Without taking this into account we are seriously delaying the development of a useful measure." (Moray, 1984, p.41). Measures of workload are themselves subject to individual differences (Meshkati & Loewenthal, 1988). In the present study, the failure to obtain a reliable predictive model from logistic regression that would entail having the same set of physiologi-

cal predictors and weights for all individuals, is evidence of the influence of individual variability in physiological workload measures. Moreover, at the group level, the graphs depicting physiological workload over time appears fairly consistent across blocks of constant taskload (e.g., Figure 26), however, the graphs of the physiological workload index for small groups of individuals (e.g., the physio-trigger groups) as well as that of individuals across blocks of constant taskload and stimuli show extreme variability which was somewhat masked in aggregate data. Though in the present study, the selected algorithm accommodated individual differences in psychophysiological responsivity to a degree, other methods that are able to model these individual differences and variability more precisely should be explored. These methods may include data mining techniques such as cluster analyses, neural networks, machine learning algorithms etc. For instance, using cluster analysis, Johannes and Gaillard (2014) identified five Autonomic Response Patterns (ARP) based on normalized eigendifference scores. These ARPs, which are influenced differently by coping styles, task demands, and individual states, were used in a scaling method to calculate an index of psychophysiological arousal that accounted for individual differences in autonomic responsivity (Johannes & Gaillard, 2014).

Practical applications

The present study is an example of how an online measure of workload can be used to drive adaptive aid through a closed-loop system. In the present context of a Human-Robot team, the construct of interest is workload or OFS during the task/mission, with the intervention being the robot's aiding behavior. In a training context, the constructs of interest may be workload and learning, which requires the neural substrates of learning and memory to be identified. Interventions in this context may be adapting information presentation format, or training curriculum to

the trainee's learning pace (e.g., adaptive tutors). In the marketing domain, physiological and neural substrates would have to be identified for the constructs of interest like emotional valence and arousal which can be indicative of purchase decisions in the same way that eye-tracking measures have been applied to advertising research (Hicks, 2014). Although there are other online measures of state such as the suite of wearable technologies available, there needs to be a better understanding of the models on which they are based. For instance, the *Apple Watch* and *Fitbit* track activity that are purportedly indicative of health and fitness (Jary, 2015), *Spire* allegedly monitors tension and stress from respiratory measures (Pilkington, 2014), *Exmocare* claims to monitor emotional state from cardiac activity, GSR, skin temperature and movement (Ostrovsky, 2008).

In essence, broader applications of this research are contingent on the extent that online physiological measures are able to capture meaningful information about the individual during a task or event. For the present context, the online workload index can help illuminate the processes that goes on during task performance and enable the relevant theories to be tested and refined. The information about these processes is also diagnostic and will be invaluable in developing strategies and interventions that can be tailored to the individual.

Future research

Limitations of the current study present opportunities for future research. Study 1 evaluated a limited number of algorithms to individualize the workload model, and the one that seemed to perform the best was selected. The designation of the cutoff for the adaptive aid was also not based on any concrete a priori criteria. An alternative is to have algorithms and cutoffs

that can be defined according to meaningful criteria. This may require other modeling techniques. In Study 2, the set of physiological workload markers for the individual was determined only from two 5-minute conditions (i.e., scenarios S and D). It was assumed that the physiological responses observed in the two conditions were representative of the individual's physiological response reflecting Low and High workload. Although this approach was based on data from a previous study that used the same tasks, results also showed that the set of markers may depend on the duration of the baseline tasks. The stability and generalizability of the workload markers across other tasks and durations remains to be determined. Furthermore, in the present study, once the physio. trigger aid was evoked, it was never revoked even if the workload index may have fallen below the cutoff value, which may indicate that the aid was no longer "needed". Although the durations of the present study scenarios were not long enough to investigate the persistent effects of providing aid that may no longer be needed, future research should examine the longer-term effects of such an aid schedule.

In order to better understand the role of workload in moderating the taskload-performance relationship, a deeper understanding of the processes involved in the operator's interaction and transactions with the task and performance is needed. Having a continuous measure such as that implemented in the present study will serve to elucidate many of the processes that occur during task performance which can provide the means to test specific theoretical postulates, specifically, the role of individual differences in influencing these processes.

Conclusion

Soldiers and human operators interacting with technology and working closely with robots will only become more of a reality in the near future. The present work demonstrated the

feasibility of using a closed-loop system based on physiological workload measures to drive adaptive robot aid. Although there are diverse applications of such an adaptive system, the findings clearly indicate a need for more research as it is not advisable to adopt a “one-size-fits-all” approach with workload measures. Future research and development on how these robots may work more collaboratively and communicate more intuitively with the human, must be driven by a deeper understanding of how individuals interact with changing taskloads, as well as the nature of the workload they experience while doing so.

APPENDIX A:
UCF APPROVAL OF HUMAN-SUBJECTS RESEARCH

University of Central Florida Institutional Review Board
Office of Research & Commercialization
12201 Research Parkway, Suite 501
Orlando, Florida 32826-3246
Telephone: 407-823-2901 or 407-882-2276
www.research.ucf.edu/compliance/irb.html

Approval of Human Research

From: UCF Institutional Review Board #1
FWA00000351, IRB00001138

To: Lauren Reinerman

Date: October 13, 2014

Dear Researcher:

On 10/13/2014, the IRB approved the following human participant research until 10/12/2015 inclusive:

Type of Review:	IRB Continuing Review Application Form
Project Title:	Enhancing the effectiveness of Human-Robot Teaming with a closed-loop system.
Investigator:	Lauren Reinerman
IRB Number:	SBE-13-09691
Funding Agency:	RDECOM-STC
Grant Title:	
Research ID:	1052585

The scientific merit of the research was considered during the IRB review. The Continuing Review Application must be submitted 30 days prior to the expiration date for studies that were previously expedited, and 60 days prior to the expiration date for research that was previously reviewed at a convened meeting. Do not make changes to the study (i.e., protocol, methodology, consent form, personnel, site, etc.) before obtaining IRB approval. A Modification Form cannot be used to extend the approval period of a study. All forms may be completed and submitted online at <https://iris.research.ucf.edu>.

If continuing review approval is not granted before the expiration date of 10/12/2015, approval of this research expires on that date. When you have completed your research, please submit a Study Closure request in IRIS so that IRB records will be accurate.

Use of the approved, stamped consent document(s) is required. The new form supersedes all previous versions, which are now invalid for further use. Only approved investigators (or other approved key study personnel) may solicit consent for research participation. Participants or their representatives must receive a signed and dated copy of the consent form(s).

All data, including signed consent forms if applicable, must be retained and secured per protocol for a minimum of five years (six if HIPAA applies) past the completion of this research. Any links to the identification of participants should be maintained and secured per protocol. Additional requirements may be imposed by your funding agency, your department, or other entities. Access to data is limited to authorized individuals listed as key study personnel.

In the conduct of this research, you are responsible to follow the requirements of the [Investigator Manual](#).

On behalf of Sophia Dziegielewski, Ph.D., L.C.S.W., UCF IRB Chair, this letter is signed by:

A handwritten signature in black ink, appearing to read "Joanne Muratori". The signature is fluid and cursive, with the first name "Joanne" and last name "Muratori" clearly distinguishable.

Signature applied by Joanne Muratori on 10/13/2014 01:57:52 PM EDT

IRB Coordinator

APPENDIX B:
INFORMED CONSENT

Enhancing the effectiveness of Human-Robot Teaming with a closed-loop system.

Informed Consent for an Adult

Principal Investigator(s): Lauren Reinerman-Jones, Ph.D.

Research Associates: Julian Abich IV, Ph.D.
Avonie Parchment, M.S.
Brandon Sollins, B.Sc.
Grace Teo, M.A.
Ryan Wohleber, M.A.
James Tyson

Faculty Supervisor: James Szalma, Ph.D.

Sponsor: Department of Defense (DOD)

Investigational Site(s): *Institute for Simulation and Training
University of Central Florida
3100 Research Parkway
Orlando, FL 32826*

Introduction: Researchers at the University of Central Florida (UCF) study many topics. To do this we need the help of people who agree to take part in a research study. You are being invited to take part in a research study which will include about 300 people at UCF. You have been asked to take part in this research study because you are a student at UCF. You must be between 18 and 65 years of age to be included in the research study.

What you should know about a research study:

- Someone will explain this research study to you.
- A research study is something you volunteer for.
- Whether or not you take part is up to you.
- You should take part in this study only because you want to.
- You can choose not to take part in the research study.
- You can agree to take part now and later change your mind.
- Whatever you decide it will not be held against you.
- Feel free to ask all the questions you want before you decide.

Purpose of the research study: The objective is to evaluate the use of a closed-loop system to enhance the effectiveness of Human-Robot teaming.

Inclusion and Exclusion Criteria: To participate in the study, you would need to meet the following criteria:

- Must be between the ages of 18 years and 65 years
- Must be in general good health
- Must not have a history of epilepsy or seizures
- Must not have metal plates in head
- Must not be pregnant
- Must not be experiencing of any of the following at a moderate or severe level: General Discomfort, Fatigue, Headache, Eye Strain, Difficulty Focusing, Increased Salivation, Sweating, Nausea, Difficulty Concentrating, Fullness of Head, Blurred Vision, Dizzy (Eyes Open), Dizzy (Eyes Closed) Vertigo, Stomach Awareness, Burping

What you will be asked to do in the study:

If you consent to participating in this study and meet all the inclusion criteria, you will be required to complete a demographics questionnaire and nutrition diary. The researcher will also fit you with the physiological sensors and a baseline measure will be taken. You will then be engaged in a computer-simulated task(s) on a microgame platform that involves completing an Intelligence and Surveillance Reconnaissance (ISR) military mission in the context of Human-Robot teaming. You physiological states will be measured by the physiological devices described below while you complete the task and respond to various questionnaires.

A range of sensors and equipment will be utilized to assess your functional states. All of the physiological equipment used in this study is noninvasive. The devices used in this experiment will be an Advanced Brain Monitoring ten-channel Electroencephalogram (EEG) cap with Electrocardiogram (ECG) sensors attached to it. Additionally, a BIOPAC sixteen-channel, four-light source Functional Near Infra-red (fNIR) strap and a bilateral two-channel Transcranial Doppler Ultrasonography (TCD) system will be used. During the task, eye movements will be recorded with an Eyetracker. Each sensor will be custom set for each individual using its respective setup procedure. The following sections provide a description of the physiological measures in use.

EEG: The EEG sensors are contained in a neoprene cap that will be placed over the participant's head and adjusted by the lab technician. The conductive gel is placed on the

sensor sponge, which allows the sensor to touch the scalp without being abrasive. For cap placement, the participant will be seated in front of the computer. The researcher will take an alcohol swab (or equivalent if allergic) and wipe the mastoid bone (behind the ears just above your neck) where the sensors will touch. The research assistant will set the cap so that the front is aligned with the nasium (brow ridge between the eyes) and inion (occipital bone at the back of the head). Once the EEG cap is in place, the research assistant will test the impedance of the sensors to assure that proper conductance is occurring.

ECG: There are two sensors that need to be placed on the right collar bone and the lower left rib bone. These sensors will be placed by the participant. The participant will take an alcohol swab and clean the areas where the sensors will be placed. The research assistant will attach the sensor to the lead and put some conductive gel on the sensor. The participant will then place the sensor in their respective place on the right collar bone or the lower left rib bone. The research assistant will turn on the device and check to see that the EEG and ECG sensors are receiving signal. The signal strength will be evaluated via software on the experimenter's computer station.

fNIR: The fNIR sensors are applied by the research assistant using a strap across the prefrontal cortex. The participant will first wipe their forehead with an alcohol swab and clean the area. Then the fNIR strap will be fitted by the researcher to the participant.

TCD: Transcranial Doppler ultrasonography (TCD) would be used to monitor blood flow velocity with pulsed Doppler effect probes attached to a head-brace on either side of the participant's head. The TCD would be used to monitor blood flow velocity throughout the scenario.

Eyetracker: A desk-mounted eye tracker will monitor eye position and movement during the scenarios. Before the actual experiment, the eye tracker would need to be calibrated to the participant's gaze.

When the task(s) are over, the researcher will help you remove all the sensors. It is most helpful to the research being gathered that you answer all questions and complete all tasks to the best of your abilities, but you are not required to answer every question or complete every task. You will not lose any benefits if you choose not to complete questions or tasks.

Location: Institute for Simulation and Training, Partnership 2, Room 338.

Time required: We expect that you will be in this research study for 4 hours. No breaks are scheduled, although they may be given at the discretion of the researcher(s).

Funding for this study: Funding for this study is provided by the Department of Defense (DOD).

Risks: There are no foreseeable risks or discomforts other than those normally encountered in the daily lives of healthy persons. As in all studies, there is a potential risk to participants; however, in this

study, those risks are minimal. Given the use of the simulated environment, there is a slight risk that participants may experience symptoms of simulator sickness (nausea, headache, etc.). All researchers are trained to recognize these symptoms, and any participants found to experience these symptoms will be given a break until they feel they are ready to continue or until they decide they cannot complete the study and will be dismissed with credit.

There are no foreseeable risks associated with the physiological sensors. Sensors *do not* alter the body in any way. The sensors do require the use of electro-conductive cream and sonar gel, both are water soluble and hypoallergenic. In order to attach the sensors, certain areas may be cleaned with alcohol swabs to increase skin conductivity. Placement of the EEG, ECG, fNIR, and TCD may require movement of the hair with the use of a fine comb.

Potential direct benefits to participants: We cannot promise any benefits to you or others from your taking part in this research. However, possible benefits include some knowledge of the use of physiological sensors in research.

Compensation or payment: Participants may expect to spend 4 hours participating in the study, for which they will receive 1 SONA credit for every hour of participation. If the experiment takes longer and the participant is able to stay and complete the experiment, they will receive extra credit in increments of 15 minutes / .25 SONA credits. Additionally if a participant is unable to complete the study for any reason, they will receive .25 SONA credits for every 15 minutes of study participation. Participants who have not been recruited through the SONA system will not receive any form of compensation.

Confidentiality: We will limit your personal data collected in this study to people who have a need to review this information. However as this study is sponsored by the Department of Defense (DoD), it is possible that representatives of the Army Human Research Protections Office may inspect the investigator's records. Data will be secured in locked cabinets at the Institute for Simulation and Training (IST) and disposed of following IRB protocol, which includes the shredding of all documents and proper deletion of electronic information.

Study contact for questions about the study or to report a problem: If you have questions, concerns, or complaints, or think the research has hurt you, talk to Dr. Lauren Reinerman-Jones at 407-882-1140 or at reinerm@ist.ucf.edu.

IRB contact about your rights in the study or to report a complaint: Research at the University of Central Florida involving human participants is carried out under the oversight of the Institutional Review Board (UCF IRB). This research has been reviewed and approved by the IRB. For information about the rights of people who take part in research, please contact: Institutional Review Board, University of Central Florida, Office of Research & Commercialization, 12201 Research Parkway, Suite 501, Orlando, FL 32826-3246 or by telephone at (407) 823-2901. You may also talk to them for any of the following:

- Your questions, concerns, or complaints are not being answered by the research team.
- You cannot reach the research team.
- You want to talk to someone besides the research team.

- You want to get information or provide input about this research.

Your signature below indicates your permission to take part in this research.

DO NOT SIGN THIS FORM AFTER THE IRB EXPIRATION DATE BELOW

Name of participant

Signature of participant

Date

Signature of person obtaining consent

Date

Printed name of person obtaining consent

APPENDIX C:
DEMOGRAPHICS SURVEY

Demographics Survey

Demographics Questionnaire	
Group	<input type="text"/> New participant
Participant #	<input type="text"/>
Date (mm/dd/yyyy)	<input type="text"/>
Major	<input type="text"/>
Age (type in the number)	<input type="text"/>
Gender	<input type="text"/>
1 What is the highest level of education you have had?	<input type="text"/>
2. When did you use computers in your education? (Check all that apply)	☐ Grade School ☐ Jr High/Middle School ☐ High School ☐ Technical School ☐ College ☐ Did Not Use
3. Where do you currently use a computer? (Check all that apply)	☐ Home ☐ Work ☐ Library ☐ Other ☐ Do Not Use
4. How often do you.....	
a) Use a mouse?	<input type="text"/>
b) Use a joystick?	<input type="text"/>
c) Use a touch screen?	<input type="text"/>
d) Use icon-based programs/software (iphone, windows)?	<input type="text"/>
e) Use programs/software with pull-down menus?	<input type="text"/>
f) Use graphics/drawing features in software packages?	<input type="text"/>
g) Use E-mail?	<input type="text"/>
h) Operate a remote controlled vehicle (car, boat, plane, or robot)?	<input type="text"/>
5. Which type(s) of computer/video games do you most often play?	<input type="text"/>
6. Which of the following best describes your expertise with computers?	<input type="text"/>
7. Are you in your usual state of health physically?	<input type="text"/>
If NO, please briefly explain:	<input type="text"/>
8. How much experience do you have with virtual environments?	<input type="text"/>
9. What category is your occupation in?	<input type="text"/>
If other, please type in:	<input type="text"/>
What is the title of your position?	<input type="text"/>
a) How many years have you been in this position?	<input type="text"/>
b) How many years since your last promotion?	<input type="text"/>

c) How many years since you won any awards?	<input type="text"/>
d) How many people directly report to you?	<input type="text"/>
14 How often do you feel eye strain?	<input type="text"/>
11. During an average work day, do you feel that you focus on near objects (about 2 meters away) more than objects that are far away (6 meters or more)?	<input type="text"/>
12. Do you speak more than one language?	<input type="text"/>
13. If you answered yes to question 12, how fluent would you rate yourself on your 2nd language?	<input type="text"/>

Current Health Status. Please indicate how you feel right now in the following areas, by selecting from the drop down menu.

General Discomfort	<input type="text"/>	Difficulty Concentrating	<input type="text"/>
Fatigue	<input type="text"/>	Fullness of Head	<input type="text"/>
Headache	<input type="text"/>	Blurred Vision	<input type="text"/>
Eye Strain	<input type="text"/>	Dizzy (Eyes Open)	<input type="text"/>
Difficulty Focusing	<input type="text"/>	Dizzy (Eyes Closed)	<input type="text"/>
Increased Salivation	<input type="text"/>	Vertigo*	<input type="text"/>
Sweating	<input type="text"/>	Stomach Awareness**	<input type="text"/>
Nausea	<input type="text"/>	Burping	<input type="text"/>

*Vertigo is a disordered state in which the person or his/her surroundings seem to whirl dizzily; giddiness

**Stomach awareness is usually used to indicate a feeling of discomfort which is just short of nausea.

[Save Questionnaire](#)

1

* Participants completed the survey on MS Access. This is a screen shot of the MS Access form.

LIST OF REFERENCES

- Aaslid, R. (1986). Transcranial Doppler examination techniques. In R. Aaslid, (Ed.), *Transcranial Doppler Sonography*, pp. 39-59. New York: Springer-Verlag.
- Abich, J. IV. (2013). *Investigating the universality and comprehensive ability of measures to assess the state of workload*. Unpublished doctoral dissertation. University of Central Florida.
- Aki, L. (2012). 7 Must-have domestic robots – R2D2 & C3PO eat your hearts out. Retrieved from <http://www.amog.com/tech/153161-domestic-robots/>
- Ayaz, H., Cakir, M.P., Izzetoglu, K., Curtin, A., Shewokis, P.A., Bunce, S., Onaral, B. (2012). Monitoring Expertise Development during Simulated UAV Piloting Tasks using Optical Brain Imaging. *Proceedings of the IEEE Aerospace Conference*, MN, USA, pp. 1–11.
- Ayaz, H., Shewokis, P. A, Bunce, S., Izzetoglu, K., Willems, B., & Onaral, B. (2011). Optical brain monitoring for operator training and mental workload assessment. *NeuroImage*, 59(1), 36–47.
- Ayaz, H., Willems, B., Bunce, S., Shewokis, P. A., Hah, S., Deshmukh, A. R., & Onaral, B. (2010). Cognitive workload assessment of air traffic controllers using optical brain imaging sensors. In T. Marek, W. Karwowski, & V. Rice (Eds.), *Advances in Understanding Human Performance: Neuroergonomics, Human Factors Design, and Special Populations* (pp. 21–32). Taylor & Francis Group.
- Backs, R.W., & Walrath, L.C. (1992). Eye movement and pupillary response indices of mental workload during visual search of symbolic displays. *Applied Ergonomics*, 23(4), 243-254.

- Bainbridge, W. A., Hart, J., Kim, E. S., Scassellati, B. (2008). The Effect of Presence on Human-Robot Interaction. *Proceedings of the 17th IEEE Symposium on Robot and Human Interactive Community, Munich, Germany.*
- Baldwin, C. L., & Penaranda, B. (2012). Adaptive training using an artificial neural network and EEG metrics for within- and cross-task workload classification. *Neuroimage*, 59 (Neuroergonomics: The human brain in action and at work), 48-56. doi:10.1016/j.neuroimage.2011.07.047
- Bartneck, C., Kulic, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robots*, 1(1), 71-81.
- Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, 91, 276-292.
- Becker, A. B., Warm, J.S., & Dember, W.N. (1994). Specific and Nonspecific Transfer Effects in Training for Vigilance. In M. Mouloua and R. Parasuraman (Eds.). *Human Performance in Automated Systems: Current Trends*, 294–299. Hillsdale, NJ: Lawrence Erlbaum.
- Billings, C. E. (1997). *Aviation automation: The search for a human-centered approach*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Byrne, E. A., & Parasuraman, R. (1996). Psychophysiology and adaptive automation. *Biological Psychology*, 42(3), 249-268. doi:10.1016/0301-0511(95)05161-9.
- Cain, B. (2007). *A review of the mental workload literature*. Report No.: RTO-TR-HFM-121-Part-II. Toronto, Canada: Defence Research and Development Canada Toronto Human System Integration Section.

- Callan, D. (1998). Eye movement relationships to excessive performance error in aviation. *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting* (pp. 1132-1136). Santa Monica, CA: Human Factors and Ergonomics Society.
- Carmody, M. A., & Gluckman, J. P. (1993). Task-specific effects of automation and automation failure on performance, workload and situational awareness. In R. S. Jensen and D. Neumeister (Eds.) *Proceedings of the Seventh International Symposium on Aviation Psychology*, pp. 167-171.
- Casali, J. G., Wierwille, W.W. (1983). A Comparison of Rating Scale, Secondary-Task, Physiological, and Primary-Task Workload Estimation Techniques in a Simulated Flight Task Emphasizing Communications Load. *Human Factors*, 25(6), 623-641.
- 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 Sciences of the United States of America*, 90(8), 3770–3774.
- Clarion, A., Ramon, C., Petit, C., Dittmar, A., Bourgeay, J., Guillot, A., & Collet, C. (n.d). An integrated device to evaluate a driver's functional state. *Behavior Research Methods*, 41(3), 882-888.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.
- Comstock, J. R., & Arnegard, R. J. (1992). The multi-attribute task battery for humanoperator workload and strategic behavior research, *Tech. Memorandum No. 104174*. Hampton, VA: NASA Langley Research Center.
- Cox-Fuenzalida, L., & Angie, D. (2005). The effects of workload history on dual task performance. *Current Psychology: Developmental, Learning, Personality, Social*, 24, 171-179.

- Cox-Fuenzalida, L., Beeler, C., & Sohl, L. (2006). Workload history effects: A comparison of sudden increases and decreases on performance. *Current Psychology: Developmental, Learning, Personality, Social*, 25, 8-14.
- Craven, P., Belov, N., Tremoulet, P., Thomas, M., Berka, C., Levendowski, D., & Davis, G. (2006). Cognitive Workload Gauge Development: Comparison of Alternative Classification Methods. *Proceedings of the Second Annual AugCog International Conference (HFES/ACI 2006)*, San Francisco, CA.
- da Vinci surgery (2013). *Changing the experience of surgery*. Retrieved January 21, 2014 from <http://www.davincisurgery.com/>
- De Waard, D. (1996). *The Measurement of Drivers' Mental Workload*, University of Groningen, Groningen.
- Dekker S.W.A., Woods, D.D. (2002) MABA–MABA or Abracadabra? Progress on human-automation coordination. *Cognition, Technology and Work*, 4, 240–244.
- Department of Defense. (2005). *Department of Defense interface standard: Common Warfighter symbology* (MIL-STD-2525B).
- Dias, M. B., Kannan, B., Browning, B., Jones, E.G., Argall, B., Veloso, M.M. & Stentz, A. (2008). Sliding Autonomy for Peer-To-Peer Human-Robot Teams", *10th International Conference on Intelligent Autonomous Systems*.
- Donmez, B.D., Nehme, C., & Cummings, M.L. (2010). Modeling workload Impact in Multiple Unmanned Vehicle Supervisory Control. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, 40, 1180-1190.
<http://dx.doi.org/10.1109/TSMCA.2010.2046731>

- DRC (2013). DARPA Robotics Challenge 2013. Retrieved Jan 10, 2013 from <http://www.theroboticschallenge.org/about>
- Duschek, S., & Scahndry, R. (2003). Functional transcranial Doppler sonography as a tool in psychophysiological research. *Psychophysiology*, 40, 436-454.
- Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors*, 37, 381–394.
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175-191.
- Fidopiastis, C.M., Drexler, J., Barber, D., Cosenzo, K., Barnes, M., Chen, J.Y.C., & Nicholson, D. (2009). Impact of automation and task load on unmanned system operator's eye movement patterns. In D. D. Schmorow et al., (Eds.). *Augmented Cognition, HCII 2009, LNAI 5638*, 229-238. Springer-Verlag, Berlin, Heidelberg: Germany.
- Field, A.P. (2013). *Discovering statistics using IBM SPSS Statistics* (4th ed.). London ; Thousand Oaks: Sage.
- Fitts, P.M. (1951). *Human engineering for an effective air-navigation and traffic-control system*. Columbus, OH: National Research Council, Division of Anthropology and Psychology, Committee on Aviation Psychology.
- Freeman, F.G., Mikulka, P.J., Prinzl, L.J., & Scerbo, M.W. (1999). Evaluation of an adaptive automation system using three EEG indices with a visual tracking task. *Biological Psychology*, 50, 61-76.
- Funke, M. E., J. S. Warm, G. Matthews, M. Riley, V. Finomore, G. J. Funke, B. Knott, and M. A. Vidulich. (2010). A Comparison of Cerebral Hemovelocity and Blood Oxygen

- Saturation Levels during Vigilance Performance. *Proceedings of the Human Factors and Ergonomics Society* 54: 1345– 1349.
- Gaudin, S. (2013). US military may have 10 robots per soldier by 2023. Retrieved from http://www.computerworld.com/s/article/9244060/U.S._military_may_have_10_robots_per_soldier_by_2023.
- Gevins, A., Smith, M. E., McEvoy, L., & Yu, D. (1997). High-resolution EEG mapping of cortical activation related to working memory: Effects of task difficulty, type of processing, and practice. *Cerebral Cortex*, 7, 374–385.
- Gevins, A., & Smith, M.E. (2003). Neurological measures of cognitive workload during human-computer interaction. *Theoretical Issues in Ergonomic Science*, 4, 1-2, 113-131.
- Gevins, A., & Smith, M. E. (2007). Electroencephalography (EEG) in neuroergonomics. In R. Parasuraman, M. Rizzo, R. Parasuraman, M. Rizzo (Eds.) , *Neuroergonomics: The brain at work* (pp. 15-31). New York, NY, US: Oxford University Press.
- Goldberg, L. R. (1999). A broad-bandwidth, public-domain, personality inventory measuring the lower-level facets of several Five-Factor models. In I. Mervielde, I. J. Deary, F. de Fruyt, & F. Ostendorf (Eds.). *Personality psychology in Europe* (Vol. 7, pp. 7–28). Tilburg: Tilburg University Press
- Gopher, D. & Donchin, E. (1986). Workload – An examination of the concept. In K.R. Boff, L. Kaufman and J.P. Thomas, *Handbook of Perception and Human Performance. Volume 2. Cognitive Processes and Performance*, pp. 41-1:41-49. John Wiley and Sons, Inc.
- Grace, R., & Stewart, S. (2001). The Co-Pilot: A Low-Cost Drowsy Driver and Driver Inattention Monitor. *Proceedings of The International Driving Symposium on Human*

- Factors in Driver Assessment, Training And Vehicle Design*, Aspen, CO, August 14-17, 2001.
- Green, S.B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research*, 26, 449-510.
- Guizzo, E., & Ackerman, E. (2012). How Rethink Robotics built its new Baxter robot worker. IEEE Spectrum. Retrieved January 21, 2014 from <http://spectrum.ieee.org/robotics/industrial-robots/rethink-robotics-baxter-robot-factory-worker>
- Hamilton, J. E., & Hancock, P. A. (1986). Robotics safety: Exclusion guarding for industrial operations. *Journal of Occupational Accidents*, 8, 69–78.
- Hancock, P.A. (1996). Effects of control order, augmented feedback, input device, and practice on tracking performance and perceived workload. *Ergonomics*, 39, 1146-1162.
- Hancock P.A. (2009) *Mind, machine and morality. Toward a philosophy of human-technology symbiosis*. Ashgate Publishing, Surrey.
- Hancock, P.A., & Chignell, M.H. (1987). Adaptive control in human-machine systems. In: P.A. Hancock (Ed.). *Human Factors Psychology*, pp. 305-345. Amsterdam: North-Holland.
- Hancock P.A., Scallen, S.F. (1996). The future of function allocation. *Ergonomics in Design*, 4, 24–29.
- Hancock, P. A., & Warm, J. S. (1989). A dynamic model of stress and sustained attention. *Human Factors*, 31(5), 519–537.
- Hankins, T. C., & Wilson, G. F. (1998). A comparison of heart rate, eye activity, EEG and subjective measures of pilot mental workload during flight. *Aviation Space and Environmental Medicine*, 69(4), 360-367.

- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *HumanMental Workload*. (pp. 139-184). North-Holland: Elsevier Science Publishers.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.
- Heerink, M., Kröse, B., Wielinga, B., & Evers, V. (2006). Studying the acceptance of a robotic agent by elderly users. *International Journal of Assistive Robotics and Mechatronics*, 7(3), 25-35.
- Heger, D., Putze, F. & Schult, F. (2010). Online Workload Recognition from EEG Data during Cognitive Tests and Human-Machine Interaction. *Advances in Artificial Intelligence: Lecture notes in computer science*, 6329, 410-417.
- Helton, W. S., T. Shaw, J. S. Warm, G. Matthews, and P. A. Hancock. (2008). Effects of Warned and Unwarned Demand Transitions on Vigilance Performance and Stress. *Anxiety, Stress & Coping* 21: 173–184.
- Helton, W. S., Warm, J.S., Tripp, L. D., Matthews, G., Parasuraman, R. & Hancock, P.A. (2010). Cerebral Lateralization of Vigilance: A Function of Task Difficulty. *Neuropsychologia* 48: 1683– 1688. doi:10.1016/j.neuropsychologia.2010.02.014.
- Hicks, J. (2014). *Eye Tracking wearables for market research: Tobii Glasses 2*. Retrieved from <http://www.forbes.com/sites/jenniferhicks/2014/05/26/eye-tracking-wearables-for-market-research-tobii-glasses-2/>
- Hilburn, B., Jorna, P.G., Byrne, E.A., & Parasuraman, R. (1997). The effect of adaptive air traffic control allocation on performance of flight-related tasks. In N. Johnston, R. Fuller, & N. McDonald (Eds.), *Aviation psychology: Training and Selection, Proceedings of the*

- 21st Conference of the European Association for Aviation Psychology (EAAP) (Vol. 2, pp. 347-353). Hants, England: Avebury Aviation.
- Hockey, G.R.J. (1997). Compensatory control in the regulation of human performance under stress and high workload; a cognitive-energetical framework.. *Biopsychology*, 45(1-3):73-93.
- Hockey, G.R.J., Gaillard, A.W.K., & Burov, O. (2003). *Operator Functional State: the assessment and prediction of human performance degradation in complex tasks*. NATO ASI Series, Vol. 335.
- Hoffman, R.R., Ford, K.M., Feltovich, P.J., Woods, D.D., Klein, G. & Feltovich, A. (2002). A rose by any other name...would probably be given an acronym. *IEEE Intelligent Systems* 17, 72–80.
- Hoover, A., Singh, A., Fishel-Brown, S., & Muth, E. (2012). Real-time detection of workload changes using heart rate variability. *Biomedical Signal Processing And Control*, 7333-341.
- Hulbert, T. (1989). A comparison of the “NASA-TLX” and “ISA” subjective workload rating techniques. *Internal Report: Cival Aviation Authority Air Traffic Control Evaluation Unit*. Bournemouth, UK.
- Israel, J., Wickens, C., Chesney, G., & Donchin, E. (1980). The event-related brain potential as an index of display monitoring workload. *Human Factors*, 22, 211-224.
- Izzetoglu, K., Bunce, S., Onaral, B., Pourrezaei, K., Chance, B.. (2004). Functional Optical Brain Imaging Using NearInfrared during cognitive tasks. *International Journal of Human-Computer Interaction* 17, 211-227.

- Jensen, T. (2013). *Cross validation*. (PDF document). Retrieved from <http://www.link.ch/wp-content/uploads/2013/10/CrossValidationFinal.pdf>
- Jary, S. (2015). *Fitbit trackers vs. Apple Watch: battle of the fitness smartwatches and activity trackers*. Retrieved from <http://www.pcadvisor.co.uk/feature/wearable-tech/fitbit-trackers-vs-apple-watch-3612954/>
- Jentsch, F., Fincannon, T., Sellers, B., Keebler, J., Ososky, S., Phillips, E., Schuster, D. (2012). *Safe operation of autonomous robot teams: impact of autonomy, teaming, and workload on operator trust and performance* (Task Order 132). Orlando, FL: University of Central Florida.
- Johannes, B. & Gaillard, A.W.K. (2014). A methodology to compensate for individual differences in psychophysiological assessment. *Biopsychology*, 96, 77-85.
- Jordon, C.S. (1992). Experimental study of the effect of an instantaneous self-assessment workload recorder on task performance. *Defense Research agency Technical Memorandum DRA TM (CADS) 92011*, DRA, Portsmouth, Hants.
- Kaber, D. B., & Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science*, 5(2), 113-153.
- Kaber, D., Omal, E. & Endsley, M. (1999). Level of automation effects on telerobot performance and human operator situation awareness and subjective workload. In M. Mouloua & R. Parasuraman, (Eds.), *Automation technology and human performance: Current research and trends*, (pp. 165-170). Mahwah, NJ: Erlbaum.

- Kahneman, D., & Treisman, A. (1984). Changing views of attention and automaticity. In R. Pasuraman & D. A. Davies (Eds.), *Varieties of Attention* (pp. 29–61). New York, NY: Academic Press.
- Kahneman, D., Treisman, A., & Gibbs, B. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, 24, 175–219.
- Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance : A historical review, a meta-analysis , and a preliminary feedback intervention theory. *Psychological Bulletin*, 119, 254–284.
- Kluger, A.N., & DeNisi, A. (1998). Feedback interventions: Toward the understanding of a double-edged sword. *Current Directions in Psychological Science*, 17, 67–72.
- Kirlik, A. (1993). Modeling strategic behavior in human automation interaction: Why an 'aid' can (and should) go unused. *Human Factors*, 35, 221-242.
- Kurimori, S., & Kakizaki, T. (1995). Evaluation of work stress using psychological and physiological measures of mental activity in a paced calculating task. *Industrial Health*, 33(1), 7-22.
- Marshall, S. P. (2002). *Index of cognitive activity: measuring cognitive workload*. Presented at the 7th IEEE Human Factors Meeting, Scottsdale, Arizona.
- Matthews, G., Campbell, S., Falconer, S., Joyner, L., Huggins, J., Gilliland, K., Grier, R., et al. (2002). Fundamental dimensions of subjective state in performance settings: Task engagement, distress, and worry. *Emotion*, 2(4), 315-340.
- Matthews, G., Emo, A. K., Funke, G., Zeidner, M., Roberts, R. D., Costa Jr., P. T., & Schulze, R. (2006). Emotional Intelligence, Personality, and Task-Induced Stress. *Journal Of Experimental Psychology. Applied*, 12(2), 96-107. doi:10.1037/1076-898X.12.2.96

- Matthews, G., & Funke, G.J. (2006). Worry and information-processing. In G.C.L. Davey & A.Wells (Eds.), *Worry and its psychological disorders: Theory, assessment and treatment* (pp. 51-67). Chichester, UK: Wiley.
- Matthews, G., Reinerman-Jones, L., Barber, D., & Abich, J. (2015a). The Psychometrics of Mental Workload: Multiple Measures are Sensitive but Divergent. *Human Factors*, doi: 10.1177/0018720814539505.
- Matthews, G., Reinerman-Jones, L., Wohleber, R., Lin, J.C., Mercado, J., Abich, J. (2015b). Workload is multidimensional, not unitary: what now? In D.D.Schmorrow & C.M. Fidopiastis (Eds.) *Foundations of augmented cognition. HCI International 2015, Proceedings*. Retrieved from http://link.springer.com/chapter/10.1007%2F978-3-319-20816-9_5
- Matthews, G., Szalma, J.L., Panganiban, A.R., Neubauer, C., & Warm, J. (2013). Profiling task stress with the Dundee Stress State Questionnaire. In L. Cavalcanti & S. Azevedo (Eds.) *Psychology of Stress: New research* (pp. 49-90), Hauppauge, NY: Nova Science.
- Maxwell, S.E. & Delaney, H.D. (2004). *Designing experiments and analyzing data: a model comparison perspective*. (2nd ed.) Mahwah, N.J.: Lawrence Erlbaum Associates.
- May, J.G., Kennedy, R.S., Williams, M.C., Dunlap, W.P. & Brannan, J.R. (1990). Eye movement indices of mental workload. *Acta Psychologica*, 75, 75-89.
- Mehkati, N., & Loewenthal, A. (1988). An eclectic and critical review of four primary mental workload assessment methods: A guide for developing a comprehensive model. In P.A. Hancock & N. Meshkati (Eds.) *Human Mental Workload*. (pp. 251-267). Amsterdam: Elsevier.
- Meister, D. (1986). *Human factors testing and evaluation*. Amsterdam: Elsevier.

- Meshkati, N., Hancock, P.A., & Rahimi, M. (1989). Techniques of mental workload assessment. In: J. Wilson (Ed.). *Evaluation of Human Work: Practical Ergonomics Methodology*. London: Taylor and Francis.
- Meshkati, N., & Loewenthal, A. (1988). An eclectic and critical review of four primary mental workload assessment methods: A guide for developing a comprehensive model. In P. A. Hancock, N. Meshkati, P. A. Hancock, N. Meshkati (Eds.), *Human mental workload* (pp. 251-267). Oxford, England: North-Holland. doi:10.1016/S0166-4115(08)62391-2
- Miyata, Y., Tanaka, Y., & Hono, T. (1990). Long term observation on Fm-theta during mental effort. *Neuroscience*, 16, 145–148.
- Moon, J., Betts, S., & Anderson, J. R. (2013). Individual differences and workload effects on strategy adoption in a dynamic task. *Acta Psychologica*, 144, 154-165. doi:10.1016/j.actpsy.2013.05.011
- Moray, N. (1984). Mental Workload. *Proceedings of the 1984 International Conference on Occupational Ergonomics*, Toronto, Canada, 41-46.
- Morrison, J.G., & Gluckman, J.P. (1994). Definitions and prospective guidelines for the application of adaptive automation. In M. Mouloua & R. Parasuraman (Eds.), *Human performance in automated systems: current research and trends* (pp. 256-263). Hillsdale, NJ: Erlbaum.
- Mulder, L.J., De Waard, D., Brookhuis, K.A.. (2005). Estimating Mental Effort Using Heart Rate and Heart Rate Variability. In: Stanton, N., dge, A., Brookhuis, K., Salas, E., Hendrick, H. (eds.), *Handbook of Human Factors and Ergonomics Methods*, pp. 20-1 -- 20-8. CRC Press: Boca Raton, 2005

- Neubauer, C., Matthews, G., Langheim, L., & Saxby, D. (2012). Fatigue and voluntary utilization of automation in simulated driving. *Human Factors*, 54(5), 734-746.
doi:10.1177/0018720811423261
- Nolan, S.A., & Heinzen, T. (2010). Confidence intervals, effect sizes and statistical power. *Essentials of Statistics for the Behavioral Sciences*, (pp. 175-200). New York: Worth Publishers.
- NSF (2013). National Robotic Initiative invests \$38 million in next-generation robotics. *Press Release* 13-179. Retrieved from
http://www.nsf.gov/news/news_summ.jsp?cntn_id=129284.
- O'Donnell, R.D. & Eggemeier, F.T. (1986). Workload assessment methodology. In K.R. Boff, L. Kaufman and J.P. Thomas (Eds.) *Handbook of Perception and Human Performance*. Vol. 2, *Cognitive Processes and Performance*., (pp. 42-1:42-49). New York: John Wiley and Sons.
- Oppermann, R. (1994). *Adaptive User Support*. Lawrence Erlbaum: Hillsdale, NJ.
- Ostrovsky, G. (2008). *Exmocare emotion detection technology: capitalizing on How R U*. Retrieved from <http://www.digitaltrends.com/fitness-apparel-reviews/spire-activity-tracker-review/>
- Parasuraman, R. (2003). Neuroergonomics: research and practice. *Theoretical Issues In Ergonomics Science*, 4(1-2), 5-20.
- Parasuraman, R., Bahri, T., Deaton, J.E., Morrison, J.G., & Barnes, M. (1992). *Theory and design of adaptive automation in aviation systems*. Technical Report No. NAWCADWAR-92033-60. Warminster, PA: Naval Air Warfare Center, Aircraft Division.

- Parasuraman, R. & Hancock, P.A. (2008). Mitigating the adverse effects of workload, stress and fatigue with adaptive automation. In P.A. Hancock and J.L. Szalma (eds.) *Performance Under Stress*. Aldershot, England ; Burlington, VT : Ashgate.
- Parasuraman, R., Mouloua, M., & Molloy, R. (1996). Effects of adaptive task allocation on monitoring of automated systems. *Human Factors*, 38(4), 665–679.
- Parasuraman, R., Mouloua, M., Molloy, R., & Hilburn, B. (1993). Adaptive function allocation reduces performance costs of static automation. *Proceedings of the Seventh International Symposium on Aviation Psychology*. Columbus, OH: The Department of Aviation, The Aviation Psychology Laboratory, The Ohio State University.
- Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39, 230–253
- 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(3), 286–297. doi:10.1109/3468.844354
- Parasuraman, R. & Wickens, C. D. (2008). Humans: Still vital after all these years of automation. *Human Factors*, 50, 511-520.
- Pilkington, E. (2014). *Spire activity tracker review: every breath you take, every move you make, Spire will be watching you*. Retrieved from <http://www.digitaltrends.com/fitness-apparel-reviews/spire-activity-tracker-review/>
- Pomranky, R.A. & Wojciechowski, J.Q., (2007). Determination of mental workload during operation of multiple unmanned systems. *Army Research Laboratory*. U.S. Army Research Laboratory Aberdeen Proving Ground, MD.

- Pope, A. T., Bogart, E. H., & Bartolome, D. (1995). Biocybernetic system evaluates indices of operator engagement. *Biological Psychology*, 40, 187-196.
- Prinzel, L. I., Freeman, F. G., Scerbo, M. W., Mikulka, P. J., & Pope, A. T. (2003). Effects of a psychophysiological system for adaptive automation on performance, workload, and the event-related potential P300 component. *Human Factors*, 4, 601-613.
- Prinzel, L. J., Freeman, F. G., Scerbo, M. W., Mikulka, P. J., & Pope, A. T. (2000). A closed-loop system for examining psychophysiological measures for adaptive task allocation. *The International Journal Of Aviation Psychology*, 10(4), 393-410.
doi:10.1207/S15327108IJAP1004_6
- Reinerman-Jones, L., Barber, D., Lackey, S., & Nicholson, D. (2010). Developing Methods for Utilizing Physiological Measures. In M. Tadeusz, K. Waldemar, & V. Rice (Eds.), *Advances in Understanding Human Performance: Neuroergonomics, Human Factors Design, and Special Populations*. Miami, Florida
- Reinerman-Jones, L., Taylor, G., Sprouse, K., Barber, D., & Hudson, I. (2011). Adaptive Automation as a Task Switching and Task Congruence Challenge. *Proceedings of the 55th Human Factors and Ergonomics Society Annual Meeting*.
doi:10.1177/1071181311551041
- Rensink, R. A. (1997). How much of a scene is seen? The role of attention in scene perception. *Investigative Ophthalmology and Visual Science*, 38.
- Rensink, R. A. (2000). Sensing, seeing, scrutinizing. *Vision Research*, 40, 1469–1487.
- Rensink, R. A. (2002). Change detection. *Annual Review of Psychology*, 53, 245–77.
- Roscoe, A. H. (1993) Heart rate as a psychophysiological measure for in-flight workload assessment, *Ergonomics*, 36:9, 1055-1062. DOI: 10.1080/00140139308967977

- Rouse, W. B. (1977). Human-computer interaction in multi-task situations, *IEEE Transactions on Systems, Man and Cybernetics*, SMC-7, 384–392.
- Rouse, W. B. (1988). Adaptive aiding for human/computer control, *Human Factors*, 30, 431–438.
- Rouse, W.B., & Rouse, S.H. (1983). *A framework for research on adaptive decision aids*. Technical Report AFAMRL-TR-83-082. Wright-Patterson Air Force Base, OH: Air Force Aerospace Medical Research Laboratory.
- Rubio, S., Diaz, E., Martin, J., Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. *Applied Psychology: An International Review*, 53(1), 63-86.
- Sabbatini, R. M. E. (1997). The History of the Electroencephalogram. *Brain and Mind Magazine*. Retrived from <http://www.cerebromente.org.br/n03/tecnologia/historia.htm>.
- Sarver, D. E., Rapport, M. D., Kofler, M. J., Raiker, J. S., & Friedman, L. M. (2015). Hyperactivity in Attention-Deficit/Hyperactivity Disorder (ADHD): Impairing Deficit or Compensatory Behavior? *Journal Of Abnormal Child Psychology*, 43(7), 1219-1232.
- Sassaroli, A., Zheng, F., Hirshfield, L. M., Girouard, A., Solovey, E. T., Jacob, R. J. K., Fantini, S. (2008). Discrimination of mental workload levels in human subjects with functional near-infrared spectroscopy. *Journal of Innovative Optical Health Sciences*, 1, 227-237.
- Satell, G. (2014). Why the Digital Revolution is really just getting started. *Forbes Magazine*. Retrieved from <http://www.forbes.com/sites/gregsatell/2014/04/05/why-the-digital-revolution-is-really-just-getting-started/>

- Sciarini, L. W., & Nicholson, D. (2009). Assessing Cognitive State with Multiple Physiological Measures: A Modular Approach. *Foundations Of Augmented Cognition. Neuroergonomics & Operational Neuroscience*, 533. doi:10.1007/978-3-642-02812-0_62
- Scerbo, M.W., (1996). Theoretical perspectives on adaptive automation. In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications*, pp. 37-63. Mahwah, NJ: Erlbaum.
- Schaefer, K. E. (2013). *The Perception And Measurement Of Human-Robot Trust*. (Dissertation). University of Central Florida.
- Scholtz, J. (2003). Theory and evaluation of human robot interactions, *Proceedings of the 36th Annual Hawaii International Conference on System Sciences*.
- Schulz, C.M., Schneider, E., Fritz, L., Vockeroth, J., Hapfelmeier, A., Wasmaier, M., Kochs, E.F., Schneider, G. (2011). Eye tracking for assessment of workload: a pilot study in an anaesthesia simulator environment, *British Journal of Anaesthesia* 106 (1): 44–50, doi:10.1093/bja/aeq307.
- Service & Entertainment Robots. (n.d.). In *Humanoid Robotics*, Idaho National Laboratory. Retrieved from <https://inlportal.inl.gov/portal/server.pt?open=512&objID=541&PageID=5993&cached=true&mode=2&userID=3338>
- Shaw, T.H., Satterfield, K., Ramirez, R., & Finomore, V. (2013). Using cerebral hemovelocity to measrues workload during a spatialised auditory vigilance task in novice and experienced observers. *Ergonomics*, 56, 8, 1251-1263. doi: 10.1080/00140139.2013.809154
- Sheridan, T. & Parasuraman, R. (2005). Human-automation interaction. *Reviews of human factors and ergonomics*, 1(1), 89-129.

- Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators*. Cambridge, MA: Cambridge University Press
- Simons, D. J. (1996). In sight, out of mind: When object representations fail. *Psychological Science*, 7, 301–305.
- Simons, D. J., & Levin, T. (1997). Change blindness. *Trends in Cognitive Sciences*, 1(7), 261–267.
- Simons, D. J., & Ambinder, M. S. (2005). Theory and consequences. *Current Directions in Psychological Science*, 14(1), 44–48.
- 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, 3, 366-380.
- Son, I. Y., Guhe, M., Gray, W. D., Yazici, B., Schoelles M. J. (2005). Human Performance assessment using fNIR. *SPIE Proceedings*, 5797, 158-169.
- Spector, R.H. (1990). Visual fields. In H.K. Walker, W.D. Hall, & J.W. Hurst (Eds.). *Clinical methods: the history, physical, and laboratory examinations*, (3rd Ed.), Chapter 116, Boston: Butterworths.
- Starkweather, J. (2011). *Cross Validation techniques in R: A brief overview of some methods, packages, and functions for assessing prediction models*. Retrieved from https://www.unt.edu/rss/class/Jon/Benchmarks/CrossValidation1_JDS_May2011.pdf.
- Stroobant, N., & Vingerhoets, G. (2000). Transcranial Doppler ultrasonography monitoring of cerebral hemodynamics during performance of cognitive tasks: A review. *Neurophysiology Review*, 10, 213-231.

- Szalma, J.L., Hancock, G.M., & Hancock, P.A. (2012). Task loading and stress in human-computer interaction: Theoretical frameworks and mitigation strategies. In J.A. Jacko (Ed.), *The human-computer interaction handbook: Fundamentals, evolving technologies, and emerging applications*, 3rd ed. (pp. 55-75). New York: CRC Press, Taylor & Francis Group.
- Szalma, J., Schmidt, T., Teo, G., & Hancock, P. (2014). Vigilance on the move: video game-based measurement of sustained attention. *Ergonomics*, 57(9), 1315-1336.
- Szalma, J. L., & Teo, G.W.L (2012). Spatial and Temporal Task Characteristics as Stress: A Test of the Dynamic Adaptability Theory of Stress, Workload, and Performance. *Acta Psychologica* 139: 471– 485. doi:10.1016/j.actpsy.2011.12.009.
- Szalma, J. L., Warm, J. S., Dember, W. N. & Parsons, K. S. (2006). Training for Vigilance: Using Predictive Power to Evaluate Feedback Effectiveness. *Human Factors* 48 (4): 682–692.
- Tabachnick, B. G., & Fidell, L. S. (2006). *Using multivariate statistics*. (5th ed.) New York, NY: HarperCollins.
- Takayama, L., Ju, W. & Nass, C. (2008). Beyond dirty, dangerous and dull: What everyday people think robots should do. *Proceedings of the Human-Robot Interaction Conference: HRI 2008*, Amsterdam, Netherlands, 25-32
- Tattersall, A. J. & Foord, P. S. (1996). An experimental evaluation of instantaneous self-assessment as a measure of workload. *Ergonomics*, 39(5), 740-748.
- Taylor, G. S. (2012). *Comparing types of adaptive automation within a multi-tasking environment*. Unpublished doctoral dissertation. University of Central Florida.

- Taylor, G., Reinerman-Jones, L., Cosenzo, K., & Nicholson, D. (2010). Comparison of multiple physiological sensors to classify operator state in adaptive automation systems. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 54, No. 3, pp. 195-199). Sage Publications.
- Taylor, G. S., Reinerman-Jones, L. E., Szalma, J. L., Mouloua, M., & Hancock, P. A. (2013). What to automate: Addressing the multidimensionality of cognitive resources through system design. *Journal Of Cognitive Engineering And Decision Making*, 7(4), 311-329. doi:10.1177/1555343413495396
- Tollner-Burngasser, A., Riley, M. A., & Nelson, W. (2010). Individual and team susceptibility to change blindness. *Aviation, Space, And Environmental Medicine*, 81(10), 935-943. Doi:10.3357/ASEM.2809.2010.
- Tripp, L. D., & Warm, J. S. (2007). Transcranial doppler sonography. In R. Parasuraman & M. Rizzo, (Eds.), *Neuroergonomics: The brain at work* (pp. 82-94). New York, NY, US: Oxford University Press.
- Tsang, P.S. & Vidulich, M.A. & (2006). Mental workload and situational awareness. In G. Salvendy (ed.) *Handbook of Human Factors and Ergonomics*, pp. 258. John Wiley & Sons: Hoboken: New Jersey.
- U.S. Army. (2007). *THAAD Theatre High Altitude Area Defense Missile System*, USA. Retrieved November 15, 2014, from www.army-technology.com/projects/thaad/
- Van Orden, K.F., Limbert, W., Makeig, S., Jung, T.P. (2001). Eye Activity Correlates of workload during a Visuospatial Memory Task. *Human Factors*, 43, 1, 111-121.

- Venables, L., & Fairclough, S. (2004). Establishing the Psychophysiological Variables That Can Identify and Predict Operator Subjective State. *Proceedings of the Human Factors & Ergonomics Society Annual Meeting*, 90-94.
- Vidulich, M. A., & Wickens, C. D. (1986). Causes of dissociation between subjective workload measures and performance. *Applied Ergonomics*, 17, 291-296.
- Warm, J. S., Finomore, V., Shaw, T. H., Funke, M. E., Hausen, M. J., Matthews, G., Taylor, P., Vidulich, M. A., Reperger, D. W., Szalma, J. L. & Hancock, P. A. (2009). Effects of Training with knowledge of results on diagnosticity in vigilance performance. *Proceedings of the Human Factors and Ergonomics Society* 53: 1066– 1070.
- Warm, J.S., & Parasuraman, R. (2007). Cerebral hemovelocity and vigilance. In R. Parasuraman & A.M. Rizzo (Eds.), *Neuroergonomics: the brain at work*. Cambridge, MA: MIT Press.
- Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomic Science*, 3(2), 159-177.
- Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors*, 50(3), 449-455.
- Wickens, C. D. (1980). The structure of attentional resources. In R. Nickerson (Ed.), *Attention and Performance VIII* (pp. 239-257). Hillsdale, N.J.: Lawrence Erlbaum. Retrieved from http://books.google.com/books?hl=en&lr=&id=oP-GHfhVX6lEC&oi=fnd&pg=PA239&dq=The+structure+of+attentional+resources&ots=LBuRt1X09f&sig=4Er_zwXXep0tJXqt9Jw9hZvAJGA
- Wickens, C. D. (1984). Processing resources in attention. In R Parasuraman & D. R. Davies (Eds.), *Varieties of Attention* (pp. 63-101). New York, NY: Academic Press. Retrieved from

http://books.google.com/books?hl=en&lr=&id=_KfaN4gKP8MC&oi=fnd&pg=PA3&dq=Processing+resources+in+attention&ots=S_z4Gj-4qw&sig=HDL-lEIrBZ7o_RvpRaad5YbgENGc

- Wickens, C. D., & Hollands, J. G. (2000). *Engineering psychology and human performance*. Upper Saddle River, NJ: Prentice Hall.
- Wiener, E. L. (1988). Cockpit automation. In E. L. Wiener and D. C. Nagel (Eds.), *Human factors in aviation*, pp. 433–459. San Diego, CA: Academic Press.
- Wiener, E.L. & Curry, R.E. (1980). Flight-deck automation: promises and problems. *Ergonomics*, 23, 10, 995-1011.
- Wierwille, W.W., Ellsworth, L.A., Wreggit, S.S., Fairbanks, R.J., & Kirn, C.L. (1994). *Research on Vehicle-Based Driver Status/Performance Monitoring: Development, Validation, And Refinement of Algorithms for Detection of Driver Drowsiness*. Final Report. (DOT HS 808 247). Washington, D.C.: National Highway Traffic Safety Administration.
- Wilson, G.F. & Eggemeier, F.T. (1991). Psychophysiological assessment of workload in multi-task environments. In D.L. Damos (Ed.), *Multiple-task performance* (pp. 329-360). London: Taylor & Francis.
- Wilson, G. F., & O'Donnell, R. D. (1988). Measurement of operator workload with the Neuropsychological Workload Test Battery. In P. A. Hancock, N. Meshkati, P. A. Hancock, N. Meshkati (Eds.), *Human mental workload* (pp. 63-100). Oxford, England: North-Holland. doi:10.1016/S0166-4115(08)62383-3
- Wilson, G. (2003). Pilot workload, operator functional state and adaptive aiding. In G. Hockey, A. Gaillard, O. Burkov (eds.) *Operator Functional State: The assessment and prediction*

- of human performance degradation in complex tasks*. IOS Press, Netherlands: Amsterdam.
- Wilson, G. F., & Eggemeier, F. T. (1991). Physiological measures of workload in multi-task environments. In D. Damos (Ed.), *Multiple task performance* (pp. 329–360). London: Taylor & Francis.
- Wilson, G.F., Russell, C.A., (2003a). Operator functional state classification using multiple psychophysiological features in an air traffic control task. *Human Factors* 45, 381–389.
- Wilson, G.F., Russell, C.A., (2003b). Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. *Human Factors* 45, 635–643.
- Wilson, G.F., Russell, C.A., (2007). Performance enhancement in an uninhabited air vehicle task using psychophysiological determined adaptive aiding. *Human Factors* 49, 1005–1018.
- Wilson, G. F. (1992). Applied use of cardiac and respiratory measures: practical considerations and precautions, *Biological Psychology*, 34, 163-178.
- Yeh, M., & Wickens, C. D. (2001). Display signaling in augmented reality: Effects of cue reliability and image realism on attention allocation and trust calibration. *Human Factors*, 43(3), 355-365. doi:10.1518/001872001775898269
- Yeh, Y., & Wickens, C. D. (1988). Dissociation of performance and subjective measures of workload. *Human Factors*, 30(1), 111-120.
- Zhang, J.H., Peng, X.D., Liu, H., Raish, J., & Wang, R.B. (2013). Classifying human operator functional state based on electrophysiological and performance measures and fuzzy clustering method. *Cognitive Neurodynamics*, doi: 10.1007/s11571-013-9243-3.