A Theoretical Adaptive Autonomy Model: Real-Time Physiological Assessment of Cognitive Workload

Dakota C. Evans

Wright State University

Follow this and additional works at: https://corescholar.libraries.wright.edu/etd_all

Part of the Operations Research, Systems Engineering and Industrial Engineering Commons

Repository Citation
https://corescholar.libraries.wright.edu/etd_all/1429

This Thesis is brought to you for free and open access by the Theses and Dissertations at CORE Scholar. It has been accepted for inclusion in Browse all Theses and Dissertations by an authorized administrator of CORE Scholar. For more information, please contact library-corescholar@wright.edu.
A THEORETICAL ADAPTIVE AUTONOMY MODEL:
REAL-TIME PHYSIOLOGICAL ASSESSMENT
OF COGNITIVE WORKLOAD

A thesis submitted in partial fulfillment of the
requirements for the degree of
Master of Science in Engineering

By

DAKOTA C. EVANS
B.S., Wright State University, 2013

2014
Wright State University
January 1, 2015


__________________________
Mary Fendley, Ph.D.
Thesis Director

__________________________
Thomas N. Hangartner, Ph.D.
Chair, Department of Biomedical, Industrial, & Human Factors Engineering

Committee on Final Examination

__________________________
Mary Fendley, Ph.D.

__________________________
Frank W. Ciarallo, Ph.D.

__________________________
Nasser H. Kashou, Ph.D.

__________________________
Robert E. W. Fyffe, Ph.D.
Vice President for Research and Dean of the Graduate School
ABSTRACT


Increases in modern-day system complexity, has led for a need to improve human performance and the interaction between the two. Three objectives: (1) to investigate physiological measures as indicators of cognitive workload, (2) to assess cognitive workload during human interaction with different autonomy levels, and (3) to develop a theoretical model for an adaptive autonomous system that changes with real-time cognitive workload measures were addressed. This effort seeks to improve human computer interaction by providing the human with the acceptable level of computer automation based on real-time cognitive state. Two experiments involved collection of measures of subject physiology, subjective survey data, and performances measures to assess cognitive workload. The first experiment involved assessment of workload during different task difficulty levels. The second experiment compared workload under different system automation levels. Fixation rate, electromyography measures, and heart rate standard deviation were found to include significant main effects for both experiments.
# TABLE OF CONTENTS

1.0 INTRODUCTION ........................................................................................................ 1
  1.1 Human Factors Engineering Overview ................................................................. 1
  1.2 Technological Issues for Human-Computer Interaction ....................................... 2
  1.3 Utilizing Technology ............................................................................................. 2
  1.4 The General Adaptive Model Details .................................................................. 4
  1.5 Complexity for HCI ............................................................................................. 5
  1.6 Designing for Dynamic System State .................................................................. 7

2.0 THESIS OVERVIEW: THE THREE PHASES: ......................................................... 9
  All phases with associated purpose are illustrated in Figure 2. .................................. 9
  2.1 Phase 1: Physiology Assessment during Varying Task Difficulty ....................... 9
  2.2 Phase 2: Physiology Assessment over Different Autonomy Levels .................... 10
  2.3 Phase 3: Real-Time Adaptive Autonomy Model ............................................... 12

3.0 PHASE 1: REVIEW OF THE LITERATURE ............................................................... 14
  3.1 Knowledge Acquisition Techniques .................................................................. 14
  3.2 Objectivity vs. Subjectivity Regarding Psychophysiological Measures ............. 15
  3.3 Subjective Knowledge Acquisition Techniques (SKATs) ................................. 16
  3.4 Objective Knowledge Acquisition Techniques (OKATs) ................................. 17
  3.5 Psychophysiological Measurements .................................................................. 19
    3.5.1 Ophthalmic Psychophysiological Measures ............................................... 20
    3.5.2 Cardiovascular Psychophysiological Measures ........................................... 23

4.0 PHASE 1 EXPERIMENT: QUICK UNDERSTANDING OF BLOCK EXTRUSION (Q.U.B.E) ............................................................................................................. 28
  4.1 Purpose and Hypotheses: Q.U.B.E ...................................................................... 28
  4.2 Experiment Overview ......................................................................................... 31
  4.3 An Introduction to Q.U.B.E. the Testbed .......................................................... 31
    4.3.1 Red Blocks ................................................................................................. 32
    4.3.2 Blue Blocks ............................................................................................... 34
    4.3.3 Yellow Blocks ............................................................................................ 35
    4.3.4 Multi-Block Puzzles .................................................................................. 37
4.4 Participant Problem Solving ................................................................. 39
4.5 Methodology ...................................................................................... 41
  4.5.2 Stimuli and Apparatus - Captiv .................................................. 43
  4.5.3 Stimuli and Apparatus - Eye Tracker Calibration ....................... 43
  4.5.4 Training for Q.U.B.E. ................................................................. 47
  4.5.5 User Testing for Q.U.B.E. .......................................................... 48
  4.5.6 Posttest Questionnaire .............................................................. 49
  4.5.7 Data Collection .......................................................................... 50
4.6 Analysis and Results Phase 1 ............................................................ 51
  4.6.1 NASA-TLX Workload Rating Analysis ...................................... 53
  4.6.2 Time to Complete Task Analysis ............................................... 56
  4.6.3 Outlier Analysis ......................................................................... 57
  4.6.4 Fixation Rate and Fixation Duration Analysis .............................. 60
  4.6.5 EMG Frontal Lobe Analysis ....................................................... 62
  4.6.6 Pupil Diameter Analysis ............................................................ 64
  4.6.7 EMG Temporal Lobe Analysis ................................................... 65
  4.6.8 Heart Rate Average and HRV Analysis ....................................... 66
  4.6.9 Heart Rate Standard Deviation Analysis ..................................... 68
  4.6.10 Hypothesis 1 Discussion ......................................................... 69
  4.6.11 Hypothesis 2 Analysis and Results .......................................... 74
  4.6.12 Hypothesis 2 Discussion .......................................................... 77
5.0 PHASE 2 EXPERIMENT: ARCANIUM .............................................. 78
5.1 Purpose and Hypotheses: Arcanium ................................................. 78
5.2 Literature Review of Autonomy ....................................................... 80
  5.2.1 Human Error ............................................................................ 81
  5.2.3 Taxonomies of Autonomy ........................................................ 83
  5.2.4 Adaptive and Adaptable Autonomy .......................................... 85
5.3 Experiment Overview ...................................................................... 87
5.4 An Introduction to Arcanium – Test Bed ......................................... 88
5.5 Methodology .................................................................................... 90
5.6 Analysis and Results ..................................................................... 93
5.6.1 NASA-TLX Workload Rating Analysis................................................................. 95
5.6.2 Time to Complete Task Analysis .......................................................................... 96
5.6.3 Outlier Analysis..................................................................................................... 97
5.6.4 Fixation Rate and Fixation Duration Analysis ...................................................... 100
5.6.5 EMG Frontal Lobe Analysis.................................................................................. 102
5.6.6 Heart Rate Standard Deviation Analysis .............................................................. 104
5.6.7 Hypothesis 1 Discussion and Results .................................................................... 106
5.6.8 Hypothesis 2 Discussion and Results .................................................................... 108

6.0 MODEL DEVELOPMENT ......................................................................................... 111

6.1 Real-time Physiological Collection ......................................................................... 111
6.2 The Theoretical Model ............................................................................................ 113
6.3 An Application of the Final Adaptive Autonomy Model ......................................... 118

7.0 APPENDIX ............................................................................................................. 123

8.0 BIBLIOGRAPHY .................................................................................................. 192
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A General Adaptive Model Based on Cognitive State Assessment</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>Thesis Phases with Associated Objective</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>Phase 1 Purpose and Contribution to Thesis</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>Red Block</td>
<td>33</td>
</tr>
<tr>
<td>5</td>
<td>Blue Block</td>
<td>34</td>
</tr>
<tr>
<td>6</td>
<td>Yellow Block</td>
<td>36</td>
</tr>
<tr>
<td>7</td>
<td>Multi-Block Puzzle</td>
<td>38</td>
</tr>
<tr>
<td>8</td>
<td>Backwards Problem Solving</td>
<td>40</td>
</tr>
<tr>
<td>9</td>
<td>Sequential Problem Solving</td>
<td>40</td>
</tr>
<tr>
<td>10</td>
<td>User Testing Station</td>
<td>44</td>
</tr>
<tr>
<td>11</td>
<td>Training Time Distribution for Q.U.B.E.</td>
<td>47</td>
</tr>
<tr>
<td>12</td>
<td>User Testing Completion Time Distribution for Q.U.B.E.</td>
<td>49</td>
</tr>
<tr>
<td>13</td>
<td>Experiment Design: Q.U.B.E. Study</td>
<td>52</td>
</tr>
<tr>
<td>14</td>
<td>NASA-TLX Workload Rating JMP 11.0 Analysis</td>
<td>55</td>
</tr>
<tr>
<td>15</td>
<td>Time to Complete Task JMP 11.0 Analysis</td>
<td>57</td>
</tr>
<tr>
<td>16</td>
<td>Mahalanobis and Jackknife Outlier Analysis</td>
<td>59</td>
</tr>
<tr>
<td>17</td>
<td>Correlation Matrix for all Response Variables</td>
<td>59</td>
</tr>
<tr>
<td>18</td>
<td>Fixation Rate Analysis</td>
<td>61</td>
</tr>
<tr>
<td>19</td>
<td>Fixation Duration Analysis</td>
<td>61</td>
</tr>
<tr>
<td>20</td>
<td>EMG Frontal Lobe Average Analysis</td>
<td>63</td>
</tr>
<tr>
<td>21</td>
<td>EMG Frontal Lobe Standard Deviation Analysis</td>
<td>63</td>
</tr>
<tr>
<td>22</td>
<td>Pupil Diameter Average Analysis</td>
<td>64</td>
</tr>
<tr>
<td>23</td>
<td>Pupil Diameter Standard Deviation Analysis</td>
<td>65</td>
</tr>
<tr>
<td>24</td>
<td>EMG Temporal Lobe Average Analysis</td>
<td>66</td>
</tr>
<tr>
<td>25</td>
<td>EMG Temporal Lobe Standard Deviation Analysis</td>
<td>66</td>
</tr>
<tr>
<td>26</td>
<td>Heart Rate Average Analysis</td>
<td>67</td>
</tr>
<tr>
<td>27</td>
<td>Heart Rate Variability Analysis</td>
<td>67</td>
</tr>
<tr>
<td>28</td>
<td>Heart Rate Standard Deviation Analysis</td>
<td>68</td>
</tr>
</tbody>
</table>
Figure 29: Nonlinear Relationship Between EMG Frontal Lobe Average and Computer Game Experience ................................................................. 73
Figure 30: Nonlinear Relationship Between EMG Frontal Lobe Standard Deviation and Computer Game Experience ................................................................. 73
Figure 31: Reduced Linear Regression Model for NASA-TLX Response .......... 75
Figure 32: Reduced Nonlinear Regression Model for NASA-TLX Response ........ 76
Figure 33: Phase 2 Objective and Contribution to Thesis ............................................... 79
Figure 34: Sheridan and Verplank’s 10 Levels of Automation (1978) ................. 84
Figure 35: Human Processor Model Steps and Automation Types (Image from M.I.T.) 85
Figure 36: Arcanium Interface for Start of Game ................................................................. 90
Figure 37: Experiment Design: Arcanium Study ................................................................. 94
Figure 38: NASA-TLX Workload Rating JMP 11.0 Analysis ................................. 96
Figure 39: Time to Complete Task JMP 11.0 Analysis ................................................. 97
Figure 40: Mahalanobis and Jackknife Outlier Analysis .................................................. 99
Figure 41: Correlation Matrix for all Response ................................................................. 99
Figure 42: Fixation Rate Analysis .................................................................................. 101
Figure 43: Fixation Duration Analysis ............................................................................ 101
Figure 44: EMG Frontal Lobe Average Analysis ............................................................... 103
Figure 45: EMG Frontal Lobe Standard Deviation Analysis ........................................... 103
Figure 46: Heart Rate Standard Deviation Analysis ....................................................... 105
Figure 47: Tested Linear Regression Model for NASA-TLX Response ................... 109
Figure 48: Tested Nonlinear Regression Model for NASA-TLX Response ............. 110
Figure 49: Xbar-s Control Chart Method ....................................................................... 113
Figure 50: Refined Theoretical Adaptive Autonomy Model ................................. 115
Figure 51: Final Adaptive Autonomy Model ................................................................. 117
Figure 52: Real-Time Evaluation of CWL Over Time Example .............................. 120
Figure 53: Example Adaptive Autonomy Model .......................................................... 121
LIST OF EQUATIONS

Equation 1: Normalization of Galvanic Skin Response………………………………..25
Equation 2: Linear Model Trained from Phase 1: Q.U.B.E. Study………………………..119
LIST OF TABLES

Table 1: Identified Physiological CWL Indicators...........................................27
1.0 INTRODUCTION

1.1 Human Factors Engineering Overview

The discipline of human factors engineering endeavors to design adequate robust systems based on human capabilities and limitations (Phillips, Repperger, & Reynolds, 2006). Many human factors works originated from the military domain. Prior to World War II, weapons, aircraft, and other military technological systems were designed with the human as the secondary consideration (Hollands & Wickens, 1999). Other engineering disciplines had progressed such that advanced military technologies were producible, but a need for more effective and usable systems was desired. An approach to this problem was to consider the human using the technology. It was discovered that many aircraft crashes were a result of poor interface designs rather than pilot disregard (Fitts & Jones, 1947). A need for a human centered design approach led to the discipline of human factors engineering as we know it today.

Technological advances are what propel a need for human factors engineering for the analysis, design, and development to optimize system performance (Phillips, et al., 2006). Progression in various disciplines of engineering leads to key problems on how a human can be integrated into the system for using new technology. The goal of this thesis it to investigate an approach for assessing the human by taking advantage of technology and capabilities that would revolutionize the methodology for evaluating
cognitive workload (CWL). This thesis involves an exploratory experimental method for investigating real-time CWL measures using physiological measures as an indicator of CWL.

1.2 Technological Issues for Human-Computer Interaction

A consequence of the advancement in technology is an increase in complexity and an increasing number of human-computer interaction (HCI) problems. Much of the increased complexity is attributed to the addition of information and information displays. This results in an increased degree of monitoring by the human (Rowe, Sibert, & Irvin, 1998, February). As computer hardware advances, more types of information can be accessible. Paradoxically, the advancements in technology can also allow for existing problems to become transparent for engineers and researchers to identify and alleviate. Technologies such as physiological measure collection devices can allow us to understand how people interact with systems or products beyond the standard interview or survey.

1.3 Utilizing Technology

Not only do these advancements provide new ways of solving problems, but technologies such as eye tracking systems and neurological electrical signal collection systems are becoming more economically obtainable (Li, Babcock, & Parkhurst, 2006, Martch; Pfeiffer, Renner, & Pfeiffer-Leßmann, 2014; Debener, Minow, Emkes, Gandras, & Vos, 2012; Rodriguez, Rey, & Alcañiz, 2013). The concept of physiological measurement hardware that is integrated into everyday consumer home products is becoming more realistic.
Research efforts should be directed at model development for HCI systems that involve the flow of information from the human to the computer beyond that of the common keyboard and mouse input. This thesis promotes the idea of information collection alternatives such as human physiology for the betterment of our HCI systems. The main inspiration for this thesis comes from the works of Byrne and Parasuraman. Byrne and Parasuraman (1996) describe two complementary roles that physiological measures play in the development of HCI systems. Firstly, physiological measures may provide product designers with information about the user and the user’s experience with a product that cannot be captured by simply asking the user. This information can help to refine operator modeling to improve overall design of user interfaces. Secondly, real-time information about the operator can be directed towards the computer “thus promoting the development of effective adaptive computational logic” (Byrne & Parasuraman, 1996). The mental state of users can be assessed real-time in order to adapt to the human. The adapting of the system can improve human performance by eliminating human error due to high CWL states.
1.4 The General Adaptive Model Details

The general adaptive model shown in Figure 1 was developed based on Byrne and Parasuraman’s (1996) works, and this model is the main inspiration for all of the work within this thesis. This model describes a general adaptive automation system such that the computer adapts to conform to the limitations and capabilities of the user. This model is comprised of 6 different modules. In the model, human cognitive state exists without being directly measured. Assessment of the human’s cognitive state through peripheral device input evaluates physiology objectively and in real-time. Adaptive artifacts are predefined and integrated into the computer system. These artifacts are designed such that different levels of task load are completed by the system and not by the human. They are also designed to support changes between task load levels. The task load of the computer helps to define the computer’s system state. The computer evaluates the user’s
cognitive state and compares the user’s task load to the computer’s system task load. If the human cognitive state indicates high cognition, the computer changes system task load levels to support the user’s high cognitive state. The distribution of task load between the human and computer would allow for high levels of performance for task completion. The level of task load support that the computer provides to the human would produce low error rate, quick completion times, and high performance descriptions.

Unfortunately, there are two pressing issues for this model to effectively work for human-computer systems. The physiology that measures cognition accurately enough to elicit changes in system state has not yet been identified (Module 2 of Figure 1). Physiology as an indicator of cognitive workload has shown promise in detecting differences among cognitive workload levels and these findings will be discussed. Secondly, an understanding of how humans interact with autonomy and when system state autonomy changes should occur has not yet been well defined (Modules 3 and 4 of Figure 1). This thesis is an attempt to progress knowledge to help investigate these issues. By the end of this thesis contributions for improving this overall model will be identified specifically for modules 2-4 in Figure 1.

1.5 Complexity for HCI

The feature “complexity” is frequently justified using a conglomerate of characteristics of a system. Complexity is invariably task dependent and requires a human’s cognitive element to empirically evaluate complicatedness (Edmonds, 1995). The human element compares the task and the degree of effort required to complete the task to their experiences. This helps them to define how difficult a task is to accomplish.
Some common characteristic of complexity include number of repetitions, disorder, and structural rules (Xing & Manning, 2005). By solely describing complexity as the large number or repetition of elements, the interrelationship and interconnections of the system are overlooked (Edmonds, 1995). Xing and Manning (2005) share an example for which number does not quantify complexity. The authors describe how putting peas in a basket as opposed to putting peas in a half basket has no complexity difference.

Disorder is another characteristic of complexity (Xing & Manning, 2005). If patterns cannot be recognized, or if a system lacks rules or standards a system tends to be described as more complex (Grassberger, 1991). Rules help to define the interrelationships of a system. A prime example is the complexity differences between chess and checkers. Chess involves more controlled game pieces and of those pieces, there are six uniquely different piece movement types. Checkers includes only one piece movement type. All spaces on the board can be occupied in chess and the more rules exist in chess than checkers. In both games the piece movements are dependent on the location of the other pieces on the game board (i.e., the interrelationship of the system).

Complexity is related to task difficulty and the human cognition required for completing the task. When working with the HCI issue of complexity, a tool called automation can be used to decrease human CWL. However, the use of automation should be used sparingly. An excessive misuse of this tool could result in user error or improper user decisions due to highly autonomous systems (Norman, 1990). Low levels of autonomy can also have an effect on performance due to the loss of situation awareness (Norman, 1990).
1.6 Designing for Dynamic System State

Another HCI problem is that computer’s user input requirements can be dynamic with high variability. For example, consider a computer system that is used to monitor and control unmanned aerial vehicles (UAVs). User input requirements can range from no user input for extended periods of time to high stress emergency situations (Endsley & Garland, 2000). Similarly, the operator may experience hours of non-activity where the user’s state can withdraw into fatigue, boredom, or complacency (Endsley, 2012). Thereafter, the operator may have to emerge back into an active alert state to manage an emergency situation.

Byrne and Parasuraman (1996) postulate, for many systems, that a static automation level design results in errors because of the dynamic user input requirements of the computer system over time. This research aims to investigate how an adaptive automated software design would work with an exceedingly dynamic system input requirement for the user. The model that was built for this thesis uses the concept of user physiological measures, which are indicators of CWL, as a catalyst for changing autonomy. A single physiological measure that is assumed to indicate cognition for an adaptive system may be affected by confounding factors, thus indicating high cognition during a low cognitive user state (Wilson, 2002). Therefore, this research effort investigates and models the use of multi-psychophysiological responses as a catalyst for an adaptively automated software design.

Additional review on autonomy including the varying taxonomies of automation will be discussed in more detail. There are three main purposes of this research. The first purpose is to investigate physiological measures as indicators of CWL, the second
purpose is to assess physiological measures during interaction with levels of automation, and the final purpose is to develop a theoretical model that uses multi-psychophysiological measures as a catalyst for adaptive automation.
2.0 THESIS OVERVIEW: THE THREE PHASES:

All phases with associated purpose are illustrated in Figure 2.

2.1 Phase 1: Physiology Assessment during Varying Task Difficulty

This thesis consists of three main phases. Each phase includes a description of the purpose along with their contribution to the overall thesis. The first phase encompasses a review of literature on knowledge acquisition techniques, physiological measures, novices vs. experts, and an experiment that looks at physiological measures as an indicator of CWL. The test-bed for phase 1 was a first-person three-dimensional puzzle game called Q.U.B.E (Quick Understand of Block Extrusion) for which players must navigate through a series of cognitively stimulating levels. NASA-TLX CWL ratings, physiology, and task time were collected for this experiment. Three levels, each varying in difficulty, were used for the study with the hypothesis that each of the physiological measure responses will differ based on CWL (related to task difficulty). Each of the physiological measures used for the study were based on literature that found relationships between CWL and specific physiology. This phase 1 experiment also provides supporting evidence that CWL differences are being detected within the tested environment.
A secondary hypothesis for phase 1 was that physiological measures could account for a significant amount of the variation of the NASA-TLX subjective survey ratings. The data from this study was used as a training set to build a linear regression model to predict the response value of NASA-TLX based on physiological predictors. Although a highly predictive model was not expected, correlations between physiologies and perceived difficulty were identified. This may lead researchers in a direction for developing a better predictive model. Another secondary hypothesis was that CWL differences are associated with level of expertise. This hypothesis was tested by administering a pretest questionnaire that asked about computer game experience. It was expected that more experienced players would yield lower CWL measurements for each of the physiological responses tested. The rationale behind this hypothesis is discussed in the literature review for the phase 1 section of this thesis.

The results of the study showed that for phase 1, NASA-TLX ratings, time to complete run, fixation duration, and fixation rate were statistically different among task difficulty levels. Results also showed that a linear model accounted for 76% of the variation of the NASA-TLX response when considering two independent physiological inputs of EMG frontal lobe standard deviation and heart rate standard deviation. These physiological measures were used in the phase 2 experiment to determine if differences could be detected among automation levels.

2.2 Phase 2: Physiology Assessment over Different Autonomy Levels

The physiological measures that were found to have correlations with difficulty levels in Phase 1 were used for the phase 2 study. Phase 2 includes a review on literature for autonomy. The modification of the software used for the study and an
overview on the different autonomy designs is discussed. Phase 2 included an experiment that evaluated physiological measures during interaction with varying automation levels. This experiment used an open-source real-time strategy (RTS) game called Arcanium. NASA-T LX CWL ratings, physiology, and task time were collected for this experiment. Because of the intricate game modifications that had to be made to support different automation levels (manual to highly autonomous), an open-source game with readily source code was used. The purpose of this study was to understand physiological measures during static automation levels so that a basis for investigating dynamic automation changes could be established. The phase 2 investigation of physiology during interaction with varying autonomy levels is directed towards understanding CWL during these levels. An understanding of CWL during autonomy levels provided a basis for the development of an HCI model for phase 3. The data from this study was used as a testing set for the linear regression model from the phase 1 Q.U.B.E study. The linear regression model was then used on the test data based on discretized sections of time. This technique provides a way for evaluating CWL as a function of time rather than as a measure of workload over the entire run (e.g. as in subjective surveys).

The results of the phase 2 study showed that NASA-T LX ratings, time to complete run, fixation rate, and heart rate standard deviation were statistically different among task automation levels. However, for hypothesis 2 the, the linear model that was created was only able to predict a small amount of the variation for the NASA-T LX response. It was determined that the physiological model that was derived in this thesis is
not universal. The results from phase 2 were used to develop a theoretical model for an adaptive autonomy system for phase 3.

2.3 Phase 3: Real-Time Adaptive Autonomy Model

Phase 3 includes development of a theoretical model for an adaptively autonomous software design. A methodology for accessing or measuring real-time CWL and using that information as a catalyst for changing autonomy levels of the software is addressed. The physiological findings from phase 1 and phase 2 and the design of autonomy conducted in phase 2 were used for the model development in phase 3. The model developed in phase 3 will serve as a basis for future work for providing sufficient automation in software design, based on the psychophysiological measure collection. All phases with associated purpose are illustrated in Figure 2.
**Figure 2: Thesis Phases with Associated Objective**

An illustration of all phases of this thesis to help describe content.
3.0 PHASE 1: REVIEW OF THE LITERATURE

3.1 Knowledge Acquisition Techniques

Subjective knowledge acquisition techniques (SKATs) such as surveys, interviews, and observations are commonly used to assess cognitive workload during tasks (Lehto, Boose, Sharit, & Salvendy, 1992). Primary application of knowledge acquisition techniques include studies involving HCI in the healthcare, military, product testing and development domains (Rivera-Rodriguez & Karsh, 2010; Voskamp & Urban, 2009; Bevan & Curson, 1997). Similarities between domains, especially with the advances in technology, include complex and dynamic tasks with cognition being “situated and shared across multiple agents, objects, and environments” (McNeese, Bautsch, & Narayanan, 1999). A major benefit of SKATs is the simplicity for orchestrating them. Data can be collected easily by administering subjective surveys or simply asking the participant for information during or after a task.

Many SKATs have been found to be a “valuable tool in obtaining insights on broad-based problems involving a large number of users in cognitively complex situations” (McNeese, et al., 1999). Although surveys, interviews, and observations are the most prevalently used techniques for evaluating cognition, the use of
psychophysiological measures as indication of cognitive functionality have also been investigated. The purpose of investigating physiology is so that an objective measure can be used to relate to cognition (Endsley & Garland, 2000).

3.2 Objectivity vs. Subjectivity Regarding Psychophysiological Measures

Psychophysiological measures are affected by a multitude of factors including the psyche such as personality traits, psychodynamic processes, and learned cognition and behavior as well as physical health, fatigue, mood, etc. (Ogborne, 2004). With so many factors that make up a user’s psychophysiology, one would tend to believe that these measures are subjective. Consider this, a patient is admitted into a hospital with an illness. He is evaluated using a heart monitor and it is found that his heart rate has increased drastically. Can one objectively say that his heart rate is an indicator of the patient’s illness?

A user’s pupil diameter during a usability test is an objective unknown diameter. It is factual that a human has a pupil diameter of some quantitative value; this value is not known until we attempt to measure it. By using an eye tracker with a certain accuracy and precision that is intended to be unbiased, an estimate for that unknown diameter can be quantified. The subjectivity arises when the quantified physiological measure is interpreted to represent or indicate some qualitative element or idea such as cognition or physical health. However, if the qualitative change can be observed (e.g., CWL) during the quantitative physiological change (e.g., pupil diameter or heart rate), It is believed that psychophysiology can objectively be used as a measure of CWL.
3.3 Subjective Knowledge Acquisition Techniques (SKATs)

By no means does this research disapprove of the use of orthodox SKATs for evaluating CWL. However, subjective assessment methods obtain information on the user’s perceived cognition which could be biased or influenced. Bertrand and Mullainathan (2001) provide findings that provoke serious doubt that subjective survey results can be used as a dependent variable for user behavior and characteristics. For example, a user may declare that they are very experienced in a specific skill in a pre-test questionnaire, be tested in that skill, perform poorly at that skill, come to realization of their poor performance, and rate or describe cognitive elements of that task unrepresentative of their actual user state. Furthermore, a user could answer a question based on social or moral norms.

Bertrand and Mullainathan (2001) conclude by stating “subject survey results may be useful as explanatory variables, but to be careful in interpretation of the results because the findings may not be causal”. Hence, the results of a pre-test survey indicating low experience for a particular skill as an independent variable for test performance may help explain test performance but may not be the cause of test performance.

Other SKATs such as structured interviews may still have faults (Hoffman, 1987). The correct information may not be obtained by the interviewed because the interviewee may not “speak of some particular subject would be impolitic, impolite, or insensitive, because they do not think to and because the interviewer does not have enough information to inquire into the matter, or because they are not able to” (Becker & Geer, 1957).
One of the most popular subject survey techniques, and the technique used for all experiments in this thesis, is the NASA-TLX CWL measurement. NASA-TLX is a subjective CWL assessment tool that uses a multi-dimensional scale to measure operator performance. The six different scales of workload include mental, physical, temporal, effort, performance, and frustration demand. Mental demand includes perceptual demand and tasks such as looking and searching type tasks. Physical demand includes tactile type tasks such as controlling or physically interacting with a system. Temporal demand consists of time stresses or how much completion pressure is exerted on the user. Effort demand is defined as how difficult or complex the task was for the user. Performance is how confident the user was at completed the task. Lastly, the frustration demand is how they felt while completing a task such as irritated or relaxed. Each of these dimensions are rated by a participant on a 0 to 100 scale, where 0 corresponds to low levels of CWL and 100 corresponds to high levels of workload. This NASA-TLX method has been validated and used in numerous studies (Hart & Staveland, 1988; Rubio, Diaz, Martin, & Puente, 2004; Cao, Chintamani, Pandya, & Ellis, 2009). A raw NASA-TLX score was used for all experiments throughout this thesis. This method involves averaging all workload ratings across all six NASA-TLX workload dimensions.

3.4 Objective Knowledge Acquisition Techniques (OKATs)

This thesis proposes that by using multiple quantitative measures such as user physiological measures in coordination with qualitative surveys, a better representation of user CWL can be assessed. The goal of this research is to identify and understand the relationship between a collection of physiological measures and CWL. This contribution
may help researchers uncover an assessment technique that can better uncover user errors and high cognition during usability testing.

OKATs can be used to assess user performance as an indicator of CWL. Performance measures are usually classified as primary task or secondary task performance. The use of primary and secondary task performance to indicate CWL is under the assumption that humans have limited resources according to the multiple resource theory of Yeh and Wickens (1988). Yeh and Wickens (1988) multiple resource theory proposes that there is a finite capacity of multiple different information processing sources that a human can access at a given instance. When two overlapping tasks require the same resource task performance begins to degrade. The method of assessing performance based cognition is to test a participant during the act of completing a specific task and measure their primary task performance measures. Then, a parallel task is added and performance on both tasks is measured. The performance of the primary task is compared across the two conditions to see how much performance on the primary task degrades in the presence of the secondary task. Task time, response or reaction time, accuracy, and error rate are examples of OKATs that are used for primary task performance when assessing CWL.

This type of OKAT is specifically for research studies with a primary task that is the central focus and a secondary task that may or may not contribute to the overall goal (Wickens, 1981). Secondary tasks can simply be thought of as distractor tasks. Secondary tasks can be further classified into loading or auxiliary tasks. Loading tasks are tasks designed to degrade performance of the primary task because they require consistent attention to the secondary. The performance differences of the primary task
are measured to see how much the addition of the secondary task degrades the primary. Alternatively, in the “auxiliary task”, this approach requires consistent performance of the primary task and a measure of secondary task performance differences is evaluated.

**3.5 Psychophysiological Measurements**

The purpose of this section is to establish a base knowledge structure for the physiology that was collected for the purpose of this thesis. Physiological measures are used to assess the functionality of the major organ systems (Ladd Prosser, C (Ed.), 1991). Physiological measurement disciplines include audiology, cardiovascular, urodynamic, gastrointestinal, respiratory, neurophysiology, and ophthalmic physiology (Gray, 1918). “Psychophysiological measures are physiological measures used to index psychological constructs (e.g., psychological states or processes)” (Blascovich, 2000). Usability testing for physiological measures includes many studies involving neurophysiology, cardiovascular, and ophthalmic measures as indicators of CWL. Some commonly used physiological measures in healthcare are blood sugar, temperature, heart-rate, etc. The importance of physiological measures from a product or HCI system design perspective is that it helps a designer uncover a user’s state. Physiological quantitative data on the user’s state can be linked to complex constructs such as mental workload, fatigue, situation awareness, health, and emotion (Endsley, 1996; Kelly, 2003). The main weak link when considering psychophysiological measures as indicators of cognition is the lack of a deep understanding for performance as it relates to workload. Much research suggests that there is a negative linear relationship between workload and performance (Cassenti & Kelley, 2006). However, recent studies have suggested that a concave nonlinear fit is more representative of the relationship (Rusnock & Geiger, 2014). This
theory adopts the idea that humans need to be engaged or tasked in some way so that performance increases. This theory also promotes the idea that there is a maximum point on the concave curve where a specific workload level results in high performance (Rusnock & Geiger, 2014).

By assessing a user’s physiological state directly, a designer will receive feedback that cannot be expressed by the user. For example, a user is not aware of which part of their brain is functioning or visually how many times they blink during a test run. However, advancements in technology have provided us with tools that can more accurately measure physiological metrics. The next few paragraphs will discuss results of studies involving various physiological measures as indicators of cognition, many of which come from the military aviation domain. Many of the measurements that are discussed were collected from both experiments within this thesis. The physiological measures that were chosen were selected based on the capabilities of the Human Performance and Cognition Laboratory at Wright State University. These measures are pupil diameter measures, fixation duration, fixation rate, electromyography (EMG), heart rate measures, and heart rate variability (HRV).

3.5.1 Ophthalmic Psychophysiological Measures

Traditionally, both paper based and computer based qualitative surveys are used in measuring CWL, but much research promotes various physiological measures such as pupil diameter average, pupil diameter standard deviation, fixation duration, and fixation count as indicators of cognition. An estimated 80% of human obtained information is collected through visual sensory input (Pulat, 1997). This expresses why ocular
physiology is researched so greatly in relation to CWL. This section discusses CWL in relation to pupil diameter, pupil diameter variation, fixation duration, and fixation count.

Both pupil diameter and pupil diameter variance are reliable estimates for cognitive load. Beatty (1982) found that in the presence of a task, human pupils begin to dilate and he named the occurrence “task evoked pupillary response”. Since this discovery, much research has been conducted to investigate this phenomenon. Gao, Li, Cai, and Sun’s (2013) paper on cognitive load modulates describes an experiment involving increasing difficulty of calculations for a human while evaluating their pupil diameter. This experiment concluded that “pupil size is affected by the cognitive load during the arithmetic task” (Gao, et al., 2013). Marshall (2002) described estimating CWL from changes in pupil diameter”. Pomplun & Sunkara (2003) conducted a study with three level of task difficulty. The results shows that as the task complexity increased the pupil diameter increased. Another study found that pupil diameter average was greater during high performance and that pupil diameter average decreased during incorrect responses for an auditory task (Tsai, Virre, Strychacz, Chase, & Jung, 2007).

The diffraction pattern of the retina is a result of pupil size and shape. The airy disk pattern made by the pupil can be described as a bright central ring that decreases in intensity from the center (Rantanen & Goldberg, 1999). The diameter of the pupil is inversely proportional to the angular size of the bright central ring (Radin & Folk, 1982). Ophthalmic research tends to stress the idea that during high mental workload the Airy disk on the retina decrease which results in a dilated pupil (Rantanen & Goldberg, 1999). Gardner et al. (1975) found that pupil dilation is more representative of recalling memory than of mental effort. Another study looked at task evoked pupillary response during an
interactive task similar to that of every day computer tasks. They found that pupil size correlates well with CWL during an interactive task (Iqbal, Xianjun, & Bailey, 2004).

Fixation count and durations are also used to understand user CWL. Longer fixation duration indicates difficulty in extracting information, or it means that the object is more engaging in some way (Just & Carpenter, 1976). Irwin (1991) postulates that “it may thus seem reasonable to assume that fixation location corresponds to the spatial locus of cognitive processing and that fixation or gaze duration corresponds to the duration of cognitive processing of the material located at fixation”. This seems reasonable when defining fixations as a brief pause in saccadic movements over informative regions of interest (Salvucci & Goldberg, 2000, November). The occurrence of multiple fixations in a specific region of an interface may be related to visual tunneling (Tsai, et al., 2007). Visual tunneling shows attention allocation of a human and is related to high CWL (Tsai, et al., 2007). This has led to the methodology of analyzing the number of fixations in an area of interest (Poole & Ball, 2006). A common method for analyzing fixation count per area is a special algorithm called the nearest neighbor index (Clark & Evans, 1954). Large fixation counts are an indication of less efficient search patterns or more human effort for completing a task (Goldbery & Kotval, 1999). Long fixation durations indicate difficulty extracting and storing information (Just & Carpenter, 1976). A fixation is commonly referred to as a brief pause in saccadic movement that exceeds 100 milliseconds (Inhoff & Radach, 1998).

Although it was not considered as a measurement for the purposes of this study, blink rate has also been extensively studied in relation to CWL. Studies have found that high CWL stimulation has resulted in suppressed blink rate (Bauer, Strock, Goldstein,
Stern, & Walrath, 1985; Davis, 1994). Many contribute this to applied focus on the task (Bauer, et al., 1985; Ryu & Myung, 2005). It has also been found that suppressed blink rate is sensitive to visual demands, especially with cockpit displays in aircraft flight (Wilson, 2002). Blink rate data has been reported to range from 8 to 30 blinks per minute during normal flight operations (Veltman, 2002). It has also been reported that more frequent blinks are indicators of fatigue or boredom (Stern, Boyer, & Schroeder, 1994). Research into blink rate as an indicator of CWL appears promising.

It should also be noted that ophthalmic measures are also sensitive to ambient illuminations changes (Kramer, 1991). Therefore, it is important to account for these changes for experimentation when collecting ocular data. For the purposes of the two experiments of this thesis, lighting conditions were kept constant. A primary benefit of using eye tracking is that there is minimal interference with the user. Many eye tracking systems are off body systems that are nonintrusive during HCI.

### 3.5.2 Cardiovascular Psychophysiological Measures

Relationships between the cardiovascular system and CWL have also been studied extensively by researchers. Heart rate is most commonly measured in beats per minute. Many studies have looked into heart rate and heart rate variability (HRV) as it relates to cognition, and results have shown that high heart rates are found during high CWL states (Lenneman, Shelley, & Backs, 2005; Mehler, Reimer, Coughlin, & Dusek, 2009). Unfortunately, the best way to score HRV as an indicator of CWL is not established as well in literature. Results will vary across different methods. One method for scoring heart rate will conclude the differences occur while other methods do not during an actual high CWL state (Roscoe, 1992). One of the simpler approaches to
measure HRV is by calculating the interbeat interval averages, standard deviations, or variances over time or for a given number of beats (Roscoe, 1992). It is mostly found in the literature that as CWL increases HRV decreases up until the task becomes too difficult for the user to complete (Rowe, Sibert, & Irvin, 1998, February; Tettersall & Hockey, 1995). At this point the stresses of not being able to complete the task begin to result in greater HRV.

Although not considered for the purposes of this study galvanic skin response (GSR) is another promising measure for indicating CWL. GSR has been investigated as an indicator of cognitive load since the mid-20th century. Early on researchers believed the changes in GSR may reflect changes in mental activity (Landis & Hunt, 1939). What led to the application of GSR to the finger tips is a study by Van der Merwe and Theron (1947) where they found a positive correlation between rates of change in finger pulse volume and emotional liability. This confirms the link between GSR and emotional activity (Van der Merwe & Theron, 1947). GSR has progressed and a recent study by Tarankar et al. (2013). Their findings indicated that respiratory responses and GSR are correlated to one another. Another breakthrough with GSR was found by Nourbakhsh, Want, Chen, & Calvo (2012). They determine that GSR is correlated to CWL measures. Nourbakhsh, Wang & Chen (2013) furthered their previous study to see if there was any correlation between blink rate, GSR, and CWL. They found that both of these physiological measures are a good indicator of CWL (Nourbakhsh, Wang, & Chen, 2013). High CWL states corresponded to high GSR frequencies and suppressed blink rates. One of the important steps in analyzing GSR is to normalize GSR as a response to account for differences among participants. The following equation can be used to do so.
\[ \text{calibrated}\_\text{feature}(i,j) = \frac{\text{feature}(i,j)}{\frac{1}{m} \sum_{j=1}^{m} \text{feature}(i,j)} \]

**Equation 1: Normalization of Galvanic Skin Response**

i is the index for number of participants, j is the index for the task, and m is the total number of tasks. Used to normalize GSR.

In this equation m is the number of tasks. This equation normalizes the data by dividing each individual data-point by the mean frequency of GSR among all participants. Another study found that GSR can be used as a reliable indicator of CWL (Shi, Ruiz, Taib, Choi, & Chen, 2007, April).

Other measures that are common in the literature on CWL but not considered for the studies of this thesis are respiratory measures. Control of respiration is modulated partly by neural factors from the respiratory cent in the hind-brain (Roscoe, 1992). Respiratory rate is a common method for collecting respiratory measures. Belts are commonly used around the waist or chest to detect inhalation peaks and exhalation valleys. Respiratory rate is how many cycles of full inhalation and exhalation occur per unit time. Mehler et al. (2009) showed that as cognitive task demands increased, respiration rate also increased. Another study by Novak, Mihelj, & Munih (2011) explained that mean respiration rate decreases as CWL increases, but increases again as the challenge becomes too much to handle.

Electromyography (EMG), also known as surface electromyography (SEMG), is collected through sensors placed on the skin surface. One of the benefits of this measure is that there are a multitude of locations that electrodes for EMG can be placed (e.g.,
forearm, neck extensor muscles, temporal or frontal head regions) (Melzer, Benjuya, & Kaplanski, 2001; Laursen, Jensen, Garde, & Jørgensen, 2002). Although this measure has not been researched as extensively as the other discussed measures with regards to CWL, research in this area looks promising. Electromyography is a measure of muscular activity in units of microvolts and it is driven by the electric potential produced by muscular cells (Buchthal, 1957). Melzer et al. (2001) found that electric potential is significantly greater for the elderly for cognitive tasks. There are studies that found electric potential differences within an individual completing tasks at varying difficulty levels (Melzer, et al., 2001). Other CWL studies have been conducted using EMG (Or & Duffy, 2007). Electromyography studies have been used more in the medical and biomedical fields for assessing muscular and neuromuscular disorders (De Luca, 1997).

Table 1 includes all physiological measures reviewed for the purposes of this thesis. The asterisks in the table indicate which physiological measures were collected in the experiments within this thesis. All measures from this table except blink rate, respiratory rate, and galvanic skin response were measured for both experiments in this thesis. The purpose for reviewing other physiological measures that were found to be indicators of CWL was to identify measures that have potential for future research.
Table 1.0: Identified Physiological CWL Indicators: Asterisks denote the use of the physiological measure in phase 1 or phase 2

<table>
<thead>
<tr>
<th>Physiological Measure</th>
<th>CWL Relationship</th>
<th>Supporting Literary Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>*Pupil dilation</td>
<td>Large pupil size or dilated pupils are detected during high CWL</td>
<td>(Gao, Li, Cai, &amp; Sun, 2013)</td>
</tr>
<tr>
<td>*Fixation rate</td>
<td>Large fixation counts are an indication of less efficient search patterns or more human effort for completing a task</td>
<td>(Goldberg &amp; Kutval, 1999)</td>
</tr>
<tr>
<td>*Fixation duration</td>
<td>Long fixation durations indicate difficulty extracting and storing information</td>
<td>(Just &amp; Carpenter, 1976)</td>
</tr>
<tr>
<td>Blink rate</td>
<td>High CWL stimulation has resulted in suppressed blink rate</td>
<td>(Bauer, Strock, Goldstein, Stern, &amp; Walrath, 1985; Davis, 1994)</td>
</tr>
<tr>
<td>*Heart rate</td>
<td>High heart rates are found during high CWL states</td>
<td>(Lenneman, Shelley, &amp; Back, 2005; Mehler, Reimer, Coughlin, &amp; Dusek, 2009)</td>
</tr>
<tr>
<td>*Heart rate variability</td>
<td>As CWL increases HRV begins to decrease up until the task becomes too difficult for the user to complete</td>
<td>(Rowe, Sibert, &amp; Irwin, 1998; Tattersall &amp; Hockey, 1995)</td>
</tr>
<tr>
<td>Respiratory rate</td>
<td>As cognitive task demands increased, respiration rate also increased</td>
<td>(Mehler, Reimer, Coughlin, &amp; Dusek, 2009)</td>
</tr>
<tr>
<td>Galvanic skin response (GSR)</td>
<td>High CWL states corresponded to high GSR frequencies</td>
<td>(Nourbakhsh, Wang &amp; Chen, 2013)</td>
</tr>
<tr>
<td>*Electromyography (EMG)</td>
<td>Electric potential will result in different within an individual completing tasks at various difficulty</td>
<td>(Melzer, Benjuya, &amp; Kaplanski, 2001)</td>
</tr>
</tbody>
</table>
4.0 PHASE 1 EXPERIMENT: QUICK UNDERSTANDING OF BLOCK EXTRUSION (Q.U.B.E)

4.1 Purpose and Hypotheses: Q.U.B.E

The purpose of this exploratory experiment in phase 1 was to understand psychophysiological responses as they relate to task difficulty. The primary hypothesis of this study was that all of the tested psychophysiological measures, time to complete task, and the subjective survey results (i.e., NASA-TLX) used in the study will vary by a statistically significant amount for varying task difficulty as supported by the reviewed literature. It should be noted that at least one of the three types of measurement techniques for CWL (e.g., physiological, performance, and subjective workload measure) were used for this experiment. The experiment will help in determining if the environment, measurement hardware, and test scenario produce CWL results consistent with the reviewed research on psychophysiology.

The findings for phase 1 physiology as an indicator of CWL is an integral component for phase 2, testing physiology during autonomy, and phase 3, building an adaptive autonomy HCI model. It is important to accurately and reliably detect differences in cognition during a controlled test when differences are explicitly expected before testing CWL differences when the workload differences are less apparent (e.g.
Cognition during autonomy levels). This will help confirm that CWL differences are detected when a task is truly more difficult. An exploration of expertise differences for both computer experience and computer game experience was also investigated. If a participant’s CWL variability is low, this could be due to their experience in computer games or computer experiences. If high variability is detected for CWL, a participant may be a novice at computer games or computers.

A secondary hypothesis is that multiple physiological measures can be used as explanatory variables for CWL. This hypothesis was tested by building a model with physiological measures as independent variables to predict CWL ratings from the NASA-TLX subjective survey results. The data from this phase 1 study was used as a training set to build a multiple linear regression model to predict the response value of NASA-TLX based on physiological predictors. The model built in phase 1 was tested on the data from phase 2 to determine how predictive the model is in a different gaming type context (see Figure 4).
Figure 3: Phase 1 Purpose and Contribution to Thesis

Illustration of the scope for phase 1.
4.2 Experiment Overview

This experiment is described in four layers. An examination of the game Q.U.B.E (Quick Understand of Block Extrusion) and the cognitive differences among the game’s levels will be described. Secondly, the design of the experiment, experimental environment, and user testing procedure will be discussed. Thirdly, the results of the study will be reported. Lastly, insight with regards to the finding and recommendations for future research will be discussed.

4.3 An Introduction to Q.U.B.E. the Testbed

For this experiment, a game called Q.U.B.E. was used to test participants on three different levels, each at a different difficulty (easy, medium, and hard). Q.U.B.E is a first-person puzzle game that requires players to navigate through a series of cognitively stimulating levels. Players are dropped into a mysterious cube environment where they advance through the stages of a level by solving a series of puzzles. The game requires players to overcome a variety of challenging puzzles and game physics to achieve success. Q.U.B.E was chosen because of the simplistic but cognitively stimulating nature of the game. Participants must complete the followings cognitive tasks throughout the levels: learn the extrusion patterns of various blocks, store the extrusion patterns within their working memory, relate extrusion patterns of blocks to each other, determine how the blocks can be oriented so that the physics of the game will allow them to continue to the next level, and use the game physics to navigate across the blocks.

This game only requires visual and motor resources from the player. Although there is an audio channel of music during the game, the music does not provide any direct advantage for completing the puzzle. The music was played during all user testing
scenarios but was not of concern in the experiment. The physics of the game allows for the player to jump one block high. In the game there are many white blocks that cannot be interacted with, and many colored blocks that have various extrusion patterns. Although the game is rather extensive, participants were only required to orient red, blue, and yellow blocks for the purposes of this experiment. The functionality of each block is described in detail in the following five sections.

4.3.1 Red Blocks

Red blocks are the first block that users must learn. The red blocks can extrude or detrude a magnitude of three blocks. Figures 5abcd are used to help visualize the first puzzle of the game from a player’s perspective. In Figure 5a the desired direction to complete the puzzle can be seen. In Figure 5b the participant is beginning to interact with the red block. The red block extrudes three blocks high and the participant can only jump one block high, therefore, the participant must detrude the block completely or orient the block one length high in order to move upon the block. Once the user is on the block they can extrude it completely, three blocks high, so that they can jump on the ledge Figure 5c, 5d, and 5e. This will require standing on the block as the block is extruded. Figure 5f is an image of when the participant has jumped on the ledge to the next level while they are looking at puzzle that they just completed.
Figure 4a: Red Block 1
Participant initial observation of puzzle

Figure 4b: Red Block 2
Begins problem solving

Figure 4c: Red Block 3
Extruded block to attempt to solve

Figure 4d: Red Block 4
Realization that puzzle required being on block

Figure 4e: Red Block 5
Orients block to jump on it

Figure 4f: Red Block 6
Solved puzzle
4.3.2 Blue Blocks

Blue blocks are the second block that the players must interact with. These blocks require the user to detrude the block into the ground. When the user orients their character on top of the detruded block the character will be sprung four blocks high. This allows users to jump on ledges similar to that seen in the red block example above. Figure 6a shows a blue block puzzle where the lower block in the image is extruded one block. Figure 6b shows the block after it has been detruded. Once the player steps upon the detruded block they will be sprung to the upper ledge. The final image, figure 6c, shows the player looking at the block that they just jumped from.

Figure 5a: Blue Block 1
Participant initial observation of puzzle

Figure 5b: Blue Block 2
Extruded block to attempt to solve

Figure 5c: Blue Block 3
Solved puzzle
4.3.3 Yellow Blocks

The final blocks that the users were required to interact with were the yellow block. This block’s functionality is similar to a step function. The yellow block type consists of three yellow linked blocks that are deducted into the ground or wall. Whichever block the user interacts with is the block that extrudes the farthest. For example, Figure 7a shows a yellow linked block type protruded into the wall. This example has the yellow blocks linked in the corner however yellow blocks can be linked adjacently. Figure 7b shows the yellow block type when the bottom block is selected, Figure 7c shows the yellow block type when the middle block is selected, and Figure 7d shows the block type when the upper block is selected. In order to solve this puzzle the player must select the upper block (Figure 7d) and climb the blocks so that they can jump to their desired destination.
Figure 6a: Yellow Block 1
Participant initial observation of puzzle

Figure 6b: Yellow Block 2
Extrudes bottom block

Figure 6c: Yellow Block 3
Extrudes middle block

Figure 6d: Yellow Block 4
Extrudes top block
4.3.4 Multi-Block Puzzles

The easiest level of difficulty only requires participants to interact with one block. For the medium and difficult level puzzle the participants must solve multiple block puzzles. This requires the participant to associate the orientation of multiple blocks such that they can physically move upon the blocks to reach the desired destination. Figure 8a shows an example three block puzzle. This puzzle includes two yellow type blocks and one red type block. The participants must observe their destination before determining how to orient the blocks to reach their destination. Figures 8b, 8c, and 8d show the proper extrusion of the blocks so that they can move upon the blocks to reach their destination. Figure 8e shows the puzzle solved after the player has reached their destination.
Figure 7a: Multi-Block Puzzle 1
Participant initial observation of puzzle

Figure 7b: Multi-Block Puzzle 2
Extrudes floor yellow block

Figure 7c: Multi-Block Puzzle 3
Extrudes red block

Figure 7d: Multi-Block Puzzle 4
Extrudes wall yellow block

Figure 7e: Multi-Block Puzzle 5
Solves puzzle
4.4 Participant Problem Solving

Q.U.B.E requires participants to consider many different variables such as how to orient blocks, what types of blocks are present, where the desired destination for finishing the puzzles is located, as well as game physics aspects such as how high and far the character in the game can jump. Participants also had to identify a decision strategy for solving puzzles. For multiple block puzzles, participants could assess their desired destination and orient blocks to reach their destination (Figure 9), or they could orient blocks sequentially (Figure 10). Participants had the options of orienting the blocks before attempting to move upon them or sequentially deciding how to orient blocks as they move upon them as illustrated in Figure 9. An interesting observation was that 44% of participants only approached the puzzles sequentially using trial and error, 39% of participants only strategized and laid out the orientation of blocks before moving upon them, and 17% mixed their strategies between puzzles. This finding was particularly interesting because the method for approaching these puzzles was not mentioned in the training session. Although this finding was not part of the scope of this thesis, research in the field of human decision making is promoted.
Figure 8: Backwards Problem Solving
Solves puzzle by determining destination and assessing elements to reach destination

Figure 9: Sequential Problem Solving
Solves puzzle by interaction with block first before considering destination
4.5 Methodology

In the experiment, participants were tasked to complete three different levels of Q.U.B.E. The three levels of difficulty are defined as easy, medium, and hard. Each level had four puzzles. The types of puzzles define the difficulty differences among levels. The easiest puzzle included single block puzzles similar to what can be seen in all figures for 5, 6, and 7. The medium difficulty includes two types of blocks per puzzle. This requires the player to associate the positioning of two different types of puzzles to determine the best orientation of blocks for completing puzzles. The hard difficulty included three types of blocks or more. The difficulty difference is based on the number of variables that players have to consider when solving the puzzles. The sequence in which each participant completed the puzzles of varying difficulty was randomized according to a Latin square randomization.

NASA-TLX CWL rating and time to complete the levels were the first steps in the analysis of this experiment. This analysis helped to strengthen the idea that the difficulty levels are indeed different for mental effort required to complete the puzzles. The controls of the game included using both the mouse and keyboard. Participants would use the standard W=forward, A=left, S=backward, D=right controls to move their character around the 3-D environment. The space bar was used to control the character to jump one block high. The mouse was used to look around the 3-D environment. Left clicking the mouse would extrude blocks and right clicking would detrude blocks.
4.5.1 Participants

Eighteen participants were tested in this study that was conducted at Wright State University in the Human Performance and Cognition Laboratory. The study was approved through the Wright State University Institutional Review Board (IRB) through expedited review. All subjects were sampled from the Wright State University faculty and student community. The sample included four female (22%) and fourteen male (78%) participants. The age range of the sample included thirteen subjects between the age of 20-29 (72%), two subjects in the 30-39 range (11%), and three subjects in the 40-49 range (17%). One subject was colorblind. However, this participant was tested on identifying the difference among the three colors of blocks and was accurately able to distinguish between them. A description of the experimental procedure for phase 1 can be simplified into six main steps. This includes pre-questionnaire, EMG electrode setup, and heart rate monitor setup, eye-tracker setup, training, user testing, and a posttest-questionnaire after each Q.U.B.E trial.

A pre-questionnaire was administered before testing to obtain demographic information, visual impairment, computer experience, and gaming experience information. The information from the collected study was used as independent variables when evaluating the dependent measure of various physiological measures. This also provides descriptive categorical data to better understand differences in psychophysiological measures between users. The pre-questionnaire can be seen in Appendix A.
4.5.2 Stimuli and Apparatus - Captiv

The Captiv T-Log system was used to collect heart rate and EMG measures. These measures include, EMG frontal lobe averages, EMG frontal lobe standard deviations, EMG temporal lobe averages, EMG temporal lobe standard deviations, heart rate averages, heart rate standard deviations, and HRV. The system is on body and requires placement of electrodes on the medial frontalis and right-unilateral as well as a heart rate monitor that attaches beneath the shirt. All Captiv sensors were placed before the eye tracking calibration began. The heart rate monitor was placed first. Participants were instructed on how to place the heart rate monitor under their shirt verbally and images were provided to ensure that the monitor was correctly placed. The two electrodes for the electromyography data collection were the next placed components. Cleansing wipes were given to the participants to rinse the electrode placement area. The experimenter placed the electrodes and applied mild pressure to ensure the adhesive affixed. The captive calibration took around five minutes on average. CAPTIV hardware samples at 16 Hz for all measures used in the study.

4.5.3 Stimuli and Apparatus - Eye Tracker Calibration

An off body eye-tracking system called Smart Eye Pro was used to collect visual physiological measures from the user. This system uses infrared illumination modules to provoke corneal reflection of the user’s eye. The system also uses infrared cameras to record the user and identify their facial features. By determining the facial features of the user, the Smart Eye Pro system can use participant facial features as a reference point to identify the eyes. The eye-tracking unit can then be used to translate the user gaze into visual scan pattern location on a two-dimensional monitor. The monitor screen is defined
in the three-dimensional world coordinate system of Smart Eye. Smart Eye can export
data on all facial features and even pitch and roll of the head. The Smart Eye system that
was used for this study was a four camera two infrared illuminator system that was
positioned around a 23 in. monitor (Figure 11). The software version that was used for
the study was Smart Eye Pro version 5.10. This system captured at a collection rate of 60
Hz with 0.5 degree typical accuracy.

![Figure 10: User Testing Station](image)

Experiment testing station. Includes mouse, keyboard, speakers, 24 in. monitor, and eye
tracking system.
There were three primary steps for setting up the eye tracking system for each participant. The first step involved creating the world model and defining where the monitor, cameras, and illuminators are located within the world. Although this was not recommended by the manufacturer due to setup time reasons, a world model was created for each participant. It was found that the world model was highly sensitive to the most miniscule actions such as vibrations from closing doors or bumping tables. Therefore, it was decided that the best method for obtaining the best possible collection was to create a world model before each participant began. Even though this is the most protracted task in setting up the Smart Eye eye-tracker, the setup can be completed prior to the participant arriving. The setup time was significantly reduced with standardizing world model creation steps and practice through repetition. This task was reduced to about a ten minute setup.

The second step was the profile creation. This involved using Smart Eye’s facial feature recognition software to collect pictures of the user to form a user profile. It was found that the best results were obtained when users with glasses removed their glasses for this step. All Captiv sensors were placed before the profile creation. Smart Eye Pro 5.10 was able to collect data adequately while having the electromyography sensors on the medial frontalis (EMG(F)) and right-unilateral (EMG(E)) on the orbicularis oculi (e.g., frontal lobe and temporal lobe electrodes). This conforms to Smart Eye’s claim that their system works while users have blockades on their face. For the profile creation the mouse cursor was manually controlled by the experimenter while the user visually followed the cursor on the monitor. The mouse cursor movement was evenly distributed over the monitor to ensure that snapshots were captured while the user locked at all
different locations of the monitor. Once twenty snapshots were recorded the profile creation was complete. 16 of the 18 participant’s profiles were created using the automated snapshot feature with minimal issues. However 2 of the 18 participant’s profiles had to be manually created and all snapshots and facial features were created by the experimenter. One of these participants was color blind and the other was blind in the left eye as indicated by the pretest-questionnaire. After manually calibrating the two participants, the SmartEye system seemed to collect accurately. This is an interesting finding that should be further investigated. However, this was not part of this experiment and will not be addressed further.

The third and final step of the eye-tracker calibration was the gaze calibration. For this step the participants that wore glasses were instructed to put their glasses on for the remainder of the study. Four blue targets that came with the Smart Eye Pro version 5.10 software were used for the gaze calibration step of Smart Eye. During this step users were instructed to look at the targets as they appear. Smart Eye determines where the users left and right eye gaze intersect to estimate where the user is looking at on the monitor. Standard deviation values are reported to the experimenter so that they can assess the variation of the users gaze in comparison to the target. All standard deviation values below one degree were accepted for this experiment. If the standard deviation was above one, the gaze calibration was conducted again until acceptable values were obtained. The profile creation and gaze calibration took around ten minutes per subject. The lighting conditions were kept constant during test runs to prevent effects on pupillary responses.
4.5.4 Training for Q.U.B.E.

All eighteen participants were given a training session for which they were informed on the controls of the game, the functionality of the blocks, and the overall goal of the game. During this time baseline physiological measures were recorded to ensure all hardware was functioning properly. Participants could practice game movements in the training session for as much time as they desired (M=2.59, SD=0.50) (Figure 12). There was one outlier in the training session that took roughly four minutes for training. This participant had very little experience with computer games as indicated in the pre-questionnaire. When the participants agreed that they were ready for testing, the training session ended and testing began. Many participants indicated that the controls of the game were rather simple and this corresponds to the short training times.

![Time Distribution](image)

Figure 11: Training Time Distribution for Q.U.B.E.

Training time distribution for learning Q.U.B.E. Mean of 2.58 minutes and standard deviation of .4957 minutes.
4.5.5 User Testing for Q.U.B.E.

The fourth step involved user testing of the test-bed Q.U.B.E. Prior to testing the participants were reminded verbally that about the controls of the game as well as the objective. The participants were not interacted with directly during the testing session unless they had an issue with the puzzle. If a participant was struggling with a particular stage of a level, they could ask for assistance and they were aided accordingly. The completion of all three levels took around fifteen minutes total to complete on average (M=5.48, SD=0.50 minutes/level) (Figure 13). Each level included four puzzles. The order that participants completed puzzles was randomized using a Latin square design (Appendix B). Only four participants required communication with the experimenter while playing Q.U.B.E. All other participants were able to complete all three puzzles with no interaction.
Figure 12: User Testing Completion Time Distribution for Q.U.B.E.

Testing time distribution for learning Q.U.B.E. Mean of 5.48 minutes and standard deviation of 2.99 minutes

4.5.6 Posttest Questionnaire

After each testing level (3 levels total) the participants completed a NASA-TLX survey that provided information on perceived cognition during testing (Appendix C). The NASA-TLX survey is based on seven point scale. Participants were required to answer questions on the six NASA-TLX dimensions of required; mental demand, physical demand, temporal demand, performance, effort, and frustration. The paper based survey method was used instead of the computer based survey.
4.5.7 Data Collection

The Smart Eye Pro version 5.10 was used to collect the visual data of pupil diameter average, pupil diameter standard deviation, fixation rate, fixation duration. Smart Eye was used concurrently with MAPPS eDx analysis software. MAPPS eDx software was used to record the video screen capture of the user testing system (16 Hz sample rate) while each participant played the three levels of Q.U.B.E. MAPPS main functionality was to sync the eye tracking data from SmartEye eDx with video screen capture and export reports. This allowed for the analyst to review where the participants were looking on the monitor.

MAPPS proprietary algorithm was used to determine fixation duration and fixation count. However, the tool was not versatile enough to produce pupil diameter in a way that could be easily analyzed. MAPPS did export raw data sheets in an .xlsx format. Unfortunately this was a manual process within MAPPS and raw data had to be saved to 18 X 4 = 72 Excel worksheets. VBA in Microsoft Excel was used to develop a pupil diameter analysis tool. The pupil diameter analysis tool was used to eliminate all pupil diameter noise collection above 0.01 meters and below 0.001 meters for the left and right eye. The average and standard deviation pupil size for each of the three Q.U.B.E levels and the baseline was calculated and stored in a final excel sheet. This tool was used to open the raw data workbook, loop through all 72 Excel worksheets and export the data to a final Excel sheet for analysis.

Captiv data collection was successful for sixteen out of eighteen of the participants. The Captiv software crashed during two of the participant’s collections, so this data was not considered for the analysis. Captiv exported raw data sheets that
included heart rate, electromyography frontal potential, and electromyography temporal potential data. VBA was again used to calculate the average and standard deviation of all measures across multiple worksheets.

4.6 Analysis and Results Phase 1

Hypothesis 1 was that all the tested psychophysiological measures, time to complete task, and the subjective survey results (i.e., NASA-TLX) that were used in the study would be found to be statistically significantly different for different task difficulty levels. Individual univariate F-test ANOVAs were conducted for NASA-TLX and time to complete task as dependent variables. Three main effects were considered for each univariate analysis: level of difficulty (easy, medium, difficult), computer experience (1-5), computer game experience (1-5). Level of difficulty was used as an independent variable for this experiment with the purpose of testing the primary hypothesis. Computer and game experience were used as independent variables in the model to determine if participant expertise differences had an effect on all tested responses. Computer and game experience were not controlled factors in the experiment so equal sample sizes did not exist among levels, and the reporting of differences for computer game experience and computer experience should be taken lightly. Game and computer experience results are exploratory with the intention of directing research efforts towards the research area of expertise vs. novices and physiology.
Figure 13: Experiment Design: Q.U.B.E. Study

Illustration of independent and dependent variables for Phase 1 experiment. Dependent variables are in circle and independent variables with associated levels are on top.
This analysis is focused mainly on the controlled independent variable task difficulty level that has equal sample sizes among levels. Subjects were contributed to the error term in the model for each univariate F-test ANOVA. Two-way interactions of the independent variables were not considered due to the lack of degrees of freedom to model the interactions. A standard least squares model personality with the restricted maximum likelihood model (REML) method was used to fit the model using the statistical software package JMP version 11 for Microsoft Windows.

4.6.1 NASA-TLX Workload Rating Analysis

Results showed that the NASA-TLX ratings were statistically significant for trial, $F(2,34) = 6.40, p < .01$, and computer game experience, $F(4,11) = 6.98, p = .01$ (Figure 15). However, video game experience was not statistically significant, $F(2,11) = 2.03, p = .16$. A post-hoc Tukey HSD was used to determine which levels of the statistically significant independent variables were found to be different. This analysis on NASA-TLX ratings revealed that high difficult level (M = 54.97) was significantly different than medium (M = 45.56) and low difficulty (M = 45.42). No statistical difference was found between medium and low difficulty for NASA-TLX ratings.

Computer experience, which was collected from the pretest-questionnaire on a scale from 1 to 5, was also statistically significant, $F(2,11) = 6.98, p = .01$ (Figure 15). Participants rated their computer experience from 3 to 5 and none of the subjects selected 1 or 2 for the pre-test questionnaire. This was not surprising because the sample was collected from college students and faculty that use computers frequently in their everyday lives. A post-hoc Tukey HSD analysis revealed that a computer experience of 3 produced the highest CWL NASA-TLX rating, (M = 70.07) and this was significantly
different than a computer experience of 4 (M = 35.41) and 5 (M = 40.46). The graph data point colors correspond to the task difficulty with purple indicating high, blue indicating medium, and green indicating a low degree of difficulty (Figure 18). All statistical tests were conducted using a 95% confidence level.

All statistical assumptions on normality and equal variance among conditions were validated. A normal Q-Q plot was used to graph the residuals of the model for the response of NASA-TLX (Appendix-D). This was to validate the F-test ANOVA assumption that the residuals of the model followed a normal distribution. The standard eye test confirmed that the residual data does not fall outside of the boundary region and that there is no evidence indicating that the residual data does not follow a normal distribution. The residual by predicted graph was used to determine if any trend could be identified within the data (Appendix-E). A characteristic of the residual by predicted plot is that the residuals from a band around zero and that the range of the residuals are similar across predicted value indicating the variance of the error terms are equal. Another character of the residual by predicted plot is that the data sets around zero confirming that the assumption of no trend is reasonable.

The last assumption that was validated was that variance among the three independent variables was similar (Appendix-F). This was validated by graphing the model residuals by each independent variable. It was found that residuals variables had roughly equivalent ranges among independent variable levels. After testing these assumptions it is believed that the model is a valid model and that the p-value from the F-test ANOVA can be trusted. This method of validating residuals plots was used for all
other statistical tests throughout this thesis, so the details on how it was conducted are only mentioned in this paragraph.

Figure 14: NASA-TLX Workload Rating JMP 11.0 Analysis

Left figure includes JMP statistical output. Right figure shows the NASA-TLX workload rating on the y-axis and trial on the x-axis. Means are plotted with standard error bars.
4.6.2 Time to Complete Task Analysis

The performance measure of time of run was found to be statistically significant for task difficulty, $F(2,34) = 6.53, p < .01$ (Figure 16). However, game experience ($F(4,11) = 1.37, p = .31$) and computer experience ($F(2,11) = .29, p = .76$) were not found to be significant. A post-hoc Tukey HSD analysis found that the time for the difficult level was significantly higher than the easy level. Results also indicated that there was no difference between the difficult and medium level and no difference between the medium and easy level for time. In minutes, the means of the Q.U.B.E levels difficult, medium, and easy were $M = 6.88$, $M = 5.54$, and $M = 4.51$ respectively. All statistical assumptions for the F-test ANOVA were validated for time run of using the same method used for the NASA-TLX analysis (Appendix G-Appendix I).
Figure 15: Time to Complete Task JMP 11.0 Analysis

Left figure includes JMP statistical output. Right figure shows the time of run workload rating on the y-axis and trial on the x-axis. Means are plotted with standard error bars.

4.6.3 Outlier Analysis

The previous analysis of NASA-TLX subjective survey and task time help to confirm that CWL difference exist in the environment as expected from the controlled Q.U.B.E task for task difficulty. This is also one of three main purposes of the phase 1 experiment. The next step was to analyze all of the physiological measures as indicators of CWL. Multivariate methods were used to correlate the physiological measures data to the NASA-TLX rating. The purpose of this multivariate method was to determine outliers in the data by using the Mahalanobis and Jackknife distances. Five outliers were found and removed from the dataset for the analysis. The data points that fell close to the
UCL for the two outlier methods were not removed from the dataset. The multivariate analysis also revealed that some collinearity existed between some of the physiological measures. Fixation duration and fixation rate ($r(54) = 0.82$), EMG frontal lobe and EMG frontal lobe standard deviation ($r(54) = 0.99$), EMG temporal lobe and EMG temporal lobe standard deviation ($r(54) = 0.92$), and heart rate average and heart rate standard deviation ($r(54)=0.98$) were found to have a collinear relationship. A univariate F-test ANOVA was conducted for each of the physiological measure responses for the independent variable task difficulty, computer experience, and computer game experience. The model was originally created by including the two-way interaction between task difficulty and game experience. However, no significance was found for any of the physiological responses and the F-test ANOVA was created only including the three main effects.
Outlier analysis to identify and remove outliers and correlation matrix for all dependent variables in the experiment.
Once the outliers were removed, the analysis of all physiological measures were completed in the same manner. The F-test ANOVA model included three independent variables of trial (easy, medium, and difficult), computer experience (1-5), and game experience (1-5). The model also included subjects in the error term to account for the source of variability due to inherent difference in human physiology. Statistically significant differences were found for heart rate standard deviation, fixation duration, fixation rate, EMG frontal lobe average, and EMG frontal lobe standard deviation for at least one of the three independent variables.

4.6.4 Fixation Rate and Fixation Duration Analysis

Fixation rate \((F(2, 28.85) = 5.77, p < .01)\) and fixation duration \((F(2, 28.82) = 3.70, p = .04)\) were found to be statistically different for task difficulty (Figure 19). A post-hoc Tukey HSD revealed that the easy difficulty level was significantly different from the medium and difficulty levels for both fixation rate and fixation duration. The medium and difficult levels analysis did not indicate a significant statistical difference. The easy difficulty level produced higher fixation rate and fixation durations for the phase 1 experiment. Game experience \((F(4, 10.73) = .57, p = .69)\) and computer experience \((F(2, 10.8) = .39, p = .68)\) were not found to be statistically significant for fixation rate. Game experience \((F(4, 10.37) = .87, p = .51)\) and computer experience \((F(2, 10.57) = 1.08, p = .38)\) was also not found to be statistically significant for fixation duration (Figure 20). All statistical assumptions, as illustrated in the NASA-TLX analysis, were validated and the p-values for the F-test ANOVA for fixation rate and fixation duration are believed to be reliable (Appendix J-Appendix O).
Figure 18: Fixation Rate Analysis

Left figure includes JMP statistical output. Right figure shows the fixation rate on the y-axis and trial on the x-axis. Means are plotted with standard error bars.

Figure 19: Fixation Duration Analysis

Left figure includes JMP statistical output. Right figure shows the fixation duration on the y-axis and trial on the x-axis. Means are plotted with standard error bars.
4.6.5 EMG Frontal Lobe Analysis

EMG frontal lobe average \( (F(4, 9.14) = 5.79, p = .01) \) and EMG frontal lobe standard deviation \( (F(4, 8.94) = 6.80, p = .01) \) were found to be statistically different for game experience (Figure 21). A post-hoc Tukey HSD revealed that a lower game experience corresponds to a higher EMG frontal lobe average. Specifically a participant rating of 1, 2, or 5 are not statistically different. However, a participant rating of 1 was found to be different than a rating of 3 or 4. EMG frontal lobe standard deviation post-hoc Tukey HSD revealed that there is a higher standard deviation for participant game experience ratings of one than all other levels. Task difficulty \( (F(2, 25.46) = 1.30, p = .29) \) and computer experience \( (F(2, 9.12) = 2.93, p = .10) \) were not found to be statistically significant for both EMG frontal lobe average (Figure 22). Task difficulty \( (F(2, 25.20) = .67, p = .52) \) and computer experience \( (F(2, 8.94) = 3.20, p = .09) \) were also not found to be statistically significant for EMG frontal lobe standard deviation. All statistical assumptions, as illustrated in the NASA-TLX analysis, were validated and the p-values for the F-test ANOVA for EMG frontal lobe average and EMG frontal lobe standard deviation are believed to be reliable (Appendix P-Appendix U).
Figure 20: EMG Frontal Lobe Average Analysis

JMP statistical output-  Game experience was found to be statistically significant at a 95% confidence interval for EMG frontal lobe average.

Figure 21: EMG Frontal Lobe Standard Deviation Analysis

JMP statistical output-  Game experience was found to be statistically significant at a 95% confidence interval for EMG frontal lobe standard deviation.
4.6.6 Pupil Diameter Analysis

The physiological measure pupil diameter average, pupil diameter standard deviation, EMG temporal lobe, EMG temporal lobe standard deviation, heart rate averages, and HRV as analyzed using an F-test ANOVA, did not result in significant differences among the three tested independent variables. No statistical significance was found for pupil diameter results for all three tested independent variables: trial difficulty \((F(2, 29.37) = 1.13, p = .34)\), computer experience \((F(2, 11.17) = 1.32, p = .30)\), and game experience \((F(4, 10.98) = 1.05, p = .43)\) (Figure 23). Pupil diameter standard deviations also did not indicate statistically significant differences: trial difficulty \((F(2, 29.17) = .09, p = .91)\), computer experience \((F(2, 11.02) = .45, p = .65)\), and game experience \((F(4, 10.88) = 2.34, p = .12)\) (Figure 24). All statistical assumptions, as illustrated in the NASA-TLX analysis, were validated and the p-values for the F-test ANOVA for pupillometry are believed to be reliable (Appendix V- Appendix AA).

Figure 22: Pupil Diameter Average Analysis

JMP statistical output- No significance was found for pupil diameter average.
Figure 23: Pupil Diameter Standard Deviation Analysis

JMP statistical output- No significance was found for pupil diameter standard deviation.

4.6.7 EMG Temporal Lobe Analysis

No statistical significance was found for EMG temporal lobe average results for all three tested independent variables: trial difficulty \( F(2, 25.15) = .95, p = .40 \), computer experience \( F(2, 9.03) = 2.06, p = .18 \), and game experience \( F(4, 9.02) = 1.15, p = .39 \) (Figure 25). EMG temporal lobe standard deviations also did not indicate statistically significant differences: trial difficulty \( F(2, 25.05) = .04, p = .96 \), computer experience \( F(2, 8.99) = .37, p = .70 \), and game experience \( F(4, 8.98) = .66, p = .63 \) (Figure 26). All statistical assumptions were verified (Appendix AB-Appendix AG).
Figure 24: EMG Temporal Lobe Average Analysis

JMP statistical output - No significance was found for EMG temporal lobe average.

Figure 25: EMG Temporal Lobe Standard Deviation Analysis

JMP statistical output - No significance was found for EMG temporal lobe standard deviation.

4.6.8 Heart Rate Average and HRV Analysis

No statistical significance was found for heart rate average results for all three tested independent variables: trial difficulty \((F(2, 24.88) = 2.75, p = .08)\), computer experience \((F(2, 8.77) = 0.84, p = .47)\), and game experience \((F(4, 8.76) = 2.00, p = .18)\)
HRV also did not indicate statistically significant differences: trial difficulty \((F(2, 24.73) = 2.87, p = .08)\), computer experience \((F(2, 8.57) = .72, p = .52)\), and game experience \((F(4, 8.56) = 2.18, p = .16)\) (Figure 28). All statistical assumptions were verified (Appendix AB-Appendix AH-Appendix AM).

**Figure 26: Heart Rate Average Analysis**

JMP statistical output- No significance was found for heart rate average.

**Figure 27: Heart Rate Variability Analysis**

JMP statistical output- No significance was found for heart rate variability.
4.6.9 Heart Rate Standard Deviation Analysis

Heart rate standard deviation showed significant statistical differences for both computer experience ($F(4, 8.20) = 10.34, p < .01$) and game experience ($F(4, 8.41) = 6.12, p = .01$) (Figure 29). A post-hoc Tukey HSD analysis found that higher heart rate standard deviations were found with those with less computer experience and less gaming experience. All statistical assumptions, as illustrated in the NASA-TLX analysis, were validated and the p-values for the F-test ANOVA for heart rate standard deviation are believed to be reliable (Appendix AN-Appendix AP).

**Figure 28: Heart Rate Standard Deviation Analysis**

JMP statistical output- Game experience and computer experience were found to be statistically significant at a 95% confidence interval for heart rate standard deviation.
4.6.10 Hypothesis 1 Discussion

This section includes a summary of the results for the phase 1 experiment hypothesis 1 and a discussion on the findings. The primary hypothesis of this study was that all of the tested psychophysiological measures, time to complete run, and the subjective survey results would be found to be significantly different for varying task difficulty as supported by the literature reviewed. The results of the study showed that NASA-TLX ratings, time to complete run, fixation duration, and fixation rate were statistically different among task difficulty levels. However, time to complete task was the only measure that indicated statistically significant differences among all three levels. The most difficult level took the longest time to complete. Differences in the NASA-TLX rating for the easy and medium level tests were not detected. However, these levels were statistically different than the hardest level of Q.U.B.E which corresponded to the highest CWL ratings.

Fixation rate and fixation duration analysis did not detect statistically significant differences between difficult and medium. Differences were statistically different for the easy level when compared to the hard and medium difficulty levels. For fixation rate and fixation duration it was found that the highest fixation rates and longest fixation durations were found at the easy level. This finding is contrary to the findings that were reviewed in the research. Research suggests that high CWL states elicit more frequent fixation rates and longer fixation durations. An explanation of this contrary finding may be related to the task itself. Q.U.B.E is a first person puzzle game that requires the participant to relate blocks to each other so that they can navigate through the physical environment. The finding of higher fixation rates and fixation durations during easier
difficulty tasks may suggest that the presence of providing users with one variable that can be manipulated may elicit longer fixation duration views on that variable more often. When a user is provided with more than one variable that can be manipulated, users fixate on them less often and for shorter periods of time.

EMG frontal lobe averages were found to be statistically different for computer game experience. The relationship between computer experience and electromyography average seems to have a nonlinear relation (Figure 109). Low experience users with the rating of 1 or 2 and very high experience users with rating of 5 have significantly higher EMG averages than users with the rating of 3 or 4. EMG is the measure of muscular activity in the units of microvolts and it is the electric potential produced by muscular cells (Buchthal, 1957). When placing an electrode on a user’s skin outside of their frontal lobe, the muscular activity can be observed when the user furrows their brow in thought. This nonlinear relationship finding from this experiment suggests that less experienced users and very high experienced users furrow their brow more often than users that are of average or slightly higher than average users. Less experienced users’ furrow their brow because of their thought processes for a task that they have little or no experience with. This could indicate a sense of confusion or struggle with a task. More experienced users’ furrow their brow possibly because of the process of relating their game experiences to the Q.U.B.E game. This could indicate a participant is applying much effort at a task that they are familiar with. It is hypothesized for future research that a nonlinear relationship exists for electromyography averages and expertise and further research on electromyography studies is promoted.
EMG frontal lobe standard deviation was found to be statistically significant for computer game experience with a low game experience of 1 indicating significantly higher measures. No difference was found between all other game experience levels. The implications of this finding are that EMG frontal lobe standard deviation measures have much more variability with individuals that have very little to no experience at a task. Although all game experiences levels besides one were not statistically significant, the nonlinear convex relationship, similar to that of the EMG frontal lobe averages, seems to exist. User computer game experience ratings of 5 seem to curve upward from a computer game experience of a lesser rating. It is duly noted that computer game experience and computer experience were not controlled in the experiment and that sample sizes for each level of these factors are not equal across. The results that are reported for differences in computer experience and computer game experience are not to conclude that differences exist but are an exploratory finding that may help direct future research on physiological differences between human experiences, skill level, and expertise vs. novices.

Heart rate standard deviation was found to be significantly different for different levels of computer experience. Significantly higher heart rate standard deviations were found at a computer experience level of 3 than at levels 4 or 5. All participants’ computer experience ratings were 3, 4, or 5 for the study and ratings of 1 and 2 did not exist. The implications of these findings are that lower experience with a type of task may elicit higher standard deviations in heart rate. Heart beat standard deviation is much more stable when subjects are exposed to a task with which they are comfortable.
All physiological measures were not statistically different among task difficulty levels except fixation duration and fixation rate. There are few interpretations of this finding from the Q.U.B.E study. A simple interpretation is that physiological measures are not as sensitive to CWL differences as NASA-TLX ratings or performance measures such as time to complete run. This interpretation would support skepticism of physiology indicating CWL at all. Another interpretation is that physiological differences are not being detected for the type of game and the types of task that the participants completed when playing Q.U.B.E. This would support the idea that CWL differences can be detected using physiology but the physiology that indicates CWL is task dependent. For instance, studies that have found pupil diameter increases as arithmetic becomes more difficult (Gao, Li, Cai, and Sun, 2013), or general computer use pupil dilation differences are a different type of task than playing a first person 3-D puzzle game. This interpretation would support the idea that physiological differences that indicate CWL differences exist for some types of tasks but not others. Regardless of the supported interpretation of the finding that you prescribe to, more research on relationships between CWL and physiological measures needs to be conducted.

The results for hypothesis one help to address the purpose of phase 1 of this thesis to test physiological measures as indicators of cognitions during varying task difficulty and evaluate expertise as a factor for CWL difference. Physiology measures fixation duration, fixation rate, EMG frontal lobe standard deviation, EMG frontal lobe averages, and heart rate standard deviation were identified as having significance during task difficulty levels of the Q.U.B.E. These measures were used in the phase 2 experiment.
Figure 29: Nonlinear Relationship Between EMG Frontal Lobe Average and Computer Game Experience

Figure 30: Nonlinear Relationship Between EMG Frontal Lobe Standard Deviation and Computer Game Experience
4.6.11 Hypothesis 2 Analysis and Results

Hypothesis two for the phase 1 Q.U.B.E experiment was that a model with physiological input variables can be used to describe much of the variation for the response of NASA-TLX CWL rating. This ideology suggests that subjective surveys, a generally excepted method for evaluation CWL, and physiology have some correlation. By testing this hypothesis it is anticipated that a set of psychophysiological measures can be identified that account for a significant amount of the variability of the NASA-TLX rating response. The importance of this type of research is to develop a robust model based on human physiology to predict CWL. If a robust model could be developed, this would revolutionize the method for researchers to collect CWL information from their participants. Studies could involve physiological measures collection instead of subjective surveys and this would provide real-time CWL measures.

The purpose of this hypothesis is to identify potential psychophysiological measures that can be used for model fitting for future research. The data from the Q.U.B.E study was used to train two different types of models. The first model is a multiple-linear regression model. The second model includes quadratic terms for all physiology. A backwards elimination method was used to reduce both models. This methodology included adding all physiology collected in the study as independent variables in the model and reducing it by removing all insignificant effects as found using JMP 11 software. The model also included participants in the error term to account for inherent differences among human physiology.
The linear model method reduced to two significant main effects that accounted for 76% of the variation of the NASA-TLX workload response (Figure 32). The two main effects included EMG frontal lobe standard deviation and heart rate standard deviation. In general, higher standard deviation for both EMG frontal lobe and heart rate corresponded to higher CWL ratings for the subject survey. The residual were analyzed to ensure that statistical biases did not exist (Appendix A0-Appendix A8). All residuals appeared to be normal and no patterns were identified across independent variables.

**Figure 31: Reduced Linear Regression Model for NASA-TLX Response**

Linear model created using two main effects of EMG frontal lobe standard deviation and heart rate standard deviation. Accounts for 76% of the variation of the NASA-TLX workload response

The quadratic model method included adding all quadratic terms to the model for each of the physiological measures data that were collected and reducing the model by removing insignificant terms. The reduced model included three quadratic term significant effects of EMG temple average, EMG temple standard deviation, and heart rate average and three significant main effects of pupil diameter average, EMG forehead
standard deviation, and heart rate standard deviation. The main effect terms for EMG temple average, EMG temple standard deviation, and heart rate average were included in the model because this was required to use the quadratic terms in the model. This model accounted for 86% of the variation for the NASA-TLX workload response (Figure 33). This is a 10% increase from only considering linear terms in the model. The residual were analyzed to ensure that statistical biases did not exist. All residuals appeared to be normal and no patterns were identified across independent variables (Appendix AT-Appendix AV).

![Response Workload](image)

**Figure 32: Reduced Nonlinear Regression Model for NASA-TLX Response**

Nonlinear model created using many different main effects and two-way interactions. Accounts for 83% of the variation of the NASA-TLX workload response.
4.6.12 Hypothesis 2 Discussion

This section includes a summary of the results for the phase 1 experiment hypothesis two and a discussion on the findings. The secondary hypothesis of this study was that that a model with physiological input variables can be used to describe much of the variation for the response of NASA-TLX CWL rating. This hypothesis was tested by training both a linear and nonlinear model to determine which physiological measures were found to be significant for the model. Results showed that the linear model accounted for 76% of the variation of the NASA-TLX response when considering two independent physiological inputs of EMG frontal lobe standard deviation and heart rate standard deviation. Further analysis showed that EMG frontal lobe standard deviation and heart rate standard deviations were positively correlated with NASA-TLX CWL ratings. This is consistent with the research reviewed on these measures as they related to CWL. The nonlinear model accounted for 86% of the total variation for the response of NASA-TLX. Three quadratic terms of EMG temple average, EMG temple standard deviation, and heart rate average along with an additional main effect of pupil diameter average were added to the model. 86% is a much larger source of variation than expected. This finding addresses the third purpose of the phase 1 experiment. The linear and nonlinear models trained in phase 1 were tested on the phase 2 Arcanum physiology data to see if any correlations translate to a different gaming context.
5.0 PHASE 2 EXPERIMENT: ARCANIUM

5.1 Purpose and Hypotheses: Arcanium

Phase 2 of this thesis can best be described in three parts. The first part includes a literature review on automation. This review helps to give insight into levels, types, and applications of autonomy by illustrating other research efforts. The second part includes an introduction of the game Arcanium that was used for this experiment. This part will give insight into the purposes of the experiment and the methodology used for the study. The third and final part will denote the findings with a discussion on physiology as an indicator of CWL and a catalyst for an adaptive interface. This part includes comparisons made between the experiment in phase 2 and the Q.U.B.E study.

The main objective of phase 2 was to investigate the physiology of fixation duration, fixation rate, EMG frontal lobe standard deviation, EMG frontal lobe averages, and heart rate standard deviation, as found as significant in phase 1, during varying static autonomy levels (Figure 34). This method helped to determine if CWL differences can be found when interacting with different degrees of autonomy, and provided information about developing a theoretical model for an adaptive HCI system. The findings of this experiment will also help to confirm that CWL is different during interaction with different levels of autonomy.
The primary hypothesis of phase 2 is that CWL is different when interacting with varying autonomy levels. Physiology, NASA-TLX subjective surveys, and time to complete task performance measures were collected during participant interaction with a freeresource real-time strategy game called Arcanium. This experiment objective was to assess CWL during participant interaction when less apparent difficulty differences existed among trials. Participants interacted with three levels of autonomy, one level for each trial. Expertise was also investigated as a factor that effects CWL similar to that of phase 1. The secondary hypothesis was that the trained model from phase 1 can be tested on the phase 2 physiological and survey data and some commonalities in both gaming contexts can be identified (Figure 34). This hypothesis will provide insight into reasonable research directions for physiology as an indicator of CWL and modeling human CWL.

**Figure 33: Phase 2 Objective and Contribution to Thesis**

Illustration of the scope for phase 2.
5.2 Literature Review of Autonomy

It is best to discuss automation by illustrating a disastrous accident due to automated issues. A prime example of autonomy issues is that of The Three Mile Island incident on March 28, 1978 in Duaphin County, Pennsylvania (Robertson, 1980). A nuclear meltdown occurred and released radioactive gases and iodine into the environment. Studies have shown that this incident has been linked to many cancer cases of residents near the nuclear power plant (Hatch, Wallenstein, Beyea, Nieves, & Susser, 1991; Talbott, et al., 2000). The incident occurred due to the lack of sufficient information provided to the operators from the computer about the coolant levels of the reactor and the high-stress user state of the operators (Collins, Baum, & Singer, 1983). The operators believed the relief valve was closed and that the water levels were too high based on their information displays, when in actuality the valve was stuck in the open position (Robertson, 1980). This occurrence has been described as a confirmation bias of the operators (Bowen, Castanias, & Daley, 1983). A confirmation bias is when people seek out information to support a particular belief and discount information that does not conform to their belief. The Three Mile incident occurred due to the lack of sufficient information acquisition by the computer (the four types of information will be discussed in a later section) (Parasuraman, Sheridan, & Wickens, 2000). Because the system did not obtain sufficient information from the reactor sensors it could not display the correct information to the operators.

Autonomy can be defined as machines carrying out functions that the human does not wish to perform or cannot perform as accurately or reliably (Parasuraman et al., 2000). Some common example functions include calculations, data mining, and repetitive
tasks. In industrial applications automation can be used to increase productivity, reduce human errors, and prevent cumulative traumas for employees (Malotke, 1985; Kroemer, 1989; Sheridan, 2008). Another perspective of autonomy looks at the elimination of the human task element and an increase of computer completed tasks (Billings, 1997). This perspective is sometimes viewed negatively because of the fear that machines or computers will replace human jobs. Automation in the medical field has advanced and improved diagnostics decision support system and highly accurate robotic surgery technologies exist (Kwoh, Hou, Jonckheere, & Hayati, 1988). With these advanced technologies some believe it is reasonable to fear automation (Everson & Tobias, 1978, March). Regardless of individual opinion about automation, a few things are undeniable; (1) automation exists in our current systems to some degree, (2) using automation in HCI design incorrectly can result in human error with catastrophic outcomes (e.g., The Mile Island), (3) as new technology is developed and complexity increases, automation issues arise.

5.2.1 Human Error

One of the common objectives of designing a HCI system is to minimize errors. The consequences of committing an error vary drastically between different HCI systems. A dire consequence could be substantial financial loss or physical harm to a user or another individual. A more mild consequence could be minimal loss of the user’s time. Weigmann and Shappell (2011) define human errors as “the mental or physical activities of individuals that fail to achieve their intended outcome”. They categorize errors as three basic error types; decision, skill-based, and perceptual errors (Wiegmann & Shappell, 2001). Decision errors are classified as either procedural errors, poor choices, or problem
solving errors (Wiegmann & Shappell, 2001). Procedural errors are those errors due to standardized sequential steps that are supposed to be followed by the user but are incorrect or inadequate for the given task (Orasanu, 1993). When the user follows the procedure as specified they commit an error. Poor decision choices are those choices made due to time constraints or outside pressures that result in errors (Orasanu, 1993).

Problem solving errors are errors that occur because the user is outside of their element or has not experienced a given situation before (Wiegmann & Shappell, 2001). Skill-based error involved memory and attention issues. In many situations users must obtain large amounts of information through sensory input and store various types of information in their working memory and these types of errors occur due to cognitive overload.

Perceptual errors occur when the senses have been degraded. This type of error also involves sensory input and an example of this type of error is misjudgment due to limited visibility at night (Wiegmann & Shappell, 2001).

Another way to classify human error is (1) errors related to learning or adaptation, (2) interference among competing cognitive control structures, (3) lack of resources, and (4) intrinsic human variability (Rasmussen & Vicente, 1989). Errors related to learning or adaptations are errors that occur due to mismatches in the human cognitive model to the computer design. For example, slight changes in new versions of Microsoft Word may lead to the user committing an error. Interferences among competing cognitive control structures involve balancing attention among multiple resources which may result in human error. Lack of resources refers to time constraints or lack of knowledge that results in an error. Intrinsic human variability refers to the variability in recall of data to make decisions or behavior changes which result in errors.
When considering the human and computer as separate elements that interact to accomplish a goal, the potential for “clumsy automation” persists (Wiener, 1989). By looking at the humans’ demands and the computer’s required input to achieve a goal separately; a system design with many human-computer mismatches will likely be developed. Sarter and Woods (1994) postulate that a better approach for human-computer interaction system design is to focus on the human and computer as one element that concurrently work to achieve a goal (e.g., A joint cognitive system (JCS)). Furthermore, the autonomy that exists within the JCS should function at a level that conforms to support human decision making and properly provides the human with accurate information to base decisions (Sarter & Woods, 1994). If the computer does not provide a desirable level of automation, automation surprises can occur.

5.2.3 Taxonomies of Autonomy

The first automation level taxonomy was originally classified into ten different levels ranging from fully manual to fully automated (Sheridan & Verplank, 1978) (Figure 35). This taxonomy described automation as more than just “all or nothing” (Sheridan & Verplank, 1978). Instead, the automation of a computer or machine can be designed in multiple ways. The highest level of automation can be described as an artificial intelligence. Artificial intelligence can be characterized as things that are done by computers that if done by a human would be considered intelligent (Brooks, 1991). An example of this would be computer rational decision making. The lowest level is described as “no computer assistance with the human making all decisions and actions” (Sheridan & Verplank, 1978).
LEVELS OF AUTOMATION

High - 10. The computer decides everything, acts autonomously, ignoring the human.
   9. The computer informs the human only if the computer decides to.
   8. The computer informs the human only if asked to.
   7. The computer executes automatically, then always informs the human.
   6. The computer allows the human a restricted time to veto before automatic execution.
   5. The computer executes a suggestion if the human approves.
   4. The computer suggests one alternative.
   3. The computer narrows the selection down to a few.
   2. The computer offers a complete set of decision/action alternatives

Low - 1. The computer offers no assistance human must make all decisions and actions.

Figure 34: Sheridan and Verplank’s 10 Levels of Automation (1978)

10 Levels of autonomy ranging from 1 (manual) to 10 (highly autonomous).

Since the original taxonomy, other taxonomies have been developed based on Sheridan and Verplank’s work. Later works expanded on the original taxonomy to include four different types of automation (information acquisition, information analysis, decision making, and action implementation) (Parasuraman et al., 2000). Each of these types of automation has four human processing states (sensory processing, perception/working memory, decision making, and response selection) (Parasuraman et al., 2000) (Figure 36). Endsley, Omal, and Kaber (1999) developed a taxonomy for automation that looked at the allocation of function between the computer and the operator. Other taxonomies have been formulated for various domains (Riley, 1989). As new technology is developed automation will need to be better understood to support
human interaction. The purpose of developing these technologies is to understand how humans interact with autonomy and therefore these works are very important for future advancements of technology.

![Human Processor Model Steps and Automation Types](Image from M.I.T.)

**Figure 35: Human Processor Model Steps and Automation Types (Image from M.I.T.)**

Types of automation from both the human and computers perspective.

### 5.2.4 Adaptive and Adaptable Autonomy

Automation is characterized as either adaptable or adaptive. The first type, adaptable automation systems, is defined as, “A system in which the flexible control of information or system performance automation resides in the hands of the user” (Miller, Funk, Wu, Goldman, & Meisner, 2005). Characteristics of an adaptable system are any feature, layout, or setting that can be changed or adjusted by the user, at the user’s discretion. An example of adaptable automation is any setting which the user can change...
to conform to user preferences, such as the layout of Microsoft Word. Any software design that allows the user to change settings is considered adaptable.

Automation can also be characterized as adaptive. Adaptive automation systems are defined as “a system for which the flexibility in information or automation behavior is controlled by the system” (Miller et. al, 2005, September). The functionality of adaptive automation is that it “can adjust its method of operation based on changing situational demands” (Scerbo, 1996). A simple example of this is the automated call center. For instance, consider a customer calling their bank about an issue. They would be asked by an automated agent to verbalize their issue. If the agent cannot recognize what word was said it will require the customer to enter a number associated with common issues on the dial pad. If there is not a response on the dial pad the automated agent will connect the customer to a human agent. This example illustrates adaptive automation based on user demand.

As best said by Kaber, Tan, Riley, Kheng-Wooi and Endsley (2001), “cognitive overload may occur when operators must perform complex, or a large number of tasks, under low levels of system automation (i.e., complete manual control). High workload can lead directly to low levels of self-awareness and task performance, as operators struggle to keep up with the dynamically changing system”. In a HCI both agents work together to complete a task. If the automation is low and the user is required to do more, especially with complex tasks, the user can experience levels of high mental workload which could lead to human error. By implementing more automation to assist with user struggles, user productivity can increase.
Although automation can increase productivity, it can also have a negative effect on user performance when applied improperly or in excess. Low levels of workload can degrade self-awareness and user performance due to boredom and complacency (Rodgers, Mogford, & Strauch, 2000). Because of the susceptibility of user error occurrence when applying too much or too little automation in software design, it is apparent why integration of a sufficient amount is important to enhance performance.

5.3 Experiment Overview

This experiment is described in four layers. An examination of the game Arcanium including the different autonomy version designs that were built for the game will be described. Secondly, the design of the experiment, experimental environment, and user testing procedure will be discussed. Thirdly, the results of the study will be reported. Lastly, insight with regards to the finding, contributions to model for phase 3, and recommendations for future research will be discussed.

For this experiment, a game called Arcanium was used to test participants on three different autonomy versions. Arcanium is a free source real-time strategy game that was found on www.sourceforge.net. It was developed at a university as a project and is now an open source game. This game was chosen due to three primary factors. The second reason that Arcanium was selected was because of the organization of the source code. Every line of source code was commented adequately and this allowed for a quick understanding of each facet of the game. This helped to define what automated changes were reasonable for Arcanium. The third reason was the type of game. Arcanium is a real-time strategy game. This means that playing Arcanium involves supervisory control of multiple resources and monitoring of many different information displays. The major
task of a real-time strategy game is to collect economic resources such that they can be spent on purchasing units, buildings, or upgrades. The goal is to build quicker than your enemy so that you can destroy the enemy. This provokes time stresses on the user and affects their decision making. This type of game closely resembles the types of tasks that are completed in military and industrial environments. Real-time strategy games such as Arcaium closely resemble unmanned aerial vehicle piloting, nuclear power plant monitoring, and other analyst or surveillance type tasks.

This game only requires visual and motor resources from the player. Although there is an audio channel of music during the game, the music does not provide any direct advantage for completing the puzzle. The music was played during all user testing scenarios and was not of concern in the experiment.

5.4 An Introduction to Arcanium – Test Bed

The objective of Arcanium is for the participant is to destroy the enemy base within a 12 minute time constraint. The complexity of the game is that the enemy base is hidden and not observable on the map by the player at the start of the game. The player must explore the fog of war, which is the hidden area of the map, so that they can identify the enemy base. Each participant was tested on three different levels and the details on the differences between will be described.

When the game begins, the participant is located at their home base as seen in Figure 37. The game has different units, which are different types of people that have different purposes. The first type of unit is a worker, which can collect resources (lumber, gold, and mana), build buildings (barracks or armories), or scout for the enemy. To collect the resources the player must direct the worker to the forest for lumber, the gold
mine for gold, or the mana crystal for mana. The two types of buildings have two crucial purposes in the game. Barracks create the second type of unit, the swordsman, which are aggressive combat units and are the only type of units that can kill enemy units (one part of winning the game). The armory creates the third type of unit, the zeppelin, which can transport the swordsman to the enemy base (once it is found) as well as scouting for the enemy base. If the participant doesn’t have enough resources to create a building, an error message appears saying something to the effect of “not enough resources, you need ___ more lumber/gold/mana”. Information displays show all current resources as well as all the enemies’ current resources.

The participant only had 12 minutes to win the game. If they didn’t finish in time, they would lose the game. The other way to lose is by the enemy destroying their base. The single way to win the game is by destroying the enemy base. Therefore, the objective of the game is to build your army faster than the enemy, find their base, and then destroy their base. If the participant does the experiment perfectly they will follow these steps:

1) Create swordsman

2) Create a zeppelin

3) Scout with the zeppelin (find the enemy base)

4) Place the swordsman in the zeppelin

5) Fly to the enemy base

6) Unload the swordsman from the zeppelin

7) Attack the enemy base with the zeppelin
Figure 36: Arcanium Interface for Start of Game

Start of the game for Arcanium with four workers, main base, and resources.

5.5 Methodology

Two different versions of the game Arcanium were developed and slight modifications were made to the original version of the game. The highly automated version, and expected to be the easiest of the three, is similar to a monitoring task for the human. The user perceived that the computer was making many decisions for them and the only task that the user was required to make was to identify and select the enemy base so that their units would eliminate the enemy base. Once the base was selected the computer autonomously routed units to eliminate the enemy base. The highly manual version required the human to direct all units with arrow keys, manually build structures, and individually direct units around the map. This included protecting the player’s base, finding the enemy, and collecting all resources. The partially automated version required
the user to place waypoints to direct units, and the computer autonomously routed units to the set waypoints. However, the resource collection was designed to be autonomous for this version to some degree. The user was only required to direct a worker to collect resources from the forest, gold mine, or mana temple one time and the system took over from there.

The levels of automation for the manual, partially automated, and highly automated were 2, 4, and 9 from Sheridan and Verplank’s 10 levels of autonomy (1978). Level ten of automation is described as the human not having override controls over the system. The highly autonomous version of Arcanium allowed for override controls but all tasks were automated. The partially autonomous version of Arcanium was the original design of Arcanium. The partially autonomous version fell closest to level four due to the computer not completely conducting any of the four different automation task types (information acquisition, information analysis, decision making, or action implementation). The manual version was level 2 because the human task load was high but the computer still provided some assistance, especially for information acquiring and action implementation task types. For Arcanium, the human would obtain information through visual sensory registry. No audio was involved for this game.

The controls of Arcanium were different among automation level designs of the game. The manual version uses the keys to move units and the mouse to select units and pan around the map. Panning is required because the full map size is larger than the screen observable area. This, the player must move the mouse towards the edge of the screen to look at other areas of the map. A combination of panning using the mouse and moving a unit using the arrow keys was required sequentially and multiple times to
successfully collect resources and explore the map. The partially automated version only used the mouse and the player would simply right click to select the unit and left click to select the movement location. The highly automated version only required interaction with the mouse. The controls were the same as the partially automated. However, if the players stopped controlling any unit for longer than 15 seconds the automation would begin controlling that unit.

Eighteen participants were tested in this study that was conducted at Wright State University in the Human Performance and Cognition Laboratory. The study was approved through the Wright State University Institutional Review Board (IRB) through expedited review. All subjects were sampled from the Wright State University faculty and student community. The sample included six female (33%) and twelve male (66%) participants. The age range of the sample included twelve subjects between the age of 20-29 (66%), four subjects in the 30-39 range (22%), and two subjects in the 40-49 range (11%). One subject was colorblind. All procedures were conducted exactly the same as the Q.U.B.E experiment from phase 1 except the training session. This identical procedure included administering a pretest-questionnaire, calibrating Captiv and Smart Eye Pro, and administering the NASA-TLX posttest-questionnaire. The differences between the training sessions of the Q.U.B.E study and Arcaniun was that, for Arcaniun, a PowerPoint slideshow was used to explain the objective and all building, units, and resources and their essentialness for meeting that objective. The controls for the game were explained right before each level. Due to the similarities in experimental procedure between this study and the Q.U.B.E study, the details will not be addressed again and
sections 4.6.1-4.6.7 can be referenced. The run order was randomized using a Latin square similar to that of the Q.U.B.E study.

5.6 Analysis and Results

Hypothesis 1 was that all the tested psychophysiological measures that were found to be significant in phase 1, time to complete task, and the subjective survey results (i.e., NASA-TLX) that were used in the study would be found to be statistically significantly different for the three autonomy levels of Arcanium. Individual univariate F-test ANOVAs were conducted for NASA-TLX and time to complete task as dependent variables. Three main effects were considered for each univariate analysis: automation level (2, 4, 9), computer experience (1-5), computer game experience (1-5) (Figure 38). Level of autonomy was used as an independent variable for this experiment with the purpose of testing the primary hypothesis. Computer and game experience were used as independent variables in the model to determine if participant expertise differences had an effect on all tested responses. Computer and game experience were not controlled factors in the experiment so equal sample sizes did not exist among levels, and the reporting of differences for computer game experience and computer experience should be taken lightly. Game and computer experience results are exploratory with the intention of directing research efforts towards the research area of expertise vs. novices and physiology.

\[ H_0: \mu_{\text{Highly Manual}} = \mu_{\text{Partially Automated}} = \mu_{\text{Highly Automated}} \]

or \[ H_0: \mu_{\text{LOA2}} = \mu_{\text{LOA4}} = \mu_{\text{LOA9}} \]

\[ H_1: \mu_{\text{Highly Manual}} \neq \mu_{\text{Partially Automated}} \neq \mu_{\text{Highly Automated}} \]

or \[ H_1: \mu_{\text{LOA2}} \neq \mu_{\text{LOA4}} \neq \mu_{\text{LOA9}} \]
Figure 37: Experiment Design: Arcanium Study

Illustration of independent and dependent variables for Phase 2 experiment. Dependent variables are in circle and independent variables with associated levels are on top.

This analysis is focused mainly on the controlled independent variable *automation* level that has equal sample sizes among levels. Subjects were contributed to the error term in the model for each univariate F-test ANOVA. Two-way interactions of the independent variables were not considered due to the lack of degrees of freedom to model the interactions. A standard least squares model personality with the restricted maximum
likelihood model (REML) method was used to fit the model using the statistical software package JMP version 11 for Microsoft Windows.

### 5.6.1 NASA-TLX Workload Rating Analysis

Results showed that the NASA-TLX ratings were statistically significant for automation level trial, $F(2,34) = 9.17$, $p < .01$ (Figure 39). However, video game experience ($F(4,11) = .11$, $p = .90$) and computer experience ($F(2,11) = .83$, $p = .54$) were not statistically significant. A post-hoc Tukey HSD was used to determine which levels of the statistically significant independent variables were found to be different. This analysis on NASA-TLX ratings revealed that low automation level ($M = 39.90$) was significantly different than partially automated ($M = 29.76$) and high automation level ($M = 23.05$). There was no difference found between partially automated and high automated for NASA-TLX ratings. All statistical tests were conducted using a 95% confidence interval. All statistical assumptions on normality and equal variance among conditions were validated similar to that of all assumptions tested in the Q.U.B.E study analysis (Appendix AW-Appendix AY).
Figure 38: NASA-TLX Workload Rating JMP 11.0 Analysis

Left figure includes JMP statistical output. Right figure shows the NASA-TLX workload rating on the y-axis and trial on the x-axis. Means are plotted with standard error bars.

5.6.2 Time to Complete Task Analysis

The performance measure of time of run was found to be statistically significant for automation level, $F(2,34) = 6.67, p < .01$ (Figure 40). However, game experience ($F(4,11) = 1.22, p = .36$) and computer experience ($F(2,11) = .05, p = .95$) were not found to be significant. A post-hoc Tukey HSD analysis found that manual autonomy was significantly different than the fully autonomous version. Results also indicated that differences were not detected between the manual version and partially autonomous version or detected between the partially automated and fully autonomous version. In
minutes, the means of the Arcanium autonomy levels manual, partially and highly were M = 10.04, M = 9.21, and M = 7.51 respectively. All statistical assumptions for the F-test ANOVA were validated for time run using the same method used for the NASA-TLX analysis (Appendix AZ-Appendix BB).

![Figure 39: Time to Complete Task JMP 11.0 Analysis](image)

Left figure includes JMP statistical output. Right figure shows the time of run on the y-axis and trial on the x-axis. Means are plotted with standard error bars.

5.6.3 Outlier Analysis

The next step was to analyze all of the physiological measures as indicators of CWL during autonomy levels. Multivariate methods were used to correlate the physiological measures data to the NASA-TLX rating. The purpose of this multivariate method was to determine outliers in the data by using the Mahalanobis and Jackknife distances. Four outliers were found and removed from the dataset for the analysis. The
data points that fell close to the upper limit for the two outlier methods were not removed from the dataset. The multivariate analysis also revealed that some collinearity existed between some of the physiological measures. Fixation duration and fixation rate were not found to collinear as found in the Q.U.B.E study (r(54) = 0.26). EMG frontal lobe and EMG frontal lobe standard deviation (r(54) = 0.94), as well as, EMG temporal lobe and EMG temporal lobe standard deviation (r(54) = 0.96) were found to be collinear. Heart rate average and heart rate standard deviation were not found to have a collinear relationship unlike the Q.U.B.E experiment (r(54)= -0.33). A univariate F-test ANOVA was conducted for each of the physiological measure responses for the independent variable task difficulty, computer experience, and computer game experience.
Figure 40: Mahalanobis and Jackknife Outlier Analysis

Outlier analysis to identify and remove outliers and correlation matrix for all dependent variables in the experiment.
The analysis of physiological measures fixation duration, fixation rate, EMG frontal lobe standard deviation, EMG frontal lobe averages, and heart rate standard deviation were completed in the same manner. The F-test ANOVA model included three independent variables of trial (manual automation, partially automated, and highly automated), computer experience (1-5), and game experience (1-5). The model also included subjects in the error term to account for the source of variability due to inherent difference in human physiology. Significant differences were found for heart rate standard deviation, fixation rate, EMG frontal lobe average, and EMG frontal lobe standard deviation for at least one of the three independent variables. Fixation duration was the only measure for which no levels, for any of the independent variables was found to be different. An analysis of each physiological measure is listed in the sections below.

5.6.4 Fixation Rate and Fixation Duration Analysis

Fixation rate \(F(2, 30.03) = 4.24, p = .02\) was found to be statistically different for trial automation level (Figure 43). A post-hoc Tukey HSD revealed that the manual automation level was found to be significantly different from the fully automated version for fixation rate. No statistical difference was found between the manual and partially automated versions as well as no difference was identified between the partially and fully automated versions. Game experience \(F(4, 10.92) = .96, p = .47\) and computer experience \(F(2, 11.01) = .72, p = .51\) were not found to be statistically significant for fixation rate (Figure 43). No statistical significance was found for fixation duration for any of the independent variables: automation level \(F(2, 29.92) = 1.09, p = .35\), game experience \(F(4, 10.48) = 1.02, p = .44\) and computer experience \(F(2, 10.81) = 1.78, p = .21\) (Figure 44). All statistical assumptions, as illustrated in the NASA-TLX analysis
from the Q.U.B.E study, were validated and the p-values for the F-test ANOVA for fixation rate and fixation duration are believed to be reliable (Appendix BC–Appendix BH).

**Figure 42: Fixation Rate Analysis**

Left figure includes JMP statistical output. Right figure shows the fixation rate on the y-axis and trial on the x-axis. Means are plotted with standard error bars.

**Figure 43: Fixation Duration Analysis**

JMP statistical output- No significance was found for fixation duration
5.6.5 EMG Frontal Lobe Analysis

EMG frontal lobe average ($F(2, 9.47) = 10.80, p < .01$) and EMG frontal lobe standard deviation ($F(2, 9.63) = 13.29, p < .01$) were found to be statistically different for computer experience (Figure 45). A post-hoc Tukey HSD revealed that a lower computer experience of level three corresponds to a higher EMG frontal lobe average that was significantly different than both computer experience of 4 and 5. No difference was found between computer experience of 4 and 5. EMG frontal lobe standard deviation post-hoc Tukey HSD revealed that there is a higher standard deviation for participant game experience ratings of three than all other levels. Task difficulty ($F(2, 26.44) = .19, p = .83$) and game experience ($F(4, 8.92) = .45, p = .77$) were not found to be statistically significant for both EMG frontal lobe average. Task difficulty ($F(2, 26.60) = .27, p = .77$) and game experience ($F(4, 9.05) = .77, p = .57$) were also not found to be statistically significant for EMG frontal lobe standard deviation (Figure 46). All statistical assumptions, as illustrated in the NASA-TLX analysis, were validated and the p-values for the F-test ANOVA for EMG frontal lobe average and EMG frontal lobe standard deviation are believed to be reliable (Appendix BI-Appendix BO).
Figure 44: EMG Frontal Lobe Average Analysis

JMP statistical output—Computer experience was found to be statistically significant at a 95% confidence interval for EMG frontal lobe average.

Figure 45: EMG Frontal Lobe Standard Deviation Analysis

JMP statistical output—Game experience was found to be statistically significant at a 95% confidence interval for EMG frontal lobe standard deviation.
5.6.6 Heart Rate Standard Deviation Analysis

Heart rate standard deviation showed significant statistical differences for automation level \( F(2, 26.53) = 4.27, p = .03 \) (Figure 47). A post-hoc Tukey HSD analysis found that higher heart rate standard deviations were found during the higher autonomous version that that of the partially autonomous version. However, no difference was found between the fully autonomous and manual automation version and the manual automation version when compared to the partially automated version. Statistical significant was not found for both computer experience \( F(2, 9.56) = .44, p = .66 \) and game experience \( F(4, 9.01) = .07, p = .99 \). All statistical assumptions, as illustrated in the NASA-TLX analysis, were validated and the p-values for the F-test ANOVA for heart rate standard deviation are believed to be reliable (Appendix BM-Appendix BP).
Figure 46: Heart Rate Standard Deviation Analysis

Left figure includes JMP statistical output. Right figure shows the fixation rate on the y-axis and trial on the x-axis. Means are plotted with standard error bars.
5.6.7 Hypothesis 1 Discussion and Results

This section includes a summary of the results for the phase 2 experiment hypothesis 1 and a discussion on the findings. The primary hypothesis of this study was that all of the tested psychophysiological measures found as significant from the phase 1 experiment, time to complete run, and the subjective survey results would be found to be significantly different for different automation levels as supported by the literature reviewed. The results of the study showed that NASA-TLX ratings, time to complete run, fixation rate, and heart rate standard deviation were statistically different among task automation levels. NASA-TLX, time to complete task, and fixation rate all indicated differences between the automation level design for Arcanium of fully automated and a more manual automation. However, heart rate standard deviation was unable to detect that difference. Differences between fully automated and partially automated were not detected for NASA-TLX, time to complete task, and fixation rate. Differences between the more manual version and the partially automation versions were not detected for time to complete task and fixation rate. In general, NASA-TLX, time to complete task, and fixation rate all seem to behave similarly in terms of detecting alike differences.

It is heart rate standard deviation that seems to follow a different trend. Rowe, Sibert, and Irwin (1998) state that heart rate variation begins to steady as CWL increases up to a certain point. When a task becomes too difficult, heart rate variation begins to increase because of exerted stresses on the human. After discussing with participants what version they prefer their feedback indicated that the fully autonomous version was boring and that they felt more as if they weren’t even playing a game at all. Whereas, the manual difficulty design was irritating and required too much input to keep track of multiple sources of information. Most of the participants preferred the partially
autonomous version and they felt that this version was the easiest to control, and to monitor information. An implication of the participant feedback as it corresponds to the physiological results is that the more highly autonomous, LOA 9, version of Arcanium may be so easy and elicit such a small amount of CWL that the participants heart rate has yet to steady. Also, for the more manual version, LOA 2, the task has reached the point of difficulty such that heart rate standard deviation begins to increase. This corresponds to the results by Rowe, Sibert, and Irwin (1998).

EMG frontal lobe averages and standard deviation were found to be statistically different for computer experience. Specifically, computer experience of level three corresponded to higher EMG frontal lobe averages and standard deviation. The level 3 computer experience was significantly different than a computer experience of 4 or 5, and no difference was detected between 4 and 5 for both EMG frontal lobe measures. These finding were different than what was found in the Q.U.B.E study. The Q.U.B.E study results indicated a difference in computer game experience, which was found to possibly nonlinear, rather than computer experience. This difference may be explained by the differences in the type of game between the types of game. Even though the game experience variable was not significantly different for the Arcanium study, the nonlinear relationship was still present.

The results for hypothesis 1 help to address the purpose of phase 2 of this thesis to test physiological measures as indicators of cognition during different automation levels and to evaluate expertise as a factor for CWL difference. Physiology measures: fixation rate, EMG frontal lobe standard deviation, EMG frontal lobe averages, and heart rate standard deviation were identified as having significance automation levels for
Arcanium. These measures are proposed as potential measures for identifying CWL differences for different types of tasks. All physiological measures identified in phase 2 are used in the model for phase 3.

5.6.8 Hypothesis 2 Discussion and Results

Hypothesis 2 for the phase 2 Arcanium experiment was that the model trained from the phase 1 Q.U.B.E. experiment could be tested on the physiological data for the phase 2 experiment and that some variation would still be explained. By testing this hypothesis it is anticipated that the physiological measures used to build a model to predict NASA-TLX can be used for two different gaming contexts. The importance of this type of research is to develop a robust model based on human physiology to predict CWL regardless of type of task. If a robust model could be developed, this would revolutionize the method for researchers to collect CWL information from their participants. Studies could involve physiological measures collection instead of subjective surveys and this would provide real-time CWL measures.

Both the linear and quadratic models that were trained in the phase 1 study were tested on the data of the phase 2 study. All models were built in JMP version 11.

The linear model method that included the main effects of heart rate standard deviation and EMG frontal lobe standard deviation only accounted for 7% of the variation of the NASA-TLX workload response (Figure 48). This finding was not satisfying and the model that was trained in the phase 1 first person puzzle game did not account for much of the variation at all in the phase 2 real-time strategy game. Furthermore, the nonlinear model accounted for 23% of the variation for phase 2 (Figure
A universal model that is able to account for a significant amount of the variation in the NASA-TLX CWL rating with physiological measure inputs was not found for this study. One of the considerations for future research is whether the model translates to the same gaming context rather than for different types of games. This hypothesis could be tested in further experimentation. Although an adequate model was not found for the purposes of hypothesis two for this experiment, this experimental approach provides insight into a method for attempting to develop a model to predict human CWL. The ideology of testing other measures that can be collected real-time, such as physiology, as a predictor of accepted CWL measures, such as NASA-TLX, is promoted.

**Figure 47: Tested Linear Regression Model for NASA-TLX Response**

Linear model created from the phase 1 using two main effects of EMG frontal lobe standard deviation and heart rate standard deviation and used to test on the data collected in the phase 2 experiment. Accounts for 3% of the variation of the NASA-TLX workload response. Not a universal model.
**Figure 48: Tested Nonlinear Regression Model for NASA-TLX Response**

Nonlinear model created from the phase 1 multiple different main effects and interactions and used to test on the data collected in the phase 2 experiment. Accounts for 3% of the variation of the NASA-TLX workload response. Not a universal model.

<table>
<thead>
<tr>
<th>Source</th>
<th>Nparm</th>
<th>DF</th>
<th>DFe</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pupil Diameter Avg.</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.3271</td>
<td>0.5765</td>
</tr>
<tr>
<td>EMG Frontal Lobe Stdev.</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.2397</td>
<td>0.6323</td>
</tr>
<tr>
<td>EMG Temporal Lobe Avg.</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.0032</td>
<td>0.9622</td>
</tr>
<tr>
<td>EMG Temporal Lobe Stdev.</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.0013</td>
<td>0.9979</td>
</tr>
<tr>
<td>Heart Rate Avg.</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4.6536</td>
<td>0.0400*</td>
</tr>
<tr>
<td>Heart Rate Stdev.</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.3318</td>
<td>0.5694</td>
</tr>
<tr>
<td>EMG Temporal Lobe Avg.*EMG Temporal Lobe Stdev.</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.0577</td>
<td>0.7964</td>
</tr>
<tr>
<td>EMG Temporal Lobe Stdev.*EMG Temporal Lobe Stdev.</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.0022</td>
<td>0.9926</td>
</tr>
<tr>
<td>Heart Rate Avg.*Heart Rate Avg.</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.2243</td>
<td>0.6407</td>
</tr>
</tbody>
</table>
6.1 Real-time Physiological Collection

One of the pressing issues is how a computer will actually adapt to the human. Specifically, if a robust model uses physiological inputs as a measurement of some overall CWL score, how will the system adapt? I propose the use of a tool, known in industrial manufacturing quality control for measuring process stability, control charting, as a computer input and a catalyst for changes in autonomy. A common quality control chart is the Xbar-S chart where Xbar is the average of a sample and S is the within sample standard deviation (Figure 50). The x axis is the number of collection assuming equal sample size across collections. The highest and lowest red horizontal lines are the upper and lower control limit of the process. The calculations for these limits are dependent on the sample size for each collection. If the subgroup standard deviation is in control, as seen in the standard deviation chart, then the Xbar chart will indicate the stability of the response on the y axis. If an average collection falls below the lower control limit or above the upper control limit it can be said that the process is unstable. Also, if any trends can be identified for the means of each subgroup an unstable process might exist.

Assume that the y-axis is a measure of CWL. Specifically, that CWL is a measure that is the response of a theoretical robust model made up of multiple physiological measures. For this method to work, time needs to be discretized into equal
size collection intervals. Assuming the data sample rate remains constant, an equal
sample size of physiological measures will be collected. These physiological measures
would be inputs to predict a CWL measure and the standard deviation and mean could be
calculated for that measure. The adaptive components that are integrated into the
software can adapt to the human based on the identified trend from the Xbar-S charts.
This methodology could exist real-time and an adaptive system that changes to support
human CWL limitations could exist.

This is an example of how an adaptive system could work. However, there are a
few details such as how many discretized collection intervals should exist to detect
differences between CWL measures. Another issue is with regards to the proper
calculation for the upper and lower control limits. The most crucial issue is determining
the proper indicators of CWL as independent variables to develop a robust model for
predicting CWL. This thesis has provided a direction for investigating measures such as
fixation rate, heart rate standard deviation, and time of completing task as indicators of
cognition. It is hopeful that research efforts will be directed towards developing an
improved CWL measure that can be sampled real-time.
Method used for assessing cognitive workload over time.

6.2 The Theoretical Model

Two main pressing issues from the general adaptive model based on cognitive state assessment (Figure 1) which was inspired by Byrne and Parasuraman’s (1996) work were to be addressed to help build a more detailed theoretical model for an adaptive
automation system. The first issue concerns the lack of physiological measures that accurately indicate CWL. This thesis describes two experiments that identified fixation rate, heart rate standard deviation, and EMG frontal lobe measures as promising measures for future model development. It was also found that time to complete task and NASA-TLX subjective surveys also indicated CWL differences among levels for both studies. It is suggested that the physiological measures identified in this thesis, along with other possible physiological measures such as GSR, respiratory rate, and others should be investigated for future research. The second issue of concern is how software can be designed based on autonomy levels for adaptive automation. This thesis provided a literature review on automation and suggested that the software design for autonomy can be designed from Verplank and Sheridan’s 10 levels of autonomy as well as considering the type of automation such as information acquisition, information analysis, decision selection, and action implementation. The design of automation levels for the phase 2 game Arcanium was used as an example for designing software for different automation levels. Another issue is how the system will be prompted to adapt. This was addressed by considering an Xbar-S control chart method for continuous monitoring of CWL measures. These elements that were addressed in this thesis are included in the refined theoretical model for an adaptive system in Figure 51.
Computer evaluates cognitive state and compares to current system state task load

Understanding automation levels and types
Identifying collection method
Design of automation levels for Arcanium

Methodology for adaptive response

Figure 50: Refined Theoretical Adaptive Autonomy Model
The contributions of this thesis include physiological measures that were found to indicate difference in CWL among levels or be correlated with NASA-TLX CWL ratings. Insight into other physiological measures that could be considered as measures of CWL was discussed in the literature review section. The collection of physiological measures is encouraged as a real-time measure of human state and, in the model; these measures are what drive system state changes. Another major contribution of this thesis is the literature review on automation levels and types of automation. This part of the model will help software designers understand the tasking components of both the human and the computer in order to develop different automation levels for their software design. Insight into methods for collecting physiological measures real-time such as eye tracking systems, EMG electrodes hardware, and heart rate monitors were used as an example for two studies. It was found that differences in some physiology were found for an experiment involving task difficult level differences and an experiment involving different automation levels. Because research suggests that physiology is an indicator of CWL and this thesis has found supporting evidence for some physiology in two different gaming contexts a method for evaluating automation real-time and adapting to the human was described. The method includes identifying CWL measures using physiology and using an Xbar-S chart to determine when the system autonomy level should adapt to the human. A more detailed model was developed from Figure 52 and it can be seen in Figure 55.
Figure 51: Final Adaptive Autonomy Model
6.3 An Application of the Final Adaptive Autonomy Model

To explain the model in Figure 52, an application of the model using data from the Arcanium study is illustrated. The human state and the computer system state exist independently as illustrated in the initial states section of the model. At the system state a baseline interface or software is developed using orthodox interface or software development methodologies. From the initial baseline design of the interface or software, other versions that include a different level of automation are developed. These versions of the interface or software may include automation in the form of information acquisition, information analysis, decision selection, and action implementation that is higher or lower in automation that the current design. The interface or software is then designed such that changes from one automation state to an increased or decreased automation state can transition fluidly. All of this is completed prior to the human interaction with the autonomous states of the automation system. The design of the three levels of Arcanium is described in section 5.5.

Once the automation levels are developed, the hardware that provides the flow of human state information to the computer must be calibrated to the human to accurately and precisely assess the human’s state. The use of the Smart Eye system and Captiv modules is described in sections 4.6.2 and 4.6.3. The calibrated physiological hardware is used to assess human CWL during the baseline condition. Although a robust model for evaluating CWL was not found in this thesis, the linear model that was trained in the Q.U.B.E study and tested in the Arcanium study will be model to demonstrate the theoretical model from Figure 52. This model can be seen in Figure 53. The CWL metric (CWM) includes two physiological inputs of EMG frontal lobe standard deviation
and heart rate standard deviation. In the model, as EMG frontal lobe standard deviation and heart rate standard deviation increase, the CWL metric increases. It is stressed that the model in Figure 53 is by no means robust. However the physiological measure independent variables do show signs of promise for indicating CWL and the idea of researching these physiological inputs, in coordination with other physiological measures, is encouraged.

\[
CWM = 10.9275 + 3.9669 \times EMG \text{ Frontal Lobe Standard Deviation} + 2.1619 \\
\times Heart \ Rate \ Standard \ Deviation
\]

*Equation 2: Linear Model Trained from Phase 1: Q.U.B.E. Study*

Equation that was trained from the phase 1 experiment to predict NASA-TLX workload responses.

For the purposes of demonstrating the theoretical model, it was assumed that the partially autonomous version was the version of Arcanium that best fit the human’s CWL model. The partially automated run for a participant was analyzed using the two physiological measures in the linear model from Figure 53. This participant’s approximately ten minute run was divided into ten discrete sections and the CWM mean and standard deviation was calculated and plotted using the S chart methodology. The X-bar chart was not used for this example because the linear model from figure 53 only includes standard deviation values and does not include means. By using this method an understanding of CWM over time can be input into the computer and the system can autonomously adapt based on human CWL levels. For this example, an arbitrary upper and control limit of one standard deviation away from the means was selected. However,
a more applicable upper and control limit need to be determined to apply this methodology. Assuming the limits are one standard deviation away from the mean, it was found that the CWM participant ten is out of control during the first and ninth discrete time interval. This computer can analyze this data real-time to determine patterns of instability for a human’s CWM over extended periods of time. This can lead to the computer adapting in autonomy or changing system state to better assist the human. For example, figure 54 indicates occurrences of high CWL outside of the control limits of CWM. If this trend continuous, it may be justifiable to increase the automation to assist with the human’s CWL. This human CWM evaluation is an iterative process that can continuously monitor human cognition and provide levels of automation adapt to better support the human. The refined model, including visuals for the tools of the model, can be seen in figure 55.

![CWM Vs. Discrete Time Intervals For A Participant's Partially Automation Version](image)

**Figure 52: Real-Time Evaluation of CWL Over Time Example**

Workload responses 10 discrete time interval from a participant’s data collection
Figure 53: Example Adaptive Autonomy Model

Levels: 2, 4, 9 of Arcanium

Levels 1-10 for the Different Adaptive Autonomy Software Design Levels

Design Process

Initial States

System States

Human Cognitive State

Create Human Physiology Profile

Human Interacts with New Level of Autonomy if High or Low Cognitive Workload was Identified

System Iterative Improvement

Design and Develop Baseline Interface or Software

Calibrate Physiological Measure Hardware

Collect Physiology of Human and Build Physiology Profile

Real-time Physiological Assessment of Cognitive State

If Cognitive Workload Indicate High Cognitive Workload or Low Cognitive Workload States, Change Automation Level to Support Human

Use Control Chart Xbar-S Method to Continuously Monitor Cognitive Workload

Compare Automation Level Provided to Cognitive Workload

Design and Develop All Autonomy Levels for Interface or Software

LOA 4

LOA 9

Determine Levels (1-10) for the Different Adaptive Autonomy Software Design Levels

CWM Vs. Discrete Time Intervals Participant 10 Partially Automation Version
7.0 APPENDIX

Appendix-A: Pretest Questionnaire

1) Please provide the following information:
   a. Please circle your age group
      • 20-29 years
      • 30-39 years
      • 40-49 years
      • 50+ years
   b. Gender
      • Male
      • Female
   c. Ethnicity
      • Hispanic
      • American Indian
      • Asian
      • African American
      • Native Hawaiian or Pacific Islander
      • Caucasian
      • Other: ______________________

2) Are you color blind?
   • Yes
   • No

3) Do you have any type of visual impairment?
   __________________________________________

4) Circle your level of computer experience (5 is the highest)
   1  2  3  4  5

5) What is your level of experience with computer games (5 is the highest)?
   1  2  3  4  5

6) Have you ever played a first person 3-d environment game before?
   • Yes
   • No

7) Have you ever played Q.U.B.E (Quick Understanding of Block Extrusion)?
   • Yes
   • No

8) Did you know about the nature of this study before participating?
   • Yes
   • No
**Appendix-B: Latin Square Randomization Q.U.B.E.**

<table>
<thead>
<tr>
<th>Participant</th>
<th>First Run</th>
<th>Second Run</th>
<th>Third Run</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participant 1</td>
<td>Easy</td>
<td>Medium</td>
<td>Difficult</td>
</tr>
<tr>
<td>Participant 2</td>
<td>Easy</td>
<td>Difficult</td>
<td>Medium</td>
</tr>
<tr>
<td>Participant 3</td>
<td>Medium</td>
<td>Easy</td>
<td>Difficult</td>
</tr>
<tr>
<td>Participant 4</td>
<td>Medium</td>
<td>Difficult</td>
<td>Easy</td>
</tr>
<tr>
<td>Participant 5</td>
<td>Difficult</td>
<td>Easy</td>
<td>Medium</td>
</tr>
<tr>
<td>Participant 6</td>
<td>Difficult</td>
<td>Medium</td>
<td>Easy</td>
</tr>
<tr>
<td>Participant 7</td>
<td>Easy</td>
<td>Medium</td>
<td>Difficult</td>
</tr>
<tr>
<td>Participant 8</td>
<td>Easy</td>
<td>Difficult</td>
<td>Medium</td>
</tr>
<tr>
<td>Participant 9</td>
<td>Medium</td>
<td>Easy</td>
<td>Difficult</td>
</tr>
<tr>
<td>Participant 10</td>
<td>Medium</td>
<td>Difficult</td>
<td>Easy</td>
</tr>
<tr>
<td>Participant 11</td>
<td>Difficult</td>
<td>Easy</td>
<td>Medium</td>
</tr>
<tr>
<td>Participant 12</td>
<td>Difficult</td>
<td>Medium</td>
<td>Easy</td>
</tr>
<tr>
<td>Participant 13</td>
<td>Easy</td>
<td>Medium</td>
<td>Difficult</td>
</tr>
<tr>
<td>Participant 14</td>
<td>Easy</td>
<td>Difficult</td>
<td>Medium</td>
</tr>
<tr>
<td>Participant 15</td>
<td>Medium</td>
<td>Easy</td>
<td>Difficult</td>
</tr>
<tr>
<td>Participant 16</td>
<td>Medium</td>
<td>Difficult</td>
<td>Easy</td>
</tr>
<tr>
<td>Participant 17</td>
<td>Difficult</td>
<td>Easy</td>
<td>Medium</td>
</tr>
<tr>
<td>Participant 18</td>
<td>Difficult</td>
<td>Medium</td>
<td>Easy</td>
</tr>
</tbody>
</table>
Appendix-C: NASA-TLX Posttest Questionnaire

**Figure 8.6**

**NASA Task Load Index**

Hart and Staveland’s NASA Task Load Index (TLX) method assesses workload on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

<table>
<thead>
<tr>
<th>Name</th>
<th>Task</th>
<th>Date</th>
</tr>
</thead>
</table>

Mental Demand  
How mentally demanding was the task?

| Very Low | Very High |

Physical Demand  
How physically demanding was the task?

| Very Low | Very High |

Temporal Demand  
How hurried or rushed was the pace of the task?

| Very Low | Very High |

Performance  
How successful were you in accomplishing what you were asked to do?

| Perfect | Failure |

Effort  
How hard did you have to work to accomplish your level of performance?

| Very Low | Very High |

Frustration  
How insecure, discouraged, irritated, stressed, and annoyed were you?

| Very Low | Very High |
Appendix-D: Normal Q-Q Plot of Residuals for NASA-TLX Workload
Appendix-E: Residual vs. Predicted for NASA-TLX Workload
Appendix-F: Residuals NASA-TLX for All Independent Variables
Appendix-G: Normal Q-Q Plot of Residuals for Time to Complete Task
Appendix-H: Residual vs. Predicted for Time to Complete Task
Appendix-I: Residuals Time of Task for All Independent Variables

![Residual Time of Run vs. Trial](image1)

![Residual Time of Run vs. Computer Experience](image2)

![Residual Time of Run vs. Game Experience](image3)
Appendix-J: Normal Q-Q Residuals Fixation Rate
Appendix K: Residual vs. Predicted for Fixation Rate
Appendix L: Residuals Fixation Rate for All Independent Variables
Appendix-M: Normal Q-Q Residuals Fixation Duration
Appendix N: Residual vs. Predicted for Fixation Duration
Appendix O: Residuals Fixation Duration for All Independent Variables

![Fixation Duration vs. Trial](image1)

![Fixation Duration vs. Game Experience](image2)

![Fixation Duration vs. Computer Experience](image3)
Appendix-P: Normal Q-Q Residuals EMG Frontal Lobe Average
Appendix Q: Residual vs. Predicted for EMG Frontal Lobe Average
Appendix R: Residuals EMG Frontal Lobe Average for All Independent Variables
Appendix-S: Normal Q-Q Residuals EMG Frontal Lobe Standard Deviation
Appendix T: Residual vs. Predicted for EMG Frontal Lobe Standard Deviation
Appendix U: Residuals EMG Frontal Lobe Standard Deviation for All Independent Variables

Residual EMG forehead SD vs. Trial

Residual EMG forehead SD vs. Game Experience

Residual EMG forehead SD vs. Computer Experience
Appendix-V: Normal Q-Q Residuals Pupil Diameter Average
Appendix W: Residual vs. Predicted for Pupil Diameter Average
Appendix X: Residuals Pupil Diameter Average for All Independent Variables
Appendix-Y: Normal Q-Q Residuals Pupil Diameter Standard Deviation
Appendix Z: Residual vs. Predicted for Pupil Diameter Standard Deviation
Appendix AA: Residuals Pupil Diameter Standard Deviation for All Independent Variables

- Residual Pupil Diameter SD vs. Trial
- Residual Pupil Diameter SD vs. Game Experience
- Residual Pupil Diameter SD vs. Computer Experience
Appendix-AB: Normal Q-Q Residuals EMG Temporal Lobe Average
Appendix AC: Residual vs. Predicted for EMG Temporal Lobe Average
Appendix AD: Residuals EMG Temporal Lobe Average for All Independent Variables

Residual EMG temple Average vs. Trial

Residual EMG temple Average vs. Game Experience

Residual EMG temple Average vs. Computer Experience
Appendix-AE: Normal Q-Q Residuals EMG Temporal Lobe Standard Deviation

Normal Quantile Plot
Appendix AF: Residual vs. Predicted for EMG Temporal Lobe Standard Deviation
Appendix AG: Residuals EMG Temporal Lobe Standard Deviation for All Independent Variables
Appendix-AH: Normal Q-Q Residuals Heart Rate Average
Appendix AI: Residual vs. Predicted for Heart Rate Average
Appendix AJ: Residuals Heart Rate Average for All Independent Variables
Appendix-AK: Normal Q-Q Residuals Heart Rate Variability

Normal Quantile Plot
Appendix AL: Residual vs. Predicted for Heart Rate Variability
Appendix AM: Residuals Heart Rate Variability for All Independent Variables
Appendix-AN: Normal Q-Q Residuals Heart Rate Standard Deviation
Appendix AO: Residual vs. Predicted for Heart Rate Standard Deviation
Appendix AP: Residuals Heart Rate Standard Deviation for All Independent Variables
Appendix AQ: Normal Q-Q Residual Plot for Reduced Linear Model
Appendix AR: Residual vs. Predicted for Reduced Linear Model
Appendix AS: Residuals NASA-TLX for All Linear Model Independent Variables

Residual Workload 2 vs. Heart Rate SD

Residual Workload 2 vs. Pupil Diameter Average
Appendix AT: Normal Q-Q Residual Plot for Reduced Nonlinear Model
Appendix AU: Residual vs. Predicted for Reduced Nonlinear Model
Appendix AV: Residuals NASA-TLX for All Nonlinear Model Independent Variables

Heart Rate SD vs. Residual Workload

EMG forehead SD vs. Residual Workload

Residual Workload vs. Heart Rate SD
Appendix-AW: Normal Q-Q Plot of Residuals for NASA-TLX Workload
Appendix-AX: Residual vs. Predicted for NASA-TLX Workload
Appendix-AY: Residuals NASA-TLX for All Independent Variables
Appendix-AZ: Normal Q-Q Plot of Residuals for Time to Complete Task
Appendix-BA: Residual vs. Predicted for Time to Complete Task
Appendix-BB: Residuals Time of Task for All Independent Variables
Appendix-BC: Normal Q-Q Residuals Fixation Rate
Appendix BD: Residual vs. Predicted for Fixation Rate
Appendix BE: Residuals Fixation Rate for All Independent Variables
Appendix-BF: Normal Q-Q Residuals Fixation Duration
Appendix BG: Residual vs. Predicted for Fixation Duration
Appendix BH: Residuals Fixation Duration for All Independent Variables
Appendix-BI: Normal Q-Q Residuals EMG Frontal Lobe Average
Appendix BJ: Residual vs. Predicted for EMG Frontal Lobe Average
Appendix BK: Residuals EMG Frontal Lobe Average for All Independent Variables

![Residual EMG Frontal Lobe Avg. vs. Trial](chart1)

![Residual EMG Frontal Lobe Avg. vs. Computer Experience](chart2)

![Residual EMG Frontal Lobe Avg. vs. Computer Game Experience](chart3)
Appendix-BL: Normal Q-Q Residuals EMG Frontal Lobe Standard Deviation
Appendix BM: Residual vs. Predicted for EMG Frontal Lobe Standard Deviation
Appendix BN: Residuals EMG Frontal Lobe Standard Deviation for All Independent Variables

Residual EMG Frontal Lobe Stdev. vs. Trial

Residual EMG Frontal Lobe Stdev. vs. Computer Experience

Residual EMG Frontal Lobe Stdev. vs. Computer Game Experience
Appendix-BO: Normal Q-Q Residuals Heart Rate Standard Deviation
Appendix BP: Residual vs. Predicted for Heart Rate Standard Deviation
Appendix BQ: Residuals Heart Rate Standard Deviation for All Independent Variables


