NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

SYSTEMS ENGINEERING CAPSTONE REPORT

TASK HANDOFF BETWEEN HUMANS AND AUTOMATION

by

Andrew J. Brown Sr., John R. Folger IV, Johnathan W. Hardin, Jean'Shay D. Moore, and Quentin Sica

December 2021

Advisor: Lawrence G. Shattuck
Co-Advisor: Robert Semmens

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			*Form Approved OMB No. 0704-0188*

1. AGENCY USE ONLY *(Leave blank)*	**2. REPORT DATE** December 2021	**3. REPORT TYPE AND DATES COVERED** Systems Engineering Capstone Report

4. TITLE AND SUBTITLE TASK HANDOFF BETWEEN HUMANS AND AUTOMATION	**5. FUNDING NUMBERS**
6. AUTHOR(S) Andrew J. Brown Sr., John R. Folger IV, Johnathan W. Hardin, Jean'Shay D. Moore, and Quentin Sica	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A	**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.	**12b. DISTRIBUTION CODE** A

13. ABSTRACT (maximum 200 words)

The Department of Defense (DOD) seeks to incorporate human-automation teaming to decrease human operators' cognitive workload, especially in the context of future vertical lift (FVL). Researchers created a "wizard of oz" study to observe human behavior changes as task difficulty and levels of automation increased. The platform used for the study was a firefighting strategy software game called C3Fire. Participants were paired with a confederate acting as an automated agent to observe the participant's behavior in a human-automation team. The independent variables were automation level (within; low, medium, high) and queuing (between; uncued, cued). The dependent variables were the number of messages transmitted to the confederate, the number of tasks embedded in those messages (tasks handed off), and the participant's self-reported cognitive workload score. The study results indicated that as the confederate increased its scripted level of automation, the number of tasks handed off to automation increased. However, the number of messages transmitted to automation and the subjective cognitive workload remained the same. The study's findings suggest that while human operators were able to bundle tasks, cognitive workload remained relatively unchanged. The results imply that the automation level may have less impact on cognitive workload than anticipated.

14. SUBJECT TERMS automation, task handoff, future vertical lift	**15. NUMBER OF PAGES** 93
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

TASK HANDOFF BETWEEN HUMANS AND AUTOMATION

MAJ Andrew J. Brown Sr. (USA), CPT John R. Folger IV (USA),
CPT Johnathan W. Hardin (USA), CPT Jean'Shay D. Moore (USA),
and CPT Quentin Sica (USA)

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN SYSTEMS ENGINEERING MANAGEMENT

from the

NAVAL POSTGRADUATE SCHOOL
December 2021

Lead Editor: John R. Folger IV

Reviewed by:
Lawrence G. Shattuck
Advisor

Robert Semmens
Co-Advisor

Accepted by:
Oleg A. Yakimenko
Chair, Department of Systems Engineering

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

The Department of Defense (DOD) seeks to incorporate human-automation teaming to decrease human operators' cognitive workload, especially in the context of future vertical lift (FVL). Researchers created a "wizard of oz" study to observe human behavior changes as task difficulty and levels of automation increased. The platform used for the study was a firefighting strategy software game called C3Fire. Participants were paired with a confederate acting as an automated agent to observe the participant's behavior in a human-automation team. The independent variables were automation level (within; low, medium, high) and queuing (between; uncued, cued). The dependent variables were the number of messages transmitted to the confederate, the number of tasks embedded in those messages (tasks handed off), and the participant's self-reported cognitive workload score. The study results indicated that as the confederate increased its scripted level of automation, the number of tasks handed off to automation increased. However, the number of messages transmitted to automation and the subjective cognitive workload remained the same. The study's findings suggest that while human operators were able to bundle tasks, cognitive workload remained relatively unchanged. The results imply that the automation level may have less impact on cognitive workload than anticipated.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

ACRONYMS AND ABBREVIATIONS

AMS	Army Modernization Strategy
C-SWAG	Continuous Subjective Workload Analysis
C3Fire	Command, Control and Communication Research Software Program
CRM	Crew Resource Management
FAA	Federal Aviation Administration
FCU	Flight Control Unit
FVL	Future Vertical Lift
HAT	Human-Automation Teaming
HSA-DM	Holistic Situational Awareness and Decision Making
HSI	Huan Systems Integration
HRV	Heart Rate Variability
LAN	Local Area Network
NASA	National aeronautics and Space Administration
NPS	Naval Postgraduate School
ORAD	On-Road Automated Driving
SA	Situational Awareness
SAE	Society of Automotive Engineer's
TLX	Task Load Index

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

The Department of Defense (DOD) Holistic Situational Awareness and Decision Making (HSA-DM) program office is responsible for determining cognitive workload drivers for future vertical lift (FVL) and developing cognitive workload management capabilities. One of the most common techniques for decreasing cognitive workload is automating tasks previously performed by human pilots. This Naval Postgraduate School (NPS) capstone contributes to HSA-DM's mission by investigating how task difficulty and the sophistication of automation affect human behavior in a human-automation team (HAT) environment. The results from the study suggest that more sophisticated levels of automation do not necessarily drive down cognitive workload as much as other factors such as scenario difficulty.

Researchers conducted a "wizard of oz" type study with twenty participants and one confederate. Each participant was teamed with the confederate and assigned to extinguish a forest fire within a software program called C3Fire. The participant population consisted of mid-grade (O3-O4) officers in the Army, Navy, and Marines pursuing graduate-level degrees at NPS. None of the participants reported any familiarity with C3Fire, but they did have ranging experiences with automation. The independents variables for this study were automation level (within; low, medium, high) and queuing (between; un-cued, cued). The confederate following a pre-scripted level of automation which "upgraded" after each iteration. The confederate followed a specified level of automation that corresponded to one of Sheridan's levels of automation (Sheridan 1978). The dependent variables included the number of messages transmitted, the number of tasks handed off to the confederate, and the subjective cognitive workload reported by the participant. The participant and confederate completed one training scenario and three (live) scenarios in approximately 70 minutes.

This study's results indicated that the level of automation did not have a significant effect on cognitive workload. The study's results did show that participants sent more tasks to automation by using features as they became available. Participants generally used the more sophisticated levels of automation to bundle tasks into single messages. Participants

handed off significantly more tasks to automation after the bundling feature became available. However, further increasing the amount of bundling did not result in a significant increase in the number of tasks passed to automation. While participants sent more tasks to the automated agent, the cognitive workload remained the same throughout each iteration.

The results suggest that developers of the next generation FVL platform should build automation features that allow humans to bundle tasks into a single transmission to enable humans to send more tasks to an automated agent. Bundling was effective in allowing participants to send more tasks to automation more quickly. But creating features that allow humans to transmit more tasks to an automated agent will not necessarily alleviate the human operator's cognitive workload.

Reference

Sheridan, T.B., & Verplank, W. 1978. "Human and Computer Control of Undersea Teleoperators." MIT, Man-Machine Systems Laboratory, Department of Mechanical Engineering, Cambridge, MA.

ACKNOWLEDGEMENTS

The results of this capstone would not have been possible without the exceptional support of our advisors. Dr. Lawrence Shattuck and Dr. Robert Semmens exhibited enthusiasm, extreme knowledge, and encouragement throughout our capstone's trial-and-error learning journey. Their attention to detail coupled with the motivation to take risks provided room for creativity in our research.

Dr. Kip Smith, based out of Washington state, contributed immensely to our efforts in providing his time to instruct our team on the employment of C3Fire software, as well as manipulating the program to suit our experimental desires.

Mr. Matt Shivers, representative and stakeholder for the Holistic Situation Awareness and Decision Making (HSADM) Team, provided direction and motivation for our research so that future automation capabilities research can advance.

Major Charlie Rowan and Assistant Professor Michael O'Neil offered up their knowledge and time to assist our research team throughout this process. With their presence and expertise, we were able to refine our research, data collection, and data interpretation methods to better identify and classify our discoveries and recommendations for future research.

Various support networks have afforded our team the resources, time, and understanding we needed to accomplish this research effort. Without our families, both immediate and academic, our progress on this endeavor would be less impactful to its purpose.

THIS PAGE INTENIONALLY LEFT BLANK

I. INTRODUCTION

A. PROBLEM STATEMENT

The Future Vertical Lift (FVL) Program is a key pillar in the Army's modernization strategy. It is an investment in next-generation air vehicles that are designed to dominate throughout the 21st Century. The creation of new initiatives and programs, such as FVL, pushes for innovation to ensure the United States' can win in an era of great power competition. One such innovation is the use of advanced automation in these future systems. Keller (2020) identifies that, as FVL becomes a reality, pilots will be working alongside automated systems on a myriad of tasks. He also determines however, that aircrew effectiveness requires better interfaces between aircrews and automation.

Developers everywhere, not just those in FVL, strive to identify the best methods for transferring tasks between humans and automation. The number of tasks controlled by either human, automation, or a combination thereof, are numerous and vary in complexity. This capstone attempts to address and make recommendations to improve existing methods of task hand-off between humans and automation. This capstone contextually relates to the FVL platform, but the results can potentially be used in other areas as well. The project scope includes different types and levels of automation that are on the market or being developed, critical themes in team interactions and situational awareness, and issues that pilots face with current automated systems.

B. OBJECTIVES

The primary objective of this research is to understand task handoff between human-automation teaming and their impact on the operator's workload. The secondary objectives explore factors of effective and ineffective human-automation teaming, as well as specific hand off methods.

This research will inform the Holistic Situational Awareness- Decision Making (HSA-DM) program office of effective methods of transitioning tasks between humans and automation to provide situational awareness, crew resource management, information processing, and decision making in the FVL initiative.

C. RESEARCH QUESTIONS

Does task difficulty and automation capability impact handoff behavior and operator workload?

1. Does higher levels of automation result in lower workload in Human Automation Teaming (HAT)?
2. Is it more effective to hand off tasks one at a time or bundle them in HAT?
3. Can task difficulty increase human to automation interaction?
4. What are some phenomena that indicate ineffective human automation teaming?

D. THESIS ORGANIZATION

The remainder of the capstone report is organized into chapters to incorporate the topic, problem, and research questions fully. Chapter II provides the literature used to explore the different levels of automation, measures of cognitive workload, and research related to human-automation team performance. The context and surrounding research of Chapters I-II introduce the methodology of the experiment in Chapter III. In Chapter IV, we present the results of the experiment. Finally, Chapter V summarizes the conclusions derived from our analysis of the results.

II. LITERATURE REVIEW

A. INTRODUCTION

Two critical questions in automation that need to be defined to determine the best way to hand off information are: What is the system doing, and how will it respond to particular interventions? The required data to answer such questions will usually be obtainable somewhere in the system, but it can be cognitively difficult to obtain the data (Sarter et al. 1997). A review of research focused on task handoffs between automation and humans is useful in determining the most effective practices. Understanding the operator's requirements, the importance of shared understanding, and what can cause an error in the use of automation is critical to the successful transition of information between humans and automation (Sarter et al. 1997).

The four main themes discussed in this literature review are automation, automation and trust, situational awareness, and automation surprise. These themes are critical to understanding the nature of this problem. To understand task handoff between humans and automation one must first know what automation is. Once an operator knows and understands the automated system the operator can begin to build trust in it. From there the operator and automation can build situational awareness, just like if the automation was another human. Finally, even with all these factors the operator can still experience automation surprise which can lead to disastrous consequences.

B. AUTOMATION

The tasks that humans do today will be carried out by automation in the future. Parasuraman and Riley (1997, 231) refer to automation as "the execution of a function that was previously carried out by a human." Shively (2018) notes that automation has transformed from simple "tools" to intelligent agents expected to function as teammates. He points out that human-automation teaming (HAT) is the idea that human agents are expected to work with automation in a similar method to how human-human teams operate. Sarter (1997) however, states that workload between the automation and operator must be slanted toward the automation, as it is able to handle more tasks. Hari et al. (2020) assess

that human-in-the-loop automation is an effective way to manage workload and task transfer between humans and automation. The automated system analyzes the data, and the human makes informed decisions based on the work the automation has produced. A collaborative effort between humans and automation in a predictive algorithm is needed to succeed, since both have complementary respective strengths and weaknesses.

No matter what the automation is tasked to do, automated systems must be categorized to understand exactly what the operator tasks are and what the automated tasks are. Sheridan (1973) recognizes 10 levels of automation and bases them on how much of the work the automation does and how much the automation tells the human. The Society of Automotive Engineers' (SAE) Taxonomy on Automated vehicles (On-Road Automated Driving (ORAD) committee 2021) defines vehicle automation into six levels. This taxonomy is based on how much responsibility the driver shares with the system and what each agent is expected to do at each level. The taxonomy identifies each agent's tasks as control of the dynamic driving task, object and event detection and response, and fallback in case of an emergency. Taxonomies like these are critical for building trust in automation and avoiding automation surprise, as the operator will know exactly what he/she has to do and exactly what the automation has to do.

C. AUTOMATION AND TRUST

One factor that improves task handoff is trust. This is the case for human-to-human interactions, as well as human to automation interactions. Castelfranchi and Rino (2010) discuss three components of building trust in a dichotomous relationship: an operator's trust in automation to complete a task; the trust an operator has that the control mechanism to intervene in an automation's task will work if the situation arises; and the trust that automation has that the operator will intervene if needed. Hoffman et al. (2013) continue this thought process and identify three different attributes: the relationship between humans and automation as incredibly complex and varies between technology and people; that trust in automation is dynamic and fluid; and, overly complex systems sometimes attract unjustified mistrust due to their complexity.

Even though a pilot trusts the copilot, the pilot must be ready to intervene when the copilot needs help. Operators also must be ready to intervene with automation. Hoffman et al. (2013) developed a new form of trust, negative trust, to explain the attitude toward an expectation of automation glitches and automated tasks operating in a degraded state. They describe negative trust as the expectation that automation will have bugs and require constant monitoring to ensure an operator can develop a workaround. This is like Castelfranchi and Rino's work in that both studies determined the best way to build trust in an autonomous relationship is for human operators to feel comfortable that they can take back control if the situation requires. Rome et al. (2002) identified an earlier version of negative trust by observing that automation, such as autopilot, is often used when the workload on the operator is low. Sarter et al. (1997) noted the same phenomenon by observing when pilots reach cruising altitude and heading in calm weather, the pilots engage the autopilot. When the flight environment is more complex, pilots assume control of the aircraft because they do not feel that the automation device will accomplish the same end state. Colebank (2008) concluded that humans do not trust automation as much as they do other humans, even if the humans are failing. Colebank conducted a "Wizard of Oz" study that showed that when the performance of the human confederate degraded, the participant tried to send more messages to the human, but when the automated confederate degraded, the participant simply internalized the problem and tried to fix it without pushing the automated agent.

Though trust is critical, having too much trust in automation often leads to catastrophe. Parasuraman and Riley (1997) found that, while previous studies showed operators lost faith in automated systems after they failed to complete a task, most operators continued to trust automation even after catastrophic failures. They determined that the ideal situation between a human-automation team is to have mutual control and intervene if the automation has low reliability of completing the task. Stilgoe (2018) backs up this assessment and indicates that some operators, especially those with little to no training, trust automation too much. Sumwalt (2021) credits this phenomenon as one of the primary reasons for a series of Tesla "autopilot" crashes between 2016 and 2021where, "the car

driver's inattention due to overreliance on vehicle automation," led to fatal traffic accidents.

D. SITUATIONAL AWARENESS (SA)

A key area that an automated system must help the operator in is situation awareness (SA). Endsley (1995) assesses that SA is essentially understanding what is happening in the environment and its influences on the present and the future. She offers the three-level model for SA and details these steps as perceiving, comprehending, and predicting. Endsley's model (Figure 1) outlines situational awareness, regarding the state of the environment, as perceived, comprehended, and applied (projected) by the human-agent with factors such as goals, objectives, and expectations in consideration. Flach explains the following on situational awareness:

> [Situational awareness] calls attention to meaning – meaning not in terms of a particular individual's interpretation but in terms of 'what matters' – that is, meaning as a measure of what could or should be known to respond adaptively to the functional task environment. In this sense, meaning is not subjective but can be objectively specified based on normative considerations of the fit or appropriateness of decisions and actions and the demands of the task environment. (Flach 1995, 152)

Researchers use this explanation of SA as a human factor and apply it to their understanding of its place in task handoff and take-back, not only between humans but also between humans and automation. One must ensure that both sides of the task transfer can attain and maintain SA as part of the process.

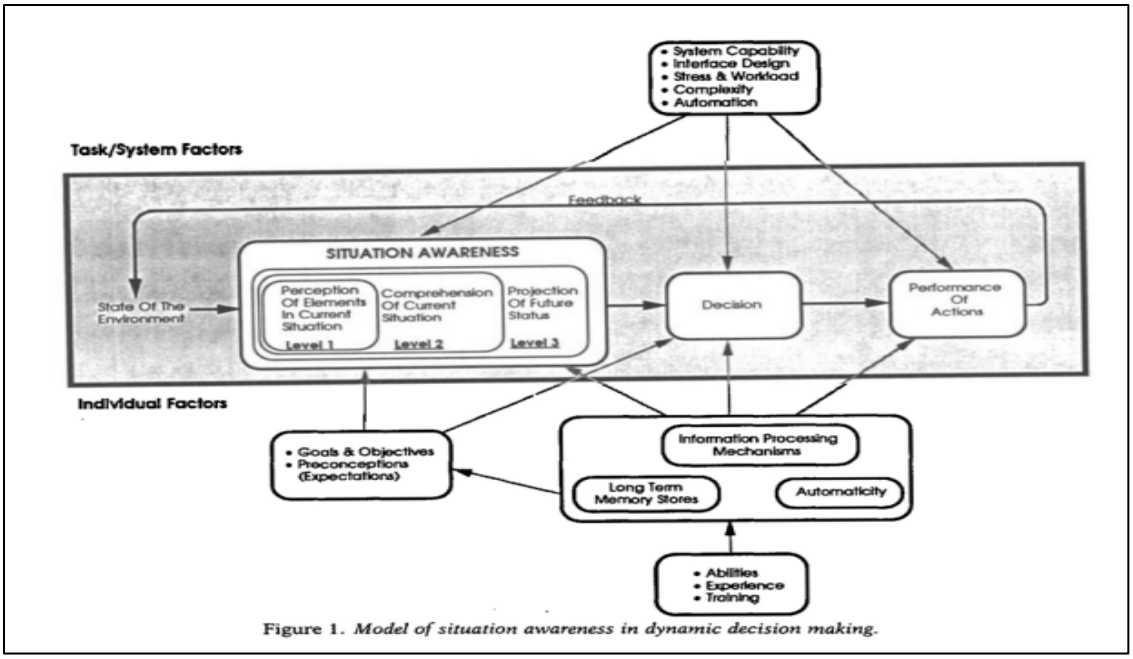


Figure 1. Model of Situational Awareness in Dynamic Decision-Making.
Source: Endsley (1995, 35).

Maintenance of SA is critical to performance. Roseman et al. (2018) steered toward SA as the common denominator for successful task transfer between humans, as well as the facilitation of subsequent decision making. Dekker (2015) speaks to the danger of losing situational awareness and refers to complacency as a catalyst for such an occurrence. He points out that "loss of situational awareness has become the favored cause for mishaps in aviation and other settings" (2015, 159). One can surmise that the utilization of situational awareness, conceptually, has become synonymous with accident prevention while the "loss of situational awareness" is chronically used to assign blame for mishaps. Parasuraman and Manzey further outline complacency as a reason for a loss in situational awareness. They state this phenomenon "results from the dynamic interaction of personal, situational, and automation-related characteristics" (Parasuraman 2010).

Just as each individual has to have SA, the entire team must have SA as well. According to Endsley, "overall team SA can be conceived as the degree to which every team member possesses the SA required for his or her responsibilities…independent of any overlaps in SA requirements that may be present" (1995, 39). The overlaps of situational

awareness between team members facilitate coordination. She suggests that the interdependency of SA in a team, whether it is human-human or human-agent, relies on all members having full situational awareness of their responsibilities, or the team's output will degrade. There are critical areas (the overlaps) that multiple or all members must be aware of for complete team performance. The total overlap point (dead center of this model) is where all the available SA is consolidated and can produce the most optimizing decision-making processes. Figure 2 further illustrates SA interdependency in a team dynamic.

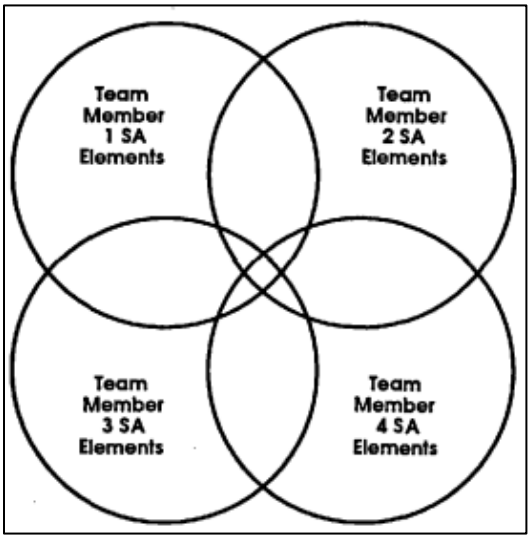


Figure 2. Venn Diagram of Team Situation Awareness. Source: Endsley (1995, 39).

Salas et al. (1995) pose a different model in Figure 3 to illustrate the dynamic of team situational awareness compared to an individual situational awareness model. Figure 3 illustrates fewer factors than Endsley's models; they portray similar constructs, highlighting the presence of pre-existing knowledge, predispositions, interdependence, and goals.

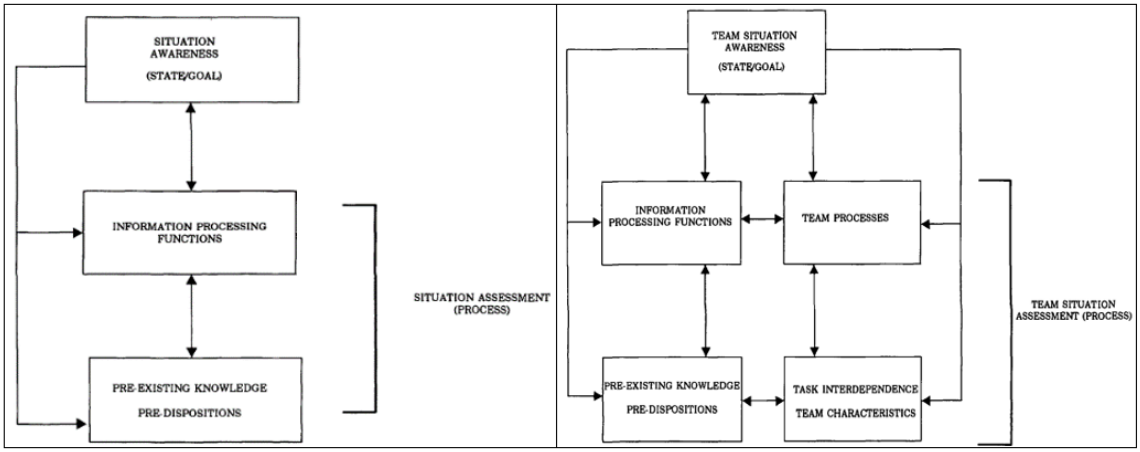


Figure 3. Figure 1: Salas' Models of Individual and Team Situational awareness. Source: Salas et al. (1995, 130).

E. AUTOMATION SURPRISE

One of the most common themes among the current research is a phenomenon called "automation surprise," where the automated device conducts an action without the operator knowing that it has taken place or that the automation can take action, such as powering itself off. The FAA (2013) notes, "A major factor in aircraft incidents and accidents is that pilots fail to keep up with technological changes, resulting in surprise and confusion. Their report suggests that insufficient crew knowledge of the automated systems is a factor in more than a third of the accidents and serious incidents." A sub-theme in automation surprise is pilot confusion (Rankin et al. 2016). If the handoff of information between the pilot and automated system is incomplete, the potential for the pilot to misinterpret the information from the system or to completely miss the information is high. Rankin et al. (2016, 623) notes, "procedures and checklists guide pilots in managing system variations and failures. However, as events unfold, such as multiple disturbances and failures, the complexity of the systems may entail difficulties in identifying subtle cues and isolating failures that, over time, may progress into serious accidents."

Compounding automation surprise is attentional impairment. Dehais et al. (2011) states that attentional tunneling causes operators to overly focus on operating the vehicle resulting in missed error or warning messages. In the Dehais experiment, the operator was given a task to perform in an automation supervisory role. During the vehicle's mission,

the display would inform the user that the battery was low on the vehicle and the car had to return to base to charge. Many operators missed the cue and overrode the vehicle to complete the mission despite the automation warnings. A similar tunneling event while flying an aircraft could lead to a collision or crash if the pilot missed the automated warning. Semmens et al. (2019) conducted a study to determine when drivers may have attentional impairment. Drivers drove a car for 45 minutes. As they drove, the car asked the driver if it was a good time to receive information. They found that steering wheel angle when turning the car and change in brake oil pressure were good predictors of the driver's desire to want to receive information. There may be similar, relatively simple, indicators of aircraft state that may help an automated system determine when is a good time to provide information to the pilot.

If an automated system is poorly designed, the pilots will not only miss the warnings, but they will also misinterpret the information presented or not receive the information at all. Pizzoli et al. (2014) states, "Automatic transitions may be notified to the pilot (via visual or aural feedback) or may be hidden. In the case of a poorly designed feedback, or because of a gap in the attention of the pilot, the feedback is likely to be missed: the transition is unseen." To test automation confusion, the pilots were given three situations where any one of the confusion events could occur. Only a small percentage of the pilots were able to catch the errors the autopilot was making. The most significant number of pilot errors were made when the autopilot chose without the pilot knowing the choice was made or that the autopilot could make a choice. Another contributing factor to missed alerts is the amount of interaction the pilots must have with the Flight Control Unit (FCU) that detract them from their ability to monitor what operations the autopilot is managing (Rome et al. 2002). During the test, an auditory warning was proven to effectively get the pilot's attention of a change to what the autopilot was managing. Regardless of the type of visual signal, the operator would miss the information that the automated system was trying to present.

F. SUMMARY

As the demand for a new generation of vertical lift systems rises, automation in "self-driving" cars soar, and higher output of task execution becomes more prevalent in our daily activities, uncertainties of the unknown increase. The delegation of tasks between humans and automation is critical in successful team cognition, mental model, and situational awareness. Although scholars and researchers have discovered methods to improve effectiveness to better interface humans and automation, there are still concerns. The problem is task handoff between humans and automation requires significant training and, in the absence of it, could lead to automation surprise that potentially results in fatal accidents.

Research suggests that, though an operator's workload might decrease with higher levels of automation, the operator's mental workload does not, as the operator still has to track what is happening around him or her, as well as what the automation is doing. Also, even though automation might be able to accomplish a majority of the tasks, an operator might not trust it to. Based on previous research, this thesis study hypothesizes that, as the scenario difficulty and automation level increase, humans will be more likely to handoff tasks to the automation confederate.

THIS PAGE INTENTIONALLY LEFT BLANK

III. METHODOLOGY

A. INTRODUCTION

The project intended to inform the Holistic Situational Awareness-Decision Making (HSA-DM) program office of effectively transitioning tasks between humans and automation for implementation in the Future Vertical Lift (FVL) initiative. Our goal is to measure the correlation between task difficulty and automation level during task hand offs, utilizing a "Wizard of Oz" type approach and the C3 Fire Program. At the end of the research, all subjects were informed of the deception through email.

B. STUDY PARTICIPANTS

In accordance with the approved Intuitional review Board protocol, all participants signed a consent form informing them of their rights as volunteer participants in the experiment. The research team solicited participants through emails, posters, flyers, and personal contacts of Naval Postgraduate School's faculty and students. The 20 participants who enrolled in the study were male and female military officers and NPS civilian employees, ages 25–55.

The research called for a participant to be the operator of the C3Fire program. The participant was tasked to fight the fire using the assets provided within the program. To help the operator, the research team provided an "automated agent," which conducted various tasks within C3fire. The "automated agent" was a research team member in an adjacent room, who served as a confederate for the study. In addition, the participants and the confederate communicated through a text chatbox in the program.

C. RESEARCH STUDY DESIGN

Our study design centered around a "Wizard of Oz" type study, in which NPS students and faculty used the C3Fire program within the Glasgow Hall HSI laboratory. We utilized a "Wizard of Oz" experiment to have more control over the levels of automation and provide a means to measure the effectiveness of task transfer. In a "Wizard of Oz" design a participant is told that they are working with a program to accomplish tasks, when

they are actually working with a human confederate in another room. This design does three things: first, it allows researchers to conduct studies without having to program or code advanced AI or software. Second, our research suggests that working with new types of automation requires extensive training for operators to use, but this design allows participants to skip that training, as the human confederate will understand what the participant is trying to do better than a rudimentary software program. Finally, our research also suggests that humans work differently with automation than they do with other humans, so for the purposes of the Future Vertical Lift (FVL) Program we needed to ensure that participants thought that they were working with automation.

The Wizard of Oz method relies heavily on effective subject deception in that they believe they are working with newly developed software to aid in task management. The deception is necessary to gain valuable data from the exercise. In addition, the subjects received the opportunity to request their data to be removed from the research.

The study used the C3Fire platform for four primary reasons. First, the program was successfully used to test human team performance and HAT in previous studies (Colebank 2008). Second, the program was readily available within the NPS Human Systems Integration (HSI) Laboratory with subject matter experts on the faculty staff. Third, the program allowed the research team to solicit participation from all NPS students and faculty compared to using a realistic aircraft simulator, limiting the participant pool to rated pilots. Finally, the C3Fire program enabled the research team to more broadly observe how humans hand off tasks to automation and take them back. Developers of C3Fire created the software program to specifically study task handoff between humans. Researchers for this capstone adopted C3Fire to study allocation of tasks between a human and a confederate acting as an automated agent. The C3 Fire Program is a computer based microworld, in which a team of human agents direct firefighting assets to extinguish a forest fire. The interface consists of a map with generic icons that represent the assets available to the team. The user controls the assets by first clicking the icon and then on the location where the user wants the asset to go. The firetrucks extinguish the fire while water trucks refill the fire trucks. Figure 4 shows a snapshot of the C3Fire interactive map.

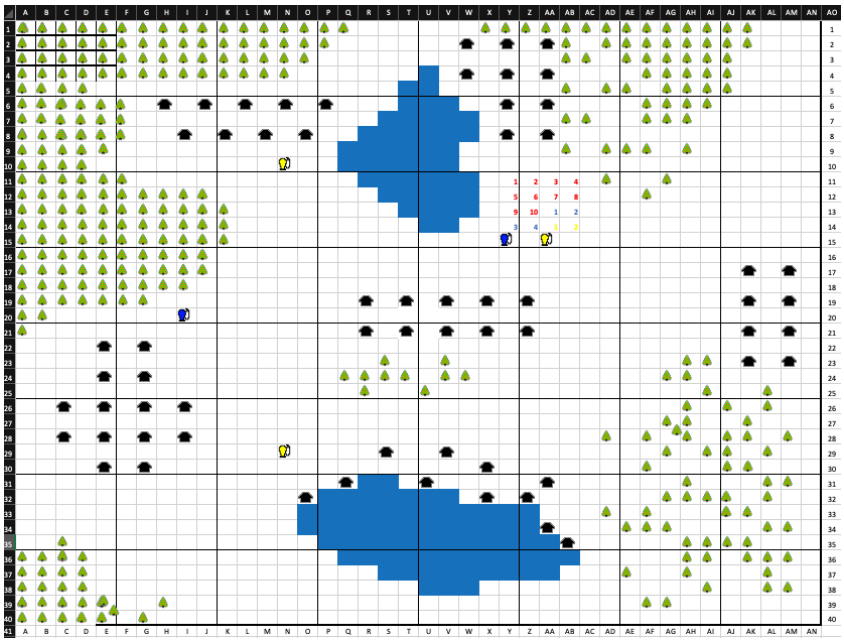


Figure 4. Snapshot of the C3Fire map

Each scenario was ten minutes long. The independent variable in each scenario was the level of automation. To create each level of automation, the team utilized Sheridan's Levels of Automation, seen in Table 1. The research team ultimately decided to base the automation levels on Sheridan's levels. The researchers' low-level automation is based on Sheridan's level one automation. It was only able to refill one fire or water truck at a time, and could only acknowledge commands. The mid-level automation was based on Sheridan's level three automation. It could refill two trucks in a row and notified the participant when a fire truck or water truck was low on water. Finally, the high-level automation was based on Sheridan's fifth level. It could refill three trucks in a row and it asked permission to refill subsequent trucks, to which all the participant had to do was reply with "y." The map, locations of assets, and the fires were in the exact place for each scenario. The only change between the scenarios is that the map was rotated 90 degrees, and the level of automation increased.

Table 1. Sheridan's Levels of Automation

Level of Autonomy	Explanation
Level 0	The computer offers no assistance: human must take all decisions and action
Level 1	The computer acquires the data from the process, and registers them without analysis (1* new level)
Level 2	The computer offers a complete set of decision / action alternatives, or
Level 3	The computer narrows the selection down to a few
Level 4	The computer suggests one alternative
Level 5	The computer executes that suggestion if the human approves, or
Level 6	The computer allows the human a restricted time to veto before automatic execution, or
Level 7	The computer executes automatically, then necessarily informs the human, and
Level 8	The computer informs the human only if asked, or
Level 9	The computer informs the human only if it, the computer, decides to
Level 10	The computer decides everything, acts autonomously, ignoring the human

For this study, NPS users used C3Fire and a chat box to communicate to the research team's confederate, who represented an advanced automation program. The confederate's computer was connected by Local Area Network Switch (LAN) from a desk in another room. During the exercise, the participants completed four scenarios (one training and three data collection scenarios) within C3Fire. During each scenario, the researcher queried the participant on their Continuous Subjective Workload Analysis Graph (C-SWAG) every minute to gauge how cognitively engaged the participant was at the point of time. After each scenario, the research team disseminated a survey to the participants to query their interaction with automation, including how they experienced handoffs, how they perceived of cognitive overload, and how training impacted their ability to perform their tasks. Table 2 shows the basis for how the human operator (NPS Student participant) communicated with the confederate.

Table 2. Script for Automation Confederate

Scenario	1	2	3
Automation Sophistication	Low	Medium	High
Sheridan Level of Automation	One	Three	Five
H to C Task handoff: Refill Fire Trucks	w# refill f#	w# refill f#	w# refill f#
H to C Task handoff: Refill 2 Fire Trucks	n/a	w# refill f#, f#	w# refill f#, f#
H to C Task handoff: Refill 3 Fire Trucks	n/a	n/a	w# refill f#, f#, f#
H to C Task handoff: Water Truck Refill	w# refill	w# refill	w# refill
C to H Task Acknowledgement of Task	acknowledged	acknowledged	acknowledged
C to H Task Notification:	n/a	f# is approaching empty	f# is approaching empty
C to H Task Request:	n/a	n/a	f# is approaching empty, refill? (y or n)

Key:	
H	Human
C	Confederate
F	Fire Truck
W	Water truck
#	number of asset

D. MATERIALS AND EQUIPMENT

1. C3Fire Software version 3.4.1.0: C3Fire presents a simulated environment that permits collaboration and controlled studies of cooperation and coordination in a dynamic environment. The program generates a task environment that consists of complex, dynamic, and opaque characteristics. We used C3Fire Software version 3.4.1.0.
2. 3 x Alienware Laptop Computer 2020 Alienware m15 R3 i9-10980HK 32GB 1.5TB SSD 15.6” FHD RTX 2070 SUPER Dark (Human Agent, Server, Confederate)
3. Separate windows-based Dell Laptop for survey completion
4. Mouse
5. Attached laptop Keyboard
6. Empatica E4 Heart Rate Monitor
7. 8 port LAN server LNX-800A
8. 2 x 10-foot ethernet cables
9. 1 x 40-foot ethernet cable (confederate laptop to server)
10. 1 x 6 Port electrical power strip

All equipment and materials listed above were arranged as shown in Research Lab Layout in Figure 5.

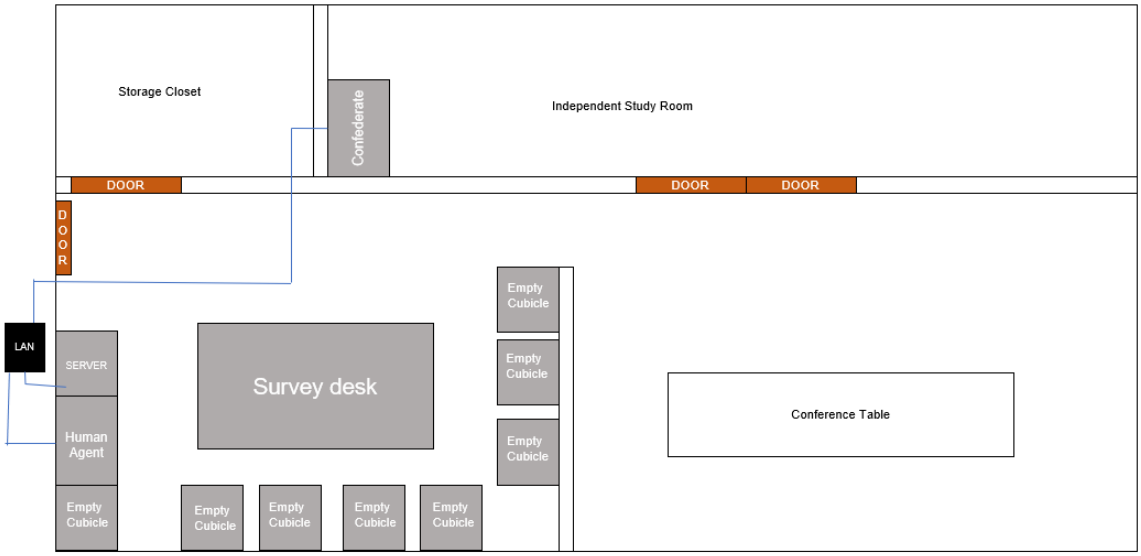


Figure 5. Research Lab Layout

E. VARIABLES

1. Independent Variables.

The independent variable in this study is the level of automation (low, medium, high). The scenario was designed to be challenging as a forcing function to motivate the participants to use the "automation agent." All participants interacted with all three levels of automation while attempting to contain the fire in three separate scenarios. The levels of automation were purposely not randomized between participants.

2. Dependent Variables

The dependent variables consisted of of NASA TLX measures, heart rate monitoring (Empatica E4 wearable device, C-SWAG, C3Fire program metrics (area burned, common inefficiencies), and character count when communicating with automation. Each measure evaluated participants behaviors in relation to the different levels of automation.

F. SEQUENCE OF EVENTS

Upon arrival, each participant filled out a consent form and completed a survey, found in Appendix A, querying their experiences with automation. Then they watched a five-minute video explaining the purpose of the research, capabilities of the automation (research confederate), and their objectives during the scenarios. During training, the participants were provided with an Empatica heart rate monitor to gauge stress during the scenarios. They received a briefing and examples of the cognitive workload questions that were asked throughout the scenarios. Next they received 10 minutes of training on C3Fire, in which each participant:

1. received instructions on the functions of C3Fire available to them
2. received instructions on the movement of assets to fight the fire
3. practiced task handoff with the automation, and
4. received instructions on general strategies for firefighting.

Once the practice scenario was complete, the participant completed the three exercise scenarios. Scenario one utilized the low level of automation, scenario two utilized the middle level of automation, and scenario three utilized the high level of automation. Each participant conducted the scenarios in this order. The data collected consisted of heart rate, C-SWAG, percentage of fire extinguished, number of tasking errors, number of tasks handed off, and survey data. The participants were asked to report their cognitive load each minute during the training and scenarios on a scale of 1 to 100:

- A C-SWAG of 1 means that the participant is not cognitively active and is most likely bored.
- A C-SWAG of 50 means that the participant is cognitively active but can handle more tasks.
- A C-SWAG of 100 means that the participant is experiencing cognitive overload and cannot handle any more tasks.

After each scenario, the participants filled out a survey to gauge their experiences with each level of automation. At the conclusion of the exercise, the participants filled out a post-exercise survey to query their interaction with the automation, to include handoffs and perception of cognitive overload. The success of the research depended on participants' belief that they were interacting with an automated agent. After the study, the participants were informed of the role of automation played by a human. In addition, the investigators confirmed that the participants continued to consent for their data to be used in the study. All surveys can be found in Appendices A through E. The results of the surveys are compiled in Appendix G. The task breakdown is shown in Table 3.

Table 3. Subject Task Breakdown

Task Number	Task	Time (Mins)
1	Receive brief from the Research Team on the exercise, Obtain Consent	5
2	Complete Automation survey	5
3	Receive C3Fire Training	15
4	Complete C3Fire Scenario 1	10
5	Complete C3Fire Scenario 2	10
6	Complete C3Fire Scenario 3	10
7	After Action Survey	10
8	Debriefing Email	5
Total		70

IV. RESULTS

A. OVERVIEW

Researchers found that participants did not send more messages to the confederate across all three scenarios. However, researchers found that participants handed off more tasks to the confederate in scenarios 2 and 3 compared to scenario 1. The difference in the number of tasks handed off to the confederate in scenario 2 was not significantly different in scenario 3. No matter the level of automation sophistication (independent variable), participants reported an increase in cognitive workload across all scenarios.

B. C3 FIRE CHAT LOG ANALYSIS

The C3Fire log files contain the messages exchanged between the participant and the confederate. The research team counted every time the participant sent a message throughout each scenario. The research team then analyzed the content of the messages sent from the participant to the confederate in order to determine the number of tasks handed off. In scenarios 2 and 3, the participant had the option of sending two or three tasks in one message. For example, in scenario 1, participants can only hand off one task to the confederate per message, such as to refill one fire truck or order a water truck refill itself. In scenario 2, the participant could request the confederate to assign a water truck to refill two fire trucks in one message. In scenario 3, the participant could request the confederate refill three fire trucks in one message and reply with a “y” or “n” in response to the confederate requesting to complete a task. The number of messages and tasks handed off by each participant are displayed in Table 4.

Table 4. Task Handoff from Participant to Automation Confederate

	Tasks Handed Off to Automation					
	Scenario 1		Scenario 2		Scenario 3	
Participant	Messages	Tasks	Messages	Tasks	Messages	Tasks
(un-cued) Study 1						
1	13	13	12	18	15	18
2	11	11	16	16	14	20
3	7	7	11	11	6	6
4	19	19	7	19	15	30
5	13	13	11	15	15	19
6	7	7	15	15	9	10
7	18	18	18	26	23	35
8	17	17	11	15	15	22
9	18	18	16	19	18	27
10	17	17	18	18	18	31
(cued) Study 2						
11	12	12	15	15	17	17
12	19	19	17	24	15	21
13	12	12	14	14	17	17
14	6	6	6	8	11	11
15	14	14	14	16	11	10
16	14	14	19	19	18	24
17	20	20	22	22	22	30
18	15	15	18	18	12	22
19	18	18	19	24	21	26
20	12	12	11	17	11	11
Median	14	14	15	17.5	15	20.5
Tasks per Minute		1.4		1.75		2.05

1. Messages Sent to Automation

There was no statistically significant difference between the number of messages sent to automation. The number of messages sent was not normally distributed, evidenced by an Anderson-Darling normality test (p=.1078). In response, the researchers conducted a non-parametric equivalent to a repeated measures ANOVA, the Friedman test, on the 20 participants to determine a statistical difference in the number of messages passed to automation across the three scenarios. Scenario 1 had the fewest tasks handed off to automation (median = 14), and Scenarios 2 and 3 had the most (median = 15). The test revealed that there was not a statistically significant difference ($Q=1.75$, $p > .416$).

2. Tasks Handed off to Automation

While there was no difference in the messages sent to automation across scenarios, there was a difference in the number of tasks handed off to automation across scenarios. Again, the data was not normal, confirmed by an Anderson-Darling normality test with a

p-value of .155. Therefore, the researchers then conducted a non-parametric Friedman test on the 20 participants to determine if there was a statistical difference in the number of tasks passed to automation across the three scenarios. Scenario 1 had the fewest number of tasks passed off to automation (median = 14), and Scenario 3 had the most (median = 20.5). The test revealed a statistically significant difference ($Q=49.668$, $p < .001$). Researchers conducted a post hoc analysis using pairwise Wilcoxson signed ranked tests, with Bonferroni correction for multiple comparisons, to determine which scenarios were different from each other. The tests revealed that the number of tasks passed off in Scenario 1 was significantly different from the number of tasks handed off in Scenario 2 ($p < .001$) and in Scenario 3 ($p < .001$). However, there was no statistically significant difference in the number of tasks passed off between Scenarios 2 and 3 ($p = .20$).

Overall, the analysis assesses the effect of scenario (1, 2, 3) and study type (cued and un-cued) on the number of tasks handed off to automation. Scenario was a statistically significant explanatory variable but not study type. Post-hoc analysis showed that the more sophisticated the automation, the more tasks were handed off to automation. Figure 6 depicts the tasks handed off to automation by scenario and by study. The statistically significant difference between scenarios 1 and 2 indicates task bundling leads to more tasks being handed off to automation. But the statistically insignificant change between scenarios 2 and 3 indicates that more bundling does not lead to more tasks being handed off to automation.

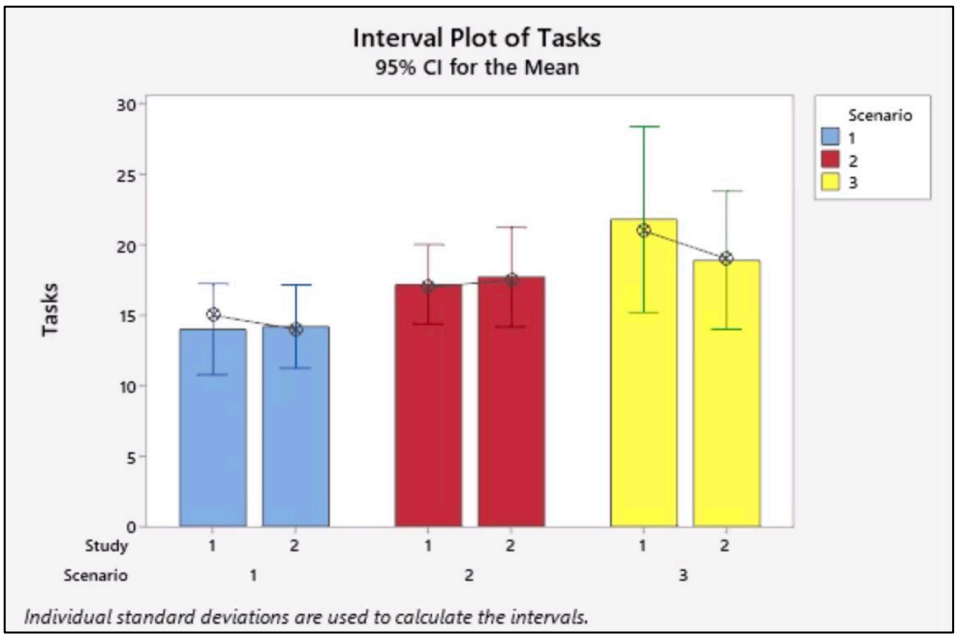
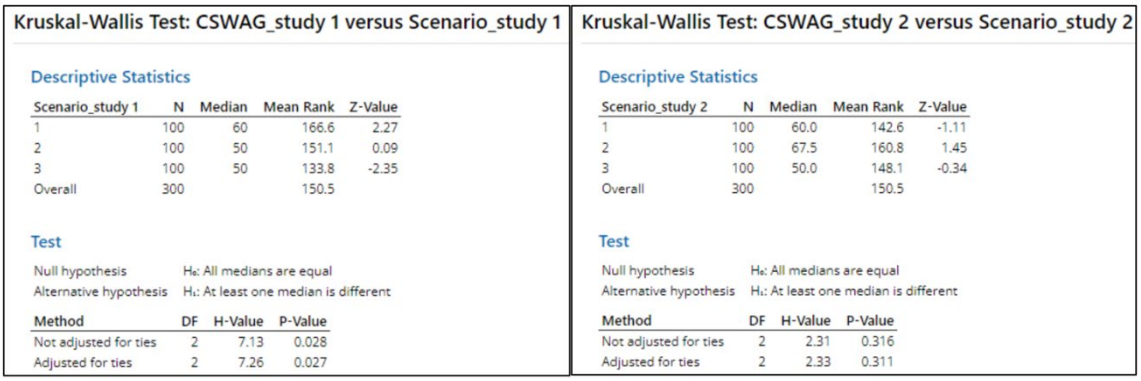


Figure 6. Interval Plot of Tasks Passed to Automation

C. COGNITIVE SUBJECTIVE WORKLOAD ASSESSMENT GUIDE (CSWAG) RESULTS

CSWAG reports in study 1 contained more variability due to some participants not noticing that the notification for the second fire at minute 3. Researchers, therefore, could not conclude if levels of automation or scenario difficulty contributed to differences in changes in CSWAG across scenarios. However, all participants in study 2 noticed or were alerted to the second fire and therefore experienced a similar scenario complexity. Study 2 participants, who reported a consistent rise in cognitive workload with no difference across scenarios, indicate that the level of automation had no significant effect on cognitive workload. The CSWAG results for scenarios 1, 2, and 3 were significantly different in study 1 but were not statistically different in study 2. A Kruskal-Wallis test for study 1 resulted in a P-value of .028, indicating that at least one scenario's CSWAG scores were significantly different. The Kruskal-Wallis test for study 2 resulted in a P-value of .311, indicating that the CSWAG scores remained unchanged between scenarios. Figure 7 depicts the Kruskal-Wallis tests for both study 1 and study 2.



Kruskal-Wallis Test: CSWAG_study 1 versus Scenario_study 1

Descriptive Statistics

Scenario_study 1	N	Median	Mean Rank	Z-Value
1	100	60	166.6	2.27
2	100	50	151.1	0.09
3	100	50	133.8	-2.35
Overall	300		150.5	

Test

Null hypothesis	H_0: All medians are equal
Alternative hypothesis	H_1: At least one median is different

Method	DF	H-Value	P-Value
Not adjusted for ties	2	7.13	0.028
Adjusted for ties	2	7.26	0.027

Kruskal-Wallis Test: CSWAG_study 2 versus Scenario_study 2

Descriptive Statistics

Scenario_study 2	N	Median	Mean Rank	Z-Value
1	100	60.0	142.6	-1.11
2	100	67.5	160.8	1.45
3	100	50.0	148.1	-0.34
Overall	300		150.5	

Test

Null hypothesis	H_0: All medians are equal
Alternative hypothesis	H_1: At least one median is different

Method	DF	H-Value	P-Value
Not adjusted for ties	2	2.31	0.316
Adjusted for ties	2	2.33	0.311

Figure 7. Kruskal-Wallis Test for Study 1 vs. Scenario (1,2,3) and Study 2 vs. Scenarios (1,2,3).

Researchers observed some subjects who overlooked the second fire in study 1 reported much lower CSWAG scores near the end of the scenario. The split population in study 1 of individuals who did and did not know there was a second fire helps explain a statistical difference among scenarios. Researchers fit a logarithmic model to each study. Researchers chose a logarithmic model because as the number of tasks increased, Researchers suspected that the workload increased logarithmically. The increase in workload from performing one task to two tasks is greater than the increase in workload from two tasks to three tasks. The resulting logarithmic models and corresponding R2 for each study and scenario are displayed in Figure 8. Note that in study 1, scenarios 2 and 3, several participants reported decreasing workloads at minutes 5 through 10. These participants did not notice that there was a second fire. In study 2, participants were made aware of the second fire at minute 5 if participants did not notice the second fire earlier. Those that did notice the second fire in study 1 are indicated with a blue line.

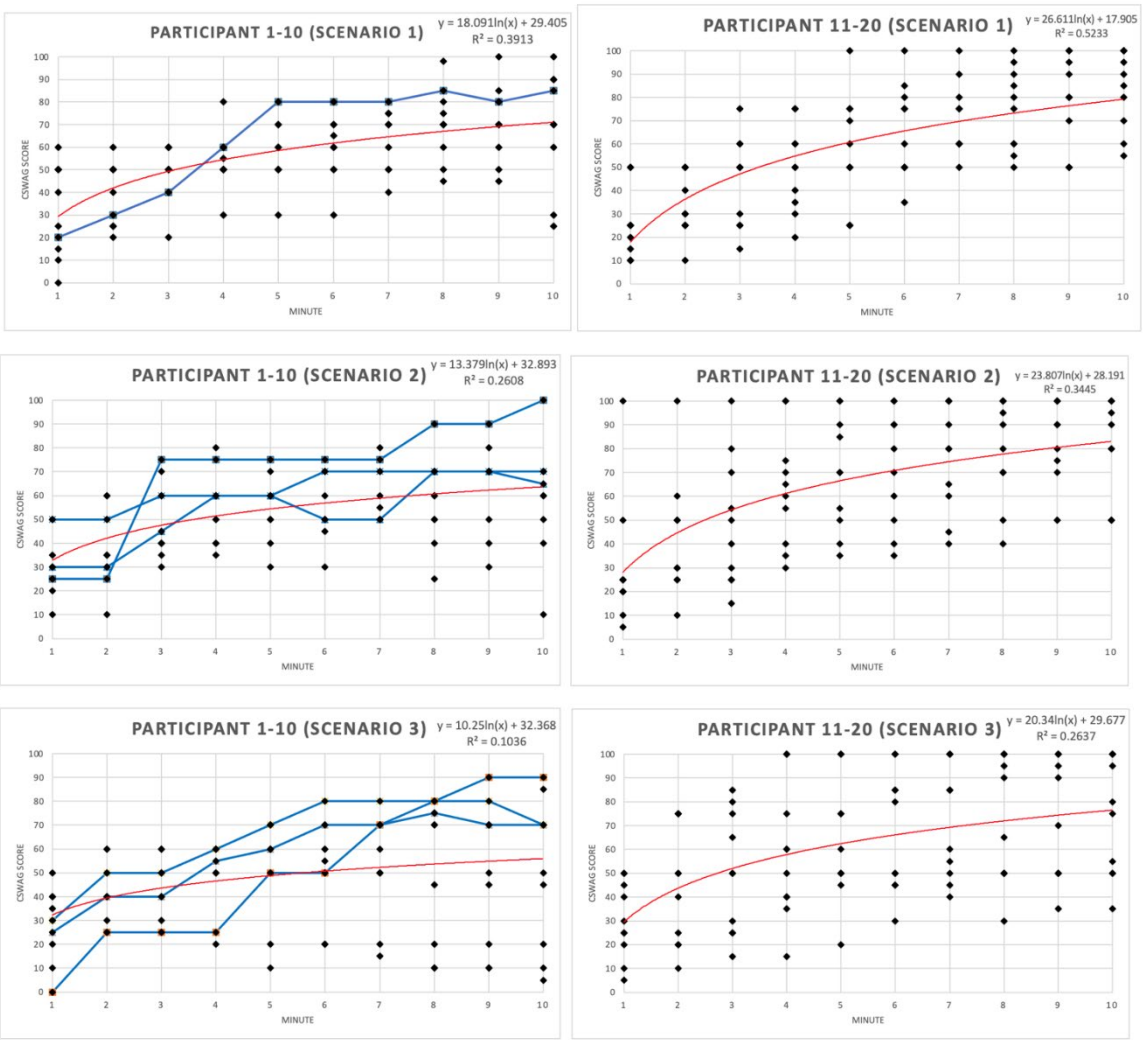


Figure 8. Study 1 and Study 2 and Associated Logarithmic Models and the Associated R Square. Blue Lines Indicate Participants Who Noticed the Second Fire in Study 1.

When the logarithmic models are overlaid on the same graph, researchers observed a decreasing CSWAG score between scenario 1, scenario 2, and scenario 3 for study 1. Study 2 displayed a decreasing score from scenario 2, scenario 1, and scenario 3. However, the range of R2 values for the mathematical models in study 1 was .10 - .39, while study 2 was .26 - .52. Thus, overall, the trends in CSWAG in study 1 were different across scenarios but with low accuracy and study 2 appeared to show very little difference across scenarios with relatively higher accuracy. The model comparison is demonstrated in Figure 9.

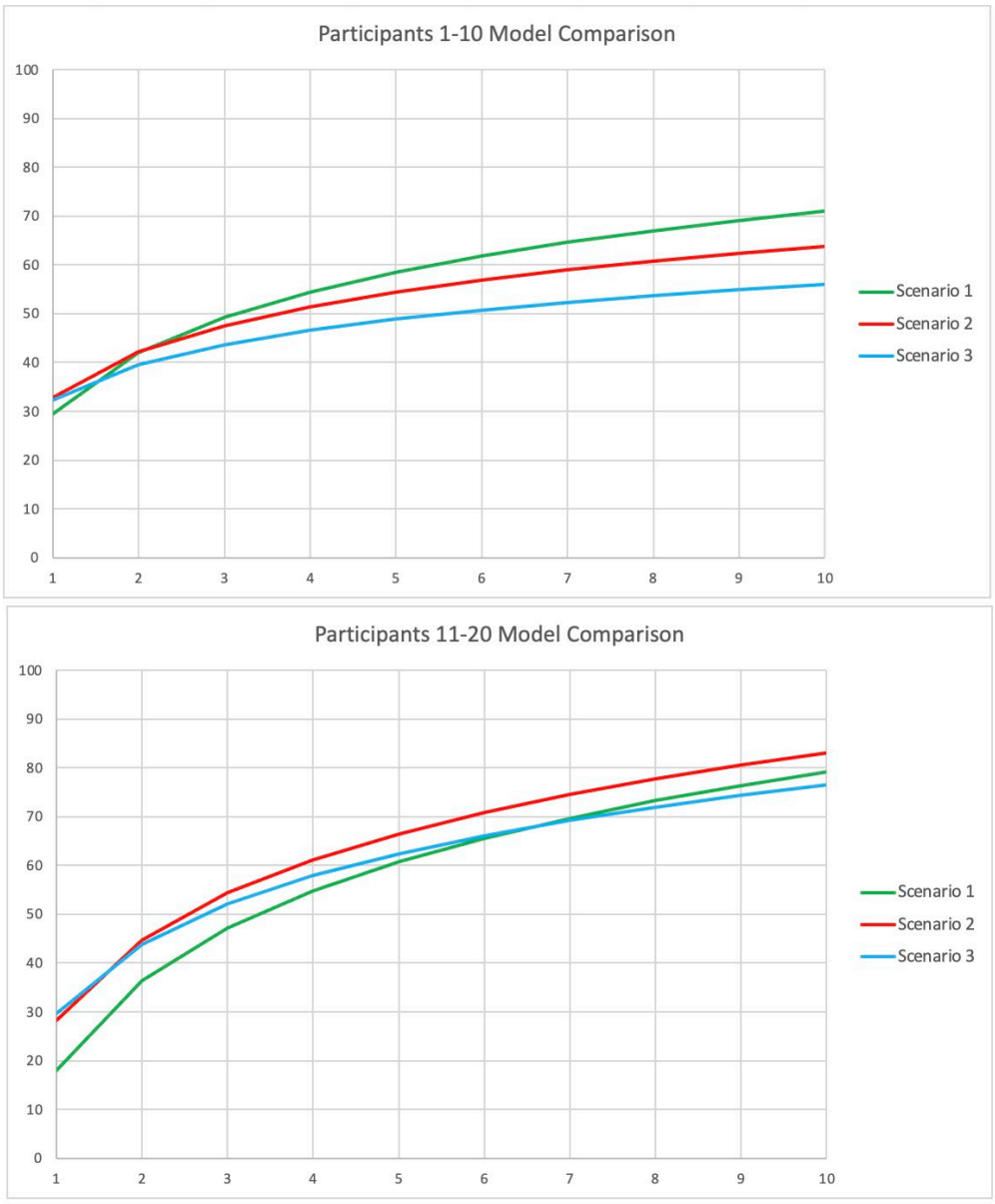


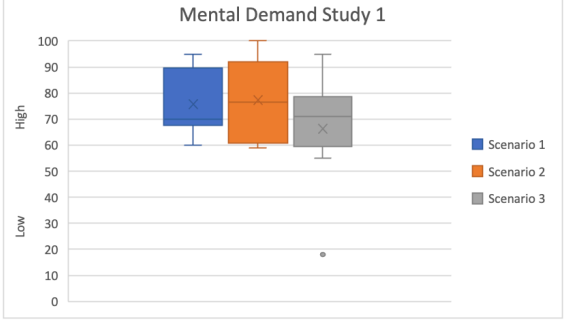
Figure 9. Comparison of CSWAG models for both study 1 and study 2

D. NASA TLX RESULTS

The NASA TLX results failed to depict statistical differences between scenarios in all the studies across each measure. However, while the mental demand question was not statistically different, there was a very slight decrease throughout the scenarios, as depicted in Figure 10. The remainder of NASA TLX is displayed in Appendix B.

Mental Demand Study 1			
Participant	Scenario 1	Scenario 2	Scenario 3
1	70	100	61
2	70	91	71
3	92	73	80
4	70	61	55
5	60	60	71
6	81	79	73
7	95	95	95
8	89	74	78
9	60	59	61
10	70	81	18
Median	70	76.5	71

Mental Demand Study 2			
Participant	Scenario 1	Scenario 2	Scenario 3
11	90	70	70
12	45	45	30
13	73	75	78
14	85	90	60
15	100	100	100
16	65	75	81
17	51	70	75
18	71	81	61
19	81	70	59
20	100	100	100
Median	77	75	72.5



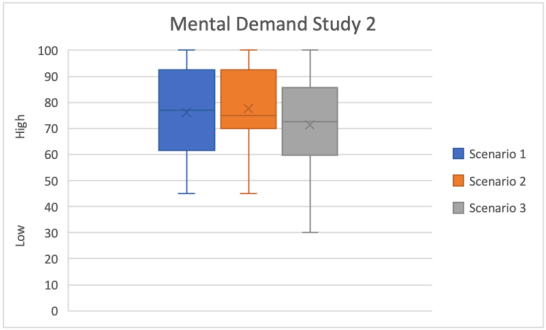


Figure 10. Mental Demand NASA TLX (Study 1 and Study 2)

The most meaningful metrics collected from this study were the C3 Fire Chat log analysis, Continuous- Subjective Workload Assessment Guide (CSWAG) scores, and NASA TLX. Researchers collected several additional metrics for this study, but analyses revealed no statistically significant results, which can be found in Appendix C. The heart rate variability for each study failed to show any collective rise or fall in workload that indicated a significant difference between scenarios or studies. NASA TLX failed to show statistically significant differences in any of the six parameters but did show a collective drop in mental demand.

V. DISCUSSION

A. PURPOSE

The study's primary purpose was to evaluate how task difficulty and automation capability impact handoff behavior and operator workload. The trends and observations from the study are intended to inform designers of the U.S. Army's Future Vertical Lift systems in developing more efficient HAT. Using a Wizard of Oz study to replicate varying levels of automation, researchers developed three ten-minute scenarios in a software-based team evaluation program called C3Fire. The participant population consisted of mid-grade (O3-O4) officers in the Army, Navy, and Marines pursuing graduate-level degrees at the Naval Postgraduate School. None of the participants reported any familiarity with C3Fire, nor did their responses regarding the frequency of use and/or trust in automation differ significantly.

B. HYPOTHESIS ONE

The higher the level of automation sophistication, the more tasks participants will handoff to the confederate

This research study showed that participants transferred statistically more tasks to automation as automation became more sophisticated, which supports the hypothesis that the levels of automation bear an effect on human behavior. Researchers defined "behavior" as the number of tasks handed off between participants and automation. Using Sheridan's levels of automation (Sheridan, 1978), researchers adapted the automated confederate's capabilities to increase the ability to transfer tasks with each scenario. Participants used the automation commands in higher levels of automation to bundle tasks into single messages. While the number of messages each participant sent did not significantly differ between scenarios, the number of tasks rose significantly between scenarios 1 and scenarios 2 or 3. The significant rise in task handoff was due to the participants' ability to bundle tasks. Participants used the capability to transfer more tasks, with a smaller handoff cost, in a standardized timeframe.

C. HYPOTHESIS TWO

The higher the level of automation sophistication, the less cognitive workload participants will report.

This study was inconclusive on the hypothesis that the level of automation would impact a human's cognitive workload. Based on the first ten participants, the analysis suggests that the average cognitive subjective workload assessment decreased as the automated capabilities increased. However, during study 1, only 40% of the participants recognized the emergence of a second fire. Those who did not notice the second fire in study 1 reported very low workload scores later in the scenario due to them not noticing automation telling them there was a second fire. The split population in study 1and subsequent analysis showed that the logarithmic models point to a difference in behavior via the unexplained variance. Therefore, the mathematical models could only be used as a general indicator. Participants 11–20 all were notified that there was a second fire due to the addition of the proctor ensuring that participants checked the log at the 5-minute mark. The mathematical models based on the CSWAG scores reported by participants 11–20 were more accurate (R2 of .52-.26) but showed no statistical difference in CSWAG between scenarios.

The research team noticed that study 2 participants reported a higher median CSWAG in scenario 2 (67.5) versus scenario 1 (60), and 3 (50). One potential interpretation comes from the experimenter's observation. In scenario 1, participants only had to send messages to the automation. In scenario 2, they had to send messages, but then also monitor the chat window for additional information. In scenario 3, they may have realized that they did not need to monitor the chat as closely because the experimenter would alert them to the second fire. While the research team cannot confirm this hypothesis, it does provide a potential explanation for the increase in scenario 2, study 2, which was not evident in study 1.

In addition to the CSWAG observations, the research experiment included post-scenario surveys based on the NASA Task Load Index (TLX) to enhance subjective workload assessments between participants. The NASA TLX results were not statistically

different between scenarios or studies, but the mental demand highlighted an indication that supports the hypothesis. Mental demand slightly decreased in scenario 3 in both pools of participants. While not statistically different, the decrease would support the hypothesis that the higher level of automation may decrease mental demand. However, such as the case of CSWAG, the decrease may be an indicator that participants were learning how to use automation and not because the level of automation was relieving them of temporal demand.

D. LIMITATIONS

By design, this research study did not account for learning effects. Researchers purposefully did not randomize the levels of automation throughout the participants. There was a high possibility for incoherent behavior if participants experienced a confederate with a high level of sophistication followed by a low level. Researchers mitigated the incoherent behavior by consistently allowing participants to experience a steadily increasing amount of automation. As a result, researchers could not fully define the benefits of automation due to learning being a confound. Learning particularly confounds any sort of downward trend in the reported CSWAG due to automation. Regardless, researchers still detected a benefit of automation in the number of tasks handed off to automation.

E. OTHER FINDINGS

The research team found several significant relationships that did not support the hypothesis; however, they warrant further discussion. First, the most surprising finding was the amount of suspected cognitive tunneling that occurred. Researchers found that the design of the communication interface with automation, chatbox with no sounds or visual cues, was so ineffective in creating real two-way communication that most participants would not read it while attempting to extinguish the fire. When individuals felt that they extinguished the fire because they did not see a second fire because they missed the automated message, their CSWAG reports decreased significantly. However, with study two, there was a consistent and almost predictable model of CSWAG reporting, which was an increasing logarithmic scale.

The research team implemented a control input in the design of the exercise when participants did not notice the second fire in study 2. The proctor alerted participants to check the chatbox if the participant did not notice the notifications from the confederate reminding the human of the second fire. In such instances, this strategy was appropriate whenever the participant overlooked the critical task of reading a fire notification. The design of the experiment can be considered as a safety measure that automatically activates to notify the participant of the second fire. A "touch on the shoulder," metaphorically speaking for a fixed application of automation, has been proven in studies to alleviate workload (Parasuraman & Riley, 1997).

F. CONCLUSIONS AND RECOMMENDATIONS

The primary research objective of the study was to analyze the relationships between objective and subjective workload measurements and the participants' behavior when working with different levels of automation sophistication. The Heart Rate Variability (HRV) is used for the objective analysis, whereas the CSWAG and NASA-TLX questionnaires are the tools selected for the subjective workload evaluation. Unfortunately, the only objective workload analysis, Heart Rate Variability, failed to show any significant change in workload between scenarios or studies. The CSWAG and NASA-TLX showed no significant difference between scenarios but did indicate a slight decrease. Participants took advantage of the different levels of automation and statistically significantly increased their rate of tasks per minute in each scenario. Study results show that perhaps the sophistication in automation levels is less important when attempting to lower user cognitive workload. Making minor changes to automation and building small efficiencies may be dwarfed by interface design and task difficulty. Even though users took advantage of bundling tasks while communicating with automation, their workload and performance remained almost constant throughout the exercises.

G. FUTURE RESEARCH

To further explore the hypothesis of whether the level of automation impacts human behavior and cognitive workload, researchers developed a few observations that improve

upon the utilized methodology. Incorporating some or all these recommendations in future research could yield more supportive findings.

Researchers recognized that participants experienced cognitive tunneling and missed or ignored automated messages regarding a second fire. This cognitive tunneling created an unanticipated confound in the experiment, splitting workload assessments between participants aware of the second fire and those who were not. There are a few ways to address this, and the researchers adapted for the second half of participants by implementing additional, automated messages (from the confederate) with a human-to-human backstop (proctor's "tap on the shoulder"). In addition, researchers should consider a software patch for the C3Fire chat box and add an audible signal, color/font variation, or "pop-out" window for system messages to mitigate the cognitive tunneling from occurring in the future experiment. Implementing a pop-out window could also include an acknowledgment requirement from the human (such as clicking an "OK" radio button to dismiss the window).

In addition to objectively assessing workloads, collecting participants' cognitive assessment could benefit from eye-tracking software. This measure could enhance observable workload data by informing researchers how the participants are devoting their attention to specific areas in the GUI. Besides, some studies proved that the increasing automation in aviation systems requires operators to monitor systems appropriately. "Operators monitoring appropriately" is defined as a method that enables them to detect automation failures and resume actions if automation fails (Hasse et al. n.d). Additionally, researchers discovered that the pupil mirrors activity in the brain (Bartels and Marshall 2012). Therefore, future researchers could use eye-tracking data to provide real-time measurements of visual and cognitive information processing. The eye tracker can determine whether the pupil's recording is precise enough to measure cognitive workload effect more accurately.

The team also believes that the research can be improved by utilizing video recordings or photographs and conducting post-experiment interviews with each participant to provide greater insight toward answering the research question. During interviews, the researchers would allow the participant to view their full metrics on C3Fire

squares burned and extinguished and a list of the commands and messages passed between the participant and the automated confederate. By giving the participant this information and asking them structured questions, the researchers would better glean their behavior during the scenarios. In utilizing video recordings or photographs of the participants, timed with CSWAG time-hacks, the researchers would incorporate structured interview questions to gauge the participants' perceptions of their workloads at those times. From there, the research team could analyze the correlation between the participants' reported workloads, during the scenarios, against their perceived workload based on facial gestures, body language, posture, etc.

Finally, additional studies should be conducted to answer the question, "When is the best time to hand off to automation?" Sub questions include, "how many times can a pilot handle task handoffs in a given time," and "is it best to hand off the task before, during, or after the task execution?" This study should utilize the modifications previously mentioned. The study should also utilize a program that is easier to measure performance in than C3 Fire. The researchers noted that success in the C3 Fire scenarios was based mainly on who identified the best strategy to fight the fire, not who worked with automation the best. Furthermore, the levels of automation should be differentiated more or based on a new taxonomy. This should yield more statistically different results.

APPENDIX A. PRE-EXERCISE QUESTIONNAIRE

Instructions: Please answer ALL questions as accurately as possible. ALL information is confidential and will be used only for research purposes.

Q1 C3 Fire is a simulation that allows a user to deploy fire trucks to fight a fire. Do you have experience with C3 Fire?

- ○ Yes, I have used C3 Fire
- ○ I may have interacted with a simulation like C3Fire
- ○ I am unsure if I have interacted with C3 Fire or another similar simulation.
- ○ To my knowledge I have not interacted with any simulation like C3 Fire.

Display This Question:

If C3 Fire is a simulation that allows a user to deploy fire trucks to fight a fire. Do you have exp... != To my knowledge I have not interacted with any simulation like C3 Fire.

Q2 Please tell us about your previous experience.

_____

_____

_____

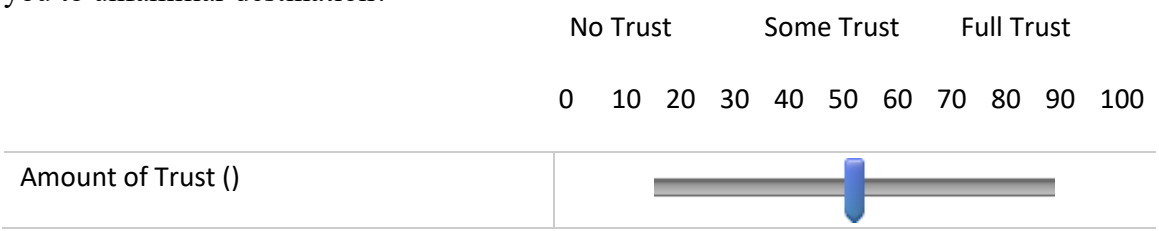
_____

_____

Q3 To what extent do you rely on automation in your daily life? (Examples of automation include GPS Navigation, Adaptive Cruise Control, Smart Home Devices, Video Games with AI Teammates)

- ○ I never use automation
- ○ I infrequently use automation
- ○ I use automation about half of the time it is available to me
- ○ I often use automation
- ○ I always use automation

Q4 To what extent do you trust navigation software (google maps, apple maps) to direct you to unfamiliar destination?



Page Break

Q5 Thank you. Please let the experimenter know that you are done with this portion. He or she will give you your next instructions before proceeding to the next survey.

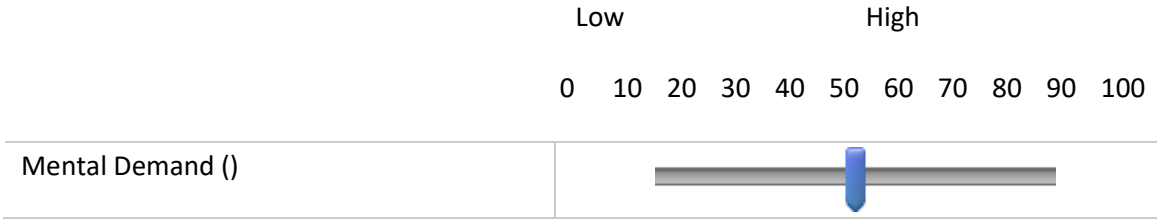
Page Break

APPENDIX B. SCENARIO 1 – POST QUESTIONNAIRE

Instructions: Please answer ALL questions as accurately as possible. ALL information is confidential and will be used only for research purposes.

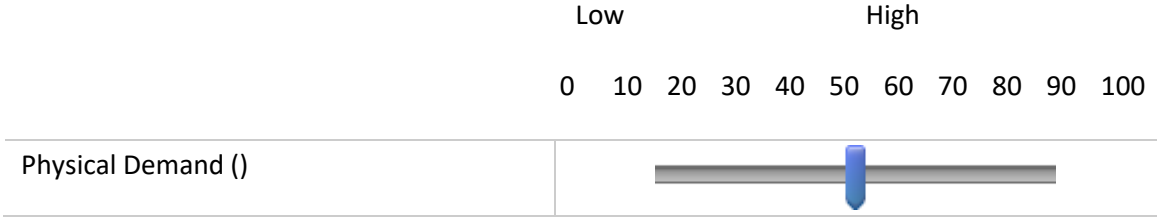
Q1 Mental Demand

How much mental and perceptual activity was required (thinking, remembering, looking, searching)?



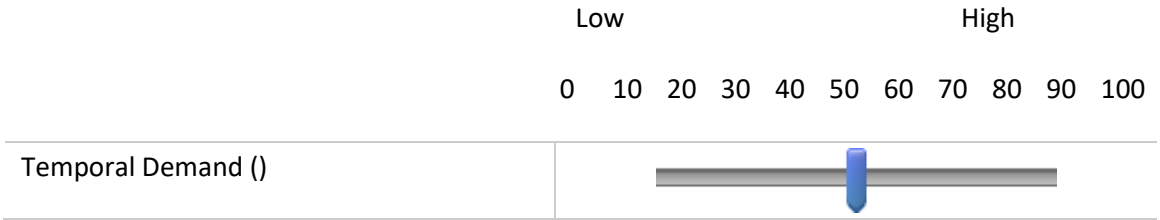
Q2 Physical Demand

How much physical activity was required (pulling, turning, controlling)?



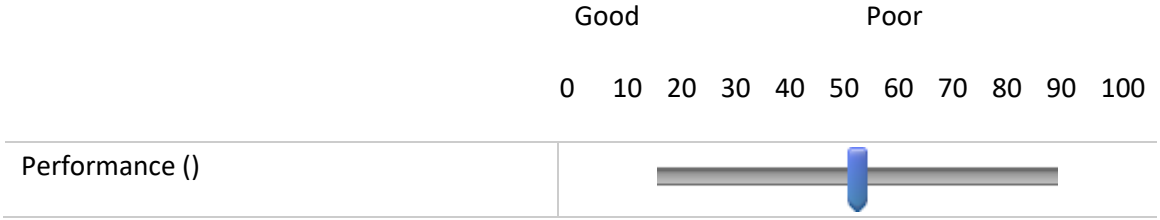
Q3 Temporal Demand

How much time pressure did you feel due to the rate at which the tasks occurred (leisurely or frantic)?



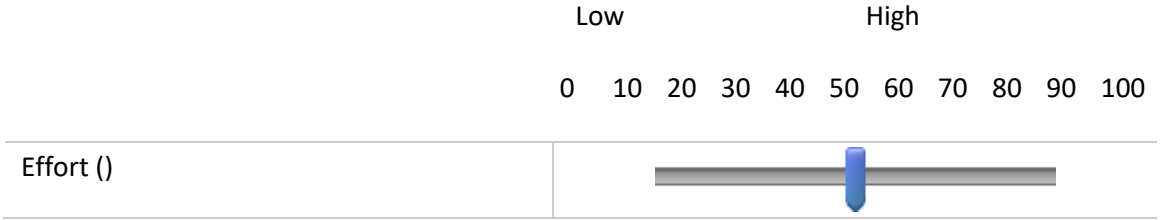
Q4 Performance

How successful do you think you were in accomplishing the goals?



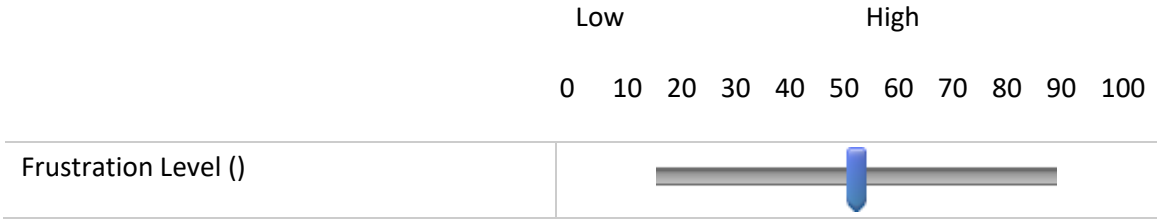
Q5 Effort

How hard did you have to work (mentally)?



Q6 Frustration Level

How insecure, irritated, stressed, or annoyed versus secure, content, relaxed did you feel?



Page Break

Q7 For the following questions, please consider the automation in the scenario that you just completed.

At what cognitive level (experience by age) did you perceive the automated device was developed to replicate?

	Toddler Adult 0 10 20 30 40 50 60 70 80 90 100
Cognitive Level ()	

Q8 How much confidence do you have that the automation managed the tasks you transferred to it effectively?

	None at all	A little	A moderate amount	A lot	A great deal
Water Truck refill Fire Truck (1)	○	○	○	○	○
Water Truck refill itself (2)	○	○	○	○	○

Q9 Think about the number of tasks you wanted to pass to automation, what percentage of those tasks were you successful in doing so?

	0 10 20 30 40 50 60 70 80 90 100
Percentage ()	

Q10 What improvements would you make to the automation program?

_____

_____

_____

_____

_____

Page Break

Q11 Thank you. Please let the experimenter know that you are done with this portion. He or she will give you your next instructions.

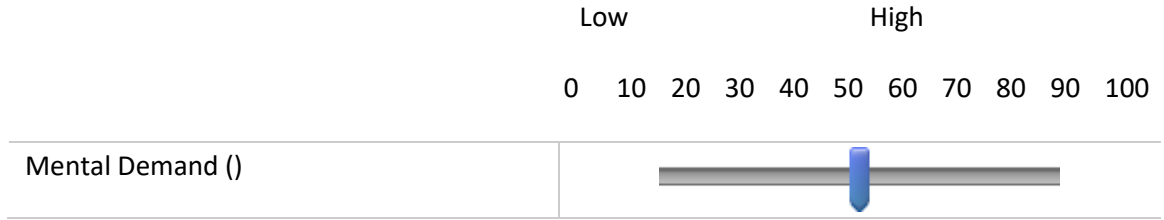
Page Break

APPENDIX C. SCENARIO 2 – POST QUESTIONNAIRE

Instructions: Please answer ALL questions as accurately as possible. ALL information is confidential and will be used only for research purposes.

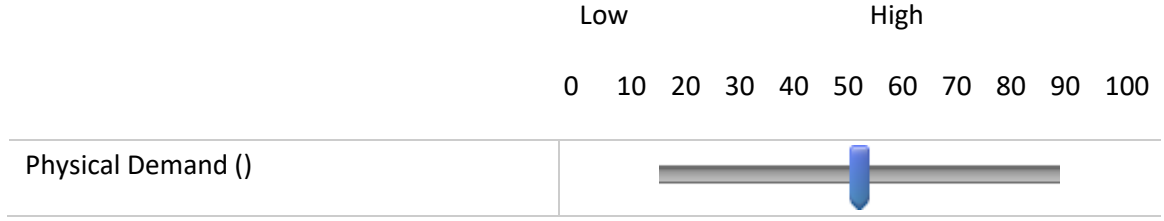
Q1 Mental Demand

How much mental and perceptual activity was required (thinking, remembering, looking, searching)?



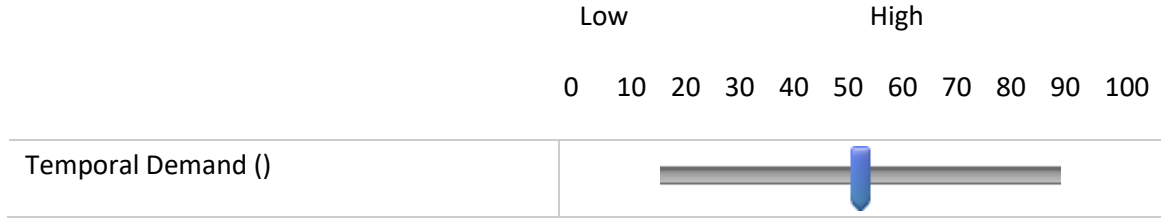
Q2 Physical Demand

How much physical activity was required (pulling, turning, controlling)?



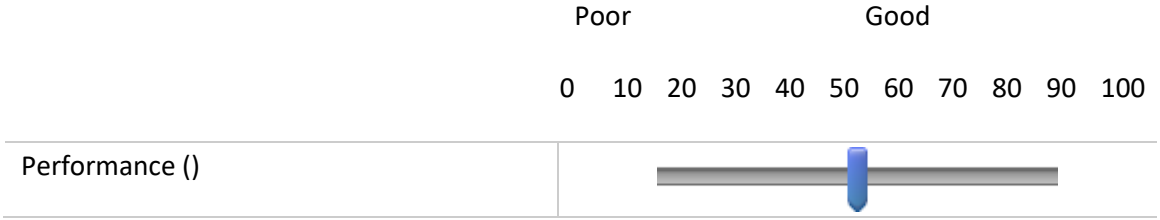
Q3 Temporal Demand

How much time pressure did you feel due to the rate at which the tasks occurred (leisurely or frantic)?



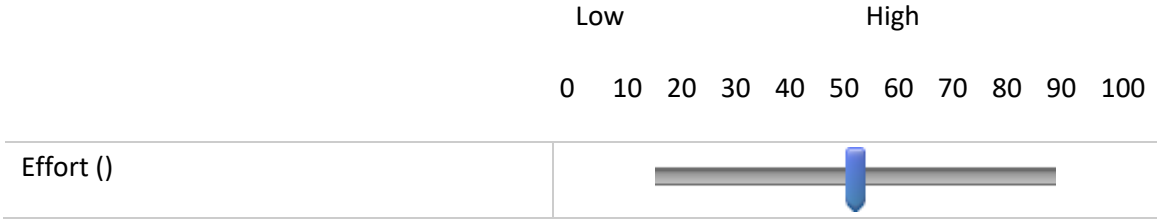
Q4 Performance

How successful do you think you were in accomplishing the goals?



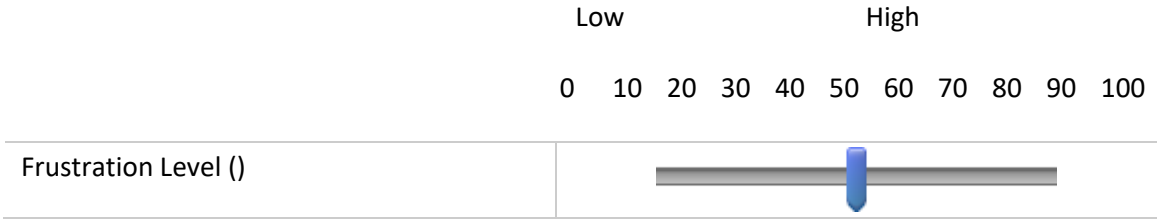
Q5 Effort

How hard did you have to work (mentally)?

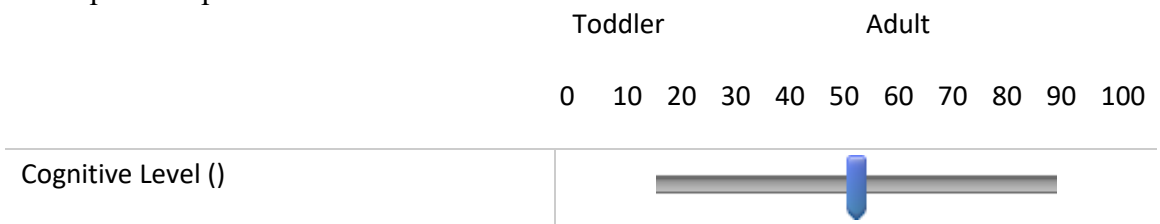


Q6 Frustration Level

How insecure, irritated, stressed, or annoyed versus secure, content, relaxed did you feel?



Q7 At what cognitive level (experience by age) did you perceive the automated device was developed to replicate?



Q8 How much confidence do you have that the automation managed the tasks you transferred to it effectively?

	None	A little	A moderate amount	A lot	A great deal
Water Truck refill Fire Truck	○	○	○	○	○
Water Truck refill itself	○	○	○	○	○

Q9 Think about the number of tasks you wanted to pass to automation, what percentage of those tasks were you successful in doing so?

	0 10 20 30 40 50 60 70 80 90 100
Percentage ()	

Q10 What improvements would you make to the automation program?

_____

_____

_____

_____

_____

Page Break

Q11 Thank you. Please let the experimenter know that you are done with this portion. He or she will give you your next instructions before proceeding to the next survey.

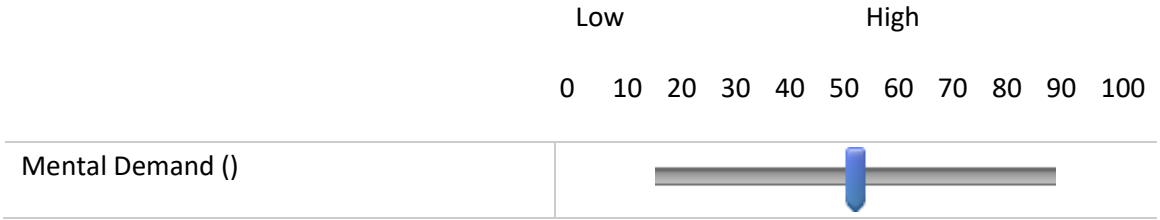
THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX D. SCENARIO 3 – POST QUESTIONNAIRE

Instructions: Please answer ALL questions as accurately as possible. ALL information is confidential and will be used only for research purposes.

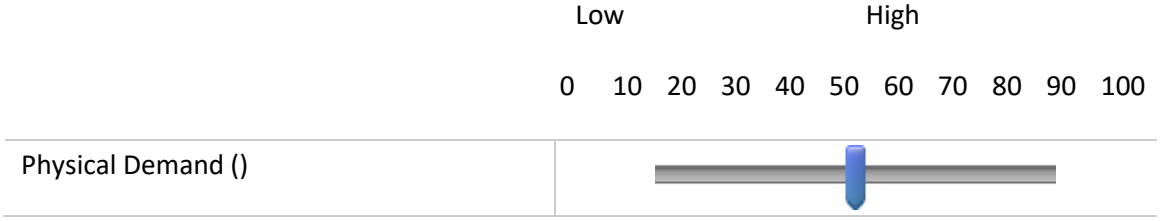
Q1 Mental Demand

How much mental and perceptual activity was required (thinking, remembering, looking, searching)?



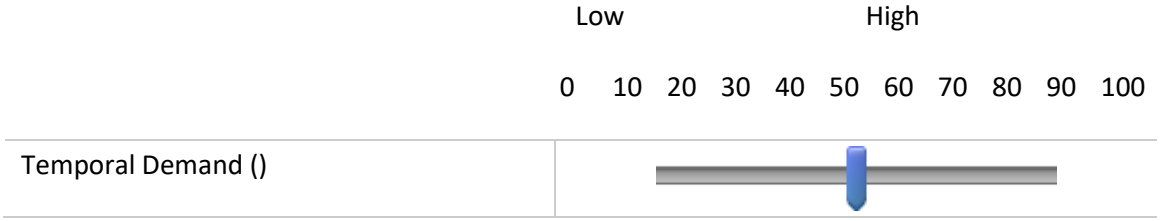
Q2 Physical Demand

How much physical activity was required (pulling, turning, controlling)?



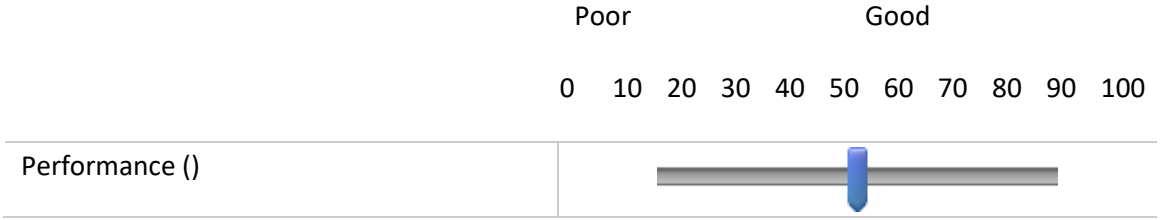
Q3 Temporal Demand

How much time pressure did you feel due to the rate at which the tasks occurred (leisurely or frantic)?



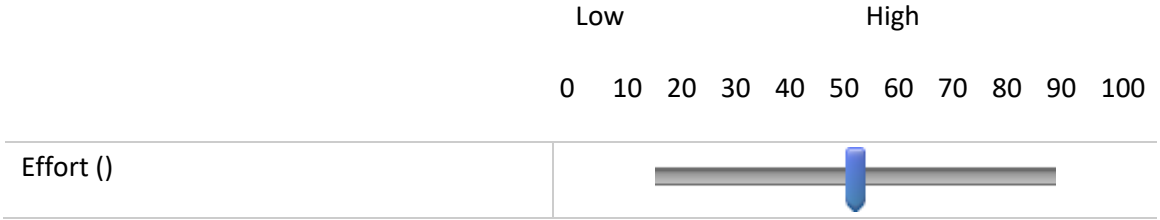
Q4 Performance

How successful do you think you were in accomplishing the goals?



Q5 Effort

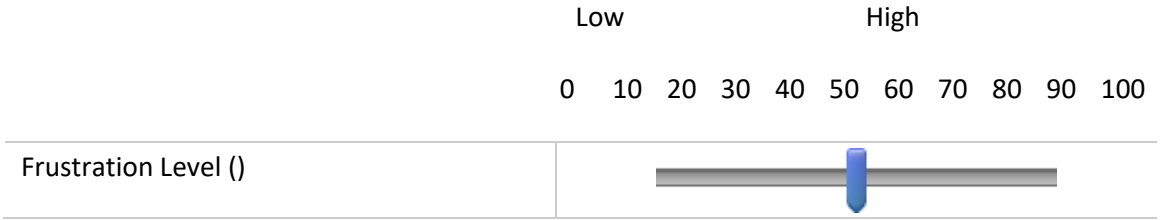
How hard did you have to work (mentally)?



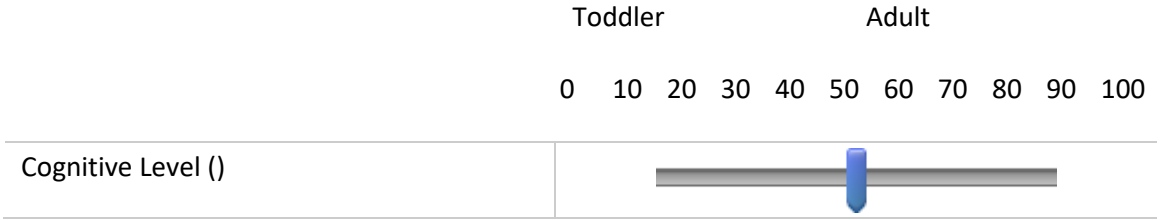
Page Break

Q6 Frustration Level

How insecure, irritated, stressed, or annoyed versus secure, content, relaxed did you feel?



Q7 At what cognitive level (experience by age) did you perceive the automated device was developed to replicate?



Q8 How much confidence do you have that the automation managed the tasks you transferred to it effectively?

	None	A little	A moderate amount	A lot	A great deal
Water Truck refill Fire Truck	○	○	○	○	○
Water Truck refill itself	○	○	○	○	○

Q9 Think about the number of tasks you wanted to pass to automation, what percentage of those tasks were you successful in doing so?

	0 10 20 30 40 50 60 70 80 90 100
Percentage ()	

Q10 What improvements would you make to the automation program?

_____

_____

_____

_____

_____

Q11 Please rate the following statements:

	Strongly agree	Agree	Neither agree nor disagree	Disagree	Strongly Disagree
I know how the automation works	☐	☐	☐	☐	☐
I found the automation to be useful	☐	☐	☐	☐	☐
The automation was easy to use	☐	☐	☐	☐	☐
I could override the automation when I needed to	☐	☐	☐	☐	☐
The automation was reliable	☐	☐	☐	☐	☐

Q12 How likely do you think the following would occur, based on your experience with this automation program:

	Extremely unlikely	Somewhat unlikely	Neither likely nor unlikely	Somewhat likely	Extremely likely
Technical issues	○	○	○	○	○
Incorrect input due to user error	○	○	○	○	○

Q13 Consider the automation that you used in this scenario compared to the previous scenarios. What was the most noticeable difference?

_____

_____

_____

_____

_____

Page Break

Q14 Thank you. Please let the experimenter know that you are done with this portion. He or she will give you your next instructions before proceeding to the next survey.

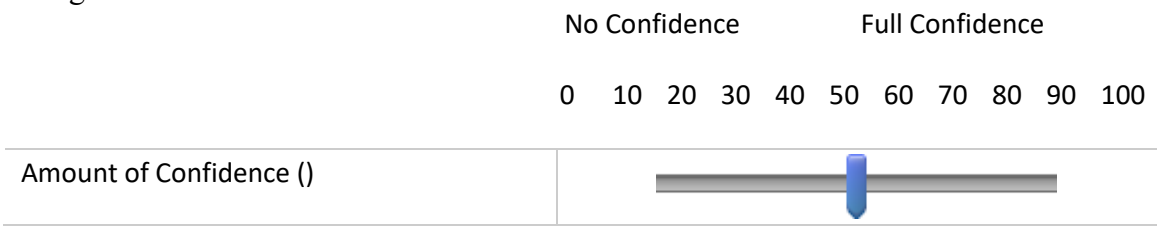
Page Break

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX E. FINAL QUESTIONNAIRE

Instructions: Please answer ALL questions as accurately as possible. ALL information is confidential and will be used only for research purposes.

Q1 Rate the level of confidence you have that the training you received prepared you to extinguish the fire with automation



Q2 In regard to the previous question, why did you report that level of confidence?

_____

_____

_____

_____

_____

Q3 Did you find any techniques that worked best when working with automation to extinguish the fire?

_____

_____

_____

_____

_____

Q4 Please rank order the scenarios based off the amount you used automation (Most to Least).
_____ Scenario 1
_____ Scenario 2
_____ Scenario 3

Q5 Please rank order the scenario that you think you worked the best with automation (Most to Least).
_____ Scenario 1
_____ Scenario 2
_____ Scenario 3

Q6 Previously you reported the scenario you worked best with automation. Please explain why you worked best with that level of automation.

_____

_____

_____

_____

_____

Page Break

Q7 Please let the experimenter know that you have completed the survey. Thank you for taking the time to participate in our experiment. If you have any questions about your rights as a research subject or any other concerns, please address them to Mr. Bryan Hudgens, 831–656-2039, bryan.hudgens@nps.edu.

End of Block: Default Question Block

APPENDIX F. MEASURES THAT FAILED TO SHOW STATISTICAL SIGNIFICANCE

Heartrate Variability Results

Overall, the heartrate data failed to provide any statistically significant stress indicators between scenarios. The heartrate variability was measured by an E4 Empatica wrist-based heartrate monitor. The proctor pressed the record button the same second the proctor clicked begin on the scenario. The research team compiled every scenario's worth of heartrate data (599 seconds each scenario) and segregated them into study 1 (participant 1–10) and study 2 (participant 11–20). To normalize the data, researchers divided the heartrate recorded every second by the median across all three scenarios for each participant. In each study, the difference in heartrate remained relatively constant and failed to show any rise in stress or workload as clearly as in other subjective measures such as CSWAG.

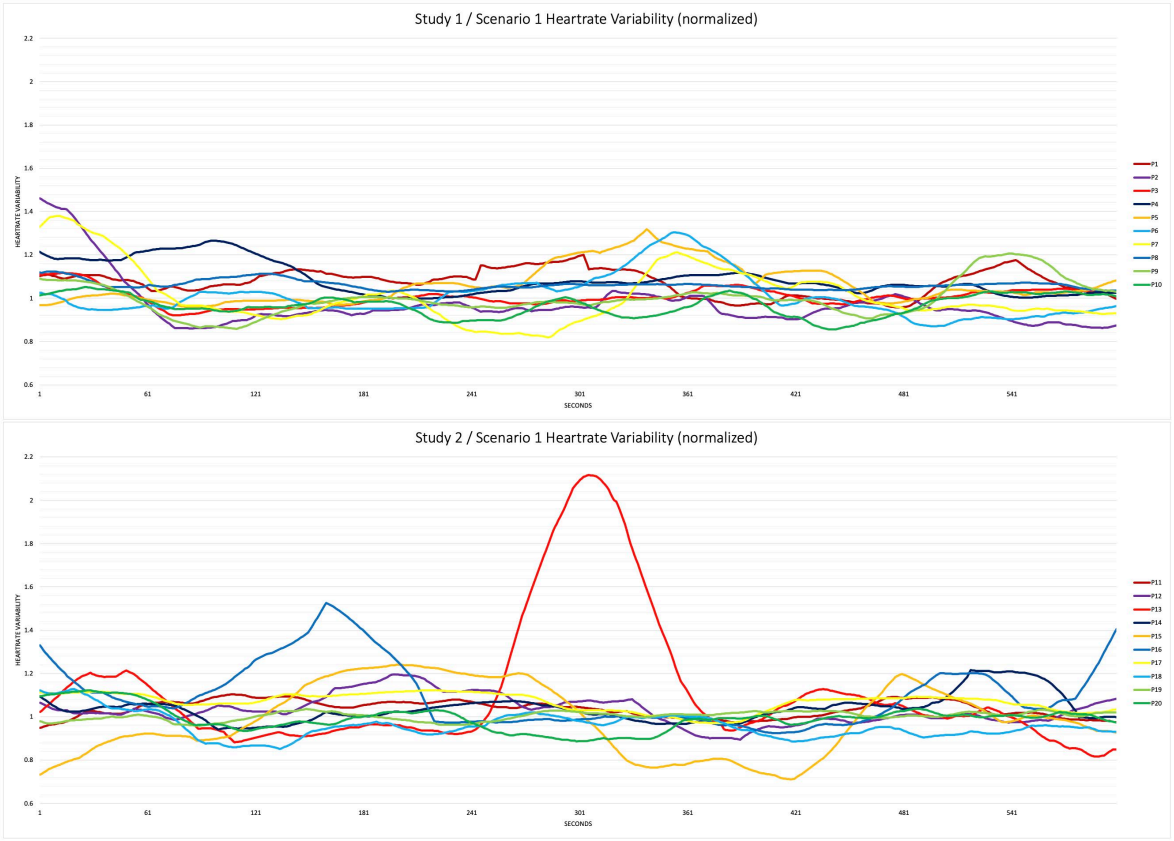


Figure 11. Study 1 and Study 2 heart rate data for Scenario 1.

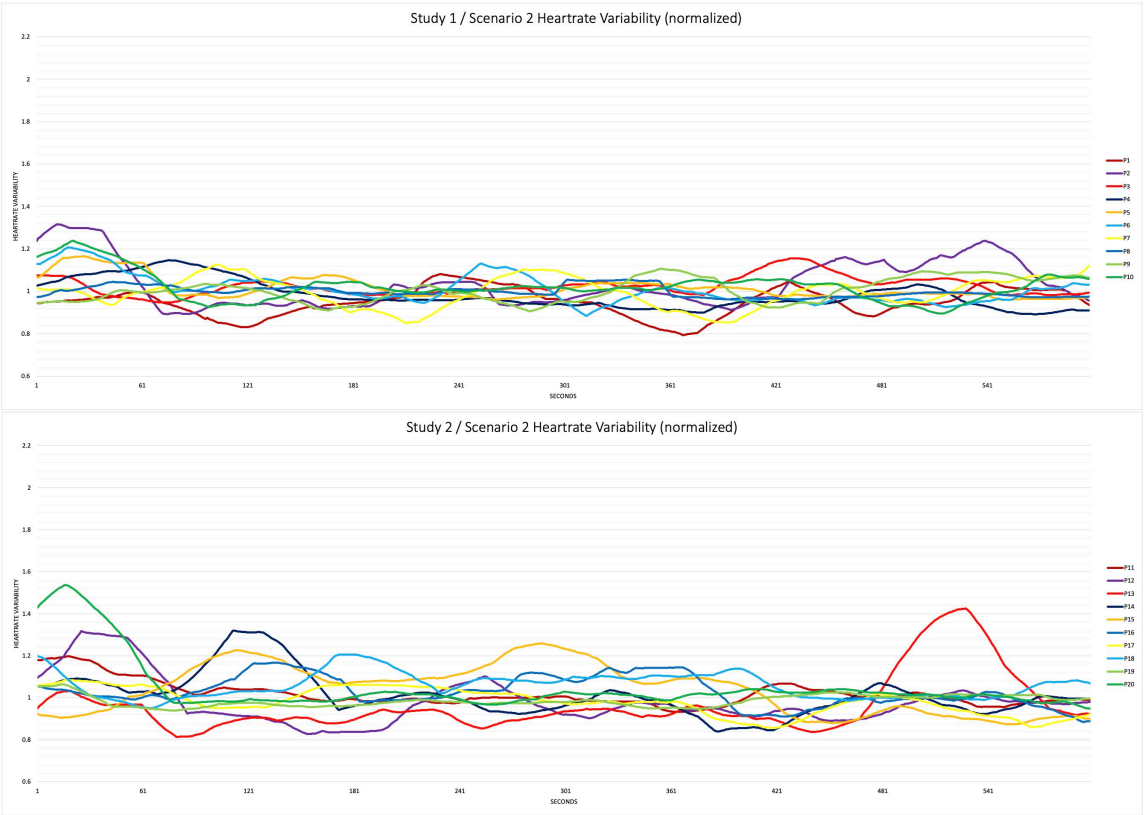


Figure 12. Study 1 and Study 2 heart rate data for Scenario 2.

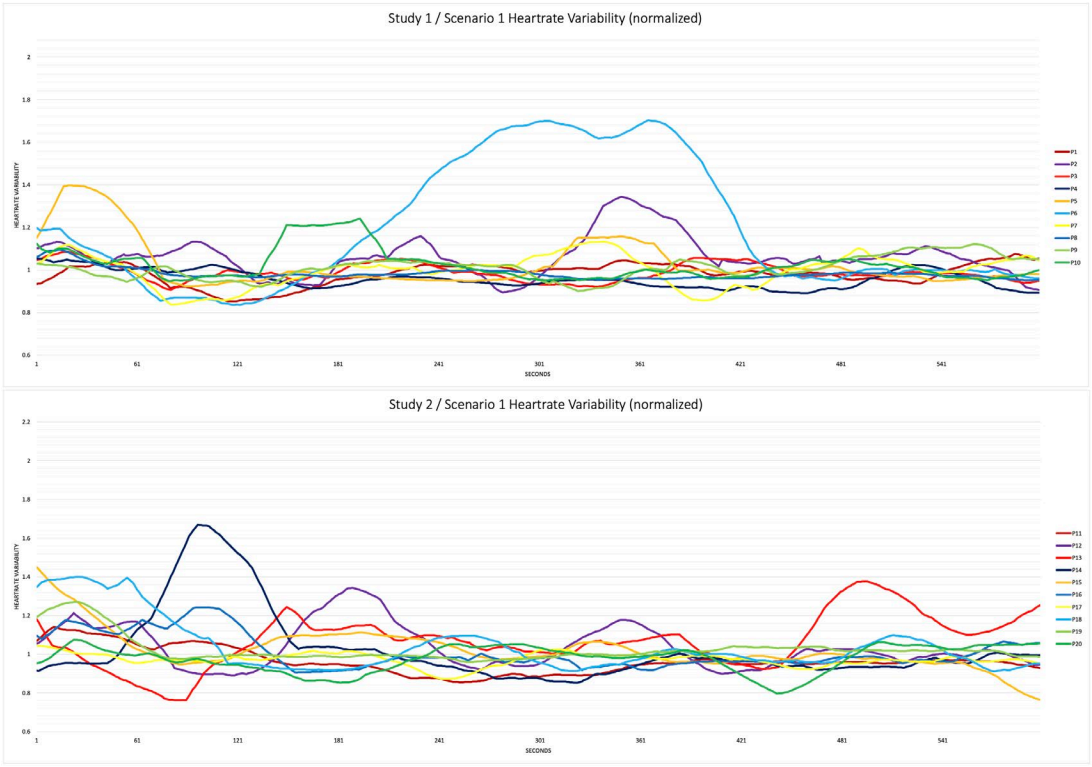


Figure 13. Study 1 and Study 2 heart rate data for Scenario 3.

APPENDIX G. SURVEY DATA

Pre-Exercise Questions:

Researchers discovered that the pool of participants varied in their baseline use of automation of automation but mostly trusted automation more then 70% of the time. All the participants claimed they have not interacted with a simulation such as C3Fire before the volunteering for the study. Study 2 participants varied in their use of automation while study 1 used automation at least half the time it was available. Figure 13 depicts the answer to "to what extent do you trust navigation software (google maps, apple maps) to direct you to unfamiliar destination." Figure 14 depicts the answer to "to what extent do you rely on automation in your daily life."

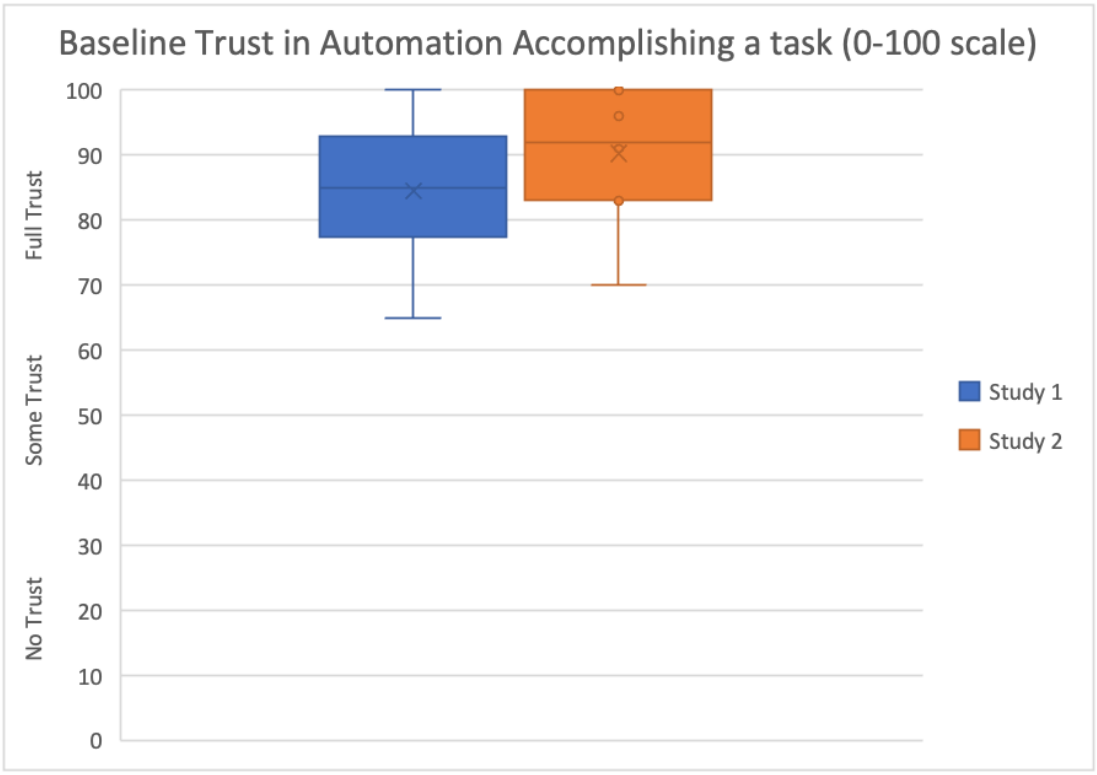


Figure 14. Participant Baseline Use of Automation

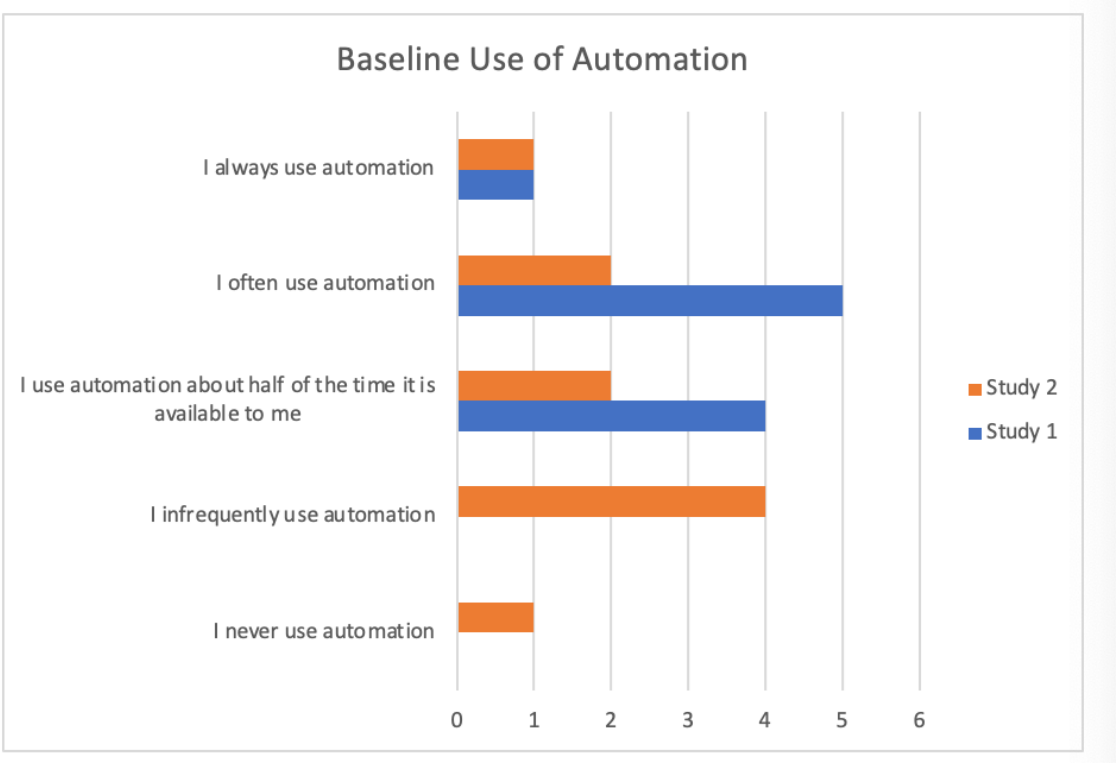


Figure 15. Baseline trust in Automation

Scenario Trust Questions

Overall, the survey questions referring to how much the participants trusted the automation to manage the task handed off were inconclusive in highlighting significant trends. However, the data shows that none of the participants felt they had zero confidence in the automation completing the tasks. Also, more participants in scenario 3 thought they had the highest level (a great deal) confidence in automation completing the tasks compared to scenario 1 and scenario 2. Figure 15 shows the results from study 1 and study 2 for the confidence in automation refilling the fire trucks and refilling water trucks. Figure 16 depicts how participants rated the automation's perceived development in relation to age. Generally, study 1 thought scenario 3's sophistication was higher than that of scenario 1 and 2. However, study 2 generally felt that scenario 3 was lower in sophistication then in scenario 1 and scenario 2. The results from both studies failed to show a statistical difference between scenarios.

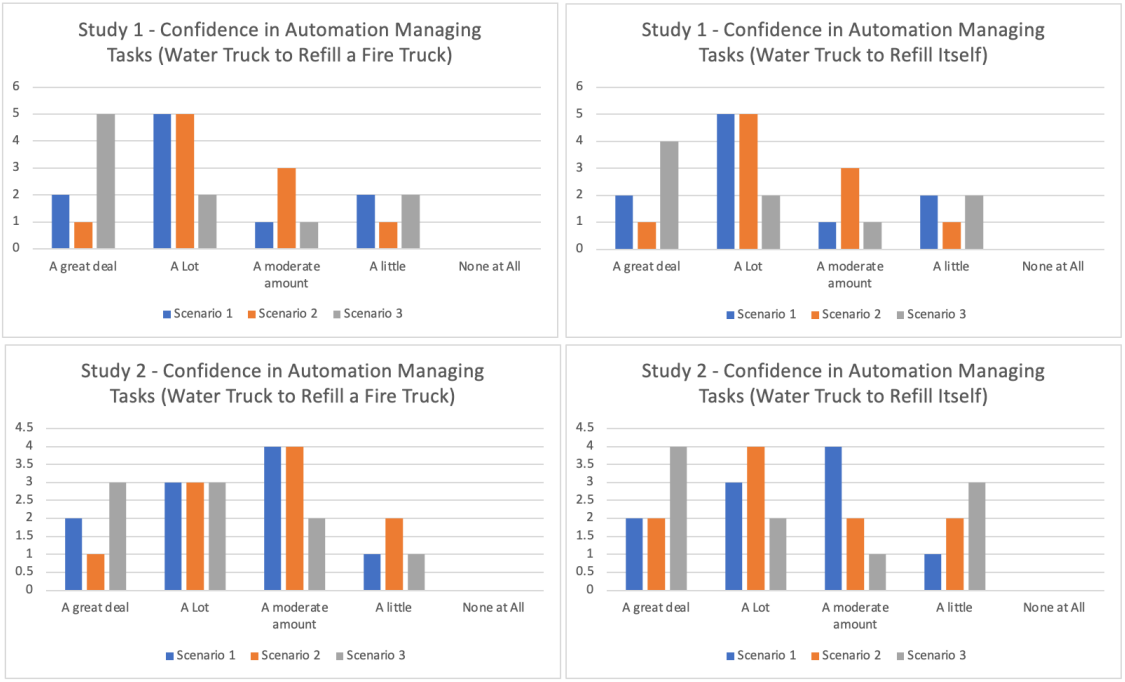


Figure 16. Study 1 and Study 2 Confidence in Automation Managing Tasks

Automation Age Study 1			
Participants	Scenario 1	Scenario 2	Scenario 3
1	60	61	61
2	50	50	91
3	41	59	69
4	35	51	63
5	19	23	32
6	81	85	89
7	20	15	15
8	29	30	30
9	71	66	79
10	10	10	10
median	38	50.5	62

Automation Age Study 2			
Participants	Scenario 1	Scenario 2	Scenario 3
1	50	50	50
2	50	50	40
3	15	20	27
4	40	40	40
5	30	30	30
6	20	25	30
7	10	19	11
8	60	67	61
9	71	62	50
10	100	80	80
median	45	45	40

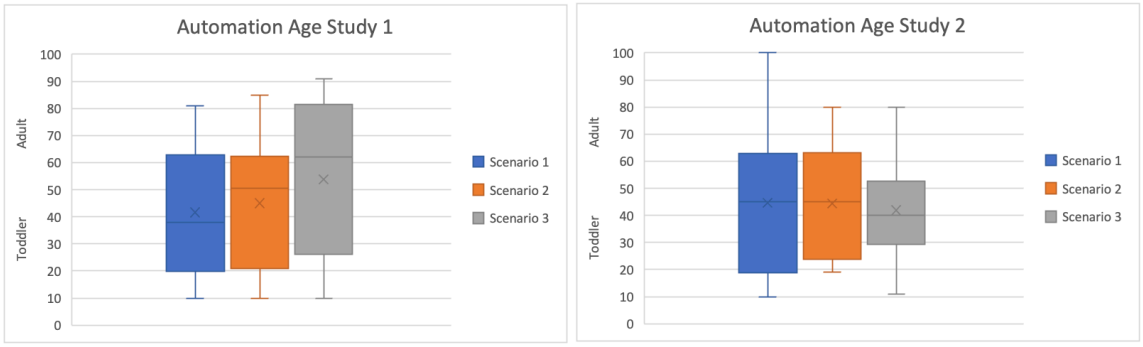


Figure 17. Study 1 and Study 2 Automation Age Comparison (At what cognitive level (experience by age) did you perceive the automated device was developed to replicate?)

Usability Survey Questions

Overall, the participants in study 1 and study 2 responded to usability questions in a relatively varied response. Due to the small size of the population, answers generally indicate that participants knew how the automation worked, could be overridden, was useful, was not user-friendly and was reliable. The ease-of-use question resulted in an equally distributed response ranging from disagree to strongly agree. While study 1 and study 2 generally responded in a wide variety of responses in the likelihood technical issues would occur in the automation, both studies indicate that there was a significant likelihood that users would incorrectly input data into the system. The majority of both studies 1 and 2 claimed high likelihood that users would input data incorrectly suggests the user interface was designed poorly and could be a good predictor of the high levels of cognitive tunneling. The results from the usability questions are referenced in Figure 17.

Figure 18. Usability Matrix Questions

Post Exercise Questions

The survey results for the amount of confidence in training was not statistically different among study 1 and study 2 participants. On a scale from 0–100, participants were

asked to rate the level of confidence they had that the training received prepared them to extinguish the fire with automation. The results from the question is displayed in Figure 18.

Figure 19. Level of Confidence That Training Prepared Them to Work with Automation

Participants in study 1 and study 2 generally ranked scenario 3 as the scenario they used automation the most, followed by scenario 2, and the least in scenario 1. Of note, study 2 had a more diverse reporting of which scenario the subject used automation more compared to study 1. The results from the question "please rank order the scenarios based off the amount you used automation (most to least)" is displayed in Figure 19.

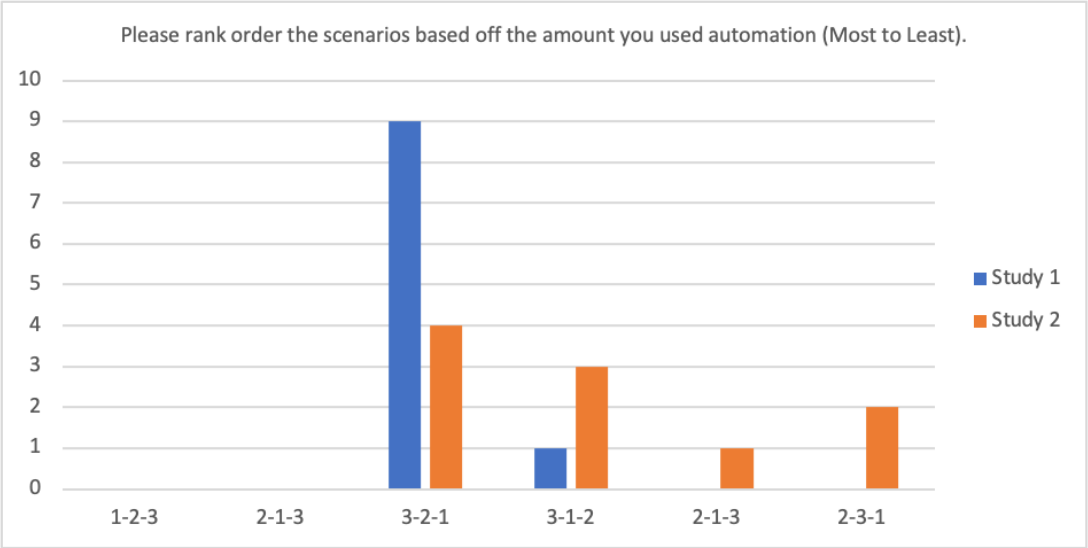


Figure 20. Reported Ranking of Scenarios Based on the Amount a Participant Used Automation (Most to Least)

Participants in study 2 reported they worked best with automation in scenario 2 the best while study 1 reported they worked best with scenario 3. Participants from study 1 and study 2 were asked "please rank order the scenario that you think you worked the best with automation (most to least)." 100% of participants in study 1 reported they worked best with scenario 3, then scenario 2, and the least with scenario 1. Those in study 2 reported to be mostly split between scenario 2–1-3 and then scenarios 2–3-1. One person in study 2 reported to work with scenario 3 the best, then scenario 1, and least with scenario 2. The results from the question regarding which scenario the participants worked best with automation is reflected in Figure 20.

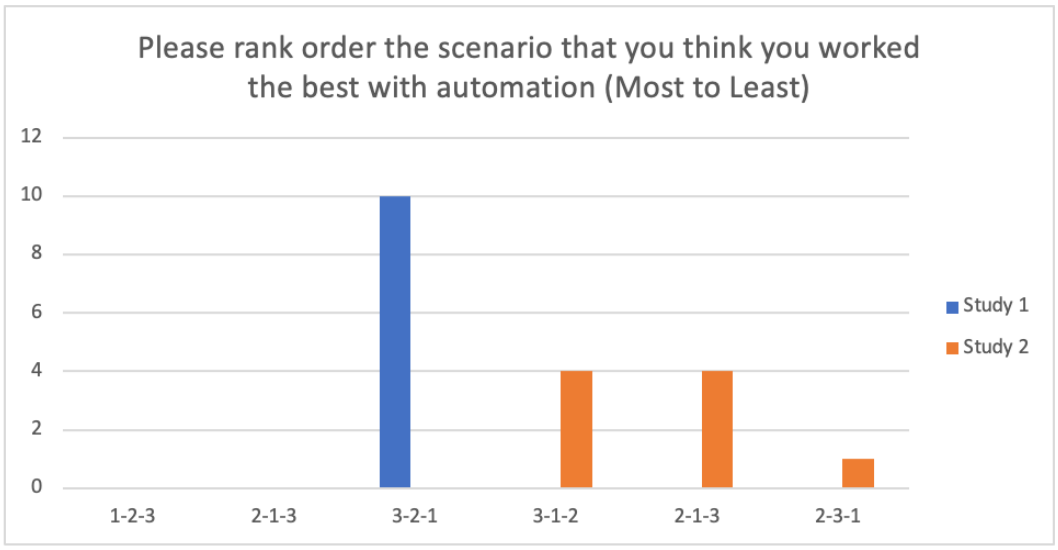


Figure 21. Reported Ranking of Scenarios based on how well they worked with automation (Most to Least)

NASA TLX

As stated in Chapter IV–Results, NASA TLX failed to show a significant difference between scenarios in both study 1 and study 2. The NASA TLX results are as follows:

Study 1 Physical Demand			
Participants	Scenario 1	Scenario 2	Scenario 3
1	20	21	43
2	50	91	80
3	10	11	10
4	10	9	15
5	10	5	6
6	30	32	32
7	80	80	80
8	0		
9	10	11	19
10	10	9	10
median	10	11	19

Study 2 Physical Demand			
Participants	Scenario 1	Scenario 2	Scenario 3
11	10	31	30
12	20	20	9
13	20	25	22
14	20	19	20
15	20	19	20
16	10	10	10
17	30	50	48
18	30	37	40
19	20	25	20
20	20	20	10
median	20	22.5	20

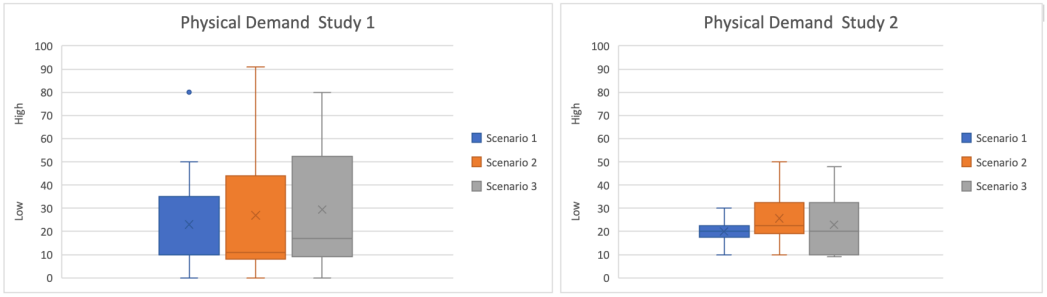
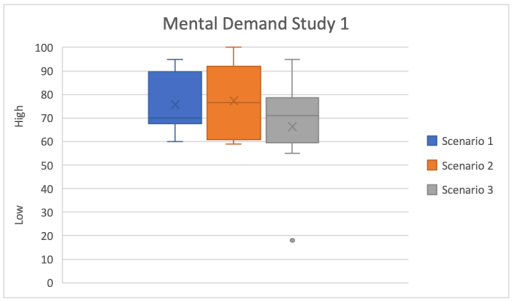


Figure 22. Figure 21: Physical Demand (How much physical activity was required (pulling, turning, controlling)?)

Mental Demand Study 1			
Participant	Scenario 1	Scenario 2	Scenario 3
1	70	100	61
2	70	91	71
3	92	73	80
4	70	61	55
5	60	60	71
6	81	79	73
7	95	95	95
8	89	74	78
9	60	59	61
10	70	81	18
Median	70	76.5	71

Mental Demand Study 2			
Participant	Scenario 1	Scenario 2	Scenario 3
11	90	70	70
12	45	45	30
13	73	75	78
14	85	90	60
15	100	100	100
16	65	75	81
17	51	70	75
18	71	81	61
19	81	70	59
20	100	100	100
Median	77	75	72.5



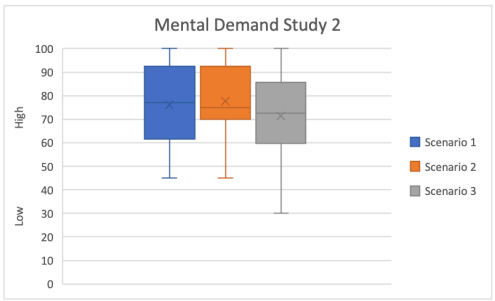
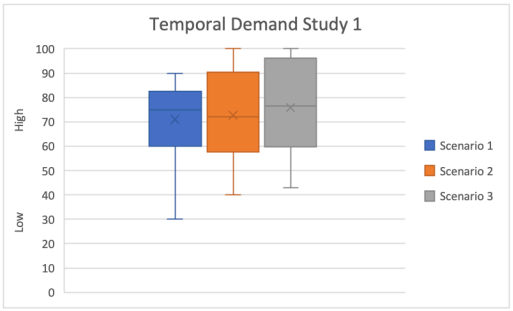


Figure 23. Mental Demand (How much mental and perceptual activity was required (thinking, remembering, looking, searching)?)

Temoral Demand Study 1			
Participant	Scenario 1	Scenario 2	Scenario 3
1	70	100	70
2	80	81	100
3	80	60	79
4	60	51	43
5	60	71	59
6	80	70	78
7	90	90	95
8	90	73	75
9	30	40	60
10	70	92	100
Median	75	72	76.5

Temoral Demand Study 2			
Participant	Scenario 1	Scenario 2	Scenario 3
11	80	40	50
12	40	34	20
13	70	76	85
14	80	91	40
15	90	90	100
16	50	76	81
17	50	61	71
18	50	65	40
19	90	79	59
20	30	90	100
Median	60	76	65



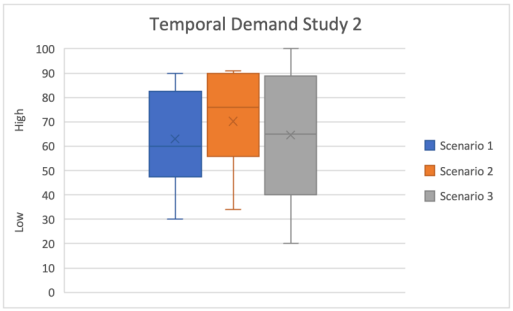
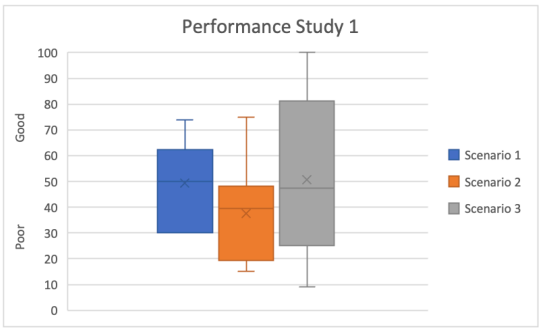


Figure 24. Temporal Demand (How much time pressure did you feel due to the rate at which the tasks occurred (leisurely or frantic)?

Performance Study 1			
Participant	Scenario 1	Scenario 2	Scenario 3
1	50	15	82
2	72	17	9
3	50	39	51
4	74	26	44
5	30	48	100
6	30	47	59
7	50	40	30
8	59	75	81
9	49	49	40
10	30	20	10
Median	50	39.5	47.5

Performance Study 2			
Participant	Scenario 1	Scenario 2	Scenario 3
11	70	60	75
12	76	72	98
13	70	12	5
14	50	50	71
15	75	85	90
16	81	71	50
17	31	20	15
18	65	71	87
19	50	38	66
20	50	20	50
Median	67.5	55	68.5



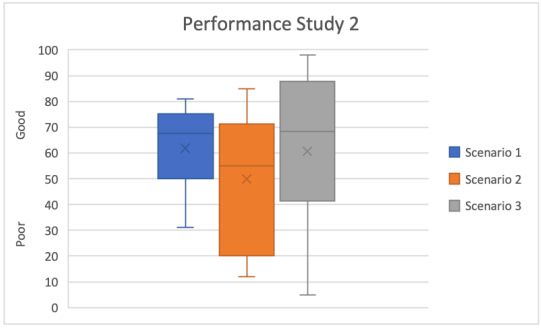
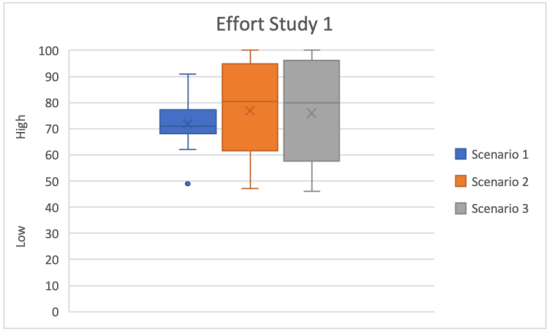


Figure 25. Performance (How successful do you think you were in accomplishing the goals?)

Effort Study 2			
Participant	Scenario 1	Scenario 2	Scenario 3
1	71	81	80
2	71	80	100
3	62	51	69
4	73	70	46
5	70	65	50
6	71	85	80
7	90	95	95
8	91	95	80
9	49	47	60
10	71	100	100
Median	71	80.5	80

Effort Study 2			
Participant	Scenario 1	Scenario 2	Scenario 3
11	91	50	75
12	45	46	24
13	64	65	68
14	80	80	60
15	91	95	100
16	60	70	85
17	40	60	73
18	50	69	51
19	71	68	50
20	70	100	100
Median	67	68.5	70.5



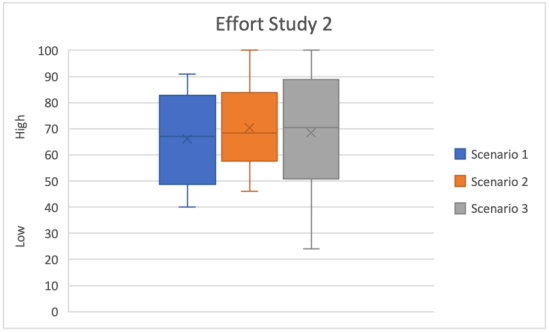


Figure 26. Effort (How hard did you have to work (mentally)?

Frustration Level Study 1			
Participants	Scenario 1	Scenario 2	Scenario 3
1	29	18	20
2	30	19	60
3	38	62	30
4	61	50	35
5	72	75	35
6	91	91	85
7	70	90	90
8	71	63	35
9	34	37	45
10	60	90	90
median	60.5	62.5	40

Frustration Level Study 2			
Participants	Scenario 1	Scenario 2	Scenario 3
11	80	50	70
12	25	38	19
13	75	84	90
14	60	50	50
15	30	15	30
16	30	60	50
17	40	71	51
18	45	48	54
19	40	55	39
20	70	60	80
median	42.5	52.5	50.5

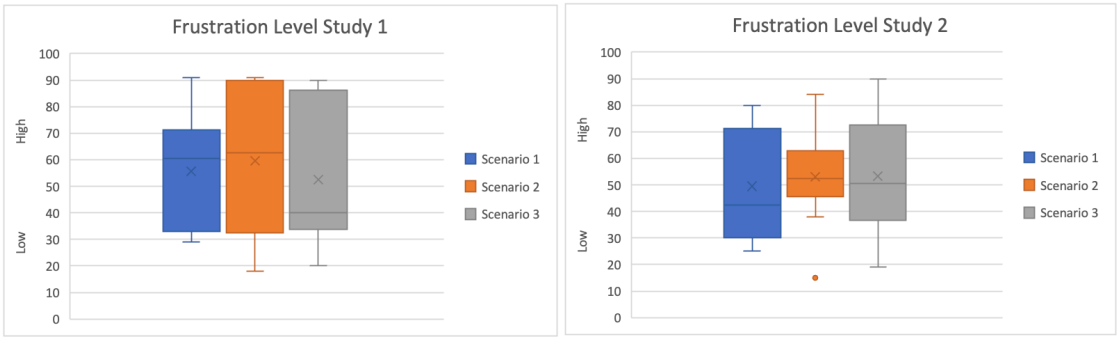


Figure 27. Frustration Level (How insecure, irritated, stressed, or annoyed versus secure, content, relaxed did you feel?

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Bartels, Michael, and Sandra P. Marshall. 2012. "Measuring Cognitive Workload across Different Eye Tracking Hardware Platforms." In *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12*, 161. Santa Barbara, California: ACM Press. https://doi.org/10.1145/2168556.2168582.

Bruder, Carmen, Hinnerk Eißfeldt, Peter Maschke, and Catrin Hasse. 2014. "A Model for Future Aviation: Operators Monitoring Appropriately." *Aviation Psychology and Applied Human Factors* 4 (1): 13–22. http://dx.doi.org.libproxy.nps.edu/10.1027/2192-0923/a000051.

Castelfranchi, Cristiano, and Rino Falcone. 2010. *Trust Theory: A Socio-Cognitive and Computational Model*. Wiley Series in Agent Technology. Chichester, West Sussex, U.K: J. Wiley. http://onlinelibrary.wiley. com/doi/.

Colebank, Jayson L. 2008. "Automated Intelligent Agents: Are They Trusted Members of Military Teams?" Master's Thesis, Naval Postgraduate School. http://hdl.handle.net/10945/3868

Dekker, Sidney W. A. 2015. "The Danger of Losing Situational Awareness." *Cognition, Technology & Work* 17 (2): 159–61. https://doi.org/10.1007/s10111-015-0320-8.

Dehais, F., M. Causse, and S. Tremblay. 2011. Mitigation of Conflicts with Automation: Use of Cognitive Countermeasures. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *53*(5), 448–460. https://doi.org/10.1177/0018720811418635

Endsley, Mica R. 1995. "Toward a Theory of Situation Awareness in Dynamic Systems." *Human Factors* 37 (1): 32–64. https://doi.org/10.1518/001872095779049543.

FAA. 2013. *Operational Use of Flight Path Management Systems*. Retrieved 9 Apr 2016, from http://www.faa.gov/aircraft/air_cert/design_approvals/human_factors/media/oufpms_report.pdfWG_Final_Report_Recommendations.pdf

Flach, John M. 1995. "Situational Awareness: Proceed with Caution." *Human Factors* 37 (1): 149–57. https://doi.org/10.1518/001872095779049480.

Hari, S. K. K., A. Nayak, and S. Rathinam. 2020. An Approximation Algorithm for a Task Allocation, Sequencing and Scheduling Problem Involving a Human-Robot Team. *IEEE Robotics and Automation Letters*, *5*(2), 2146–2153. https://doi.org/10.1109/LRA.2020.2970689

Hoffman, Robert R., Matthew Johnson, Jeffrey M. Bradshaw, and Al Underbrink. 2013. "Trust in Automation." *IEEE Intelligent Systems* 28 (1): 84–88. https://doi.org/10.1109/MIS.2013.24.

Keller, John. 2020. "Army Asks Industry for Open-Systems Avionics Technologies for Future Attack and Reconnaissance Helicopters." Military & Aerospace Electronics. Last modified 20 Apr, 2020. https://www.militaryaerospace.com/computers/article/14174361/opensystems-avionics-helicopters.

On-Road Automated Driving (ORAD) committee. n.d. "Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles." SAE International. Accessed 13 May, 2021. https://doi.org/10.4271/J3016_202104.

Parasuraman, Raja, and Dietrich H. Manzey. 2010. "Complacency and Bias in Human Use of Automation: An Attentional Integration." *Human Factors* 52 (3): 381–410. https://doi.org/10.1177/0018720810376055.

Parasuraman, Raja, and Victor Riley. 1997. "Humans and Automation: Use, Misuse, Disuse, Abuse." *Human Factors* 39 (2): 230. https:// doi.org/10.1518/001872097778543886

Pizziol, S., C. Tessier, and F. Dehais. 2014. "Petri Net-Based Modelling of Human–Automation Conflicts in Aviation." *Ergonomics*, *57*(3), 319–331. https://doi.org/10.1080/00140139.2013.877597

Rankin, A., R. Woltjer, and J. Field. 2016. Sensemaking Following Surprise in the Cockpit—A Re-framing Problem. *Cognition, Technology & Work*, *18*(4), 623–642. https://doi.org/10.1007/s10111-016-0390-2

Rome, F., G. Adam, J. Condette, M. Causse, and F. Dehais. 2012. "Go-around Manoeuver: A Simulation Study." European Association for Aviation Psychology Conference, Sardinia, Italy, September 22–26.

Rosenman, Elizabeth D., Aurora J. Dixon, Jessica M. Webb, Sarah Brolliar, Simon J. Golden, Kerin A. Jones, Sachita Shah et al. 2018. "A Simulation-Based Approach to Measuring Team Situational Awareness in Emergency Medicine: A Multicenter, Observational Study." *Academic Emergency Medicine* 25 (2): 196–204. https://doi.org/10.1111/acem.13257.

Salas, Eduardo, Carolyn Prince, David P. Baker, and Lisa Shrestha. 1995. "Situational Awareness in Team Performance: Implications for Measurement and Training." *Human Factors* 37 (1): 123–36. https://doi.org/10.1518/001872095779049525.

Sarter, N. B., D. D. Woods, and C. E. Billings. 1997. "Automation Surprises," *Handbook of Human Factors & Ergonomics.* 25.

Semmens, Rob, Nikolas Martelaro, Pushyami Kaveti, Simon Stent, and Wendy Ju. 2019. "Is Now A Good Time?: An Empirical Study of Vehicle-Driver Communication Timing." In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. Glasgow Scotland Uk: ACM. https://doi.org/10.1145/3290605.3300867.

Sheridan, T.B., and W. Verplank. 1978. "Human and Computer Control of Undersea Teleoperators." Cambridge, MA: Man-Machine Systems Laboratory, Department of Mechanical Engineering, MIT.

Shively, Robert J., Joel Lachter, Robert Koteskey, and Summer L. Brandt. 2018. "Crew Resource Management for Automated Teammates (CRM-A)." In *Engineering Psychology and Cognitive Ergonomics*, edited by Don Harris, 215–29. Cham: Springer International Publishing.

Stilgoe, Jack. 2018. "Machine Learning, Social Learning and the Governance of Self-Driving Cars." *Social Studies of Science* 48, no. 1: 25–56. https://doi.org/10.1177/0306312717741687.

Sumwalt, Robert. Letter to U.S. Department of Transportation. 2021. "National Transportation Safety Board Review of 'Framework for Automated Driving System Safety." Washington, DC: Office of the Chairman, National Transportation Safety Board.

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California