2022

Testing the Lumberjack Analogy: Automation, Situational Awareness, and Mental Workload

Justin W. Morgan
Wright State University

TESTING THE LUMBERJACK ANALOGY:

AUTOMATION, SITUATIONAL AWARENESS, AND MENTAL WORKLOAD

A Thesis submitted in partial fulfillment of the
requirements for the degree of

Master of Science

by

JUSTIN W MORGAN

B.S., Eastern Kentucky University, 2018

2022

Wright State University

WRIGHT STATE UNIVERSITY

GRADUATE SCHOOL

July 20, 2022

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Justin W Morgan ENTITLED Testing the Lumberjack Analogy: Automation, Situational Awareness, and Mental Workload BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science

Assaf Harel, Ph.D.
Thesis Director

Scott N. J. Watamaniuk, Ph.D.
Graduate Program Director

Debra Steele-Johnson, Ph.D.
Chair, Department of Psychology

Committee on Final Examination:

Ion Juvina, Ph.D.

Greg Funke, Ph.D.

Barry Milligan, Ph.D.
Dean of the Graduate School

ABSTRACT

Morgan, Justin W., M.S. Department of Psychology, Wright State University, 2022. Testing the Lumberjack Analogy: Automation, Situational Awareness, and Mental Workload.

This study examines the effects of automation on the human user of that automation. Automation has been shown to produce a variety of benefits to employees in terms of performance and a reduction of workload, but research in this area indicates that this might be at the cost of situational awareness. This loss of situational awareness is thought to lead to "out-of-the-loop" performance effects. One way this set of effects has been explained is through the "lumberjack" analogy, which suggests these effects are related to degree of automation and automation failure. This study recreates the effects of automation on mental workload, performance, and situational awareness by altering the characteristics of automation in a UAV supervisory control environment; RESCHU was chosen because of its complexity and the ability to manipulate levels of control within the task. Afterwards, it will be discussed whether the effects align with the predictions of the lumberjack analogy. Participants were assigned to one of two automation reliability groups, routine or failure, and all participants experienced all three degrees of automation – manual/low, medium, and high. Scores collected for mental workload, situational awareness, and performance were compared across groups and conditions. Results indicated differences in performance for both degree of automation and reliability, but no interaction. There was also a main effect of degree of automation on raw NASA-TLX scores, with a few main effects reported for individual subscales.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

TABLE OF FIGURES

Figures Page

LIST OF TABLES

ACKNOWLEDGEMENTS

This thesis would not have been possible without the support of my advisor and thesis chair, Dr. Assaf Harel, and my thesis committee, Dr. Ion Juvina and Dr. Gregory Funke. I am also grateful for the support from all the faculty and graduate students in the Wright State Psychology Department, especially those who helped with this project.

A special thanks to my wife, Samantha, and my family for being there every step of the way.

Testing the Lumberjack Analogy: Automation, Situational Awareness, and Mental Workload

Introduction

Society has made massive strides in technological advances the past few decades, so it is not surprising that more automation is being introduced into the workplace. Automation has expanded from more “concrete” mechanical systems to abstract and complex work environments that involve human-computer interactions. Businesses/organizations may try to improve their functionality by investing in the latest technology with the goal of improving productivity, the quality of products or services, organization, reduce operational costs, etc. There is no doubt that automation is invaluable when tackling these problems, but these new advances do not come without their own issues. When integrating these systems, people must consider the limitations of technology and the human operator to be effective. To do this, a designer must consider the effects (negative and positive) of the automation they are implementing and how this will alter the operator’s behavior while performing the task. The current study will investigate human-automation teaming to better understand how automation may influence humans.

Automation (IV)

Although many systems are automated, not all systems are automated equally. To study automation properly, early researchers needed a way to classify automation. One of the more well-known classification systems from that time was proposed by Sheridan and Verplank (1978). They classified automation from 1 (all human, no automation assistance) to 10 (autonomous automation, with no human). This attempt at classifying automation focuses on the “level of authority” that the automation has in completing its tasks. This is something that can still be seen in modern classification models of automation.

In fact, one of the most popular classification systems today stems from this earlier model. In this newer classification system, degree of automation was defined in terms of *stages* and *levels* (Wickens et al., 1998; as described in Parasuraman et al., 2000). According to this model of automation, the degree-of-automation (DOA) is affected not only by the level of authority (levels), but also by the type of activity (stages) that the automation supports. Stages refers to how the automation can support four stages of information processing (sensory processing, perception/working memory, decision making, action selection), and these stages are referred to as information acquisition, information analysis, decision selection, and action implementation, respectively (Wickens et al., 1998). This was a notable improvement from the earlier model (the earlier model only focused on the decision-making stage), whereas Wickens and colleagues’ model considers how automation supports the full perception-action cycle. The addition of stages was the largest change, as the idea of “level of authority” within each stage remained the same. Each stage can be rated along the same dimension as the original Sheridan and Verplank (1978) model. What makes this model even more robust though, is that DOA can increase by increasing the level of a single stage AND/OR by increasing the level of a “later stage” of processing–rather than an earlier one (i.e., automating decision making to level 5 instead of sensory processing to level 5 would result in a higher DOA). This model not only expands the possible automation paradigms used for research but can be used as a tool to help designers implement automation.

Performance (DV)

This is not to say that these designers should attempt to maximize automation at all four stages. Endsley and Kiris (1995) found that an operator's situational awareness was affected by level of operator control, which was affected by the level of automation in the decision stage (at the time of the study this was based on Sheridan & Verplank's, 1978, model). While the automation did improve performance, the loss of situational awareness was thought to be caused by operator complacency brought on by highly automated, reliable systems. Problems that occur during the failure of these systems are often referred to as out-of-the-loop performance issues. Rovira et al. (2007) also found that unreliable decision selection automation (60% reliability) led to worse performance than not having automation at that stage. Wickens et al. (2010) also confirms the existence of these out-of-the-loop issues.

Yet, these performance effects are not always as clear as researchers would like to believe. Several studies have failed to report similar effects between DOA and failure performance, or even DOA and situational awareness (Lorenz et al., 2002a; Lorenz et al., 2002b; Kaber & Endsley, 2004). Shaw et al. (2010) used a method of adaptive automation to help improve overall performance and response time to abnormal events when operators had to retake manual control of the task. Therefore, perhaps there are ways of implementing automation to a degree that may mitigate these types of effects, or maybe they are not as robust as earlier research may suggest. In either case, there is a need for more research in this field to determine the kinds of tasks/automation that may lead to these out-of-the-loop deficits.

While the benefits of using automation are obvious, these studies demonstrate that automating a new system is something that must be considered with caution. Parasuraman et al. (2000) created a guide/model which emphasized the importance of understanding what needs to be automated, determining the current and proposed stages and levels of automation for the system, re-evaluating your initial assessment, looking at risks/costs of automation, etc., to help reduce/avoid the unintended consequences that can be brought on by higher levels of automation. Recommendations such as these recognize that automation changes human behavior, and that task constraints play a large role in how automation should be implemented. Models such as these are incredibly important for system designers and should be used to help alleviate potential out-of-the-loop performance decrements, while maximizing the benefit to consumers.

To understand potential automation decrements further, it is important to take a deeper look at the constructs that consistently come up in the literature, such as situational awareness, mental workload, and performance.

Situational Awareness (DV)

Situational awareness (SA) describes someone's "internal model of the world around them at any point in time" (Endsley, 1988). According to Endsley's model of situational awareness, SA can be further divided into three levels: perception, comprehension, and projection. *Perception* deals with the things that an operator might notice with their senses (i.e., a sound, light, color, etc.). *Comprehension* requires integrating these different components/things from the world to deduce a certain state. *Projection* is the ability to determine the proper course of action to follow based on the assessment from level two. For an operator to perform a task successfully, they need to be able to perceive the appropriate things, integrate them appropriately, and then take the appropriate actions.

This progression is presented as cyclical because an active operator is constantly evaluating a scenario to monitor changes. If the automation is performing these functions instead, then the operator is not maintaining this cycle and is only passively involved. This could lead to a vigilance decrement, as the operator’s job has shifted from active control to passive monitoring, and to a loss of operator skill over time (Endsley & Kiris, 1995; Endsley, 1996). Kaber and Endsley (2004) supported this notion by finding a decrease in SA during fully automated conditions. In addition, it has been shown that people are better at recalling information that they help generate (Slamecka & Graf, 1978). Thus, the loss of active participation removes the “generation effect,” which may also explain the negative relationship prior research has found between DOA and SA. However, Fuchs et al. (2013) showed that their adaptive automation display for UAV operators decreased workload while also increasing situational awareness, so, an increase in automation does not always mean a break in this cycle.

From the model proposed by Endsley, SA depends on the operator’s ability to recall important information, to integrate it, and then develop a course of action. If the automation takes over most of this process, it stands to reason that the operator’s own SA would decrease and increase the likelihood of out-of-the-loop performance issues, especially if the automation fails. This does not necessarily mean that all automation will decrease SA, or that automation will never increase SA, but that the detriment to SA is most likely in systems with high DOA.

Mental Workload (DV)

In the same way that increasing DOA is believed to reduce SA, using automation to aid in tasking is thought to reduce mental workload (MWL). In fact, MWL is often grouped together with SA because they are so closely related (e.g., Vidulich & Tsang, 2015). MWL refers to the relationship between the resource demands of a task and the mental resources available to the operator, and so it is commonly associated with attention (Wickens et al., 2013, p. 347-438). This is because the operator is allocating these attentional resources to cope with task demands. SA is typically associated with memory, as it requires the operator to gather and recall information about the current state of the system. On top of these theoretical differences, Endsley (1993) notes that while these constructs are related, they are not dependent on each other.

Once this distinction has been made, it is important to look at the various ways of measuring MWL. First, it is important to note that MWL cannot be captured by task performance alone (Wickens et al., 2013, pg. 350). For example, a task can be more difficult, and so a person may put in more effort (increasing their workload) to compensate and maintain their current performance level. In this case, performance would not show that the operator has a greater mental workload. Disassociations between different measures of MWL have consistently been reported (e.g., Yeh and Wickens, 1988). This can mean a lack of correlation between MWL self-report scores and neurophysiological indices of MWL, significant findings for one measure but not the other, etc. This is not to say that MWL is not useful, but that MWL can be complex, and it often requires multiple measures to ensure the entire construct has been captured. Two major types of these measures are self-reported (explicit) measures and implicit measures.

One way to capture subjective MWL is with a measure provided at the end of a trial, like the NASA-TLX (Hart & Staveland, 1988). The NASA-TLX is a survey instrument that is administered post-trial and requires participants to rate six dimensions of workload (mental demand, physical demand, temporal demand, performance, effort, and frustration) from 0-100 (Appendix B). Since this method relies on self-report, and therefore the operator's awareness of their own MWL, it is considered an explicit measure of MWL. In terms of subjective MWL and DOA, Roettger et al. (2009) found an "almost linear" relationship between the decrease in subjective MWL as DOA increased. Kaber and Endsley (2004), Rovira et al. (2007), and Lin et al. (2019) have found similar trends in subjective measures of MWL, but not as strong. Yet, other studies reported null findings for subjective MWL and DOA (Endsley & Kiris, 1995; Lorenz et al., 2002a). Differences in the findings could stem from differences in the tasks being completed (i.e., automation helping with detecting faulty products, vs ai helping with decision making), which further emphasizes the importance of considering task constraints. A benefit to using a measure like this is that people have a general impression of how easy or hard a task is. A major criticism of subjective measures like these is that these measures are administered after the task, so participants may not be fully aware of their workload during the entire task. In addition, it cannot capture the progression of MWL throughout the task, and so is often seen as a coarse, more fallible measure of workload.

As with SA and DOA, the relationship between MWL and DOA is not always as clear as researchers would like it to be, but the general trend is that increasing DOA tends to decrease MWL (Wickens et al., 2010). This does not mean that operators are more equipped to handle automation failures. Hancock (1989) found that operators experienced greater workload during automation failure than when they were operating manually. There is also a chance that employers may see this decrease in operator workload as an opportunity to provide workers with additional tasks to complete. In this case, an automation failure could result in even greater operator overload, leading to greater performance decrements as the operator now has more tasks to manage.

The Lumberjack Analogy

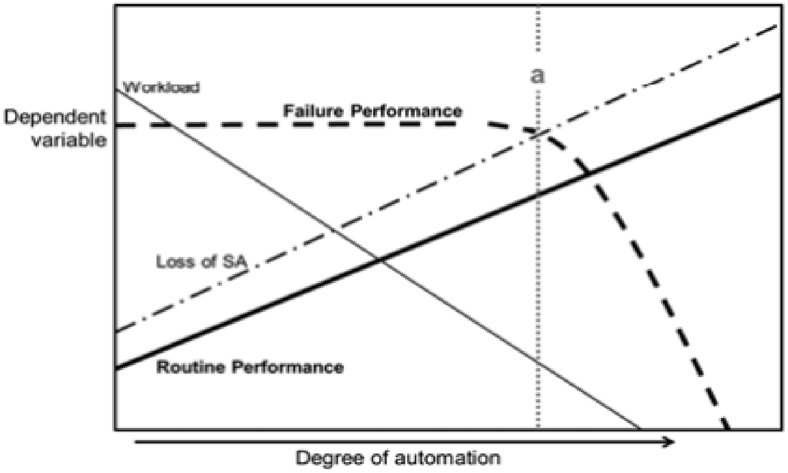
When comparing two systems, or two versions of the same system, it's easy to see that they can differ in the amount of automation within them. What might be harder to see is how these differences in DOA affect the people operating the system. The studies mentioned above, in addition to numerous other studies in the area, led to a meta-analysis of studies involving human-automation interaction. This meta-analysis reported that higher levels of automation are associated with critical performance deficits during an unexpected automation failure (Onnasch et al., 2014). This same article goes on to elaborate that these deficits can be predicted by the lumberjack analogy ("the bigger they are, the harder they fall") described in Onnasch et al. (2014). This analogy arose from the results of earlier studies on human-automation interaction and investigations into real-life automation failures (many involving aircraft).

Although real-world automation failures are less common than their experimental counterparts, they are very important to consider because many people only see the benefits of automation and do not consider the risks involved with making a process more automated. In certain high-cost, time-sensitive industries (i.e., aviation or nuclear energy), automation failure in a highly automated environment could lead to a catastrophic failure, such as a crash or meltdown. Despite these risks, it also does not make sense to be so afraid of automation to never enjoy its benefits. In fact, Endsley (1996) discusses the possibility of automation actually increasing SA by aiding perception and comprehension of incoming stimuli. In these situations, the automation is being implemented at a lower stage of processing and keeps the operator in the loop.

Overall, this set of effects has led to the aforementioned "lumberjack analogy," as discussed by Onnasch et al. (2014; Figure 1). This model predicts that systems with higher DOAs, which are functioning properly, should only increase the operator's routine performance. Conversely, systems with higher DOAs may also create larger deficits in performance, compared to manual control, primarily when the automation fails (failure performance). This DOA model, or "lumberjack analogy," is aligned with the DOA trade-off between MWL and SA that were reported in the Wickens et al. (2010) meta-analysis.

Figure 1

Lumberjack Analogy Diagram



Note: This figure shows the relationships between degree of automation, mental workload, situational awareness and performance as outlined in the text (from Onnasch et al., 2014).

As with any model there are skeptics and critics that seek to test its boundaries. Jamieson and Skraaning (2019) attempted to do this by expanding the DOA model into a complex work environment (compared to more simplistic, controlled lab studies) but was unable to replicate its predictions. This is significant because one of Jamieson and Skraaning’s reasons for doing this was the lack of complex, real-world work environments being used in these studies. In fact, they reported an increase in SA, no effect on workload or task performance, and a decrease in out-of-the-loop performance at greater DOAs. Wickens et al. (2019) rebuked their findings stating that it was not a proper test of the lumberjack analogy.

Some of Wickens et al.’s (2019) major points were that their study did not include a comparison of routine performance, an actual measure of SA, and that the failure performance was based on failure of the underlying system and not the automated system. Without a comparison to routine performance, there is no way to determine the difference in performance from routine to failure. Similarly, it is inappropriate to operationalize SA with a measure of task knowledge and say that the SA is increasing. The last point listed is an important distinction, as the lumberjack analogy refers to automation failure and not how automation helps manage system failures. Jamieson and Skraaning (2020) conceded some of these points and argued others but concluded that this DOA model needs additional research before it can be applied to complex work settings. The current study addresses the concerns that Wickens et al. had with the Jamieson and Skrannings’ study, while also pushing into the more complex work environments that Jamieson and Skranning mentioned were lacking in this research area.

Rationale

Although there have been numerous studies looking at the effects of automation on workload and situational awareness (see above), the lumberjack analogy is a relatively recent model. The effects proposed by this model, although generally supported, have had mixed results in the literature and this model has recently been criticized for not living up to its predictions in complex environments. If the generalizability of the model is in question, it is important to study the model in the context of these more complex environments so that it can be applied to real-world problems. Although it can be difficult to imitate real-world work environments in a laboratory setting, some tasks are naturally more complex/realistic than others. Given the constraints of the study, using a testbed like the Research Environment for the Supervisory Control of Unmanned Aerial Vehicles (RESCHU; Donmez et al., 2010), which imitates a display that a UAV operator may be using, over the MAT-B (Comstock & Arnegard, 1992), which emulated a small two-engine aircraft of the 1970s but no longer resemble a real-world display, could serve as a closer approximation to a more futuristic, complex work environment.

The current study aims to target these deficits in the literature by testing the predictions of this model using multiple versions of a complex system (each with a different DOA; see Methods). Testing the predictions of this model requires this study to show that increasing DOA should benefit performance, as automating a task should inherently reduce the demands of that task. This in turn would reduce the operator's MWL, while also reducing the operator's engagement with the system (resulting in a loss of SA). Conversely, if the automation does not reduce the operator's engagement, there is a possibility that SA could actually increase. The loss of SA that comes from "over-automating" a system should lead to greater performance decrements when the automation unexpectedly fails (is taken away or suddenly becomes unreliable). The hypotheses that were generated from these predictions and deficits are to be tested in the following experiments.

Hypotheses

Hypothesis 1a: Participant performance will be highest in the high DOA condition, lower in the medium DOA condition, and lowest in the manual/low DOA condition, and their performance will be higher in routine conditions than in failure conditions.

Hypothesis 1b: Participant performance will be highest under high DOA, and lowest under low DOA for the routine reliability conditions only, but vice versa for the failure conditions.

Hypothesis 2: Participant SA will be lowest at the highest DOA and highest at the lowest DOA.

Hypothesis 3: Participant MWL will be highest under the lowest degree of automation and lowest under the highest degree of automation, and the failure conditions will be higher in MWL than the routine conditions.

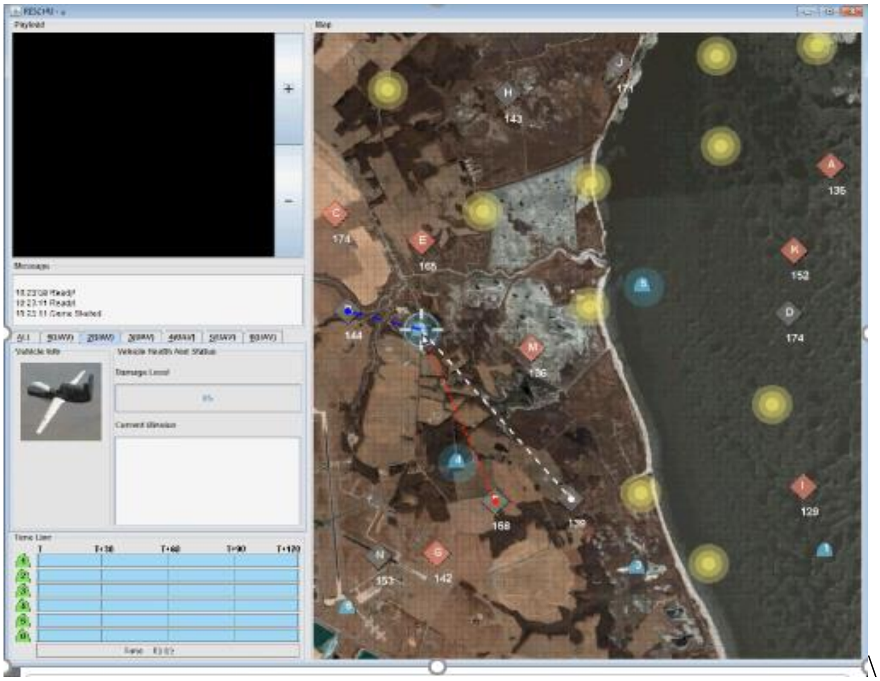
Methods

Participants

Participants (*n* = 28) were recruited from undergraduate psychology courses via the SONA research system. Each participant reported their gender (71.4% women, 25% men), race (78.6% white, 17.9% black/African American), and age (*M* = 20.25, *SD* = 6.35). Each student was awarded SONA credits for their participation in the study.

Figure 2

RESCHU



\

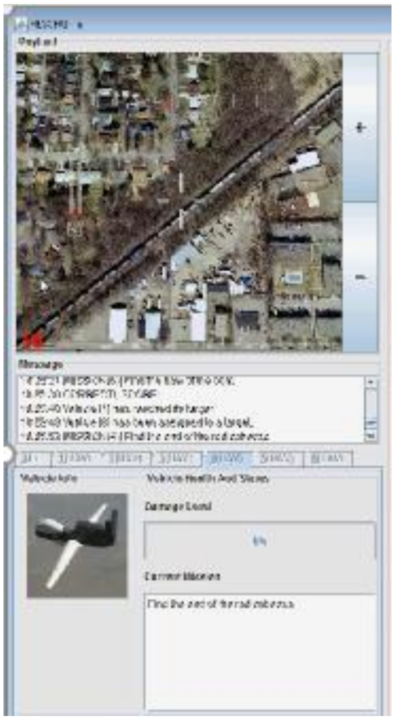
Note: The entire RESCHU task environment. The right side is where UAV paths are set to target locations, and the left is where mission details are located and payload targets are identified.

Experimental testbed: RESCHU

The testbed for this experiment is the Research Environment for the Supervisory Control of Unmanned Aerial Vehicles (RESCHU; Donmez et al., 2010; Figure 2). This program is a futuristic example of an interface that simulates a single operator controlling multiple unmanned aerial vehicles. Currently, it takes multiple operators to control a single UAV, which is not as efficient as a single operator controlling multiple UAVs. This type of environment offers greater complexity than simple laboratory experiments and should provide a more realistic test of this DOA model. While there are a variety of programs that are meant to imitate these types of environments (e.g., ALOA multi-UAS research test bed developed by Johnson et al., 2007; as cited by Lin et al., 2019), this is a program that is meant to imitate future UAV control and has been shown to be viable in past research (e.g. Donmez et al., 2010, Cummings et al., 2019). The purpose of using an environment like this is because the complexity of these types of tasks creates a demanding environment that is useful for testing these DOA effects (lumberjack analogy).

Figure 3

RESCHU Payload Task and Info



[illegible]
[illegible]
[illegible] Vehicle [1] has reached its target.
[illegible] Vehicle [6] has been assigned to a target.
[illegible]

Find the end of the rail [illegible]

Note: The payload/search task is pictured above with the UAV and target description listed below. Participants pan through the image on top to locate the specified target.

In this program, the operator is in control of multiple UAVs (6 total) and they must set a path for the UAVs to reach “targets” while also avoiding “hazard zones.” Once a UAV reaches a target, the operator must complete a search task (Figure 3) within 30 seconds. A target description (e.g., “yellow train car”) appears in the system log and the operator must pan and zoom in an overhead image (e.g., an aerial view of a city) to find and then click the target (top of Figure 3). After clicking, the operator receives feedback in the system log (success or failure) and then they must reassign the UAV to a new target. In addition, the RESCHU interface features a status window for the UAVs which includes their health and another window showing the timelines for the UAVs (bottom of Figure 3).

Measures

Performance. Task performance was measured using two primary performance metrics, 1) total score (a composite score based on UAV arrivals, UAV damage, and success/failure in finding payload targets) and 2) UAV utilization (proportion of time UAVs spent traveling to a target or engaged in a payload task). UAV utilization was an average of the utilization rates for all six UAVs used during a trial. See Crandall and Cummings (2007) and Mancuso et al. (2015) for additional approaches to operationalize task performance in supervisory control environments.

Mental Workload. This construct was measured after each trial with the NASA-TLX questionnaire (Hart & Staveland, 1988; Appendix B) and UAV utilization (Cummings et al., 2019). The NASA-TLX involves two parts. First, there is a pairwise comparison of the six subscales (Mental Demand, Physical Demand, Temporal Demand, Performance, Frustration, Effort) which is used to create weights for the task in question. Once all comparisons are made, each dimension is tallied up to create the weight. Then, following each trial, participants rate each subscale from very low (0) to very high (100). These post-trial scores on each subscale are multiplied by the corresponding weight, combined, and divided by 15 (the number of comparisons) to create the weighted scores. Participants completed the pairwise comparisons after the last practice trial, and the subsequent subscales were each rated using a sliding bar on a line (slider) after each experimental trial.

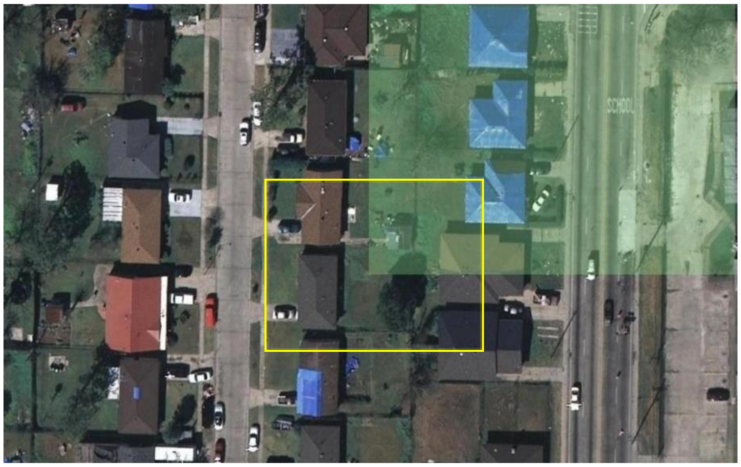
Situational Awareness. Situational awareness was measured with the situational awareness rating technique (SART; Taylor, 1990; Appendix C). There are ten questions that are structured similarly (i.e., "How changeable is the situation? Is it very stable and straightforward (low) or is the situation highly unstable and likely to change suddenly (High)?"). These questions are rated from 1 (low) to 7 (high). This was administered after each trial, after the NASA-TLX. The overall score for the self-assessment was computed by taking the sum of the items for all three dimensions (Understanding, Demand, and Supply) and then using those in the following formula: SA = U - (D - S).

Procedure

To examine the effects of Automation Reliability (between) and DOA (within) further, each participant was placed into either the routine automation group or the failure automation group. We also used three task configurations of varying degrees of automation (low, medium, high), but each participant – regardless of group – experienced all three of these configurations. In the first system, users searched for targets without any assistance (low DOA/manual). In the second, the quarter of the map containing the target was shaded light green, but the participant still had to rely on the description to find the target (medium DOA; Figure 4 – the yellow box indicates roughly how much of the image the participant can see in the window). In the third, the target was surrounded by a smaller, green circular region so that participants only needed to locate this circle (high DOA; Figure 5). For participants in the routine automation condition, the target was always located within the shaded region. In the automation failure conditions, the target was in the shaded region 80% of trials. 80% reliability was chosen to make the automation failures more like unexpected failures than a persistent flaw in the automation itself.

Figure 4

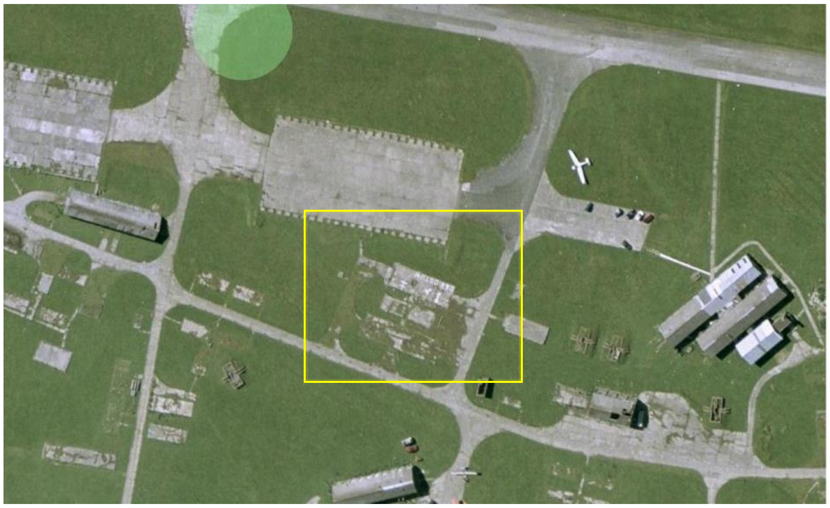
Medium DOA Condition



Note: In the medium DOA condition, participants were told that the target was located within the green area. The yellow box is all the image that the participant could see at any given time.

Figure 5

High DOA Condition



Note: In the high DOA condition, participants were told that the target was located within the green circle. The yellow box indicates the amount of the image a participant could see at any given time.

In the first part of the session, participants completed demographic information (Appendix A) and then went through a PowerPoint slideshow introducing the RESCHU system (including a brief tutorial video). Afterwards, they practiced the task for each DOA for 5 minutes each (starting with the lowest DOA then moving up). For both reliability groups, the automation during training was 100% accurate. This was important in establishing the expectation that the automation works properly so that the automation failures were true failures and unexpected. At this point, participants completed the pairwise comparisons of the NASA-TLX and were offered a chance to take a break before beginning the experimental rounds. The order of the experimental conditions was randomized for each participant. After every trial each participant answered the NASA-TLX and SART questionnaire (a total of three time). The entire session took no longer than 2 hours.

Results

To test the three hypotheses for the study, the data from 28 participants was analyzed using multivariate ANOVAs. These models were created using degree of automation (low, medium, and high) and reliability (routine or failure) as the independent variables and the measures of performance, situational awareness, and mental workload as the dependent variables.

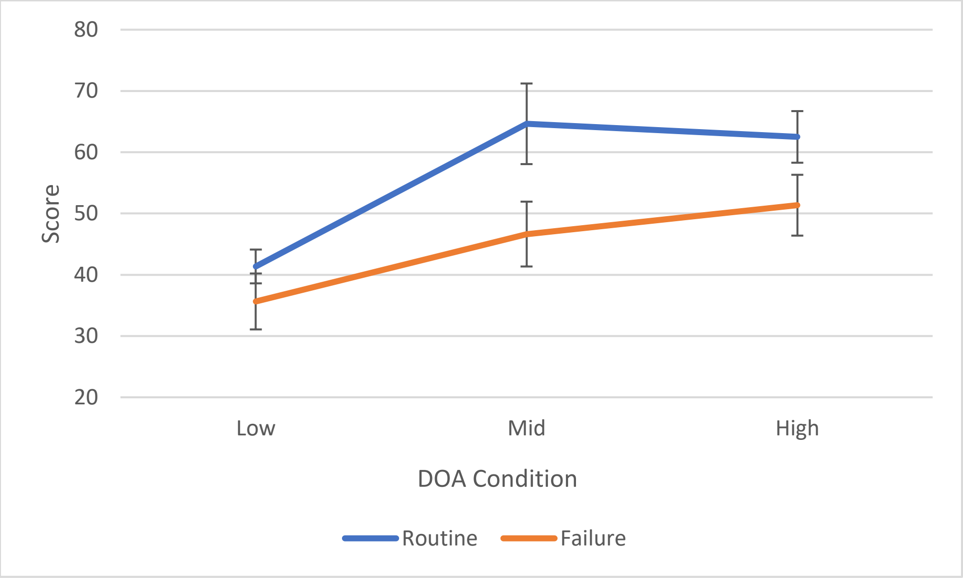
Performance

The first hypothesis was tested by entering the calculated participant performance score into the model with degree of automation and reliability condition (routine vs failure). Since there are three levels to the repeated measure, the Greenhouse-Geisser adjustment was used on the degrees of freedom (Grieve, 1984). Although there were several metrics of performance available, the correlation matrix (Table 1) showed that many of these were highly correlated with the composite performance score (as many of them were included in this score). Average utilization and accuracy were not, so these were also explored as potential performance metrics in separate models. The mean participant performance scores (Figure 6) were significantly different across DOA, $F(1.82, 47.36) = 29.71$, $p < .001$, ηp^2=.2, and reliability conditions, $F(1, 26) = 8.85$, $p < .01$, ηp^2= .16. There was no significant interaction between DOA and condition, $F(1.82, 47.36) = 2.66$, p = 0.09, ηp^2=.02. A post-hoc pairwise comparison using the conservative Bonferroni correction found that participants in the medium and high DOA conditions had significantly higher performance scores than participants in the low DOA condition, but there was no significant difference between the medium and high DOA conditions.

This supports Hypothesis 1a for the main effects of DOA and reliability on performance scores but does not support the interaction between DOA and reliability in Hypothesis 1b.

Figure 6

Participant Performance Scores by Degree of Automation Across Reliability Groups.



Note: Score is not out of 100 but rather represents the composite performance score calculated from correct/incorrect target identification, UAV arrivals, and UAV damage.

Table 1

Correlation Matrix for Performance Metrics

	DOA	Trial	Score	Condition	Util	Errors	Correct	Arrivals	DmgEvents	Attempts	Accuracy
Trial	0										
Score	0.38*	0.14									
Condition	0	0	(-0.29)*								
Avg. Utilization	-0.04	0	0.1	-0.1							
Errors	-0.08	0.02	-0.28*	0.30*	0.25*						
Correct	0.40*	0.16	0.98*	-0.33*	0.07	-0.26*					
Arrivals	0.31*	0.15	0.94*	-0.27*	0.13	-0.23*	0.94*				
DmgEvents	0.09	0.11	-0.15	-0.23*	-0.03	0.08	0.04	-0.003			
Attempts	0.08	-0.06	0.16	-0.05	0.14	-0.01	0.19	0.17	0.13		
Accuracy	-0.04	-0.07	0.06	-0.03	0.16	0.16	0.09	0.09	0.14	0.59*	
TBP	-0.27*	-0.39*	-0.57*	0.23*	0.05	0.07	-0.60*	-0.58*	-0.09	-0.08	-0.03

Note: Condition = Reliability (Routine vs Failure); TBP = Time Between Payloads; Util = Average UAV Utilization Rate; Trial = Order of trials; Damage Events occur when a UAV travels through a hazard zone

* $p < 0.05$

The other metrics of performance for this study were also analyzed with ANOVAs using the Greenhouse-Geisser adjustments. Average participant UAV utilization (Figure 7) was entered with DOA and reliability, and there were a significant main effects for DOA, *F*(1.94, 50.49) = 7.66, $p < .001$, ηp^2=.09, and reliability condition, *F*(1, 26) = 6.01, $p < 0.05$, ηp^2=.13, but no interaction, *F*(1.94, 50.49) = 2.68, $p = 0.08$. A post-hoc test using the conservative Bonferroni adjustment found lower UAV utilization in manual DOA conditions than in medium DOA conditions, while the failure reliability group had the lowest utilization amongst group. Accuracy (Figure 8) produced no significant main effects for DOA, *F*(1.79, 46.42) = 0.06, $p = 0.79$, reliability, *F*(1, 26) = 0.08, $p = 0.93$, or their interaction, *F*(1.79, 46.42) = 0.54, $p = 0.57$. These non-significant findings provide no additional support for Hypothesis 1a or 1b.

Figure 7

Average UAV Utilization by Degree of Automation Across Reliability Groups

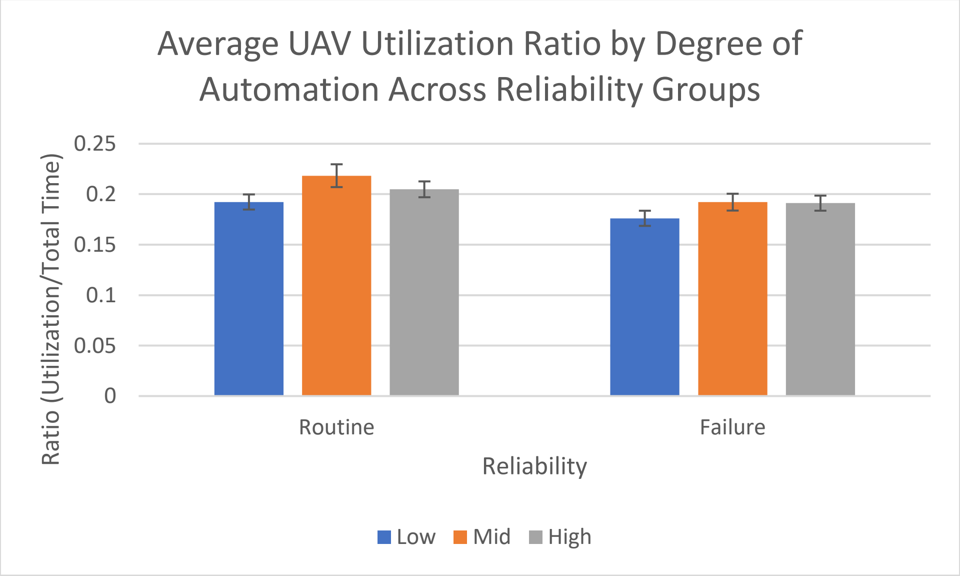
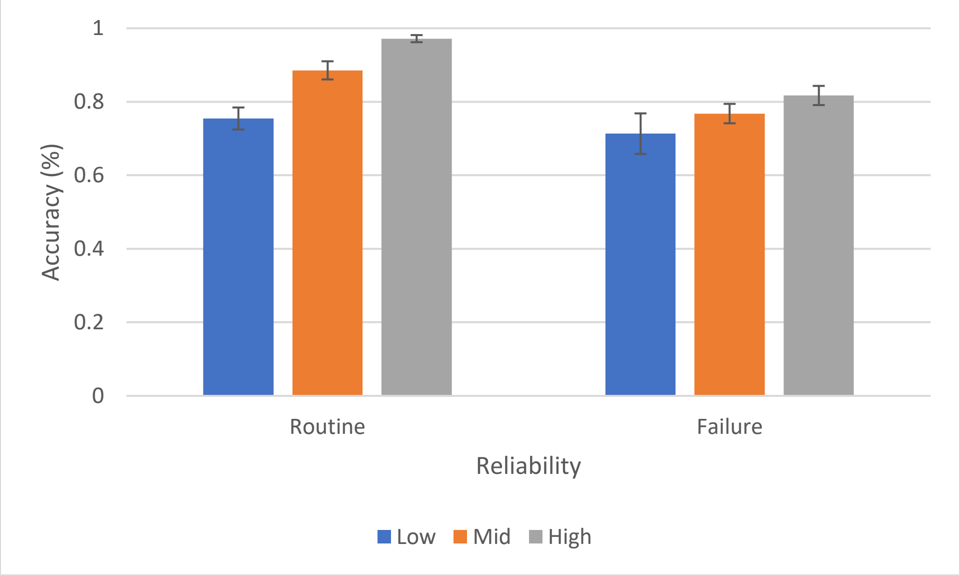


Figure 8

Participant Accuracy by Degree of Automation Across Reliability Groups

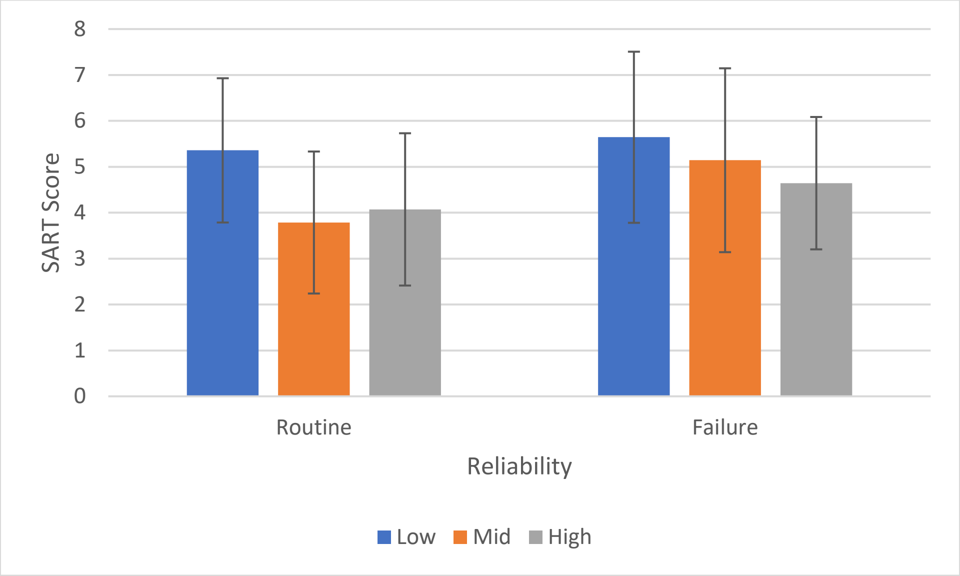


Situational Awareness

To test Hypothesis 2, another ANOVA was used to examine the difference in SART scores using the same method as above. Composite SART scores (Figure 9) obtained from the survey following each task were entered into a model with DOA and Reliability. There were no main effects for DOA, *F*(1.99, 51.8) = 0.55, *p* = .58, reliability, *F*(1, 26) = 0.16, *p* = .69, or the interaction between the two independent variables, *F*(1.99, 51.8) = 0.11, *p* = .9.

Figure 9

Participant Situation Assessment Rating Technique (SART) Scores by Degree of Automation Across Reliability Groups



Mental Workload

Finally, Hypothesis 3 was tested using the NASA-TLX scores and sub-scores obtained from the survey following each task. The initial test of hypothesis 3 was conducted using a standard metric of MWL, which is the weighted NASA-TLX scores (Figure 10). These were entered into the ANOVA model with DOA and Reliability. There were no main effects for DOA, F(1.89, 49.26) = 2.7, p = .08, Reliability, F(1,26) = 0.02, p = 0.88, or the interaction between the two independent variables, F(1.89, 49.26) = 0.46, p = 0.62. This provides no support for Hypothesis 3. Nevertheless, some additional analyses were conducted, based on past research on the NASA-TLX measure. Some researchers (e.g., Moroney, 1992) suggest it is preferable to just use the raw scores, while others note potential issues that can arise while using the traditional NASA-TLX weights (e.g., Virtanen et al., 2021). Therefore, to examine mental workload further, raw NASA-

TLX scores were utilized (Figure 11) and a new model was ran. There was a main effect for DOA, $F(1.62, 42.11) = 7.24$, $p < 0.01$, ηp^2=.05, but not for reliability, $F(1, 26) = 0.08$, $p = 0.78$, ηp^2=.00, or the interaction between DOA and reliability, $F(1.62, 42.11) = 0.78$, $p = 0.44$, ηp^2=.01. A post-hoc pairwise comparison using the conservative Bonferroni correction found that participants in the low DOA trials reported significantly higher raw NASA-TLX scores than the high DOA groups ($p < 0.05$). When using the raw scores instead of the weighted scores, there is partial support for Hypothesis 3.

Figure 10

NASA Task Load Index (NASA TLX) Weighted Scores by Degree of Automation Across Reliability Groups.

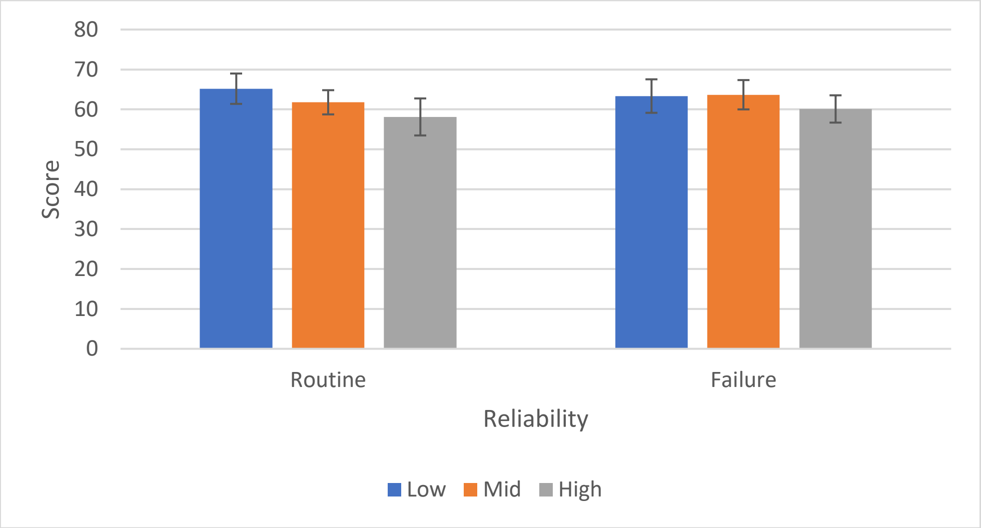
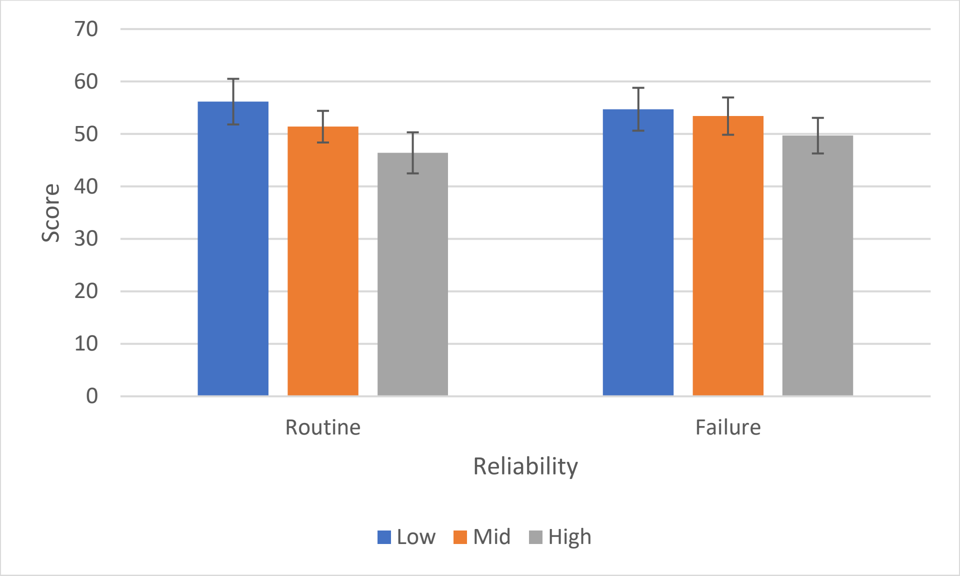


Figure 11

NASA Task Load Index (NASA TLX) Raw Scores by Degree of Automation Across Reliability Groups



Another way to investigate MWL is to examine the raw scores of the subscales of the NASA-TLX (e.g., Galy et al., 2018). To reduce familywise error, a 2 (Reliability) x 3 (Degree of Automation) x 6 (Subscale) MANOVA was conducted to test for difference in reliability groups and DOA conditions across subscale scores. There was a main effect for DOA, Pillai's Trace = 0.52, F(2, 52) = 2.8, p < .01, but not for Reliability, Pillai's Trace = 0.39, F(1, 26) = 2.3, p = .08 or the interaction between DOA and reliability, F(2, 52) = 0.44, p = 0.43. Individual univariate ANOVAs, with Greenhouse-Giesser adjustments, were conducted to examine each of the subscales across DOA conditions while any post-hoc analyses included the Bonferroni adjustment. Mental Demand had a significant main effect for DOA, $F(1.7, 46) = 6.86$, $p < .01$, ηp^2=.03, A post-hoc analysis showed that the average mental demand subscale rating was significantly higher in low DOA conditions than high DOA conditions. Physical Demand had a significant main effect for DOA, $F(1.74, 47.07) = 5.04$, $p < .05$, $\eta p^2=.03$, A post-hoc analysis showed that the average physical demand subscale rating was significantly higher in low DOA conditions than medium or high DOA conditions, but there was no significant difference in medium and high. Temporal Demand did not have a significant main effect for DOA, $F(1.98, 53.48) = 1.54$, $p = .22$, $\eta p^2=.02$. Performance had a significant main effect for DOA, $F(1.81, 48.88) = 12.2$, $p < .001$, $\eta p^2=.15$, A post-hoc analysis showed that all three conditions were significantly different from each other. Effort had a significant main effect for DOA, $F(1.69, 45.58) = 8.13$, $p < .01$, $\eta p^2=.09$, A post-hoc analysis showed that the average effort subscale rating was significantly higher in low DOA conditions than medium or high DOA conditions, but there was no significant difference in medium and high. Frustration also had a significant main effect for DOA, $F(1.71, 46.14) = 9.27$, $p < .001$, $\eta p^2=.06$, A post-hoc analysis showed that the average mental demand subscale rating was significantly higher in low DOA conditions than medium or high DOA conditions, but there was no significant difference in medium and high.

Discussion

This study examined the effects of characteristics of automation - degree of automation (DOA) and reliability – on performance, situational awareness (SA), and mental workload (MWL). These effects have been previously captured by the lumberjack analogy, which demonstrates not only the benefits of automation but also potential dangers of using too much automation (Onnasch et al., 2014). From this analogy, increasing automation not only increases task performance, but also reduces mental workload. The caveat here is that this comes at a cost of situational awareness. While this may not be an issue when the automation is functional, the lumberjack analogy goes on to explain that when a highly automated system fails there will be drastic performance deficits as users struggle to regain their lost situational awareness. While the lumberjack analogy provides a good lens to summarize the set of effects within this body of research, the current research sought to better understand the effects of automation on the users of automation by assessing the extent to which the lumberjack analogy would also hold in a complex, supervisory control environment such as the RESCHU. At the same time, the current study employed conditions similar to those used in previous studies of the lumberjack analogy. Next, the results of the study will be compared against the predictions of the lumberjack analogy itself, and the notable differences will be discussed.

Performance: *Hypothesis 1*

Regarding performance, participants earned higher scores in the routine reliability condition than in the failure condition and they scored better with assistance from the automation (medium and high DOA) than without it (manual). Yet, participants in the failure group didn’t experience the effects of DOA any differently than the routine group.

These results are not totally in line with the Hypothesis 1 predictions for the lumberjack framework. Although both main effects were supported for Hypothesis 1a, the lumberjack analogy would predict an interaction between these two factors. For the routine automation group, it was predicted that performance would be best in the high DOA condition and worst in the low DOA condition, but vice versa for the failure automation group. The difference in score between the two reliability groups arguably came from the effects of these failures, which led to inaccurate responses, more time spent on payloads, and missed points. However, scores from the high DOA condition were the highest and the ones from the low DOA condition the lowest, regardless of group. Hypothesis 1b predicted that in the failure group, the scores would be the highest in the low DOA condition and lowest in the high DOA condition. In other words, having unreliable automation assistance was found to be more beneficial than no automation assistance at all.

The composite performance score captured many aspects of the participants' performance, but the lack of a correlation between this score and UAV utilization rate and accuracy led to these metrics being analyzed separately. UAV utilization rate showed similar significant findings as the overall composite score, but the same cannot be said for accuracy. Despite these significant results, these results would not only indicate that the lumberjack analogy is not as robust as it might seem, but that the overall risks to performance from faulty automation on any DOA are minimal compared to manual performance. Although these results are not aligned with studies reporting negative effects of unreliable automation (Rovira et al, 2007), they strengthen the findings of other studies that failed to find a relationship between failure DOA and performance (e.g., Lorenz et al., 2002a, Lorenz et al., 2002b). These results are very reassuring for users of automation as they show that even unreliable automation can help improve performance over no automation.

Situational Awareness: *Hypothesis 2*

SART scores did not support the predictions of hypothesis 2 for the lumberjack analogy. It was predicted that SA would be lowest when DOA is highest, and highest when DOA is lowest. Not only were these scores highly variable, but there was no clear distinction in SART scores between reliability groups or DOA conditions. The lumberjack analogy would predict that SA should be lower in interfaces with higher DOA. The lack of a pattern may indicate several things. First, it is possible, but unlikely, that SA is unaffected by changes in DOA and reliability. A second possibility is that the measure of SA was not sensitive enough for this task, which might be evidenced by the variability in the responses. Thirdly, it is also possible that the manipulations were not strong enough to reduce SA enough to produce the deficit in performance that were predicted by the lumberjack analogy model. In either case, the results for hypothesis 2 are not aligned with the predictions of the lumberjack analogy.

Although some prior research has failed to find a lack of connection between degree of automation and situational awareness (e.g., Fuchs, 2013), there is a lot of evidence to support a connection between situational awareness and degree of automation (e.g., Kaber & Endsley, 2004). Thus, our finding that SA was not substantially impacted by the automation was one of the more surprising findings of the study. It therefore seems likely that the lack of significant results stems from certain issues with how SA was operationalized, how automation was manipulated, or both. Endsley et al. (1998) compares SART to the SAGAT (a more implicit measure of SA) and highlights differences in how they capture SA. The SART measures the participant's overall perception of their own situational awareness after the task, and it is possible these participants were unaware of their own SA during the task. In that way, it may not have been sensitive enough to capture the changes in SA that might have occurred during automation failure.

However, another possibility is that the automation may have been "adaptive" enough to keep participants in-the-loop and preserve SA in the same way the automation did in Fuchs' (2013) study, or perhaps participants had enough time to regain SA after experiencing the automation failure. In either case, this could also explain the lack of an interaction for hypothesis 1b, but that is much more difficult to prove without testing again with modifications to the design. Future studies attempting to capture SA should ensure their measures of SA are appropriate for the task context. However, assuming that our results were correct, it takes stronger manipulations in automation to bring out significant changes in SA and it is extremely important to verify the DOA for the entire system and perform manipulation checks while trying to examine changes in SA.

Mental Workload: ***Hypothesis 3***

The final hypothesis that was tested was the last component of the lumberjack analogy - mental workload. It was predicted that MWL would be lowest in high DOA conditions, and highest in low DOA conditions. The weighted scores produced by the NASA-TLX were used in the analysis to explore the relationship between MWL, DOA, and reliability. Surprisingly, the analysis of these variables revealed that participants had similar weighted scores between reliability groups, across all three interfaces. This was unexpected because participants were performing better but they did not report a decrease in MWL. This is not in line with the lumberjack analogy either, because the lumberjack analogy proposes that increasing the DOA should decrease MWL. Furthermore, it is interesting that intermittent automation failures did not result in any notable changes in MWL.

Although the scoring guide for the NASA-TLX outlines creating weighted scores for the final product of the NASA-TLX, some researchers believe the weights are unnecessary and that researchers can use “raw” scores for analyses (e.g., Moroney, 1992; Nygren, 1991). To explore mental workload further Raw NASA-TLX scores were also analyzed. It was found that participants had higher raw TLX scores on manual condition than the high DOA condition, which supports hypothesis 3. Discrepancies between raw and weighted NASA-TLX scores are not uncommon (e.g., Virtanen et al., 2021), which was one of the reasons they were also analyzed, and the results of analyzing the raw scores did show effects that support the predictions of the lumberjack analogy.

Another proposed way of examining the NASA-TLX, and thus another way to explore mental workload further, is to look at values of the individual subscales, each targeting a specific component of workload (e.g., Galy et al., 2018). Perhaps unsurprisingly, by adopting this approach, I found that most sub-scales supported the significant findings from the raw scores. All the subscales, except temporal demand, had significant differences between DOA condition means. Post-hoc tests showed that the significant differences were between the manual condition and medium or high conditions (except mental demand which was just manual and high, and performance which had differences for each pair). The caveat to the scores on this performance subscale though, is that participants could see their personal score for each trial in the lower left of the window.

These analyses further examined the relationship between Mental Workload, DOA, and Reliability and the findings from the raw scores and subscales are better aligned with hypothesis 3 and the predictions of the lumberjack analogy. The significant findings for mental workload, and the lack of significant findings for situational awareness, emphasizes that mental workload and situational awareness may not change together and that they should both be measured when studying automation.

Virtanen et al. (2021) explains that weighted scores should be used whenever it is believed that the dimensions for a task are not equally important. That was the belief with the RESCHU task, but perhaps the significant differences in raw scores and subscales indicate that these dimensions could be equally important in this environment, or that the weights actually are unnecessary (Nygren, 1991). If so, then these results are much more supportive of the hypothesis 3 than expected, and the analysis of the subscales may be more meaningful and reflective of prior research on MWL and automation (e.g., Wickens et al., 2010). Conversely, if the weighted scores should be preferred, then these findings indicate participants did not report any difference between automation groups or conditions which is not in itself a unique finding (e.g., Endsley & Kiris, 1995, Lorenz et al., 2002a). Considering the meta-analytical work of Wickens et al. (2010) and support for alternate scoring methods for the NASA-TLX, I would say that automation does affect mental workload and that it is likely that the weights were unnecessary in this context.

The Lumberjack Analogy

Overall, the results went mostly against the predictions of the lumberjack analogy that were captured by the hypotheses. In fact, the crux of the lumberjack analogy was not supported in this context at all as participants did not experience any negative effects from the automation. Instead, it was found that participants only benefited from the automation. This is great news for many developers and consumers of automation as it indicates that only certain contexts may lead to these catastrophic performance deficits. In fact, the current results demonstrate the importance of the exchange/discussion between Jamieson and Skraaning and Wickens, as these relationships are much more complex than initially portrayed.

This exchange implies a disconnection between what designers of automation are learning from studies on automation. Although warnings such as the lumberjack analogy could be useful in many contexts, especially in high-risk contexts like aviation, they could be overgeneralized to environments that do not carry the same consequences. For example, for our study there were very minimal consequences for automation failure besides the inconvenience of manually having to search for the target and potentially losing a point. Tatasciore et al. (2020) also argues that the temporal demands of the task can moderate the lumberjack analogy. In less time-sensitive situations, the lumberjack analogy is less impactful as operators have time to regain control of the situation. The lack of significant changes in participant NASA-TLX temporal demand subscale scores across conditions and groups may indicate there may have not been enough temporal demand during automation failures.

While it is important to caution developers and users of automation, it is also important to understand which contexts may allow for greater automation to maximize benefits to consumers. All research in this area should take the utmost care to ensure proper operationalization of their variables (manipulations and measures) from their corresponding theoretical constructs - as discussed in Wickens et al. (2019).

Limitations

There are several limitations to this research that warrant further investigation. First, participants had all their training and the experimental sessions in the same day. These participants were inexperienced psychology undergraduates, rather than trained UAV operators, who were likely driven to try harder by the novelty of the task and it is likely that they were still improving their scores during their experimental session. It is possible these participants experienced a loss of engagement after this novelty wore off, and their lack of experience may have played a role in some of our findings. Attempts were made to minimize these risks such as requiring participants to type a statement about “trying their best” and “answering honestly,” going through a PowerPoint training, brief tutorial video, and practicing all three conditions before being offered a break.

Second, there may not be sufficient power in the manipulations. There were only 14 participants per reliability group, and any decrements suffered from automation failure might have had limited impact (it was only one portion of the overall task; a good tip for minimizing risk with automation) on the overall system, especially in the limited time they had for each trial. Furthermore, of the three DOA conditions, the highest DOA condition may have not been automated enough to produce the benefits/decrements that you would expect to see. It would have been ideal if participants were prompted to select “yes” or “no” for the target instead of having to locate the identifying circle, but this smaller circle still identifies the target for the user which is more than the green rectangle. Unfortunately, limitations in the customizability of the testbed made some changes impractical to implement.

Lastly, no additional, implicit measures for the subjective measures of MWL or SA were explored in the current work. MWL was measured using the explicit measure of the NASA-TLX only. UAV utilization could serve as a rough secondary task performance metric for MWL, but it is dependent on the same mental resources as the primary task, so it is not ideal. The NASA-TLX is not an implicit measure of MWL, such as those generated from EEG data. There is a similar issue with the measure of situational awareness, SART. This is an explicit measure of situational awareness that is administered post-task. Although the SART was selected because it is a valid measure and less disruptive to the task (compared to a measure like the SAGAT; Endsley, 1988), it may be inferior to a more implicit, invasive measure of SA such as the SAGAT (Endsley et al., 1998). This makes it much more likely that something extraneous could have affected how the participants responded to these survey items. To reduce chances of misunderstandings, instructions were provided on every measure each time they were presented, and the researcher was available if the participants needed clarification on anything.

Future Research

Future research on automation should focus on using implicit measures to capture mental workload and situational awareness and should begin by testing extremes in the degree of automation being implemented. This would ensure more real-time/reliable data for these variables than the self-reported, end-of-task surveys, and by using extremes it should lead to the greatest difference in responses.

To further explore the lumberjack analogy more tasks/testbeds/experimental designs need to be tested. Ideally, future tests will involve complex, real-world environments. Jamieson and Skraaning (2019) discuss the inapplicability of the lumberjack analogy in real-world environments, and although there were some holes in their methodology, the results of this study indicate a need for more clarification on the generalizability of these effects. This includes having multi-stage tasks, multiple tasks, longer sessions, trained operators, etc., so that the results can be applied to any work context. Most importantly though, there should be a distinction between tasks with severe consequences and time constraints, and those without. Furthermore, this suggests the need to study the temporal aspects of the lumberjack analogy surroundings these automation failures. Automation can be adjusted in many ways and future results need to not only focus on extreme changes in automation but the finer levels between those extremes. This would allow us to better understand how a variety of work environments benefit from various types of automation that would allow us to better fit automation to the situation. To Wickens et al (2019) point, it will be nearly impossible to make these kinds of decisions though without high-quality measures of the dependent variable and true manipulations of the underlying constructs.

Conclusion

Overall, the goal of this study was to better understand the interaction between humans and automation by further examining the relationship between automation, mental workload, situational awareness, and performance, addressing the aptness of the lumberjack analogy. The current findings indicate an overwhelmingly positive response of people to automation that does not match the “doom and gloom” that can be associated with the lumberjack analogy. These results indicate that, at least in this particular task environment, the benefits of automation on user performance greatly outweighs the risks of the automation.

References

Crandall, J. W., & Cummings, M. L. (2007, March). *Developing performance metrics for the supervisory control of multiple robots*. In Proceedings of the ACM/IEEE international conference on Human-robot interaction, 33-40.

Comstock Jr, J. R., & Arnegard, R. J. (1992). *The multi-attribute task battery for human operator workload and strategic behavior research* (No. NAS 1.15: 104174).

Cummings, M., Huang, L., Zhu, H., Finkelstein, D., & Wei, R. (2019). The impact of increasing autonomy on training requirements in a UAV supervisory control task. *Journal of Cognitive Engineering and Decision Making*, *13*(4), 295-309.

Donmez, B., Nehme, C., & Cummings, M. L. (2010). Modeling workload impact in multiple unmanned vehicle supervisory control. *IEEE Transactions on Systems Man And Cybernetics Part A-Systems And Humans*, *40*(6), 1180–1190. https://doi-org.ezproxy.libraries.wright.edu/10.1109/TSMCA.2010.2046731

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society Annual Meeting*, *32*(2), 97-101. Sage CA: Los Angeles, CA: SAGE Publications.Endsley, M. R. (1993). Situation awareness and workload: Flip sides of the same coin. In R. S. Jensen & D. Neumeister (Eds.), *Proceedings of the Seventh International Symposium on Aviation Psychology*, 906-911. Columbus, OH: Department of Aviation, The Ohio State University.

Endsley, M. R. (1995). Measurement of situation awareness in dynamic systems. *Human Factors*, *37*(1), 65-84.

Endsley, M. R. (1996). Automation and situation awareness. *In R. Parasuraman & M. Mouloua (Eds.)*, *Automation and human performance: Theory and applications.* 163-181. Mahwah, NJ: Lawrence Erlbaum.Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors*, *37*(2), 381-394.

Endsley, M. R., Selcon, S. J., Hardiman, T. D., & Croft, D. G. (1998). A comparative analysis of SAGAT and SART for evaluations of situation awareness. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *42*(1), 82.

Fuchs, C., Ferreira, S., Sousa, J., & Gonçalves, G. (2013). Adaptive consoles for supervisory control of multiple unmanned aerial vehicles. In *International Conference on Human-Computer Interaction*, 678-687.

Galy, E., Paxion, J., & Berthelon, C. (2018). Measuring mental workload with the NASA-TLX needs to examine each dimension rather than relying on the global score: an example with driving. *Ergonomics*, *61*(4), 517-527.

Grieve, A.P. (1984). Tests of sphericity of normal distributions and the analysis of repeated measures designs. *Psychometrika*, *49*(2), 257-267.

Hancock, P. A. (1989). The effect of performance failure and task demand on the perception of mental workload. *Applied Ergonomics*, *20*(3), 197-205.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139-183). North-Holland.

Holm, A., Lukander, K., Korpela, J., Sallinen, M., & Müller, K. M. (2009). Estimating brain load from the EEG. *The Scientific World Journal*, *9*, 639-651.

Jamieson, G. A., & Skraaning, G. (2019). The absence of degree of automation trade-offs in complex work settings. *Human Factors*, 0018720819842709.

Jamieson, G. A., & Skraaning, G. (2020). The harder they fall? A response to Wickens et al.(2019) regarding the generalizability of lumberjack predictions to complex work settings. *Human Factors*, 0018720820904623.

Johnson, R., Leen, M., & Goldberg, D. (2007). *Testing adaptive levels of automation (ALOA) for UAS supervisory control (AFRL-HE-WP-TR2007– 0068)*. Wright-Patterson Air Force Base, OH: Air Force Research Laboratory.

Kaber, D. B., & Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science*, *5*(2), 113-153.

Lin, J., Matthews, G., Wohleber, R. W., Funke, G. J., Calhoun, G. L., Ruff, H. A., Szalma, J., & Chiu, P. (2019). Overload and automation-dependence in a multi-UAS simulation: task demand and individual difference factors. *Journal of Experimental Psychology: Applied*. Advance online publication. http://dx.doi.org/10.1037/xap0000248

Lorenz, B., Di Nocera, F. N., Röttger, S., & Parasuraman, R. (2002a). Automated fault-management in a simulated spaceflight micro-world. *Aviation, Space, and Environmental Medicine*, *73*(9), 886-897.

Lorenz, B., Di Nocera, F., & Parasuraman, R. (2002). Varying types and levels of automation in the support of dynamic fault management: An analysis of performance costs and benefits. *Human factors in Transportation, Communication, Health, and the Workplace*, 517-524.

Mancuso, V. F., Funke, G. J., Strang, A. J., & Eckold, M. B. (2015). Capturing performance in cyber human supervisory control. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 59,* 317-321.

Moroney, W. F., Biers, D. W., Eggemeier, F. T., & Mitchell, J. A. (1992). A comparison of two scoring procedures with the NASA task load index in a simulated flight task. In *Proceedings of the IEEE 1992 National Aerospace and Electronics Conference@ m_NAECON 1992* (pp. 734-740). IEEE.

Nygren, T. E. (1991). Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload. *Human Factors, 33*(1), 17-33.

Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human performance consequences of stages and levels of automation: An integrated meta-analysis. *Human Factors*, *56*(3), 476-488.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *30*(3), 286-297.

Roettger, S., Bali, K., & Manzey, D. (2009). Impact of automated decision aids on performance, operator behaviour and workload in a simulated supervisory control task. *Ergonomics*, *52*(5), 512–523. https://doi-org.ezproxy.libraries.wright.edu/10.1080/00140130802379129m

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. *Human Factors*, *49*(1), 76-87.

Sebok, A., & Wickens, C. D. (2017). Implementing lumberjacks and black swans into model-based tools to support human–automation interaction. *Human Factors*, *59*(2), 189–203.

Shaw, T., Emfield, A., Garcia, A., de Visser, E., Miller, C., Parasuraman, R., & Fern, L. (2010). Evaluating the benefits and potential costs of automation delegation for supervisory control of multiple UAVs. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 54(19), 1498-1502.

Sheridan, T. B., & Verplank, W. L. (1978). *Human and computer control of undersea teleoperators*. Massachusetts Inst of Tech Cambridge Man-Machine Systems Lab.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human learning and Memory*, *4*(6), 592.

Tatasciore, M., Bowden, V. K., Visser, T. A. W., Michailovs, S. I. C., Loft, S. (2020). The benefits and costs of low and high degree of automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *62*, 874–896

Taylor, R. M. (1990). Situation awareness rating technique (SART): the development of a tool for aircrew systems design. In Situational Awareness in Aerospace Operations (Chapter 3). France: Neuilly sur-Seine, NATO-AGARD-CP-478.

Taylor, G., Reinerman-Jones, L., Cosenzo, K., & Nicholson, D. (2010). Comparison of multiple physiological sensors to classify operator state in adaptive automation systems. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 54(3), 195.

Vidulich, M. A., & Tsang, P. S. (2015). The confluence of situation awareness and mental workload for adaptable human–machine systems. *Journal of Cognitive Engineering and Decision Making, 9*(1), 95-97.

Virtanen, K., Mansikka, H., Kontio, H., & Harris, D. (2021). Weight watchers: NASA-TLX weights revisited. *Theoretical Issues in Ergonomics Science*, 1-24.

Wickens, C. D., Li, H., Santamaria, A., Sebok, A., & Sarter, N. B. (2010). Stages and levels of automation: An integrated meta-analysis. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 54*(4), 389-393.

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2013). *Engineering Psychology and Human Performance* (4th ed.). Upper Saddle River, NJ: Pearson Education Inc.

Wickens, C. D., Mavor, A. S., Parasuraman, R., & McGee, J. P. (1998). The Future of Air Traffic Control: NRC Report.

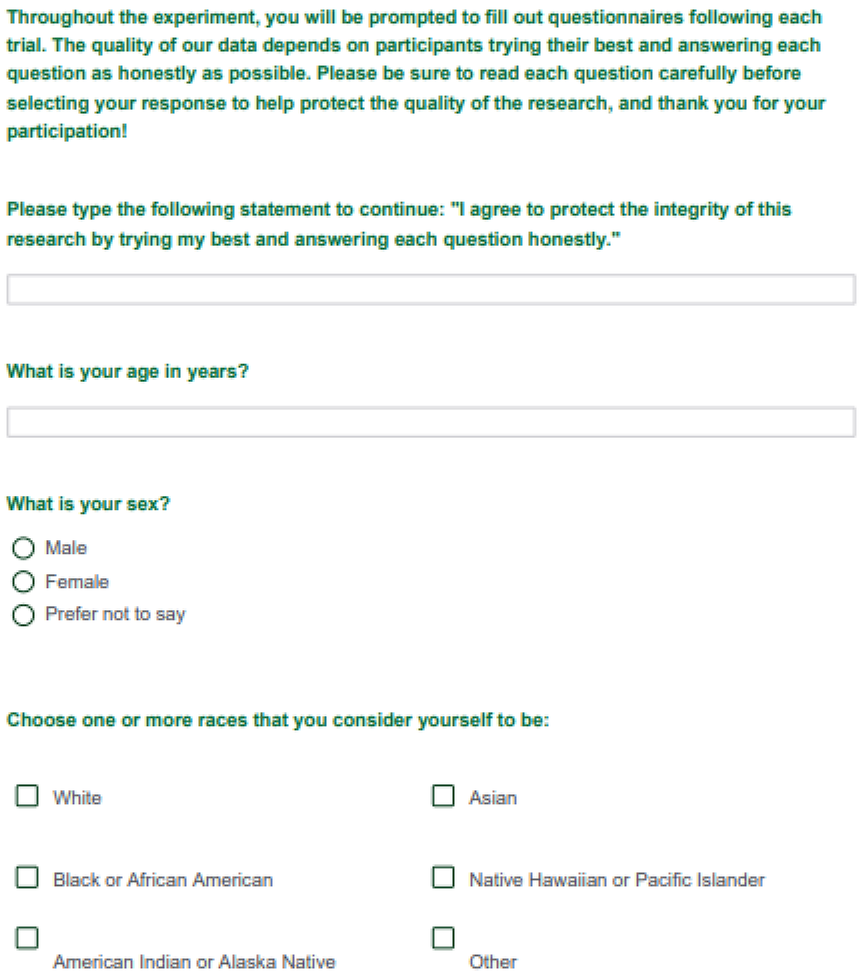
Wickens, C. D., Onnasch, L., Sebok, A., & Manzey, D. (2019). Absence of DOA effect but no proper test of the lumberjack effect: A reply to Jamieson and Skraaning (2019). *Human Factors*, 0018720820901957.

Yeh, Y. Y., & Wickens, C. D. (1988). Dissociation of performance and subjective measures of workload. *Human Factors*, *30*(1), 111-120.

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: mental workload in ergonomics. *Ergonomics*, *58*(1), 1–17. https://doi-org.ezproxy.libraries.wright.edu/10.1080/00140139.2014.956151

Appendix

Appendix A: Participant Demographic Forms and Questionnaires



Throughout the experiment, you will be prompted to fill out questionnaires following each trial. The quality of our data depends on participants trying their best and answering each question as honestly as possible. Please be sure to read each question carefully before selecting your response to help protect the quality of the research, and thank you for your participation!

Please type the following statement to continue: "I agree to protect the integrity of this research by trying my best and answering each question honestly."

What is your age in years?

What is your sex?

○ Male
○ Female
○ Prefer not to say

Choose one or more races that you consider yourself to be:

□ White

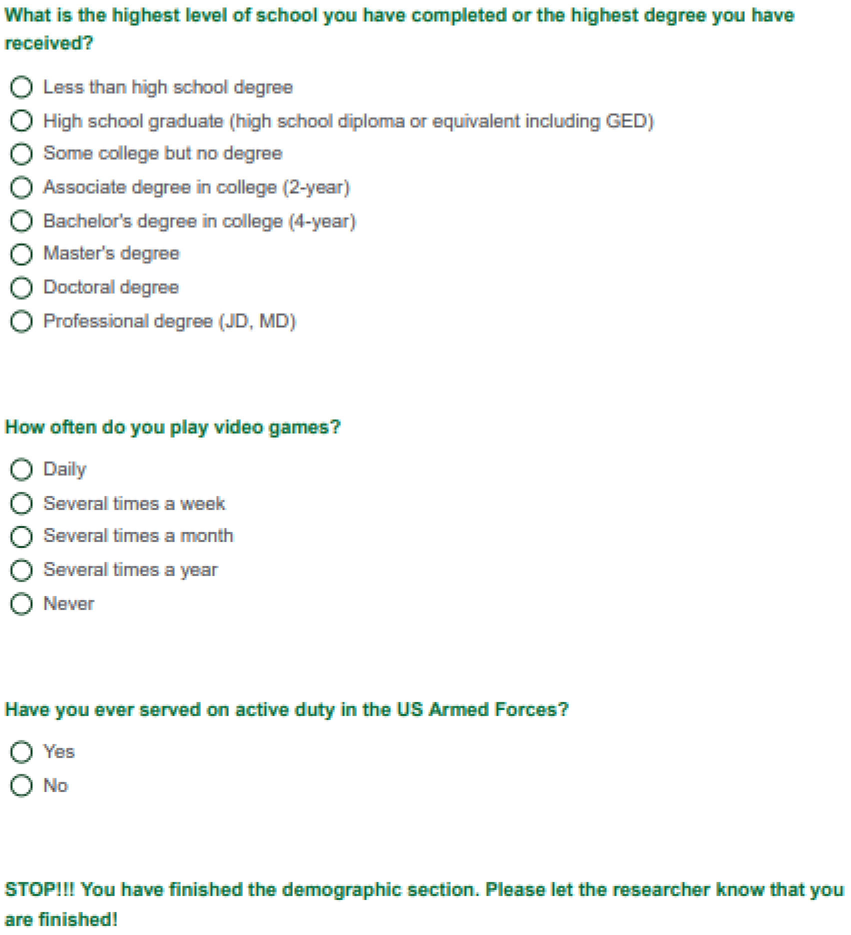
□ Black or African American

□ American Indian or Alaska Native

□ Asian

□ Native Hawaiian or Pacific Islander

□ Other



What is the highest level of school you have completed or the highest degree you have received?

○ Less than high school degree
○ High school graduate (high school diploma or equivalent including GED)
○ Some college but no degree
○ Associate degree in college (2-year)
○ Bachelor's degree in college (4-year)
○ Master's degree
○ Doctoral degree
○ Professional degree (JD, MD)

How often do you play video games?

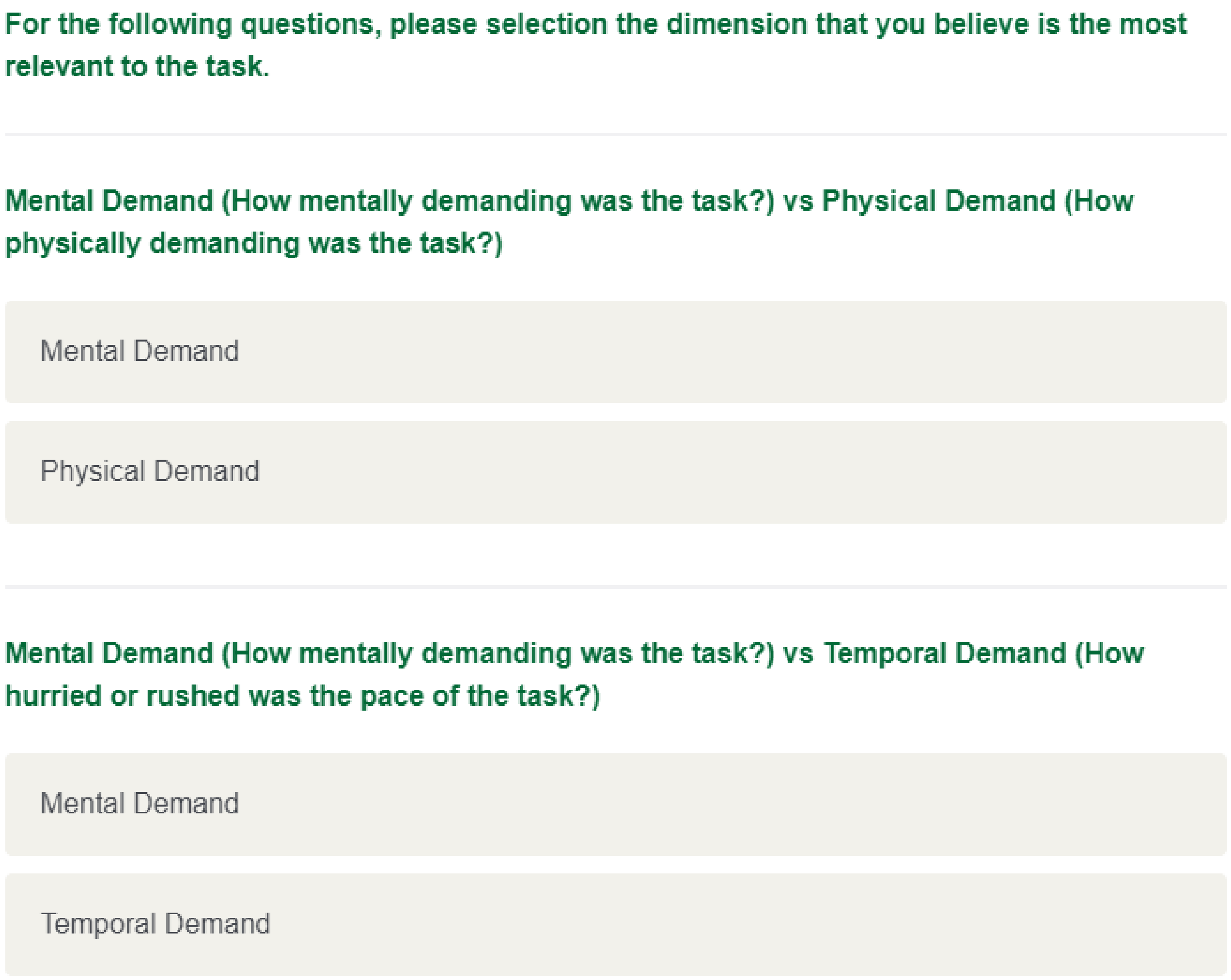
○ Daily
○ Several times a week
○ Several times a month
○ Several times a year
○ Never

Have you ever served on active duty in the US Armed Forces?

○ Yes
○ No

STOP!!! You have finished the demographic section. Please let the researcher know that you are finished!

Appendix B: ***NASA-TLX Pairwise and Post-Task Survey***



For the following questions, please selection the dimension that you believe is the most relevant to the task.

Mental Demand (How mentally demanding was the task?) vs Physical Demand (How physically demanding was the task?)

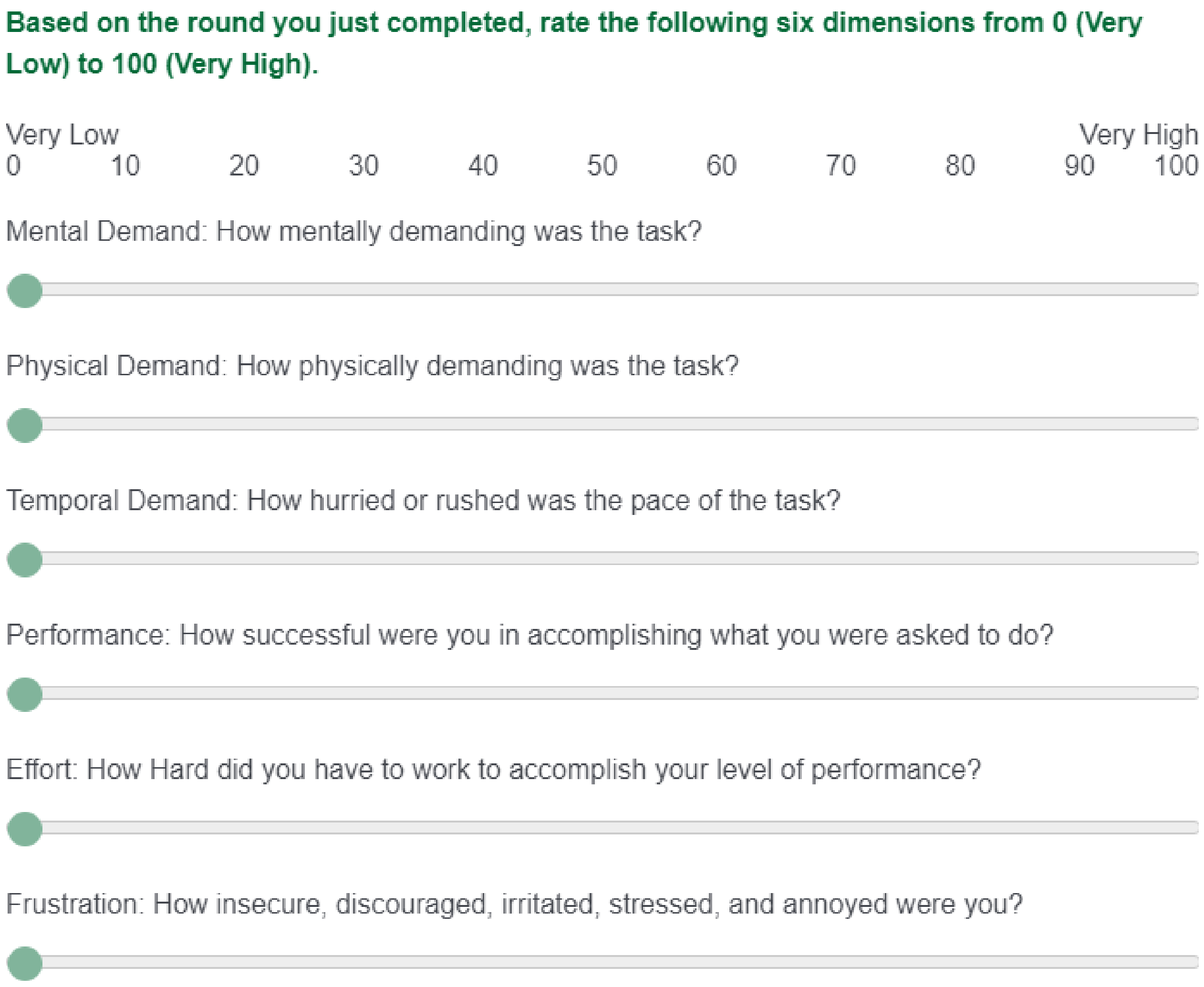
Mental Demand

Physical Demand

Mental Demand (How mentally demanding was the task?) vs Temporal Demand (How hurried or rushed was the pace of the task?)

Mental Demand

Temporal Demand



Based on the round you just completed, rate the following six dimensions from 0 (Very Low) to 100 (Very High).

Very Low   Very High
0 10 20 30 40 50 60 70 80 90 100

Mental Demand: How mentally demanding was the task?

Physical Demand: How physically demanding was the task?

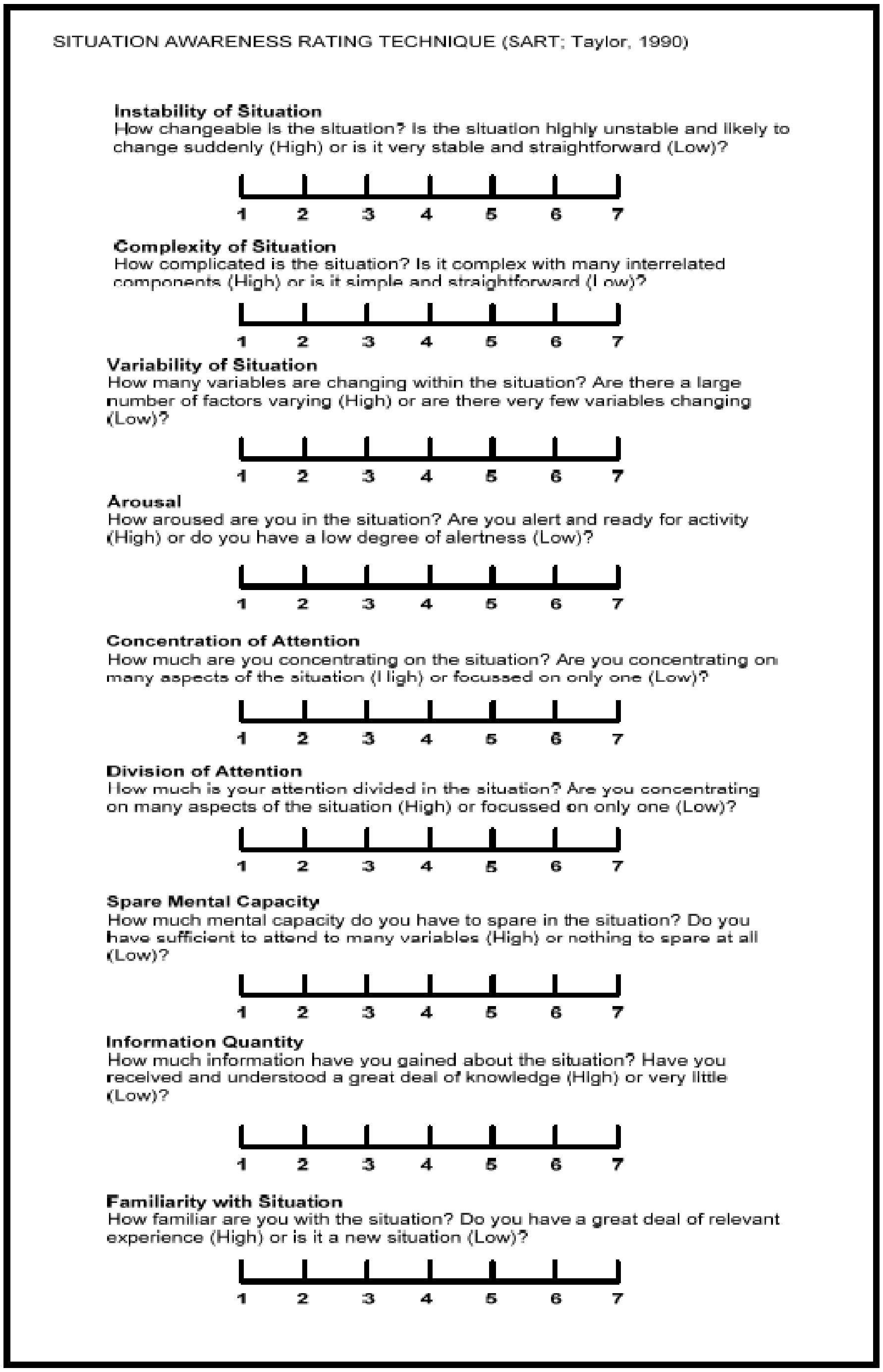
Temporal Demand: How hurried or rushed was the pace of the task?

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

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

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

Appendix C**: *SART and Qualtrics Example***



SITUATION AWARENESS RATING TECHNIQUE (SART; Taylor, 1990)

Instability of Situation
How changeable is the situation? Is the situation highly unstable and likely to change suddenly (High) or is it very stable and straightforward (Low)?

1 2 3 4 5 6 7

Complexity of Situation
How complicated is the situation? Is it complex with many interrelated components (High) or is it simple and straightforward (Low)?

1 2 3 4 5 6 7

Variability of Situation
How many variables are changing within the situation? Are there a large number of factors varying (High) or are there very few variables changing (Low)?

1 2 3 4 5 6 7

Arousal
How aroused are you in the situation? Are you alert and ready for activity (High) or do you have a low degree of alertness (Low)?

1 2 3 4 5 6 7

Concentration of Attention
How much are you concentrating on the situation? Are you concentrating on many aspects of the situation (High) or focussed on only one (Low)?

1 2 3 4 5 6 7

Division of Attention
How much is your attention divided in the situation? Are you concentrating on many aspects of the situation (High) or focussed on only one (Low)?

1 2 3 4 5 6 7

Spare Mental Capacity
How much mental capacity do you have to spare in the situation? Do you have sufficient to attend to many variables (High) or nothing to spare at all (Low)?

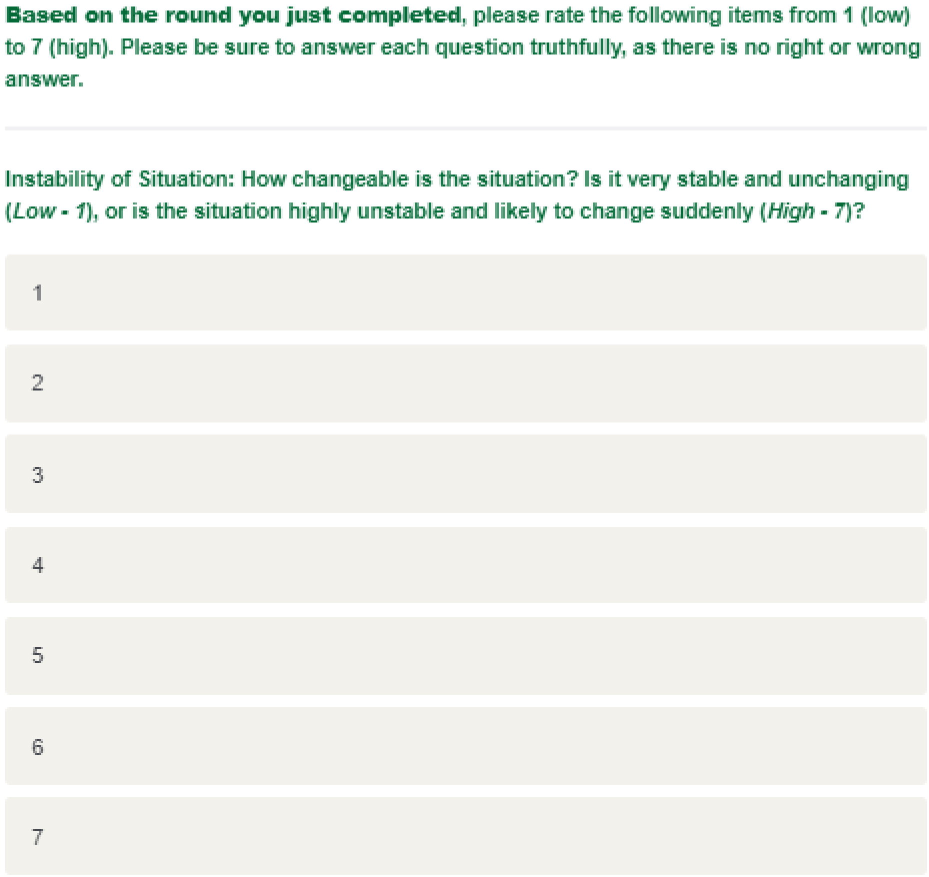
1 2 3 4 5 6 7

Information Quantity
How much information have you gained about the situation? Have you received and understood a great deal of knowledge (High) or very little (Low)?

1 2 3 4 5 6 7

Familiarity with Situation
How familiar are you with the situation? Do you have a great deal of relevant experience (High) or is it a new situation (Low)?

1 2 3 4 5 6 7



Based on the round you just completed, please rate the following items from 1 (low) to 7 (high). Please be sure to answer each question truthfully, as there is no right or wrong answer.

Instability of Situation: How changeable is the situation? Is it very stable and unchanging (*Low - 1*), or is the situation highly unstable and likely to change suddenly (*High - 7*)?

1

2

3

4

5

6

7