Summer 2019

A Comparative Study of the Influence of Level of Automation and Reliability of IDS Systems on Cyber Situation Awareness

Ian Anderson Cooke
San Jose State University

A COMPARATIVE STUDY OF THE INFLUENCE OF LEVEL OF AUTOMATION AND RELIABILITY OF IDS SYSTEMS ON CYBER SITUATION AWARENESS

A Thesis

Presented to

The Faculty of the Department of Psychology

San Jose State University

In Partial Fulfillment

of the Requirements for the Degree of

Master of Arts

By

Ian Cooke

August 2019

The Designated Thesis Committee Approves the Thesis Titled

A COMPARATIVE STUDY OF THE INFLUENCE OF LEVEL OF AUTOMATION AND RELIABILITY OF IDS SYSTEMS ON CYBER SITUATION AWARENESS

by

Ian Cooke

APPROVED FOR THE DEPARTMENT OF PSYCHOLOGY

SAN JOSÉ STATE UNIVERSITY

August 2019

Dr. David Schuster	Department of Psychology
Dr. Sean Laraway	Department of Psychology
Dr. Evan Palmer	Department of Psychology

ABSTRACT

Computer network defense (CND) protects organizations and individuals against cyber threats by monitoring, identifying, analyzing, and defending network infrastructure from infiltration. Network defenders must maintain high levels of cyber situation awareness (CSA) in order to correctly identify and act on threats to the network. Intrusion detection systems (IDSs) are automated systems designed to assist network defenders in building CSA by sifting through network traffic and flagging potential threats. These systems are plagued by high false alarm rates that inhibit the ability of network defenders to build CSA. More capable IDSs have been developed that are capable of increasing the hit rate and lowering the false alarm rate by analyzing gathered network information. The influence of these IDS technologies on CSA has yet to be explored. 172 San Jose State University psychology students performed a signal detection task for intrusion detection to examine whether integrated automation with a multilayered analysis incorporating both liberal and conservative response criteria leads to better CSA than less-integrated, yet liberally responding automation (high hit rates and high false alarm rates) or conservatively responding automation (with low hit rates and low false alarm rates). The IDS condition was manipulated at three levels (liberal, conservative, both). The reliability of the IDSs was manipulated at three levels (60%, 80%, 95%). This study was unable to observe any differences in task performance or CSA for any of the conditions.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Introduction

Statement of the Problem

Computer network defense (CND) protects organizations and individuals against cyber threats by monitoring, identifying, analyzing, and defending network infrastructure from infiltration. Network defenders are required to distinguish actual threats to the network from normal network traffic, determine the nature of the threat, and decide how to respond (Sawyer et. al. 2014). Success at CND requires an extensive amount of goal-relevant awareness derived from information in the task environment, which has been labeled situation awareness (SA; Endsley, 1988). SA has been further adapted into the domain of cyber security and labeled *cyber situation awareness* (CSA; Champion, Rajivan, Cooke, & Jariwala, 2012; D'amico, Whitley, Tesone, O'Brien & Roth, 2005; Gutzwiller, Hunt, & Lange, 2016; Mahoney et. al, 2010; Onwubiko, 2009; 2016). Building and maintaining CSA is essential for recognizing and responding to network threats (Jajodia, Liu, Swarup & Wang, 2010; Onwubiko & Owens, 2012).

Intrusion detection systems (IDSs), a form of automation, have been designed to assist the network defender in monitoring network traffic. IDSs can help defenders build and maintain the awareness necessary to defend networks. For example, an IDS will run a scan on the number of users that have logged in and/or out at unusual times and present those data to the network defender. The data output might take different forms depending on the type of system. These forms will be described in further detail. Human-machine relationships are especially important to consider in the domain of CND because it is impossible for a human network defender to perform the task completely unaided, which

makes it impossible to completely divorce the human and machine components in CND from one another.

Misuse-based IDS systems can only detect the threats they are programed to discover. The advantage of a Misuse-based system is that it is very effective at detecting known threats (Werlinger, Hawkey, Muldner, Jaferian & Beznosov, 2008). One of the disadvantages of a Misuse-based system is that it is incapable of detecting novel threats (Werlinger, Hawkey, Muldner, Jaferian & Beznosov, 2008).

Anomaly-based systems monitor a network and are calibrated to the normal flow of network traffic on that specific network. If an element of user traffic deviates from the norm, it is flagged as a potential threat to the system. Anomaly-based systems are advantageous because it can potentially detect novel threats (Kemmerer & Vigna, 2002). However, not all abnormal traffic is malicious. Anomaly-based IDSs are incapable of making that distinction (Werlinger, Hawkey, Muldner, Jaferian & Beznosov, 2008), which leads to the primary disadvantage of an Anomaly-based system; it produces a vast number of false alarms that must be reconciled later by a human cyber defender (Mukkamala, Sung & Abraham, 2005).

Efforts have been made to develop new, more capable IDSs that increase the hit rate and lower the false alarm rate by analyzing gathered network information. A popular approach is to integrate both misuse-based and anomaly-based systems together to combine the strengths of each type of system while compensating for their weaknesses (Kim, Lee & Kim, 2015). The approach is called a *hybrid IDS*. A hybrid IDS increases the level in which the IDS is involved in the task of CND. A hybrid IDS may increase or

reduce the CSA of cyber defenders depending on how these systems are designed and implemented. Research on how the hybrid approach influences the CSA of human network defenders would inform the design of these new systems to optimize performance outcomes. As hybrid systems are relatively new and have limited deployment in the field, research in this area has yet to be conducted.

As IDSs are a form of automation, previous research examining how level of automation (LOA) influences the SA and decision making of human operators may be applicable in the cyber domain. SA can benefit from the use of diagnostic aiding tools (Goodrich et. al. 2007; Horrey & Wickens, 2001; Rudisill, 2000). However, under conditions in which the human and automation operate more independently, increasing the LOA can lead to detriments in performance outcomes (Kaber & Endsley, 2000; Ruff, Narayanan, and Draper, 2002).

It is imperative that we improve the CSA of cyber defenders and help them overcome the challenges inhibiting the discrimination of actual threats from false alarms when using an IDS. We must understand whether integrated automation with a multilayered analysis incorporating both liberal and conservative response criteria leads to better CSA than less-integrated, yet liberally responding automation (high hit rates and high false alarm rates), or conservatively responding automation (with low hit rates and low false alarm rates).

Situation Awareness and Cyber Situation Awareness

Situation awareness (SA) arose as a scientific construct when researchers began studying the source of pilot errors, but it is a broadly applicable skill (Harwood, Barnett,

& Wickens, 1988). SA is an essential skill in the effective operation of a complex, dynamic system (Endsley, 1995). SA can be described as task-relevant knowledge (Endsley, 1988). Endsley (1988) highlighted SA as occurring in three stages. Stage one is perception, in which information in the task environment is perceived by the operator via the senses and diagnostic tools. Stage two involves the integration and organization of the information into a meaningful structure. Stage two is called information comprehension. Stage three is projection, in which the operator uses the information structure to project a future outcome and plan future behavior (Endsley 1988). Failures can occur within any of these stages, resulting in poor SA and subsequent task performance. Sources of these failures can be attributed to high workload, poor sensing, and unreliability in automated tools used to aid task performance (Parasuraman, Sheridan & Wickens, 2008).

SA is a domain-specific construct best studied in the context in which it operates (Flach, 2015). As CND is a highly complex dynamic system, the construct of SA has been adapted to this domain. The components of SA as they pertain to CND have been identified and labeled as CSA. The three-stage model still applies, but the task-relevant knowledge is specific to CND (Onwubiko, 2009; 2016). CSA has been refined more recently into cyber cognitive situation awareness (CCSA; Gutzwiller, Hunt, & Lange, 2016). According to this model, defenders must maintain awareness of a multitude of network attributes (network health, status, network architecture, typical traffic), their role in the team structure (their task, their teammates task, and how they relate to the superordinate goal), and knowledge of the global security landscape (previous hacks, active attackers, attack profiles, political relations) to be proficient in CND (Champion,

Rajivan, Cooke, & Jariwala, 2012; D'amico, Whitley, Tesone, O'Brien & Roth, 2005; Mahoney et. al, 2010).

Reliability of Automation

Reliability is commonly defined as the percentage to which an automated system can perform a desired task successfully (Singh, Tiwari & Singh, 2009). In general, automation reliability and human performance tend to have a positive relationship, wherein highly reliable automation leads to better task performance. This effect has been found frequently in the literature in regard to signal detection tasks similar to that in CND (Madhavan & Phillips, 2010; Dixon, Wickens, & McCarley 2006). Conversely, as reliability declines, so does human performance. Wickens and Dixon (2007) identified that performance declines because the human must expend extra cognitive effort to sort through erroneous information to successfully complete the task.

Reliance on automation and task demand have been found to be moderating variables between reliability and task performance (Wickens & Dixon 2007). Wickens & Dixon point out that operators in positions of high task demand rely heavily on automation despite their awareness of its low reliability level. In CND, due to the volume and speed of data, human network defenders are in a position in which they cannot perform the task unaided. As task demands and reliance on the IDS are high in CND, network defenders may be likely to rely on automation regardless of whether it is reliable or not.

Levels of Automation and SA in CND

Automation is the execution of a task by a machine agent (Parasuraman, 1997). Levels of automation (LOA) describe how and to what extent the automation is involved

in a task (Parasuraman, Sheridan & Wickens, 2000; Sheridan & Verplank, 1978). There are multiple taxonomies of LOA (Kaber & Endsley, 1999; Sheridan & Verplank, 1978). This study focuses on the levels of diagnostic aiding (Parasuraman, Sheridan & Wickens, 2000); Wickens & Dixon, 2007). Of these levels, the first is information acquisition. At the first level, the automation gathers information about the task environment and presents it to the human agent. The second level is information analysis. The second level refers to when the automation performs an analysis of the data and presents the results to the human agent. The two levels of diagnostic aiding also correspond to the first two levels of Endsley's model of SA, in which information acquisition and information analysis are similar to perception and comprehension (Schuster, 2013; Horrey et. al. 2009). In the perception level of SA, information about the task environment is gathered via the senses. The task information is then integrated with existing knowledge structures to form a complex analysis of the problem space at the comprehension level. Level 1 automation (information acquisition) gathers information from the task environment and organizes it for presentation. At the information analysis level, the automation will cross-check this new information with old information and present the information in an integrated form. For example, a hybrid IDS will examine the network architecture for anomalous traffic (information acquisition), compare it with what it is programed to consider normative network behavior, and flag any deviations as potential threats (information analysis). A hybrid IDS system that implements anomaly-based and misuse-based detection in succession would essentially be performing both information acquisition and information analysis.

Effects of Level of Automation on SA and Performance

SA is typically shown to increase as LOA increases from level 1 (information acquisition) to level 2 (information analysis) by reducing the cognitive workload of the human at the expense of the richness of the information provided (Horrey & Wickens, 2001). At higher stages, the human is further abstracted from relevant data about the problem, which leads to situations in which increasing the LOA decreases SA as opposed to raising it further (Ruff, Narayanan, and Draper, 2002). However, layering the automation so that level 1 information is provided along with level 2 analysis leads to better performance and SA (Dexter, Willemsen-Dunlap, & Lee, 2007).

However, operators only have better performance in conditions in which the automation does not make an error. When the automation does make an error, the SA of the human operator may be reduced (Dexter et. al. 2007). Human operators rely on the automation and trust that the information they acted on was correct, which leads to an increase in errors if the information provided by the automation was incorrect. This effect suggests that not only is SA sensitive to the LOA, but it is also sensitive to the reliability of the information provided by the automation. This relationship is especially important to consider in CND where network defenders are highly dependent on the IDS automation to obtain information about network threats.

Reliability and Level of Automation

Automated systems that incorporate diagnostic aiding have been shown to improve performance of human operators on tasks (Goodrich et. al, 2007). This effect is strengthened when the automation is reliable (Madhaven & Philips, 2010). Furthermore,

research suggests that unreliable automation can have different effects on operator SA depending on the level of diagnostic aiding being implemented. When compared to automation employing solely information acquisition, automation employing information analysis has been shown to have more negative impact on SA when information is unreliable (Rovira, McGarry, & Parasuraman, 2007; Sarter & Schroeder, 2001). Similarly, anomaly-based IDSs that only flag potential threats and present them to the network defender (information acquisition) have the potential to reduce the CSA of a network defender when false alarm rates are very high. False alarms have been shown to be more damaging to SA than misses (Dixon, Wickens & McCarley, 2006). This effect is particularly impactful for anomaly-based IDSs because they are calibrated with a more liberal response criterion to capture more anomalous network traffic, but produce more false alarms.

Unreliability in IDSs can better be explained in terms of signal detection theory (SDT; Swets & Pickett, 1982). IDSs are essentially alarm systems. To most effectively build SA, a decision criterion must be chosen for the IDS that optimizes how liberal or conservative it is when determining the alerting threshold (Kuchar, 1996). Ensuring performance means that the probability that an alarm reflects an actual attack must be increased as much as possible (Parasuraman, Hancock & Olofinboba, 1997).

Signal Detection in Cyber Security

Signal detection theory (SDT) has applications in cyber security. Intrusion detection systems perform signal detection when scanning network traffic and flagging alerts (Mukkamala, Sung & Abraham, 2005). These systems attempt to differentiate potentially

malicious network traffic (signal) from the mass flow of normal network traffic (noise). In the context of CND, a hit would be flagging an actual threat, a false alarm would be flagging a non-threat as a threat, a miss would be failing to flag an actual threat, and a correct rejection would be not flagging a non-threat.

Anomalous network traffic is flagged and presented to a cyber network defender. However, not all abnormal network traffic is malicious. Once the flagged traffic has been presented to the cyber defender, he/she must make the distinction between the true alerts and false alarms. This task requires enough information from the task environment and other sources to help the cyber defender produce a keen awareness of the situation in order to make the correct decisions of how to respond (Sawyer et. al. 2014).

Intrusion Detection Systems

One aspect of CND involves continuous monitoring of network traffic to discriminate anomalous patterns from nominal traffic throughout complex computer networks. This is an impossible task for a human to perform unaided. IDSs have been developed to assist in making these discriminations. Network defenders use information provided by IDSs to make decisions about how to respond to potential attacks.

Misuse-based IDSs are the oldest of these systems (Kemmerer & Vigna, 2002). These systems are programed to search for specific threats to the network. These types of systems use pre-determined search criteria to evaluate whether or not patterns of traffic in the network characterizes an attack. Thus, this type of IDS performs information analysis (Wickens & Dixon, 2007). This tool is characterized by low hit rates because it is only capable of detecting attacks it is designed to search for. However, this programming also

means misuse-based systems have low false alarm rates because they do not often flag non-threats. Misuse-based systems are efficient at detecting known threats they are programed to detect. However, these types of systems are incapable of detecting previously-unseen threats. As attacks are always evolving, this makes the effective miss rate very high (Werlinger, Hawkey, Muldner, Jaferian & Beznosov, 2008). When novel attack types occur, network defenders must identify and respond to the threat manually. Afterward, network defenders must update the IDS to search for that attack type in the future. As hackers are constantly finding novel ways to circumvent these systems, network defenders are always a step behind the attackers.

Anomaly-based IDSs were developed in response to the limitations of misuse-based IDSs. These systems sort through network data, flag potentially malicious network events, and report these events to the human network defender for interpretation (McHugh, Christie, & Allen, 2000). This automation performs level 1 diagnostic aiding (information acquisition; Wickens & Dixon, 2007). These serve to make the information required to complete the task more salient and available to the network defender. As anomaly-based IDSs are effective at differentiating between normal and abnormal network traffic, these systems can detect a greater variety of threats to the network at the cost of high false alarm rates. This trade-off illustrates that in the eyes of a network defender, even though the miss rate for this type of IDS is very low, the high false alarm rate obscures the actual threats (Mukkamala, Sung & Abraham, 2005).

Hybrid systems are an approach to balance out the strengths and weaknesses of misuse-based and anomaly-based systems (Aydin, Zaim & Ceylan, 2008). Some of these

systems attempt to layer misuse-based and anomaly-based systems in succession (Bronte, 2016; Tesfahun & Bhaskari, 2015). Other attempts involve machine learning or statistical analysis to narrow in on what attack vector anomalous code represents (Peddabachigari, Abraham, Grosan & Thomas, 2007). These systems have lower false alarm rates and higher hit rates than using either of the previous systems individually (Aydin, Zaim & Ceylan, 2008). As these systems integrate both anomaly-based and misuse-based systems, hybrid systems can be considered to perform both information acquisition and information analysis (Wickens & Dixon, 2007).

Research Need

Although research has been conducted on SA in automated systems, the extent to which the level of diagnostic aiding and reliability of the automation interact to influence the SA of the operator in CND has yet to be examined. The present research varied levels of reliability as well as levels of diagnostic aiding to explore how the effectiveness of diagnostic aiding changes with respect to different levels of reliability.

Although research has been conducted on how reliability and level of automation influence the SA of human operators (Goodrich et. al, 2007; Madhaven & Philips, 2010; Rovira, McGarry, & Parasuraman, 2007; Sarter & Schroeder, 2001), currently no studies have examined this relationship in the domain of cyber security. This study tested these constructs in this new domain and examined how they can be modified for application in cyber systems while adding to knowledge of human-automation interaction generally.

Although IDS systems are commonly used, an insufficient amount of empirical testing has been conducted on how people use and interact with these systems (Werlinger

et. al. 2008). This research provides empirical evidence for how people interact with these systems in controlled conditions by varying the reliability of the IDS and the extent to which the IDS is involved in the decision making.

Literature on IDS development has proposed hybrid IDSs as a viable method for detecting threats (Aydin, Zaim & Ceylan, 2008). However, this proposition has only been validated by assessing the hybrid IDS's increased ability to detect certain attack types over previous systems (Tesfahun & Bhaskari, 2015). Furthermore, the literature has yet to examine how human network defenders perform using these types of systems.

The most recent methodology for developing a Hybrid IDS layered the misuse-based IDS (information analysis) prior to the anomaly-based IDS (information acquisition) in the processing stream (Kim, Lee & Kim, 2015). Based on previous literature on levels of automation and a novel definition of hybrid IDSs (information acquisition followed by information analysis), this study tested whether this method would improve the CSA of network defenders beyond what other systems can achieve.

Purpose of this Research

The purpose of this study was to explore the moderating effects of level of diagnostic aiding and automation reliability on the human ability to build CSA in CND. The strengths and limitations of IDSs led to the conditions in this applied study. Specifically, the study examined whether *hybrid* (both information acquisition and information analysis) automation with a multilayered analysis incorporating both liberal and conservative response criterion leads to better CSA than both *anomaly-based* (only information acquisition) systems with liberal response criterion (high hit rate, and high

false alarm rates) and *misuse-based* (information analysis) systems with conservative response criterion (low hit rates and low false alarm rate)].

Hypotheses

Based on previous literature with respect to diagnostic aiding and its influence on SA, there was reason to expect that SA would increase with more machine agent assistance in instances of high reliability (Goodrich et. al, 2007; Madhaven & Philips, 2010). In these instances, human agents correctly make their decisions regarding task completion based on the information provided by the automation. Since the reliability of the IDS is sufficiently high in these situations, the human network defender should be able to rely more heavily on the accuracy of the information provided by the automation. At the same time, it was reasonable to expect that human network defenders would refer to the automation more often in situations of high reliability and would have higher CSA when performing the task.

Furthermore, because the analysis information provided by the hybrid IDS can be verified by examining log files, it was reasonable to expect that higher levels of diagnostic aiding would lead to higher CSA. This relationship would translate into more successful discrimination between true security alerts and false alarms, which led to the hypothesis that hybrid IDS systems would lead to higher CSA.

Because the stages of diagnostic aiding (Wickens & Dixon, 2007) map onto Endsley's first two levels of SA (Schuster, 2013), diagnostic aiding that performs both information analysis and information acquisition would allow the human network defender to develop significantly more CSA than information acquisition without

analysis (Schuster, 2013). As hybrid IDSs are characterized by performing both information acquisition and analysis, the following hypotheses were formed:

H1a. Hybrid IDS (information acquisition then analysis) will lead to higher CSA than anomaly-based (information acquisition),

H1b. Hybrid IDS (information acquisition then analysis) will lead to higher CSA than misuse-based (information analysis).

When automation is imperfect, information relevant to the task environment that is presented to the human agent contains errors, which makes it more difficult for the human to make correct decisions about how to proceed with the task. In these situations, it was reasonable to expect that the CSA of human defenders would be significantly decreased because they would have to expend more time and cognitive resources sifting through errors to make correct decisions. As misuse-based systems tend to present the results of their analyses without the raw data used in their analysis, it was reasonable to expect that participants would rely on the automation more heavily in this condition and make more frequent errors along with the automation. Below are the following hypotheses.

H2a: Misuse-based IDS (information analysis) will lead to higher levels of CSA of the human network defender than anomaly-based IDS (information acquisition).

H2b. Misuse-based IDS (information analysis) will lead to lower CSA than Hybrid IDS (information acquisition and analysis).

False alarms have been shown to be more damaging to SA than misses (Dixon, Wickens & McCarley, 2006). This effect is particularly impactful for anomaly-based

IDSs, as they are typically calibrated with a more liberal response criterion that captures more anomalous network traffic, but produces more false alarms. As anomaly-based IDSs present raw information to the defender without any form of analysis to sift through false alarms, anomaly-based systems will produce the lowest levels of CSA because human defenders will not have the assistance of the automation in any capacity. Below are the following hypotheses.

H3a. Anomaly-based IDS (information acquisition will lead to lower CSA than Hybrid IDS (information acquisition and analysis).

H3b. Anomaly-based IDS (information acquisition) will lead to lower CSA than misuse-based IDSs (information analysis).

Method

Participants

Participants consisted of 172 individuals, 64 males and 106 females, aged 18 to 50 years. Two participants chose not to disclose their gender. The average age of participants was 18.79 (*SD* = 2.76). Participants were recruited from the population of General Psychology students at San José State University via SONA systems. A power analysis conducted in G*power for a 3 x 3 factorial ANOVA with a medium effect size *f* = .25, α = .05, and power = .8 revealed that 158 participants were required to have sufficient power for analysis. Participants were compensated with course credit.

Materials

The mock IDS (Figure 1) was a program written in Visual Basic .NET that displayed simulated network traffic to the participant. It served as a low-fidelity proxy for an IDS and was used to simulate how network defenders interact with these systems. On the left side of the screen, simulated network traffic scrolled at a constant rate. This content was the same across all conditions. On the right side of the window, simulated alerts from one of three automation conditions were presented. The number of simulated alerts was the same across conditions. However, the alerts changed depending on the reliability of the automation because the automation made errors. Sometimes simulated alerts were flagged distractors instead of threats. In other instances, threats remained in the simulated network traffic screen and not flagged to the simulated alerts screen.

Network traffic log events contained distractors in the form of user activity such as: “User# login.”, “User# logout.”, “User# sent message to user#.”, and “New user registration.” Log events also contained threats. There were three types of threats with corresponding log events. A virus corresponded to the log event “Item deleted.” A worm was associated with the log event “Message sent to user.” A brute force attack could be recognized by the log event “Unknown user login.”

Participants were tasked with distinguishing threats from distractors and specifying the attack type. Participants made this decision by selecting one of the three options on a button array above the log viewer. Participants could click the brute button to indicate a brute force attack, the worm button for a worm attack, and the virus button for a virus attack. Participants had to make this decision as soon as they observed the log event. Log events were 2-3 seconds apart. Once a new log was presented, the click would be in reference to that newest log event. Once a log has passed, participants were unable to flag that event as a threat. The intent of this constraint was to create a sense of urgency to the task and encourage focus.

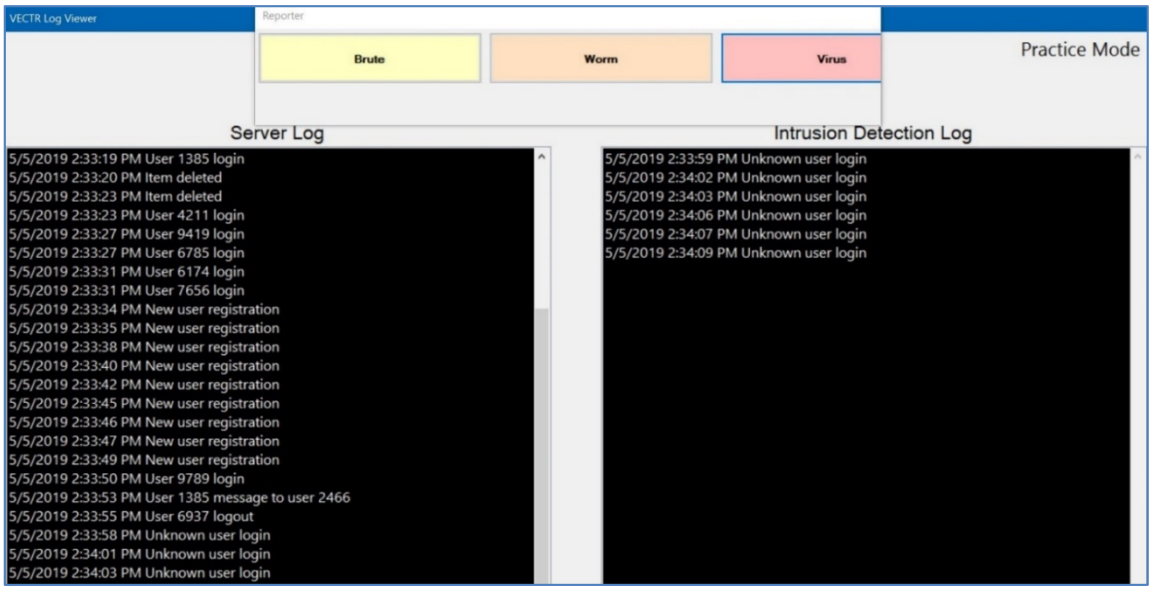

Figure 1. An image of the mock IDS

Manipulations

Intrusion detection system. As response criterion is an intrinsic property of each IDS (Aydin, Zaim & Ceylan, 2008; Mukkamala, Sung & Abraham, 2005; Werlinger, Hawkey, Muldner, Jaferian & Beznosov, 2008), it varied along with IDS, which resulted in three conditions: a hybrid IDS with a multilayered use of liberal and conservative response criterion, a misuse-based IDS with a conservative criterion (low hit rate, low false alarm rate), and an anomaly-based IDS with a liberal criterion (high hit rate, high false alarm rate).

Network log events could be displayed in one of three ways, depending on condition. In the anomaly-based condition, the network log event was displayed by itself (item deleted). In the misuse-based condition, the corresponding attack type was displayed by itself (virus). In the hybrid condition, both were displayed (item deleted – virus). The

presentation of the attacks was consistent within condition whether the log event was displayed in the simulated network traffic screen or the simulated alert screen.

Reliability. The reliability of the IDSs was manipulated at three levels (60%, 80%, 95%). We chose 60% to represent a low level of reliability because it has been determined to be the lower bound of a 95% confidence interval in which automation can still be effective (Wickens & Dixon, 2007). The reason that 95% was chosen was to represent a high, yet imperfect, level of reliability, which was expected to prevent participants from adopting a strategy of completely trusting the automation. 80% was chosen as a median for the reliability level factor.

The IDS could make three types of errors. It could flag a distractor to the simulated alert screen, fail to flag a threat to the simulated alert screen, or perform an incorrect analysis in the form of misattributing an attack to the distractor log instead of the proper corresponding log. For example, since the correct log event for a virus is “item deleted” the IDS could flag this as a worm in error. These error types were distributed equally across reliability conditions.

The reliability represented the proportion of log entries that were incorrectly interpreted by the automation in the trial. The conditions were created by calculating the ratio of errors to total number of log events and making sure that ratio corresponded to the three reliability levels. In this study, the total number of events in all conditions including both threats and distractors was 542. Of those events, there were 153 attacks (51 worms, 51 viruses, 51 brutes). This breakdown was consistent across all conditions. In the 60% reliability condition, 217 of 542 events (40%) were misinterpreted by the

automation; that is, if the event was anomalous, it would be labeled as normal. If the event was normal, it would be labeled as anomalous. The process was repeated for each reliability level across all levels of the IDS variable.

Measures

SAGAT. The situation awareness global assessment technique (SAGAT) is a global measure designed to access the cognitive elements of SA within a task environment (Endsley, 1988; Endsley, 2000). During the task, the simulation would freeze at unexpected moments in which the participant was asked random questions from a battery about the task environment. Participants' responses were based on knowledge obtained from the task environment prior to the freeze. SA was assessed via the type and correctness of these responses in order to obtain what task relevant knowledge might be available to the participant at that time and how it is organized in their minds (Endsley, 1995). In this study, questions were only asked at the perception and comprehension levels of SA because the task had no projection components that were expected to affect performance. The task did not require participants to project to future outcomes because they were only required to distinguish threats in real time.

Although participants were not able to predict when the freezes would occur, all participants were asked questions at the same fixed points in the simulation. This part of the procedure was to ensure that questions had similar correct answers at those points to serve as effective comparisons between participants. Participants were not told that those questions would be asked, or when. However, participants were instructed to pay attention to the content on which they would be quizzed.

Task performance. For this study, task performance was defined as the ability of participants to escalate potential network threats. The performance of the participants was determined by calculating *d*' from the ratio of hits to false alarms. As for the evaluation criteria for performance success, a hit was a successful escalation of a malicious network traffic string. A miss was a failure to escalate an actual threat. A false alarm would be the escalation of a network anomaly that is not malicious. A correct rejection was successfully not flagging normal network traffic.

Level of confidence. Between the training and actual task, participants were asked to rate their level of confidence with the task and their potential for success on a scale of 1 – 10 (1 being not very confident and 10 being very confident). Participants were asked, "On a scale of 1 – 10 (1 being not very confident and 10 being very confident) how would you rate your level of confidence in your performance on the task?" The purpose of this measure was to ascertain participants' expectations of their performance on the upcoming task.

Level of difficulty. After completing the task, participants were asked to rate the perceived level of difficulty of the task on a scale of 1 – 10 (1 being very easy 10 being very difficult). Participants were asked, "On a scale of 1 – 10 (1 being very easy 10 being very difficult) how would you rate the level of difficulty of the task?" This questions was intended to determine how difficult the task was for them to complete.

Demographic questionnaire. A short survey was administered prior to the task to obtain demographic information about the participants, as well as ascertain their perceived aptitude with usage of technology. These included age, gender, and self-report measures on a 1 – 9 scale (1 being not knowledgeable and 9 being very knowledgeable) about computer competency, familiarity with mobile technology, knowledge of internet usage, and basic knowledge of cyber security. For computer competency, participants were asked the following question: "On a scale of 1-9 (with 1 being not knowledgeable and 9 being very knowledgeable), how would you rate your level of aptitude with computers?" For aptitude with mobile technology, participants were asked: "On a scale of 1-9 (with 1 being not knowledgeable and 9 being very knowledgeable), how would you rate your level of aptitude with mobile technology?" To assess self-reports of internet usage, participants were asked: "On a scale of 1-9 (with 1 being not knowledgeable and 9 being very knowledgeable), how would you rate your level of aptitude with internet usage?" Finally, to assess perceived competence with cyber security, participants were asked the following: "On a scale of 1-9 (with 1 being not knowledgeable and 9 being very knowledgeable), how would you rate your level of confidence in cyber security?"

Procedure

Participants were recruited through SONA systems and brought into a lab space in the Department of Psychology at San José State University. The entire duration of the study was 30 minutes. In the first 10 minutes, participants were provided with an informed consent form and instructed on the procedure of the experiment. Participants were seated in front of the computer and asked to complete a brief demographic questionnaire

administered on Qualtrics. Next, participants were briefed on the details of the task. Participants were instructed to escalate potential attacks using their best judgement and information provided by the IDS. Participants were asked to complete a 5-minute training exercise to get familiar with interacting with the IDS identifying anomalies in the network traffic, flagging them and escalating. Once complete, the participants were asked to rate their level of confidence with their understanding of the task from a scale of 1 – 10 (1 not confident and 10 being extremely confident). Next, participants were informed about the functional properties and reliability of the IDS systems they were to be working with. Each participant completed one 15-minute session featuring a particular IDS system (anomaly based, misuse based, or hybrid based). Three times per session, the program was paused at predetermined points during the session and participants were presented a question about the current state of the task environment. This question was randomly selected from a pre-written battery of task-relevant questions about the current state of the system. Participants were scored on correct responses to the questions. Lastly, participants were debriefed.

Design

The independent variables (IV) in the study were type of automation (anomaly-based IDS, misuse-based IDS, and hybrid IDS) and reliability (60%, 80%, 95%). The primary dependent variables were CSA and task performance. CSA was represented by the percentage of correct SAGAT scores. Task performance was measured by *d*'. Data were analyzed using SPSS version 25. Two 3 x 3 factorial Analysis of Variance (ANOVA)

tests were conducted, one with SAGAT scores as the dependent variable (DV) and the second with *d*' as the DV. All factors were between-subjects factors.

Results

Self-Reported Measures

Across the demographic survey, one response was left blank on the mobile technology knowledge question and the cyber security knowledge question by one participant. This omission was due to an error in the survey design that allowed participants to continue without responding to a question. The descriptive statistics for the self-reported measures are presented below (Table 1).

Table 1

Descriptive Statistics for Self-report Measures of Knowledge of Mobile Technology, Computers, Internet Usage, and Cyber Security

Variable	*M*	*SD*	skewness/*SE*	kurtosis/*SE*	*n*
Mobile	6.61	1.42	-0.72/.18	.013/.37	171
Computer	5.6	1.61	-0.49/.18	-0.06/.37	172
Internet	6.42	1.49	-0.16/.18	-0.43/.37	172
Cyber Security	4.2	1.8	0.24/.18	-0.52/.37	171

When participants were asked about their knowledge of mobile technology and its usage, participants tended to rate themselves near 7, forming a negatively skewed distribution (M = 6.61, SD = 1.42, skewness = -0.72, SE = .18, kurtosis = 0.13, SE = .37, n = 171). The negative skewness indicates that participants tended to perceive themselves to be relatively knowledgeable about mobile technology. When asked about their knowledge of computer usage, participants rated themselves near 6 (M = 5.6, SD = 1.61,

skewness = -0.49, *SE* = .18, kurtosis = -0.06, *SE* = .37, *n* = 172). The negative skewness indicates that participants generally view themselves as moderately knowledgeable about computer usage. Most participants asked about their familiarity with internet usage rated themselves above a score of 5. The mode of this scale was 6 at 49 responses with a negative skew (*M* = 6.42, *SD* = 1.49, skewness = -0.16, *SE* = .18, kurtosis = -0.43, *SE* = .37, *n* = 172). The negative skewness indicates that participants generally tended to view themselves as knowledgeable about internet browsing and usage. However, participants tended to rate their knowledge of cyber security below a rating of 5. The mode of this scale was 5 at 37 responses. The average scores here dip compared to the others with a positive skew (*M* = 4.2, *SD* = 1.8, skewness = 0.24, *SE* = .18, kurtosis = -0.52, *SE* = .37, *n* = 171). The positive skewness indicates that participants tended to view themselves as less knowledgeable about matters related to cyber security. These combined results indicate that participants had a relatively firm grasp about knowledge of mobile technology, computers in general, and internet usage. However, participants tended to report themselves as much less knowledgeable about cyber security.

For the results related to the self-report measures of confidence and difficulty, it is important to mention that the following statistics are reported with missing data. The confidence scores are missing 4% of responses and have an *n* = 168. Difficulty scores are missing 8% of responses with an *n* = 158. The relatively large amount of missing data justified a test for whether the confidence and difficulty data was missing completely at random, as the missing data could impact the results. A Little's MCAR test (Little, 1988) was conducted for confidence and difficulty. The null hypothesis was not rejected for

confidence (24, n = 165) = 29.1, p = .22, or for difficulty χ^2(24, n = 158) = 29.1, p = .22, indicating that missing data were missing at random. Participants tended to report their confidence level with the task above a score of 5. Most participants rated themselves as 5, 6, or 7 with frequencies of 34, 34, and 37, respectively (M = 6.0, SD = 1.56, skewness = -0.34, SE = .19, kurtosis = -0.36, SE = .38, n = 165), indicating that participants viewed themselves as relatively confident they would perform well on the task after the training, and understood the nature of the task. After the task was completed, participants tended to rate the task difficulty above a score of 5. Most participants rated the difficulty a 7 or 8 out of 10 with recorded responses being 45 and 31, respectively. The descriptive statistics indicate that once the task was completed, participants tended to look back on the task as relatively difficult (M = 6.5, SD = 1.82, skewness = -0.64, SE = .19, kurtosis = 0.46, SE = .38, n = 158).

Alternatively, these results suggest a Dunning-Kruger effect (Schlösser, Dunning, Johnson & Kruger, 2013). This effect is observed when a novice overestimates his or her ability to perform a task when learning a new skill because novices are incapable of grasping the nuances that mastery requires. Once novices obtain a higher level of skill, confidence decreases. As the novice gains enough task knowledge to reach intermediate level, he or she begins to understand how much more knowledge is required to attain true mastery. This results in a parabolic confidence curve in which confidence is high at beginner level, decreases at intermediate level, and returns once mastery is attained. In the context of this study, participants underwent a 5 min training course during which the experimenter walked them through the task. Afterwards, participants rated their

confidence as high. As participants became more familiar with the task and the true difficulty of the task was revealed, their confidence may have decreased. This relationship may explain the high reports of perceived task difficulty after the study was completed.

SAGAT Measure of SA

A 3 (anomaly-based, misuse-based, hybrid) x 3 (60%, 80%, 95%) factorial ANOVA on IDS type and reliability was conducted for the SAGAT data percent correct. Descriptive statistics are presented in the table below (Table 2). The bar chart below depicts the differences in means between conditions for percent correct of SAGAT scores (Figure 2). The analysis revealed no significant main effects for type of automation, $F(2, 163) = .15$, $p = .86$, partial $\eta^2 = .002$ $\beta = .06$, or reliability $F(2, 163) = 0.04$, $p = .96$, partial $\eta^2 = .001$, $\beta = .07$. These main effects were not qualified by a significant interaction, $F(4, 163) = 0.87$, $p = .48$, partial $\eta^2 = .02$, $\beta = .27$. These results indicate that this study cannot conclude that any observable difference in impact to participant CSA exists between the different intrusion detections systems tested, their affiliated criterions, or the reliability settings.

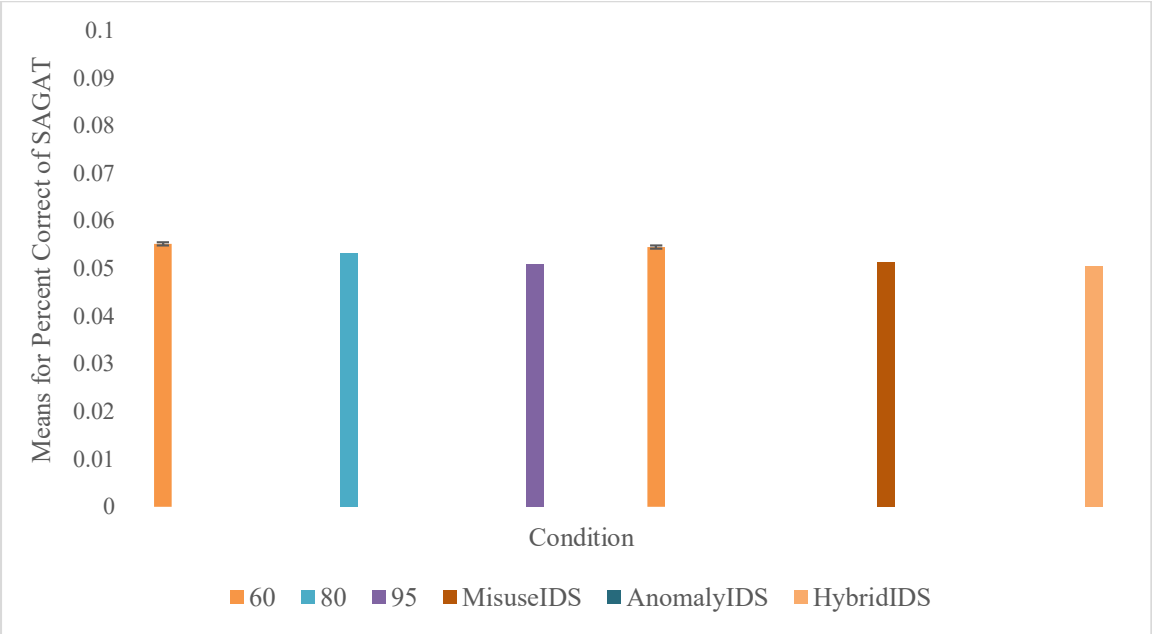

Figure 2. Means for percent correct of SAGAT

Table 2

Descriptive Statistics for Percent Correct of SAGAT Scores

Variable	*M*	*SD*	skewness/*SE*	kurtosis/*SE*	*n*
Type of Automation	1.0	0.13	2.15/.19	3.59/.37	172
Reliability	78.43	14.39	-0.18/.19	-1.51/.37	172
Age	18.79	2.76	9.21/.19	98.73/.37	172
Percent Correct	0.05	0.13	2.15/.19	3.59/.37	172

Performance (*d*')

A 3 (anomaly-based, misuse-based, hybrid) x 3 (60%, 80%, 95%) factorial ANOVA between IDS type and reliability was conducted for *d*'. Descriptive statistics are presented in the table below (Table 3). The bar chart below depicts the differences in means between conditions for task performance (Figure 3). The analysis revealed no significant main effects for type of automation, $F(2, 163) = 1.03$, $p = .36$, partial $\eta^2 = .01$, $\beta = .23$, or reliability, $F(2, 163) = 0.12$, $p = .89$, partial $\eta^2 = .001$, $\beta = .07$. These main effects were not qualified by a significant interaction, $F(4, 163) = 1.46$, $p = .22$, partial $\eta^2 = .035$, $\beta = .45$. These results indicate that this study cannot conclude that any observable difference in impact to participant CSA exists between the different intrusion detections systems tested, their affiliated criterions, or the reliability settings.

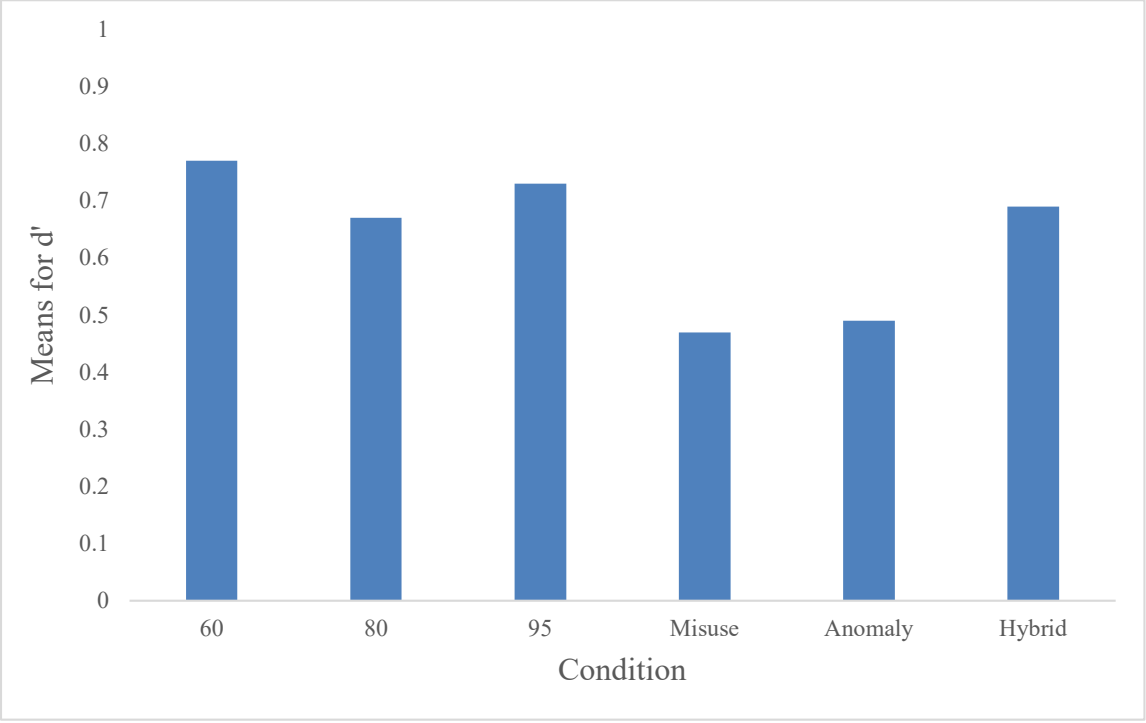

Figure 3. Means for task performance

Table 3

Descriptive Statistics for Task Performance

Variable	*M*	*SD*	skewness/*SE*	kurtosis/*SE*	*n*
Type of Automation	1.0	0.82	0.0/.18	-1.5/.37	172
Reliability	78.34	1.09	-0.18/.18	-1.5/.37	172
d’	.72	1.13	-0.37/.19	-0.83/.37	172

Below are the means for the performance calculations (Table 4) and a projection of expected performance of participants across conditions (Table 5). As the hybrid IDS layers misuse-based analysis on top of anomaly-based information gathering, multiplying the miss rates of each will give us a probability of how hybrid systems are expected to perform. The miss rates for misuse and anomaly-based systems were multiplied to obtain a projection for what the hybrid criterion values should be in an ideal state. These could be compared to the observed values from participants. As displayed, the actual observed miss rates for the hybrid condition are higher than the projected value. The small negative values suggest participants underperformed from the expected. This finding is not surprising as the ANOVA results were not significant.

Table 4

Means for Sensitivity, Bias, Criterion, Hit Rates, and False Alarm Rates Across Conditions.

Variable	d'	Beta	C	Hit	FA
60%	.77	2.91	-0.86	.39	.15
80%	.67	2.72	-0.85	.37	.18
95%	.73	2.59	-0.96	.36	.16
Misuse	.47	2.54	-1.11	.27	.15
Anomaly	.49	3.15	-.72	.45	.16
Hybrid	.69	2.52	-.87	.38	.17

Table 5

Projected Criterion for Hybrid Condition Compared with Actual Observed Values.

Variable	*M* Misuse misses	*M* Anomaly misses	Projected *M* Hybrid misses	Observed *M* Hybrid misses	Difference
60%	0.65	0.56	0.36	0.63	-.027
80%	0.77	0.56	0.43	0.55	-0.12
95%	0.71	0.54	0.38	0.67	-0.29

Average criterion for the participants was calculated between the reliability conditions to see if their own individual criterion would be affected by reliability. The

equation used was the normal distribution function -.5*(NORMSINV(hits) + NORMSINV(false alarms)). This calculation was performed on the 60%, 80%, 95% reliability conditions. The outputs for these calculations were as follows: for 60% the output was 0.67; for 80% the output was 0.63; for 95% it was 0.67. These results suggests that the average participant criterion was a bit lower for the 80% condition.

Discussion

The purpose of this study was to explore the moderating effects of level of diagnostic aiding and automation reliability on the human ability to build CSA in CND. Specifically, this study examined whether hybrid (information acquisition and information analysis) automation with multilayered analysis incorporating both liberal and conservative response criterion lead to better CSA than anomaly-based (information acquisition) systems with liberal response criterion (high hit rate, high false alarm rates) or misuse-based (information analysis) systems with conservative response criterion (low hit rate, low false alarm rates). As the results are inconclusive, this study was unable to achieve this goal. Although this research did not reveal relationships between the variables tested, it can inform how limitations of the study may have affected the results. Finally, an experimental redesign that may mitigate these limitations is proposed.

Limitations of this Study

One of the limitations of this study was having insufficient power to achieve statistical significance. Given the obtained effect sizes, it is possible that the effect size was overestimated during the power analysis and the actual effect is much smaller, meaning that the sample size used in this study was too small to observe the actual effect. It is also possible that the participant pool itself may have been mismatched to the practitioner-focused area of cyber security examined in this study. Intrusion detection and threat escalation are professional cyber security activities. The participants in this study were not cyber security practitioners. This incongruence was deliberate, yet expected to yield interpretable results. SJSU General Psychology students were chosen as the primary

participant pool for this study instead of cyber security practitioners because unlike practitioners, the participants were readily available. Cyber security practitioners are more inaccessible due to the nature of their profession. Some of these reasons include: the secrecy and security limitations of their work, the time it would take to participate, and the lack of resources for appropriate compensation. Using SJSU students afforded the ability to bypass some of these limitations.

The partial eta squares reported indicate relatively small effect sizes. Combined with the relatively high self-reported task difficulty scores and low self-reported ratings of cyber security domain knowledge, the findings suggest that a floor effect may have obscured the ability to observe differences between conditions. Steps were taken to make the task more appropriate for this population. Extensive piloting resulted in the development of contingencies in the form of more explicit instructions, longer and more rigorous training exercises, reduced workload, and simplified SAGAT questions. The ultimate task that was developed shares little overlap with the professional task. Thus, simplified task simulations may not be effective. Despite efforts to combat task difficulty, the results suggest that the task was too difficult for participants to understand and complete with their available domain knowledge. Future iterations of the study should include stronger manipulations in the form of more robust SAGAT questions. Longer and more rigorous training may be required unless the task is simplified or modified to be more familiar to participants from this population.

Another possible explanation for the small effect size could be that the constructs examined in this study do not operate in the same manner in the domain of cyber security

as they do in other domains in human factors. It is possible that the decisions in CND and the CSA required to make them differ from other human-machine systems examined in previous literature. The likely floor effects would obscure observation of this possibility. Future research that reduce floor effects could provide more insight into this possible problem by allowing better observation into human decision making in CND.

As most participants rated themselves relatively low on cyber security domain knowledge, it is possible that participants were not lacking in understanding of cyber security. Perhaps, how end users conceptualize cyber security was misaligned with the task they were asked to perform. That being said, it is possible to examine the influences of level of automation and automation reliability on human CSA in this domain with this participant pool. These constructs could be transformed and applied in a way that is not based on intrusion detection. Rather, a task that is more grounded in the day to day lives of end users, which might include being safe in the mobile, computer, and online worlds. Participants reported being relatively confident in their knowledge on these subjects. That knowledge could be applied to a typical behavior that end users engage in everyday, such as checking emails and evaluating the strategies involved in avoiding exploitation in the form of phishing attempts.

Proposed Modified Study Design

To adjust the study and compensate for the misalignment, the study could be revised by changing the nature of the task to something more relevant to the daily tasks of the population. For example, participants could perform a phishing email detection task with a Gmail simulation in conjunction with an automated feature that helps flag phishing

attempts and bring them to the attention of the user. In this instance, the level of automation could be varied by the extent to which the automation is involved in decision making around determining whether or not a suspicious email is a phishing attempt. In one condition it could send a small notification indicating that the email seems different than one the user typically receives. For example, an email from a sender that is not recognized in your contacts. In another, it could send it to a spam folder without notifying the user. In the last condition, it could send to the spam folder and notify the user. Reliability could be manipulated similarly to the current study, in that it varies the errors the automation makes by flagging non-phishing emails at different rates (60%, 80%, 95%). CSA would be evaluated from a performance-based measure similar to the current study with signal detection *d*' calculated from a ratio of hits (ability to detect a phishing email) to false alarms (flagging a non-phishing email). Responses would be recorded by clicking a button to flag as phishing and notify the service provider. Responses would be attached directly to the email that inspired the click by aligning the timestamps, which will ensure that the response can be coded with a specific signal to improve accuracy of the analysis.

A simulated Gmail account could be created through a user interface prototyping tool. These tools create a series of interconnected static screens that can simulate interactivity similar to a flipbook. When these prototypes are fully constructed and completely interconnected, a simulation of fully interactive task environments can be created. Participants will be able to click on any email or folder in this environment as though it were a real account. Participants could be given a hypothetical scenario for the study.

Participants could be told that over summer vacation their SJSU email account has received hundreds of messages they had neglected to check. The first email in the inbox could be from the SJSU email administrator informing participants that a new automated phishing email detection system has been implemented on SJSU email accounts. To test this new system, fake phishing email attempts increased in prevalence over the summer. All students should go through their inbox to identify these attempts along with this system.

A SAGAT could also be performed examining the perception and comprehension elements of CSA about the phishing attempts. While participants click through the simulated inbox, the simulation could freeze after a certain number of clicks. This number of clicks could be randomized through the programing of the simulation. The click could bring participants to one of the interconnected nodes. These would be a blank screen with a SAGAT question and a continue button. Participants could log their response and click continue to bring them back to the inbox screen.

SAGAT questions might include perception level elements of end user CSA. For example, these questions could ask: how many phishing emails have been detected so far? What is the current status of your inbox? Questions could also include comprehension elements, which would require participants to integrate multiple perception level elements together or explain in more detail how they are distinguishing phishing emails from non-phishing emails. Questions of this nature might include: how many phishing emails have been detected since the last non-phishing email you found?

At what time was the most recent phishing email detected? What clue tipped you off to the last phishing email?

This new experimental design should address some of the issues with task difficulty and construct validity in this current study. Changing the nature of the task from intrusion detection to identifying email phishing techniques will ground the detection of cyber threats in something familiar to the end user. SJSU students should be able to conduct this task more effectively and accurately, which will likely dissipate floor effects and allow for more visibility into how the level of automation and automation reliability interact to moderate CSA in end user cyber hygiene practices.

Conclusion

As reported, the analyses were not significant. This study could not observe any main effects or interactions effects across type of automation or the reliability conditions for both percent correct of SAGAT scores and performance scores. This study could not ascertain the extent to which intrusion detections systems and reliability moderate CSA in CND. The scores for confidence and difficulty suggest a Dunning-Kruger effect in which participants felt confident going into the task, but found it very difficult by the time it was completed. This finding suggests that the training program may have been insufficient to prepare participants for a task that was too difficult for them to perform.

The relatively high self-reported knowledge ratings for mobile, computer and internet usage when compared with lower ratings for cyber security related knowledge suggest that participants may have had insufficient domain knowledge to perform the task successfully. However, it is also possible that participants' knowledge was mismatched

with the practitioner focused cyber security task in this study. Modifying the task to be more aligned with the knowledge base of the participant pool may lead to more successful observations in future iterations of this study.

References

Aydın, M. A., Zaim, A. H., & Ceylan, K. G. (2009). A hybrid intrusion detection system design for computer network security. *Computers & Electrical Engineering*, *35*(3), 517-526.

Bronte, R. N. (2016). A Framework for Hybrid Intrusion Detection Systems. (Unpublished master's thesis). Kennesaw State University, Kennesaw, GA.

Champion, M. A., Rajivan, P., Cooke, N. J., & Jariwala, S. (2012). Team-based cyber defense analysis. In *Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 2012 IEEE International Multi-Disciplinary Conference,* 218-221. IEEE.

D'Amico, A., Whitley, K., Tesone, D., O'Brien, B., & Roth, E. (2005). Achieving cyber defense situational awareness: A cognitive task analysis of information assurance analysts. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 49(3), 229-233. Sage CA: Los Angeles, CA: SAGE Publications.

Dexter, F., Willemsen-Dunlap, A., & Lee, J. D. (2007). Operating room managerial decision-making on the day of surgery with and without computer recommendations and status displays. *Anesthesia & Analgesia*, *105*(2), 419-429.

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2006). How do automation false alarms and misses affect operator compliance and reliance? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(1), 25-29. Santa Monica, CA: Human Factors and Ergonomics Society.

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the independence of compliance and reliance: Are automation false alarms worse than misses?. *Human Factors*, *49*(4), 564-572.

Endsley, M. R. (1988). Situation awareness global assessment technique (SAGAT). *IEEE 1988 National Aerospace and Electronics Conference*, (*3)*, 789-795. doi: 10.1109/NAECON.1988.195097

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. (1995). Toward a theory of situation awareness in dynamic systems. Human Factors: *The Journal of the Human Factors and Ergonomics Society*, 37(1), 32-64. doi:10.1518/001872095779049543

Endsley, M. R. (2000a). Direct measurement of situation awareness: Validity and use of SAGAT. In M. R. Endsley & D. J. Garland (Eds.), *Situation Awareness Analysis and Measurement, (*pp.147-173). Mahwah, NJ: Lawrence Erlbaum Associates.

Endsley, M. R. (2000b). Theoretical underpinnings of situational awareness: A critical review. In M. R. Endsley & D. J. Garland (Eds.), *Situation Awareness Analysis and Measurement,* 3-32. Mahwah, NJ: Lawrence Erlbaum Associates.

Flach, J. M. (2015). Situation awareness: Context matters! A commentary on Endsley. *Journal of Cognitive Engineering and Decision Making*, *9*(1), 59-72.

Goodrich, M. A., McLain, T. W., Anderson, J. D., Sun, J., & Crandall, J. W. (2007). *Managing autonomy in robot teams: Observations from four experiments*. Paper presented at the meeting of the ACM International Conference on Human-Robot Interaction, Arlington, Virginia.

Gutzwiller, R. S., Fugate, S., Sawyer, B. D., & Hancock, P. A. (2015). The human factors of cyber network defense. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 59(1), 322-326. Sage CA: Los Angeles, CA: SAGE Publications.

Gutzwiller, R. S., Hunt, S. M., & Lange, D. S. (2016). A task analysis toward characterizing cyber-cognitive situation awareness (CCSA) in cyber defense analysts. In *Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 2016 IEEE International Multi-Disciplinary Conference on* (pp. 14-20). IEEE.

Harwood, K., Barnett, B., & Wickens, C. D. (1988). Situational awareness: A conceptual and methodological framework. In *Proceedings of the 11th Biennial Psychology in the Department of Defense Symposium* (pp. 23-7).

Horrey, W. J., Wickens, C. D., Strauss, R., Kirlik, A., & Stewart, T. R. (2009). Supporting situation asessment through attention guidance and diagnostic aiding: The benefits and costs of display enhancement on judgment skill. In A. Kirlik (Ed.), *Adaptive Perspectives on Human-Technology Interaction: Methods and Models for Cognitive Engineering and Human-computer Interaction* (pp. 55-70). Oxford, England: Oxford University Press.

Horrey, W. J., & Wickens, C. D. (2001). *Supporting battlefield situation assessment through attention guidance and diagnostic aiding: A cost-benefit and depth of processing analysis* (Report No. ARL-01-16/FED-LAB-01-1). Savoy, IL: Aviation Research Lab.

Jajodia, S., Liu, P., Swarup, V., & Wang, C. (2010). *Cyber situational awareness*, 14. New York, NY: Springer.

Kaber, D. B., Omal, E., & Endsley, M. (1999). Level of automation effects on telerobot performance and human operator situation awareness and subjective workload. *Automation Technology and Human Performance: Current Research and Trends*, 165-170.

Kaber, D. B., Onal, E., & Endsley, M. R. (2000). Design of automation for telerobots and the effect on performance, operator situation awareness, and subjective workload. *Human Factors and Ergonomics in Manufacturing*, *10*(4), 409-430.

Kemmerer, R. A., & Vigna, G. (2002). Intrusion detection: a brief history and overview. *Computer*, *35*(4), l27-l30.

Kuchar, J. K. (1996). Methodology for alerting-system performance evaluation. *Journal of Guidance, Control, and Dynamics*, *19*(2), 438-444.

Kim, G., Lee, S., & Kim, S. (2014). A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. *Expert Systems with Applications*, *41*(4), 1690-1700.

Little, R. (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association, 83*(404), 1198-1202. doi:10.2307/2290157

Madhavan, P., & Phillips, R. R. (2010). Effects of computer self-efficacy and system reliability on user interaction with decision support systems. *Computers in Human Behavior*, *26*(2), 199-204.

Mahoney, S., Roth, E., Steinke, K., Pfautz, J., Wu, C., & Farry, M. (2010). A cognitive task analysis for cyber situational awareness. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting,* 54(4), 279-283. Sage CA: Los Angeles, CA: SAGE Publications.

McHugh, J., Christie, A., & Allen, J. (2000). Defending yourself: The role of intrusion detection systems. *IEEE Software*, *17*(5), 42-51.

Mukkamala, S., Sung, A., & Abraham, A. (2005). Cyber security challenges: Designing efficient intrusion detection systems and antivirus tools. *Vemuri, V. Rao, Enhancing Computer Security with Smart Technology*, 125-163.

Onwubiko, C. (2009). Functional requirements of situational awareness in computer network security. In *Intelligence and Security Informatics, 2009. ISI'09. IEEE International Conference,* (pp. 209-213). IEEE.

Onwubiko, C., & Owens, T. (2012). Review of situational awareness for computer network defence. *Situational Awareness in Computer Network Defense: Principles, Methods and Applications*, 1-9.

Onwubiko, C. (2016). Understanding Cyber Situation Awareness. *Intl. Journal on Cyber Situational Awareness*, 1(1).

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230-253.

Parasuraman, R., Hancock, P. A., & Olofinboba, O. (1997). Alarm effectiveness in driver- centred collision-warning systems. *Ergonomics*, *40*(3), 390-399.

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.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *Journal of Cognitive Engineering and Decision Making*, *2*(2), 140-160.

Peddabachigari, S., Abraham, A., Grosan, C., & Thomas, J. (2007). Modeling intrusion detection system using hybrid intelligent systems. *Journal of Network and Computer Applications*, *30*(1), 114-132.

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.

Rudisill, M. (2000). *Crew/automation interaction in space transportation systems: Lessons learned from the glass cockpit*. Hampton, VA: NASA Langley Research Center.

Ruff, H. A., Narayanan, S., & Draper, M. H. (2002). Human interaction with levels of automation and decision-aid fidelity in the supervisory control of multiple simulated unmanned air vehicles. *Presence: Teleoperators and Virtual Environments*, *11*(4), 335-351.

Sarter, N. B., & Schroeder, B. (2001). Supporting decision making and action selection under time pressure and uncertainty: The case of in-flight icing. *Human Factors*, *43*(4), 573-583.

Sawyer, B. D., Finomore, V. S., Funke, G. J., Mancuso, V. F., Funke, M. E., Matthews, G., & Warm, J. S. (2014). Cyber Vigilance Effects of Signal Probability and

Event Rate. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 58(1), 1771-1775.

Schlösser, T., Dunning, D., Johnson, K. L., & Kruger, J. (2013). How unaware are the unskilled? Empirical tests of the "signal extraction" counterexplanation for the Dunning–Kruger effect in self-evaluation of performance. *Journal of Economic Psychology*, *39*, 85-100.

Schuster, D. (2013). The effects of diagnostic aiding on situation awareness under robot unreliability. (Unpublished doctoral dissertation). University of Central Florida, Orlando, FL.

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

Singh, A. L., Tiwari, T., & Singh, I. L. (2009). Effects of automation reliability and training on automation-induced complacency and perceived mental workload. *Journal of the Indian Academy of Applied Psychology*, *35*(2009), 9-22.

Swets, J. A., & Pickett, R. M. (1982). *Evaluation of diagnostic systems: methods from signal detection theory*. New York, NY: Academic Press.

Tesfahun, A., & Bhaskari, D. L. (2015). Effective hybrid intrusion detection system: A layered approach. *International Journal of Computer Network and Information Security*, *7*(3), 35.

Werlinger, R., Hawkey, K., Muldner, K., Jaferian, P., & Beznosov, K. (2008). The challenges of using an intrusion detection system: is it worth the effort?. In *Proceedings of the 4th Symposium on Usable Privacy and Security* (pp. 107-118). ACM.

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science*, *8*(3), 201-212.

Appendix A

SJSU SAN JOSÉ STATE UNIVERSITY

Office of Research
TEL: 408-924-2272
Washington Square
Academic Affairs San

SanJosé State University
Division of One
officeofresearch@sjsu.edu
José, CA 95192-0025
sjsu.edu/research

SAN JOSE STATE UNIVERSITY

HUMAN SUBJECTS INSTITUTIONAL REVIEW BOARD

IRB Notice of Approval

Date of Approval: 1/8/2018

Study Title: A Comparative Study of the Influence of Level Automation and Reliability of IDS Systems on Cyber Situation Awareness

Primary Investigator(s): Dr. David Schuster

Student(s): Ian Cooke

Other Team Members: Pilar Bianchi, Kristina Devtyan, Elizabeth Shallal

Funding Source: National Science Foundation

IRB Protocol Tracking Number: S17162

Type of Review

☒ Exempt Registration: Category of approval §46.104(d)(2ii)

☐ Expedited Review: Category of approval §46.110(a)()

☐ Full Review

☐ Modifications

☐ Continuing Review

Special Conditions

☒ Waiver of signed consent approved

☐ Waiver of some or all elements of informed consent approved ☐ Risk determination for device: N/A ☐ Other:

Continuing Review

☒ Is not required. Principal Investigator must file a status report with the Office of Research one year from the approval date on this notice to communicate whether the research activity is ongoing. Failure to file a status report will result in closure of the protocol and destruction of the protocol file after three years.

☐ Is required. An annual continuing review renewal application must be submitted to the Office of Research one year from the approval date on this notice. No human subjects research can occur after this date without continuing review and approval.

Approved by Dr. Pamela C. Stacks

Associate Vice President

Institutional Official

Office of Research

San Jose State University

IRB Contact

Alena Filip

Human Protections Analyst

408-924-2479

Alena.Filip@sjsu.edu

Primary Investigator Responsibilities

- Any significant changes to the research must be submitted for review and approval prior to the implementation of the changes.
- Reports of unanticipated problems, injuries, or adverse events involving risks to participants must be submitted to the IRB within seven calendar days of the primary investigator's knowledge of the event.
- If the continuing review section of this notice indicates that continuing review is required, a request for continuing review must be submitted prior to the date the provided.

Appendix B

REQUEST FOR YOUR PARTICPATION IN RESEARCH

TITLE OF THE STUDY

The Effect of Level of Automation and Reliability of IDS Systems on Cyber Situation Awareness

NAME OF THE RESEARCHER

Principle Investigator: Ian Cooke San Jose State University graduate student

Faculty Advisor: Dr. David Schuster PhD.

PURPOSE

The purpose of this research is to identify how and/or to what extent the level of automation and the reliability of intrusion detection systems influences the ability for human computer network defenders to successfully recognize cyberattacks.

PROCEDURES

Upon signing this agreement, you will be asked to conduct a quick survey about your basic characteristics (gender, age, etc.) and your knowledge of computers and cyber security. After that, you'll sit in front of a laptop computer and step into the role of a network defender and search through network traffic and identify cyberattacks with the help of our intrusion detection system simulator. You'll start by performing a 10 minute training exercise to help you get familiar with the task. When you're ready, you'll then perform the main task in one 10 minute section. Finally, you'll be debriefed and given credit upon completion of the experiment.

POTENTIAL RISKS

There are no potential risks beyond those incurred during normal interaction with a computer.

POTENTIAL BENEFITS

There are no direct benefits to you by participating in this research beyond course credit. However, your participation will contribute to our knowledge of cyber security and may lead to safer networks in the future.

COMPENSATION

Should you choose to participate, you will be compensated with one hour of Sona credit toward your Introduction to Psychology course requirement. You also have the option of completing an alternative assignment for credit in supplement of or in conjunction with this study. The details of this assignment are located on the San Jose State research pool website at the following link:
http://www.sjsu.edu/psych/Undergraduates/subjectpool.html

In addition to earning credit for your course, research participation gives you hands-on experience with the psychological research process and introduces you to techniques that may be useful for a career in psychology. It also allows you to contribute to the scientific study of mind, brain, and behavior.

CONFIDENTIALITY

You will be represented by a number throughout the study. The only record we will have of participation will be the signed informed consent document. The informed consent document will not be linked to your data once it has been collected.

Informed consent documents will be kept in the lab behind a locked door and/or in David Schuster's locked office. Data without identifying marks will be shared within the research team.

PARTICIPANT RIGHTS

Your participation in this study is completely voluntary. You can refuse to participate in the entire study or any part of the study without any negative effect on your relations with San Jose State University. You also have the right to skip any question you do not wish to answer. This consent form is not a contract. It is a written explanation of what will happen during the study if you decide to participate. You will not waive any rights if you choose not to participate, and there is no penalty for stopping your participation in the study.

QUESTIONS OR PROBLEMS

You are encouraged to ask questions at any time during this study.

• For further information about the study, please contact ***Ian Cooke*** [email: iacooke@sjsu.edu]

• Complaints about the research may be presented to ***Lynda Heiden*** [email: Lynda.heiden@sjsu.edu]

• For questions about participants' rights or if you feel you have been harmed in any way by your participation in this study, please contact Dr. Pamela Stacks, Associate Vice President of the Office of Research, San Jose State University, at 408-924-2479.

SIGNATURES

Your signature indicates that you voluntarily agree to be a part of the study, that the details of the study have been explained to you, that you have been given time to read this document, and that your questions have been answered. You will receive a copy of this consent form for your records.

Participant Signature

_____
_____ Participant's Name (printed)
Participant's Signature Date

Researcher Statement

I certify that the participant has been given adequate time to learn about the study and ask questions. It is my opinion that the participant understands his/her rights and the purpose, risks, benefits, and procedures of the research and has voluntarily agreed to participate.

_____

_____ Signature of Person Obtaining Informed Consent Date.

Appendix C

Protocol for Ian's Thesis Study

Setting up a SONA researcher account

Email someone about this and she'll get you signed up.

She should give you a username and password to login. You can change this when you get in.

Your student email should be your login and you'll have to make a password.

Once you're in your account, I'll put you as admins on my study so you should be able to see it.

Putting up study sessions to SONA

Once you have your account and you're on my study, you should be able to see it under the "my studies" page.

You can click on my study name and it will give you details about the study.

To add timeslots: (you'll have some liberty here)

1. Click the timeslot button on next to my study name on the "my studies" page. You'll be able to see a list of all the available previous, scheduled timeslots and who is running them when.

2. You can add one or multiple slots (tabs up at the top.)

3. Specify number of slots

4. write the date of the study

5. Specify the time

6. free time between slots is for multiple (just if you want a break in between)

7. Participants per slot will always be 3 (number of computers we have to run)

8. Specify the location of your lab space you selected so your participant knows where to go

9. click add and you're done!

On the day…Set up the stuff (BEFORE THE PARTICIPANT ARRIVES)

Make sure you bring all three laptops and the master flash drive with you. It will always be in a backpack on the back shelf. DOUBLE CHECK TO MAKE SURE IT"S ALL THERE. There are three laptops, chargers, and one master flash drive.

1. Turn on all laptops. Make sure they are all plugged in to the wall.

2. Get online and set up the Qualtrics preliminary survey. This is in the VECTR Lab qualtics account. (let me know if you don't have access) The survey will be called: Ian's Thesis Demographic Survey

3. Open the folder on the desktop called Ian's Experiment.

Run the log viewer program for the training condition (check the participant sheet excel document to see the participant number and version ID. Make sure these all match up and load the proper version)

Make sure that the participant number you put in corresponds to what version you're supposed to open on the sheet.

4. Have consent forms set up for each participant.
Write the participant number and the condition ID on the top right corner of each one.
Cross reference the spreadsheet to make sure everything is right.

5. Open the Participant list and test condition study spreadsheet and write down the participant's name, gender, age etc. next to the participant number and test condition they're in. DOUBLE CHECK AGAIN TO MAKE SURE EVERYTHING IS RIGHT AND LINES UP.

Running the study

1. Bring participant into the shared data collection space (room specification will be up to you)

2. Have the participant read and sign the written consent form and make sure they understand everything. Be sure to ask them if they have any questions.

3. Explain the basics of the experiment tell them you'll be happy to discuss any of the details of the study and/or answer any questions in the debriefing session at the end of the experiment. Say this:

"Thanks for participating in the study today. You will be put in the role of a cyber defender. You'll be tasked with monitoring a network and identifying potential attacks for about 15 minutes. I'll explain in more detail momentarily and if you have any questions about the nature of the study, I'll be happy to answer them at the end. But first,

please take our preliminary survey so we can learn a bit about you and your familiarity with computers and technology."

4. Have the participant take the preliminary survey and log their responses (it should do this automatically when they're done, but double check)

5. Explain the instructions of the participant.

"OK, thanks for doing the survey, now we're going to get you trained on the task. You'll be watching a log simulation of network traffic. Your task is to correctly identify the potential hacks to the network. Highlight the most recent log presented. If it's an attack, hit notify. If it isn't, hit pass. It's important to keep track of the attacks and distractors to distinguish them, so try to pay close attention to the timestamps of the attacks and distractors, and how many have occurred.

There are 3 types of attacks: Viruses, Worms, and Brutes. Viruses are indicated by a log called Item deleted, Worms are indicated by message sent to user #, and brutes are indicated by unknown user login. We'll do a little practice session to get you used to everything. Any questions?"

Once you're done explaining load the training session and tell them to click begin whenever they're ready. Inform them to ask questions and to tell you when they're done with the training.

Stay with them while they're completing their training and guide them through the process. Do your best to answer any questions they have about the task or the program.

When they're done, ask if they have any questions. Also, ask them how they feel about everything and gauge their confidence. There's nothing we can do, but it would just be good to know.

6. Load the actual trial. Double check the sheet to make sure it's the right condition. **In the analysis only condition, cover the data log on the lefthand side with some paper.**

Input the proper participant number in the box and hit enter.

Tell the participant to click OK begin whenever he/she is ready.

7. At three times during the session, the program will freeze and leave to another screen. When this happens, randomly choose one of the SAGAT question printed out and hand it to the participant.

Give them a minute to write a response or give up.

Make sure you write the participant number on the top right of the page and set it aside.

8. Be sure to pay attention to any peculiar things about the study (e.g. did the participant have a problem with something? Did the logviewer glitch? Was someone extra confused? Any idiosyncratic things)

When you observe anything like this, be sure to mention it in the comments.

Open the Participant list and test condition excel sheet where you logged the participant's name age gender etc. and write anything strange in that comment section.

Finishing the study

1. When the participant has finished, ask them "on a scale of 1 – 10 1 being extremely easy and 10 being extremely difficult how would you rate the difficulty of this task. Have them write their response on the spreadsheet.

2. When participants finish the survey, ask them if they have any questions. If they don't, tell them you will give them credit as soon as they leave and tell them they are free to go. Thank them for their time.

3. For those that are curious, say the following:

"We looking to improve the situation awareness of computer network defenders. We're varying the reliability of the automation and the extent to which the automation is involved in the task to see if we can optimize which combination will increase SA the most and produce the most successful outcome."

4. Ask them if they have any further questions. Do your best to answer them, but "I'm not sure" is an acceptable response if you don't know.

Finishing the studies for the day

When you're all done with running your sessions there's going to be some clean up work to do…

1. The data is all stored locally on each computer. You're going to have to take the master flash drive and download all of the recent sessions off of the "experiment logs" data folder.

2. Clean up any mess you or your participants made while you were there and make sure the door is locked when you leave. We need to leave these shared data spaces in better conditions than we found them. Remember that you're representing VECTR lab.

3. Grab everything you brought with you and bring it back to the lab

4. Plug the flashdrive into one of the computers and upload all of the data to my master data folder in my Ian's thesis 2018 folder on the VECTR team drive.

5. Put all equipment (computers, chargers, and master flashdrive) back to the backpack where you found it on the shelf.

6. make sure you return whatever key to whatever room back to the psych office and the lab key when you're done.

Appendix D

Ian's Thesis Demographic Survey

Start of Block: Default Question Block

Q10

Hi there!

Thanks for your participation in thus study! The following preliminary survey is intended to gather basic demographic information about you and your level of comfort with computers, technology and security. When you're ready, please click below to begin.

Have fun!

Page Break

End of Block: Default Question Block

Start of Block: Default Question Block

*

Q4 Age

_____

Q5 Sex

- Male (1)
- Female (2)
- Choose not answer (3)

Q6 On a scale of 1-9 (with 1 being not knowledgeable and 9 being very knowledgeable), how would you rate your level of aptitude with computers?

Computer Aptitude (1)	

Q7 On a scale of 1-9 (with 1 being not knowledgeable and 9 being very knowledgeable), how would you rate your level of aptitude with mobile technology?

Mobile Technology Aptitude (1)	

Q8 On a scale of 1-9 (with 1 being not knowledgeable and 9 being very knowledgeable), how would you rate your level of aptitude with internet usage?

Internet Usage Aptitude (1)	

Q9 On a scale of 1-9 (with 1 being not knowledgeable and 9 being very knowledgeable), how would you rate your level of confidence in cyber security?

Cyber Security Confidence (1)	

End of Block: Default Question Block

Appendix E

SAGAT Questions

- How many worms have been detected at this point?
- How many brutes have been detected at this point?
- How many viruses have been detected at this point?
- At what time was the most recent Virus detected?
- At what time was the most recent Brute detected?
- How many new users have registered since the last virus was detected?
- How many users have logged in since the last brute was detected?
- How many user messages have been sent to other users since the last worm was detected?
- At what time was the most recent worm detected?
- How many users have logged out since the last brute was detected?