3-23-2018

Estimating Defensive Cyber Operator Decision Confidence

Markus M. Borneman

ESTIMATING DEFENSIVE CYBER OPERATOR DECISION CONFIDENCE

THESIS

Markus M. Borneman, Captain, USAF

AFIT-ENG-MS-18-M-013

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

DISTRIBUTION STATEMENT A.
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government. This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

AFIT-ENG-MS-18-M-013

ESTIMATING DEFENSIVE CYBER OPERATOR DECISION CONFIDENCE

THESIS

Presented to the Faculty

Department of Electrical and Computer Engineering

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Cyber Operations

Markus M. Borneman, BS

Captain, USAF

March 2018

DISTRIBUTION STATEMENT A.
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AFIT-ENG-MS-18-M-013

ESTIMATING DEFENSIVE CYBER OPERATOR DECISION CONFIDENCE

Markus M. Borneman, BS

Captain, USAF

Committee Membership:

Dr. Brett J. Borghetti
Chair

Dr. Gregory J. Funke
Member

Dr. Kristen K. Liggett
Member

AFIT-ENG-MS-18-M-013

Abstract

As technology continues to advance the domain of cyber defense, signature and heuristic detection mechanisms continue to require human operators to make judgements about the correctness of machine decisions. Human cyber defense operators rely on their experience, expertise, and understanding of network security, when conducting cyber-based investigations, in order to detect and respond to cyber alerts. Ever growing quantities of cyber alerts and network traffic, coupled with systemic manpower issues, mean no one has the time to review or change decisions made by operators. Since these cyber alert decisions ultimately do not get reviewed again, an inaccurate decision could cause grave damage to the network and host systems. The Cyber Intruder Alert Testbed (CIAT), a synthetic task environment (STE), was expanded to include investigative pattern of behavior monitoring and confidence reporting capabilities [1]. By analyzing the behavior and confidence of participants while they conducted cyber-based investigations, this research was able to identify a mapping between investigative patterns of behavior and decision confidence. The total time spent on a decision, the time spent using different investigative tools, and total number of tool transitions, were all factors which influenced the reported confidence of participants when conducting cyber-based investigations.

Acknowledgments

I would like to express my sincere gratitude to my faculty advisor, Dr. Brett Borghetti, for his guidance and support throughout the course of this thesis effort. The insight and experience was certainly appreciated. I would, also, like to thank Dr. Gregory Funke, Dr. Kristen Liggett, and the 711 Human Performance Wing for their support.

Markus M. Borneman

Table of Contents

List of Figures

Page

List of Tables

Page

ESTIMATING DEFENSIVE CYBER OPERATOR DECISION CONFIDENCE

I. Introduction

1.1 General Issue/Motivation

Cyber operators, the colloquial term for humans engaged in cyber defense activities on the Air Force Enterprise Network, are tasked with making judgements and decisions about the correctness of machine decisions, including how to remedy network or host-based threats. For the purposes of this research, only network-based threats are of importance, as they must be delivered over a monitored and defended network to a target machine. The Cyber Intruder Alert Testbed (CIAT) synthetic task environment (STE) mimics a real-world security information and event management (SIEM) system, allowing for cyber-based alerts to be displayed and analyzed by the user [1]. Human operators interact with the system by identifying, validating, and tracking network-based security threats to the network. Thus, operations and training require humans to excel in the understanding of their task, such that they can make informed and correct decisions even if the tools and sensors may not always be correct. Typically it takes 6-12 months to become comfortable and confident on these weapons systems based on personal subjective levels of analysis. Once a human is certified on a system, they must maintain currency and proficiency on a month-to-month basis with yearly evaluations to ensure they are properly prepared to handle their job requirements. A "one size fits all" method of training is not necessarily tailored to individual operator’s areas needing improvement, so those lacking in experience or confidence in select areas may or may not receive the most effective training regimen. It is impractical, if not impossible, to prepare these cyber operators for every task or scenario they may encounter, thus they will have to rely on their own independent reasoning and problem solving skills based on their training and experience. The investigative process for each alert is dependent on the information available and the operator's expertise and experience. These investigations ultimately lead to the operator making a decision with some level of confidence. The level of decision confidence may be measured using behavioral indicators, subjective indicators, and even electrophysiological indicators. Decision confidence, defined by Insabato et al., is the feeling of having done something correctly or incorrectly, which is an important aspect of subject experience during decision-making as this increases for correct decisions and decreases for error decisions [2]. With the ability to identify cyber operators in low confidence situations, they can be augmented with increased attentiveness by other cyber operators, which in effect would be a tailored and specific usage of quality control to improve operations. Additionally, these low confidence situations, if detectable, would allow for tailored training to remedy these otherwise lower confidence situations. In worst case scenarios, trends may be established to identify when a cyber operator is in their normal state of decision confidence, be it normally high or low, and flag or alert the operator to decisions made outside their normal threshold.

1.2 Problem Statement

By observing the behavior and estimating the decision confidence of human subjects while they make decisions in a cyber-defense task environment, we may be able to identify when an operator needs assistance. Assistance may then be provided in the form of investigation review, training, and operational work using the new decision confidence information. Using machine learning-based behavior pattern classifications, we may be able to map changes in confidence levels to address variations in tool compatibility, analyst skillsets or experience, and workload.

Previous human research in decision confidence has primarily focused on interview and survey type experiments where participants were asked to self-assess their decision confidence. The objective is to expand on past human research into decision confidence, specifically in the domain of cyber-defense, by observing the operator's investigative patterns of behavior. Decision confidence will be estimated using decision performance measures such as time to decision, accuracy/correctness, and the participant's self-reported confidence. Physiological data will be recorded from electroencephalogram (EEG), electrocardiogram (ECG), and electrooculography (EOG) equipment, for association with mental and physical behaviors related to decision confidence. The behavior observed while participants investigate cyber-alerts in the CIAT STE will carry over to real-world cyber-based alert investigations, as the environment and tools resemble what cyber-defense analysts would use. Understanding the investigative patterns of behavior and estimating decision confidence will lead to a better understanding of how decisions are made.

1.3 Research Questions/Hypotheses

RQ1: What does the pattern of behavior, exhibited while investigating an event, tell us about operator confidence in the formulation of a decision?

Hypothesis: Investigative behavior has an effect on operator confidence.

RQ2: What investigative and evidence collection techniques does the operator use to make a decision?

Hypothesis: Differences in decision confidence will be evident in both patterns of investigative behavior and differences in the operator's electrophysiology.

RQ3: What are the behavior patterns associated with a confident decision?

Hypothesis: Operator behavior patterns associated with higher confidence will be reflected in faster decision-making and quantifiable electrophysiological metrics.

RQ4: What are the behavior patterns associated with a correct and confident decision?

Hypothesis: Operator behavior patterns associated with high confidence and correct decision selection, will exhibit electrophysiological metrics which are quantifiably different from decisions made in lower confidence.

1.4 Research Focus

The focus of this research is to estimate decision confidence during a cyber defense investigation. While investigating the effect of alert difficulty on the investigative patterns of behavior, decision confidence will be determined by mapping the patterns of behavior to self-reported factors and recorded physiological information. If the investigative workflow and behavior patterns can be mapped to known electrophysiological indicators of the formulation of a decision and the associated decision confidence, then the more readily available non-physiological measurements can be used to estimate human decision confidence in order to provide feedback for efficiency and performance enhancement.

1.5 Methodology

The methodology is composed of three distinct parts, because each portion allows for observations which can then be identified and correlated across the other parts in order to model the formulation of a decision. Collecting self-reported confidence scores for each investigation is the easiest to obtain and review, thus it will be the first focus.

The self-reported confidence scoring was done by presenting the participant's with Likert scales. Likert-type scales are frequently used in medical education research and clinical studies to measure self-reported data such as anxiety or self-confidence [3], [4]. The typical Likert scale is a 5- or 7- point ordinal scale used by respondents to rate the degree to which they agree or disagree with a statement [3], [5]. The reason a Likert-type scale was selected was to benefit from the ordinal scale. A 3-point ordinal scale of "not confident", "somewhat confident", and "very confident" was created. An ordinal scale allowed for distinct answer choices, but made comparing raw values difficult since the scale is not necessarily equidistant. The Likert-type scale used in this experiment was set to a scale of 0-100 values, using 3 subjective anchor words. The CIAT STE would display the Likert scale to the participant during each cyber-alert investigation. After a certain amount of alerts, the participant would be asked to rank the alerts, in order from top-to-bottom, as highest-to-lowest confidence, respectively. This ordering task forces any ties to be broken, should any of the alerts have identical Likert scale values. The numeric confidence scores are not available to the participant while they complete this ordering task. Since the participants will not have access to their confidence scores, they will have to rely on their notes and short-term memory. The ordering task acts as a validation control, ensuring the participants understood the confidence scores and the relation of alerts when comparing them to each other.

The second method is a behavioral analysis, made possible by observing and analyzing the workflow of each investigative and decision-making choice. Behavioral analysis involves recording the timing and value of every mouse click and keyboard input. By cataloging and reconstructing this data, a workflow and timeline can be generated for each participant. This workflow will replay the investigation of every alert, including every tool accessed and how long each tool was accessed. In addition, the recorded workflow can identify when tools were skipped or avoided. Skipping or avoiding tools could suggest learning effects or mistakes, dependent on other behavioral features. The intent of the behavioral analysis is to determine whether certain actions cause changes to confidence when reviewing the accuracy of alerts.

The final method to investigate is the relationships of the previous two methods with the participant's physiological measurements. Various sensors will record the electrophysiological activity of each participant as they complete the cyber-alert investigations. Similar to patterns of behavior, the physiology of each participant will allow for an analysis of the evidence accumulation process when conducting the cyber-alert investigations. Additionally, certain physiological patterns manifest during decision-making, dependent on confidence [2].

1.6 Assumptions and Limitations

1.6.1 Assumptions

All of these methods are susceptible to the learning effect. The learning effect explains accelerated improvement to new or unfamiliar tasks, which would otherwise be negligible if someone was experienced with a task. All participants, especially those without any formal cyber security experience, will be learning and improving their cyber alert investigation process during the experiment. Because human subjects continuously absorb information about their surroundings, they cannot be expected to treat each alert as independent. Tool and process familiarity must be accounted for outside of the experiment, in order to minimize the effects of workflow improvement during the experiment. Therefore, a 2-hour training phase was created to reduce the learning effect for participants. The training phase involved interface and tool familiarization, as well as a hands-on tutorial with a step-by-step investigation walkthrough using several example alerts. The training phase also included a complete round of alerts, where participants were allowed to practice without assistance. The assumption is that the participant will know enough about how to conduct a cyber-based investigation and make a decision based on the evidence they collect. The 2-hour training phase occurred prior to the experiment. Participant selection assumed that participants would understand how to operate a computer, and be willing to undergo training in order to understand and practice the cyber-alert investigative process.

1.6.2 Limitations

Participants were recruited from students, faculty, and staff of the Air Force Institute of Technology. Because the backgrounds of participants in this experiment may be different from the background of a typical cyber operator, it may be necessary to conduct additional experiments to validate whether the findings are similar for participants with a background in cyber security who are more familiar with cyber security tools and concepts.

The CIAT STE uses one computer screen, meaning that all of the options and actions available to the participant were presented all at once. The STE differs from real-world scenarios and situations, in that all of the tools are available in one display window and in one location. Real-world systems typically require multiple tools, systems, and computer monitors in order to access relevant information while conducting a cyber-based investigation. In order to eliminate additional timing factors, such as window switching between tools, the design choice of one main window with all tools and alert information was made. Because the tools and interface were only on one screen, as the experiment's results are limited to environments with similar limitations. The modular nature of this STE allows for relatively easy changes to be made to mimic other capabilities or tools, if that becomes the focus of future research.

1.7 Contributions

This study refines other work on cyber decision-making and decision confidence, with the inclusion of physiological measurements. In addition, the cyber-defense focus on the patterns of behavior provides empirical evidence of otherwise subjective measurements for decision-making and decision confidence.

Although more analysis is necessary, especially in the realm of EEG, the patterns of behavior deduced from the trove of user-provided mouse clicks and keyboard input suggests that certain activity is repeated throughout the investigative process, up to and including when a decision is made. The usage of tools, and the order at which they were used, provides key insight into the workflow and process each participant uses when gathering information to make an informed decision. The participant’s tendency to alternate between tools, time on a tool, and creation of notes indicates a degree of confidence which may be isolated and compared between participants and across one participant’s completion of 30 alerts. Furthermore, consistency in the participant’s subjective decision confidence and the experiment’s estimated alert difficulty, as well as the average time to complete each investigation and selection of a decision, enables various data features to be analyzed and compared across participants, ensuring the consistency and validity of the intended alert difficulty.

1.8 Preview

The rest of the document will be divided into four chapters. In Chapter II, the Literature Review will define several definitions and concepts which led to the formulation of this research. The Literature Review identifies gaps in understanding the investigative decision-making process and decision confidence in cyber defense. Chapter III greatly expands upon the methodology and intricacies specific to the setup and creation of the experiment. Chapter IV will describe the compilation and analysis of the data recorded from the experiment, and present the results. Finally, Chapter V will conclude with a discussion of the results and recommendations for future work.

II. Literature Review

2.1 Chapter Overview

The purpose of this chapter is to provide a synopsis of the known research in the area of cyber defense decision-making. Computer science concepts, relationships, and psychology ideas relevant to the pursuit of this research, will be defined. The major themes of decision-making research are confidence, certitude, and self-confidence. With an understanding of the previously completed work in the realm of cyber defense, the reasons behind the pursuit of researching cyber defense decision confidence should become clear.

2.2 Definitions, Themes, and Concepts

The Merriam-Webster Dictionary defines confidence as a feeling or consciousness of one’s powers or of reliance on one’s circumstances [6]. In the field of cyber defense, analysts and operators rely on their computer systems and skill, in order to make decisions. These decisions may be confined by information availability and the time remaining to make a decision. The feeling of confidence is subjective. Feelings cannot accurately be captured within the bounds of numerical measurements, and feelings can change spontaneously.

Decision confidence describes how confident a person feels when considering how they feel about their decision. Confidence is difficult to measure if the information available to make the decision, or if the scale used to represent the measuring of confidence, is misunderstood. Therefore a measuring scale for decision-making tasks, which can record confidence, is required. This scale is known as a decision self-efficacy scale [7].

Self-efficacy is confidence in one's ability to achieve an intended result [8]. The process leading to the result must be scrutinized for validity, as various effects of bias can corrupt the decision-making process. Bias describes a person's tendency to view something from a particular perspective. Biases may prevent or impede a person from being objective and impartial [9].

Several key biases, which this research needs to be aware of during the experimentation process, will be highlighted in this chapter. Biases, with respect to a participant's decision-making could lead to greatly skewed results. For example, the way in which information is presented to participants could prime or bias them towards this information should they come across it again later during the experiment. Methods for controlling these biases will be expounded upon in the Chapter III, Methodology.

Pfleeger separates biases as status quo, framing effects, optimism, control, confirmation, and the endowment effect [9]. Status quo is simply the resistance of an individual to change their behavior without a reason or incentive. Feedback and repercussions for actions can be used to address and reduce status quo biases. Framing effects bias involves the presentation or manner in which information is presented. The efficacy of a trial can be framed in terms of gains, rather than losses, or by appealing to particular characteristics. This method of information presentation, e.g. ordering or words used, can influence and dramatically affect the decision. Similar to framing effects, priming or anchoring also leads to biases, as information presented earlier is easier to rely on than information presented later.

Optimism bias is the belief that a person will perform or be presented with a higher likelihood of positive events. This bias is an over or under estimation of the likelihood of positive and negative events occurring. Optimism bias may, for example, induce people to ignore preventive care measures, such as patching software, because they believe they are unlikely to be affected [9]. Similar to the optimism bias, control bias is the tendency of people to believe they can control or influence outcomes they clearly cannot.

Confirmation bias is the tendency of favoring or interpreting information based on previously held “confirmed” beliefs. When looking at a situation, a person affected by confirmation bias will tend to place a higher emphasis on confirming and aligning with their previously held beliefs than reviewing the situation across all facets. This short-circuit of the decision-making process can become evident due in part to the speed at which a decision is made, or by creating situations or presenting evidence in a way to catch those who do not review all pertinent areas of the information. The endowment effect bias describes the fact that people usually place a higher value on objects they own than objects they do not own [9]. This may lead people to react more strongly to a loss than to a gain. For example, when an action is expressed as a loss of privacy, rather than a gain in capability, people tend to act negatively.

For the pilot community, Holland and Freeman explored mishaps involving the loss of situation awareness of F-16 pilots, and deemed the occurrences due to channelized attention [10]. Channelized attention is similar to a confirmation bias, in that the human subject’s focus may make them miss or completely dismiss other relevant information due to their preconceived notion or fixation on other elements of information. Cyber defense and piloting aircraft can involve much of the same sorts of tasks, such as accurately gauging and maintaining situation awareness of the environment. Graphic user interface (GUI) construction, specifically those involving various colors that users must react to, can lead to channelized attention. Users may focus the majority of their time and effort on visual information which is color coded by priority, for example, and disregard relevant but different colored information.

Slight nuances, such as the ordering of cyber alerts, can illicit different behaviors and responses. Network events are typically ordered from newest-to-oldest, due to how host, network, and intrusion detection systems detect and report traffic. Investigating traffic out of order can hide malicious payloads or cause traffic to look benign. This can be dangerous in the cyber defense environment, because the standardization of protocols and traffic may make many things look almost identical. An awareness of participant's reliance on past performance or behavior indicators for decision-making, especially when they are new to a task, is of valid concern when reviewing participant behavior.

Outside of biases, there are other concepts that influence decision confidence, such as choice certainty, cues to action, and situation awareness. Decisions are usually accompanied by a degree of certainty or confidence, which reflects a graded belief about the likelihood of different outcomes [11]. Choice certainty facilitates adaptive regulation of behavior by furnishing a basis for learning from outcome, and supports decision-making in complex environments where subsequent decisions depend on the predicted outcome of recent decisions before the actual consequences are known [12]. Cues to action are events that trigger or remind an individual to take an action they either forgot or were not originally intending to take, such as a reminder about the return date for a DVD to a service like Redbox, or an overdue library book. By using cues to action, one can influence a user to make a decision or ignore the new information [13].

Situation awareness is a broad theme that can be applied in the cyber defense domain [14]. Without situation awareness, it can be difficult to decide on a course of action. If network defense actions are required, a lack of situation awareness could drastically limit or impede the necessary network actions from taking place to contain a threat and maintain services. Situation awareness in cyber defense ordinarily requires not just an understanding of the local machines and environment, but the context of machines geographically separated and isolated from the defender. This makes it difficult to assess the problem, and this detachment from the real-world environment affects the perceived risk-versus-reward for the operator, as they at least rarely have physical repercussions to worry about due to a decision.

Tyworth and his colleagues assert that the greater research community tends to focus analytical attention on new technologies instead of understanding and improving the underlying socio-cognitive work performed by human cyber security professionals [14]. Their solution argues for distributing situation awareness across human and technological agents, thereby re-focusing and enhancing the human-centric approach needed in cyber defense analysis. Typically, the human resource is the hardest to recruit, train, and maintain, thus technological solutions seem more valuable in the short-term to cover these gaps by producing rapid and consistent data analysis. Yet, a human is involved in all cases, either as the creator of the hardware and software solution or in-the-loop deciding whether to follow the guidance of the technology. Humans and technology end up not working in tandem, as the technology is still reliant on the human to program or tell it how to carry out the analysis task. Technology enables the human to create or become aware of a situation, through the use of visual or other sensory cues. Endsley's experiments on measuring cognitive perspective of human operator's understanding of an environment at a particular point in time, artificially controlled through the use of freeze-probe measurements techniques, brought about a well-valued theory of situation awareness in dynamic systems [15]. Tyworth suggests that the situation awareness technique proposed by Endsley, is unable to distinguish between situation awareness based on knowledge and experience of the operator or from the underlying technologies which support the insight alone.

Cyber defense analysts struggle with low situation awareness due in part to the speed and rigor they are required to categorize incoming and outgoing traffic. These analysts may not know why something is or is not worth paying special attention to, because of their limited situation awareness. This situation awareness gap is due in part to policy, but mostly due to the vastness of the threat landscape which analysts are expected to patrol. Cyber defense organizations are typically structured into separate teams or tiers, with increasing levels providing further insight into the network through tools and capabilities. The cyber defense analyst in this research is typically located at the lowest level in a cyber organization, where they monitor and react to near real-time network alerts. This lowest level is the first, and sometimes only, chance to identify and react to potentially malicious network activity. The goal of this new research is to identify when and how decision confidence plays a role in the formation of decisions by human analysts, such that the correct areas can be focused on for improvement.

The last concept, taken from behavioral science literature, is that recognition is significantly easier than recall [9]. Cyber defenders are tasked with rapidly cataloging, identifying, and responding to potentially malicious traffic. This visually dominated work involves reading and surveying complex text, pictures, and numbers. Recognition tasks should increase confidence, as there is little investigating needing to be done outside of recalling a situation. Therefore, information representation must be uniform throughout the interface in order to produce consistent situation awareness.

2.3 Decision-Making and Behavior

Several papers proposed strategies and models for investigating human decision-making. Whereas one strategy involved comparing and contrasting two popular theories of decision-making strategies, notably Long Term Working Memory (LTWM) and Take-The-First (TTF), a significant exception was a paper which recommended the need to account for and test whether evidence was reliable, as conjecture shows this can affect decision-making and confidence [16], [17]. Yeung and Summerfield explore the "post-decisional locus model" and the findings on how decisions occur and what makes people "change their mind" once a decision is cast. The drift-diffusion model illustrates decisions as an accumulation of evidence over a period of time, until either one of two thresholds, θ or – θ, is met or exceeded. By including the metacognitive process known as error monitoring, humans are able to adapt both their short- and long-term actions based on outcomes observed prior to their next decision. Mapping this to the drift-diffusion model, future outcomes based on accumulation of evidence to the decision point of one decision may lead a human to either maintain the decision into future situations, based no additional information, or instead opt to choose the other decision based on the accumulation of time and evidence. Thus, observation and modeling of the underlying biological components of the brain is necessary, due to the subjective and malleable decision-making process of humans. Error monitoring seems similar, if not identical, to the process of learning, which is of critical importance in human subject experiments, as it is one of the many biases from which the experimental design intends to negate or minimize the effects. Additionally, when coupled with trust, error monitoring relies on accurate information gathering, which certain tools and systems in the experiment could be modulated to either accurately or inaccurately provide feedback on what course of action to take. Furthermore, a lack of feedback may also have the potential to affect the way in which error monitoring is carried out by the human.

The two decision-making strategies proposed by Belling et al., LTWM and TTF, are likewise of importance due to one of the biggest assumptions of this research, namely the recruitment of human subjects who are not necessarily cyber defense experts [16]. The LTWM theory suggests that experts rely on stored knowledge when placed in a new environment, and the TTF heuristic relies on taking the first action that comes to mind. Their experiment involved several trials with human-subjects, to determine whether time and the number of options generated by participants affected participant accuracy in prediction and response trials. The procedure involved recreational-level soccer players viewing video clips of live soccer matches. The players were tasked to determine the next course of action of the recorded player, when the clip ended or occluded at a critical decision point. In the trials involving prediction, participants illustrated options for any combination of players, actions, movements, and ball position, under the focus of being a defender. In the response trials, the participant rated how likely they were to pursue each generated course of action. Additionally, only half of the trials involved time constraints. The results challenged the hypothesis on whether time constraints lead to increased use of TTF strategies. Contrary to their expectations, LTWM strategies were employed when participants were under time constraints.

Ward and colleagues reviewed decision-making strategies in various other disciplines, conducting experiments in competitive chess gameplay [18]. The competitive chess gameplay results pointed to no evidence of performance differences, under time-constraints. In contrast, less skilled chess players showed a significant performance decrement. Extending this research to the cyber domain, future research could compare whether skilled cyber defense analysts maintain effectiveness given varying degrees of time-constraints.

Because real-world cyber defense analysts must make rapid and accurate decisions in order to not become inundated by the volume of alerts, imposing time constraints on decision-making could identify when decisions become hampered by limited time. Likewise, the number of alerts presented to operators is tunable based on the broadening or constricting the signature base matching and heuristic settings of network alert sensors. Flooding operators with alerts and requiring a set amount of decision actions to be taken over a period of time could also affect the decision-making process, but this approach would not align with the results featured from the competitive chess players study.

2.4 Measuring Decision Confidence and Decision-Making

In order to be able to estimate decision confidence, operator behavior has to be observed. Behavioral observation allows for passive analysis, which does not cause interruptions such as those experienced when compiling self-reported measurements, or the operator to don cumbersome equipment such as those needed for measuring physiological signals. This research augments behavioral analysis with self-reporting by participants and physiological measurements. The physiological measurements are included as they benefit and enhance the behavioral observations, and because it allows a mapping between the physical and mental parts of the body during decision-making. With this combined understanding of the underlying decision-making process, the behavior observation can then be the focus of monitoring and reacting, as this can be passively observed with minimal evasiveness in a cyber-based environment.

2.4.1 Self-Assessment and Reporting

Survey-question based human analysis dominates psychological literature and the vast majority of cyber effects studies involving human subjects [19]–[24]. Survey-questions are reliant on various factors, including the subject's experience and willingness to honestly self-assess. The timing of the survey questions is the single most influential variable in effecting the outcome. A subject's perception is ever-changing during an experiment, therefore the timing of a survey question could be heavily influenced by when and how interruptive a question is. As discussed by He, et al., comparing and understanding surveys for cross-study comparisons proved very difficult due to inconsistent, confusing, or misunderstood measurements [25]. In this case, the research attempted to not only compare and group the various definitions used by each study, but also the dimensions of the questioning, further supporting their argument that cross-study comparisons were a difficult undertaking. Survey questioning should occur in distinct phases during an experiment: pre-, intra-, and post-trial. Using surveys during only one or two of these critical phases would bear insufficient holistic information, which is critical in maintaining the consistency in the whole experiment allowing for normalizing of data, and should ease identification of outliers or inconsistent users. Additionally, any result can be questioned in a follow-up interview to further delineate and quantify the results [24].

Several assumptions must be presented. A Likert scale was chosen, as it provided for a method to garner ordinal feedback from participants. Likert scales can be made in a variety of different configurations, but the most common tend to be 5- or 7- point scales. Questionnaires involving more than 5 options were seen as too difficult to accurately align with, by participants. For instance, one study involved a 100-point Likert scale with 10 point increments, effectively making it a 10 option scale, but this was no more effective than asking respondents to rate on a scale of 1 to 5 for how strongly they agreed or disagreed with a presented choice [26], [27]. Likert scales are seen as a way of forcing a choice on the responder, who may not have a definitive answer, but is forced to answer anyway [28]. The NASA Task Load Index (NASA-TLX) questionnaire format closely resembles the Likert scale, although respondents are given the choice of making an in-between measurement, such as a decimal value [29]. Additionally, emoticons were not advised, but an example sentence or example situation for each option of the Likert scale is strongly encouraged in order to help establish the scoring mindset in the responder [24].

Humans are bad at self-assessment, mainly because of various biases which contribute to the inability to objectively evaluate skill [30], [31]. Compeau and Higgins postulated that extraneous factors may bias participant responses due to the nature of the event being measured [32], [33]. Repeated questioning of the same information may lead to fatigue, or disengagement, affecting the authenticity and validity of response.

Another limitation of survey-based questionnaires are the questions themselves. The ordering of questions can have an effect on the outcome of later questions. Since the training, and experiment conditions, and questions will be identical for all of the participants ordering will not play a role when comparing across participants. A demographic and computer-usage survey will be administered in order to determine if frequent usage of computers in participant's lives and job influenced their ability to perform the cyber defense based task. The computer-usage survey will aid in identifying trends, or the need for calibration when comparing reported confidence.

Another avenue of procuring self-assessments is through interviews. Interviews allow for the assessor to focus on and examine qualitative features that a self-metered survey will not accurately record. For example, the decision time and accuracy of a decision can be examined, through questioning and ascertaining the exact reasoning for a decision or behavior, if the responder is conscious of the action in question. Unlike the survey methodology that will include questioning during pre-, intra-, and post-trial, the interviews work best before and after the trials or the entire experiment. This allows for minimal distraction, but requires the assessor to maintain notes or logs of the responder's actions so that they can be discussed by referencing if necessary. Structured interview questions, concerned with analyzing the subject's time, accuracy, and threshold for decision-making, leads to a better understanding of their capabilities and expertise [9], [34], [35]. Unstructured, loosely controlled question and answer interviews can make it difficult to conduct cross-study comparisons.

Assuming the structure of the interview can be repeatable across subjects, the compatibility of the results during comparisons and tabulation should be straightforward. Along with the survey information, the interviews will aid in stratifying and separating situations in which the same values or ranks were provided. Since a five point Likert scale is recommended for surveys, further granularity can only be achieved through interviewing the responder about their answers and comparing trials. Typically, interviews are seen as more favorable by participants, since they allow for more flexibility, compared to the strict numerical representation of their answers in a survey-based questionnaire, including the affordance to explain why or how a certain response is given. Observations and logs will help allow the assessor and responder to share situation awareness of the experiment, allowing for easy recall and play-by-play analysis of decision points and junctions.

As was previously mentioned, consistency can be difficult to guarantee if the structure and rigor of the interviews is not maintained. Additionally, only one interview during each phase should be the limit, as continuous subject interviewing, similar to repeat survey questioning, will lead to frustration and fatigue in the participant. Lastly, another limitation is the timing of the interview. An interview following a trial or experiment should be conducted as soon as possible, as to take advantage of the short-term memory of the participant and to question decisions and actions while they are still fresh on the minds of the participants.

Since most of the benefits attributed to interviews can be contained in a survey questionnaire constructed to allow the participant to rank order their selections during the task and compare their decision based on groupings, an actual interview will be relegated to future experiments whereby it is more feasible or practical to illicit feedback in this manner.

2.4.2 Behavioral Analysis

Workflow and process observation are the crux of this experimental analysis. Pfleeger identified the behavioral aspects of security, as the concept of leveraging what is known about people and their perceptions in order to provide more effective security [9]. Behavioral science literature generally supports and demonstrates that recognition is significantly easier than recall, possibly explaining why LTWM seemed to eclipse TTF in experimentation [16]. Biases also play a significant influence in human behavior, illustrated by the numerous constraints and assumptions imposed on this experiment in the methodology section. The psychology behind these biases help explain why technological enhancements may not always provide the expected result or effect.

By using both subjective and objective metrics, the state of the human can be estimated. Human cognition is measured through physiological measurements, but associating the subjective measurements taken from investigating alerts may allow for an understanding of how decision confidence and decision-making affect each another. Knowing what is taking place cognitively, by way of physiological measurements, and associating this with the subjective correlation of the alerts, should allow for an understanding of how decision confidence influences and determines the decision-making process. With a greater understanding of the underlying mechanics of decision confidence, the ability to provide near-real-time help for operators in low-confidence decision situations is possible. Likewise, prioritizing the review of decisions made under low confidence situations would allow for quality assurance mechanisms to aid in the verification and checking of decisions made under subpar standards. Lastly, by feeding this information back into training and the user interface design, the focus can be placed on the areas and types of decisions most often associated with low confidence.

2.4.3 Physiological Measurements

Electrophysiological measurements are recorded by the observer and are non-self-reported, objective measurements of brain, heart, and muscle activity as well as other body states. Coupled with self-report based results taken from surveys, electrophysiological measurements such as EEG and ECG provide a general observation of the physical and mental actions taken by a participant [36], [37].

2.4.3.1 Electroencephalography

Lateral Intraparietal Cortex (LIP) neuron measurements have been shown to represent the accumulation of evidence by subjects, leading to the formation of decisions and degrees of certainty [36]. With an EEG measuring apparatus, brain activity can be monitored during the evidence accumulation phase, through the decision-making phase, and into post decision-making phases. This capability will augment our understanding of the stresses experienced by participants. In addition, using the participant’s response times, coupled with network traffic and cognitive workload, it becomes possible to understand how decisions are formulated from the decision-making process.

2.4.3.2 Electrocardiography

ECG will aid in identifying stress points or workload strains of the operator. Biometric monitoring involving EEG and ECG enables human performance-based attributes involving physical and mental manifestations of mobility and thought to be coupled with mental self-assessments inherent in self-reported measurements in order to support otherwise purely subjective-based measurements. Whereas surveys are a subjective assessment by the human subject, the objectivity of the biometric measurements is directly characterized by the subconscious mechanics of the human body. Biometric measurement analysis may be coupled with the subjective measurements to determine and characterize what is occurring in the mind and body of the human participant.

2.4.3.3 Electrooculography

Lastly, measuring eye movement and fixation is another non-self-reported element that monitors the subject's visual field and to what degree they are attention-switching. Visual recognition utilizes the same aforementioned LIP neurons in measuring the formation of decision confidence and degrees of certainty [36]. Cyber defense analysts and operators conduct a visually focused examination of Intrusion Detection System (IDS) alerts, which involves recognition and memory recall. Attentiveness and situation awareness require focused and directed responses to visual stimuli. Visual stimuli in computer programs are typically presented to the user through graphical user interfaces. These interfaces may lessen or enhance the burden of a user attempting to gain situation awareness. Overloading operator cognitive resources causes performance decrement [38].

This previous work investigated modifying trust in a cyber security tool, by increasing and decreasing the accuracy and reliability of the tool. Tool accuracy was measured by how much information was displayed about an alert. When the tool was more accurate the screen was filled with more information, showing the user what was being detected and acted upon, but this limited and inversely affected the performance of the human user in charge of agreeing or disagreeing with the computers analysis. Applying this work to confidence measurements, the focus on information that is pertinent and relevant to making a decision may not always lead to the most appropriate or correct decision. Thus, it is important to follow the process of information acquisition through the primary means of information presentation in cyber defense, which is visually through an aggregator or correlation platform that is fed alerts from IDS devices. Focus on a part of the screen and a tool, is supporting evidence of fixation and may hint at a cue to action, prior to the activation of the subject's fine motor skills that are the result of some decision.

Eye-tracking may enable the measurement of tool usage, prioritization of information, and other cognitive attributes related to identifying cyber investigative workflow [27]. Coupled with mouse movements, graphical user interface window focus, and keyboard input, eye-tracking provides insight into workflow, but not decision confidence. This methodology shows what information was reviewed and for how long, based on fixation, but with the assumption that the interface is simple enough in order to differentiate between different graphic elements and windows.

One of the biggest limitations to eye-tracking data collection is that it is only valid inside the context of the training environment, i.e. what can be measured from the user looking at the computer screen and not outside the bounds of the computer screen [39].

Additionally, as mentioned earlier, eye movements map to the participant's accumulation of evidence and leads to a decision. Thus, this is only another data point to use when analyzing the behavior of the participant when conducting an investigation, which needs to fit into the broader analysis for a holistic view of how and when an operator makes a decision.

2.5 Conclusions

In summary, one of the biggest challenges evident from the literature review is the need to augment the subjective, user-provided information from the survey comparisons with the objective, physiological data. Bridging these two paradigms will provide a greater understanding of the actions humans take when given information and constraints in which to make a decision, as well as objective performance data. The biggest merit to survey questionnaires is the relative ease in performing measurements, but their consistency and validity can vary as the human participants become fatigued - because of the duration of the task or because of frequent surveying - which can have a negative effect on task attention. EEG, ECG, and other electrophysiological measurements are novel approaches, extended from the medical and psychology domain, to review and analyze reasoning and decision-making. Although they may prove to be impractical outside of baseline tool configuration and workflow analysis, the operator's decision confidence expresses whether information presentation, user skillset, and physiological effects have any measurable effect on job performance.

Past research has shown that humans are better at resolving ambiguity and providing contextual mission relevant information to automated security systems, rather than handling large amounts of information and weeding out false alerts, which is not the way humans are often employed in operational units [40]. Trusting the system through accuracy, timeliness, and consistency, allows for human operators to focus their efforts on review and analysis of ambiguous decisions. This may lead to benefits such an improved culling of the seemingly endless alerts present in current cyber defense aggregation and correlation platforms, and an improved prioritization of alerts and situations outside the norm that cause operators to lack confidence in their assessments.

Finally, as was pointed out by the various biases, the design of the experiment and the analysis of the participant data will need to be account for the effects of these biases, as they would affect the findings. The biases which can be controlled will be identified in the methodology chapter.

III. Methodology

3.1 Chapter Overview

The purpose of this chapter is to establishing the research questions and outline how the experiment will be carried out. The various factors and variables which will be changed, as well as recorded for analysis will be defined. The makeup of the participants, the required assumptions, and the analysis plan will also be covered in this chapter. Additionally, the CIAT STE will be showcased, with examples and pictures of how the tool was configured for the participants.

3.2 Background

Cyber defensive operations continue to be human-intensive activity. While many researchers try to improve detection mechanisms, ultimately human operators will make judgements about the correctness of the machine decisions and how to resolve the alerts. Thus, research in the human component of decision-making during cyber analysis remains vital. This experiment supports research which seeks to identify and characterize the influences of decision confidence on information gathering and investigative processing as it relates to the job of an Air Force Cyber Defense (ACD) Operator.

The study investigates decision-confidence relationships between self-reported confidence, behavior, and psychophysiological signals collected when a participant makes a decision – specifically in the domain of cybersecurity defense. By modeling the relationships between self-reported confidence, physiological measurements, and observed data from operators conducting decisions on cyber network traffic samples, this study investigates the relationships between behavior patterns and decision confidence.

The study determines the key attributes and behaviors, exhibited by cyber defense operators, which affect the accuracy and decision confidence of cyber triage. Correlating the self-reported confidence, physiological measurements, and observed behaviors patterns from human subjects engaged in cyber triage of traffic samples should allow for an understanding which can be represented by model and pattern analysis.

EEG, ECG, and EOG signals will be collected and used to determine what techniques and behavior an operator uses to make a decision. Combined with decision accuracy and self-reported confidence results taken from the alert presentation and analysis software, electrophysiological measurements will provide another lens into of the physical and mental actions taken by a participant in order to analyze the associated behavior [1][2].

3.2.1 Research Questions

Using the cause and effect relationships for modeling decision confidence from observing behavior patterns, recording self-assessed confidence, and measuring physiological measurements, the goal is to identify factors which correlate with confidence.

Investigative Question 1: What does the pattern of behavior, exhibited while investigating an event, tell us about operator confidence in the formulation of a decision?

Hypothesis: Investigative behavior has an effect on operator confidence.

How a participant investigates an alert, identified through their pattern of investigative behavior, indicates how confident they are in their decision. Behavioral cues, such as repeat visits to certain tools or shorter time spent researching, may indicate a level of confidence related to investigations handled in a similar manner. For this experiment the operator will be asked to report their confidence after making each decision selection. By identifying the patterns of behavior for each investigation, an estimation of operator decision confidence can be inferred.

Investigative Question 2: What investigative and evidence collection techniques does an operator use to make a decision?

Hypothesis: Differences in decision confidence will be evident in both patterns of investigative behavior and differences in the operator's electrophysiology.

Survey-question based human analysis dominates psychological literature and the vast majority of cyber effects studies involving human subjects. In order to understand how an investigation occurs, it is prudent to observe the behavioral and psychological process, in order to identify patterns. An investigation workflow handout, see Appendix F, will be given to each participant during both the training and experiment. Even with an investigation workflow handout and the associated training day, participants may "cut corners" or rely on tools more than others, which may affect the reported confidence. These investigative behavior patterns will be used to determine when a cyber alert causes the participant to change their behavior to overcome the difficulties of investigating a more difficult alert.

Investigative Question 3: What are the behavior patterns associated with a confident decision?

Hypothesis: Operator behavior patterns associated with higher confidence will be reflected in faster decision-making and quantifiable electrophysiological metrics.

The degree of confidence in a decision provides a probabilistic assessment of the expected outcome. Higher confidence would assert that there is a higher probability of the decision being correct. It is generally thought that certainty is informed by a neural representation of evidence at the time of a decision [36]. Results have shown that decision certainty was inversely correlated with reaction times and directly correlated with motion strength, suggesting that speedy decisions are coincident with lower confidence [11]. The time to a decision and the associated behaviors which led to the formulation of the decision, are expected to have a ceiling or maximum set of actions which, being quantifiable, would allow for comparing between decisions made with a higher reported confidence.

Investigative Question 4: What are the behavior patterns associated with a correct and confident decision?

Hypothesis: Operator behavior patterns associated with high confidence and correct decision selection, will exhibit electrophysiological metrics which are quantifiably different from decisions made in lower confidence.

Experience, a trust of the tools, an understanding of presented information, and habitual work all play a role in improving the confidence of operators [41].

3.3 Experiment

Human subject performance studies on decision-making often rely on self-reported mechanisms, such as surveys and interviews – and rarely involve interpreting confidence from physiological measurements and behavior during the decision-making process [13], [21], [42]. This study intends to augment self-reported subjective results, by incorporating both behavioral and physiological measurements. The combination of self-reported results and physiological measurements will inform an understanding of decision-making behavior patterns. Through understanding how self-reported results and physiology correlate with behavior patterns, real-world operations could possibly be augmented by only observing human behavior. Behavior can be directly observed and correlated to decision confidence. Observing and analyzing human behavior is the only viable measuring technique during actual real-world cyber defense operations, as self-reported and physiological measuring would be impractical and cumbersome in environments where cyber defenders operate.

Physiological measures included EEG, ECG, and EOG signals. These measurements were recorded throughout the experiment with the intent to be mapped to the behavior and self-reported results, in order to better understand what lead to decision confidence in cyber defense operators.

3.3.1 Variables

3.3.1.1 Independent Variables

The independent variables, which will be manipulated during the experiment, are listed in Table 1. The variability of the difficulty for the alerts will allow for identification and correlation of purposeful low-confidence situations and situations where a higher-confidence should be achieved.

Table 1: Independent Variable Summary

Control variable	***Measurement precision***	***Proposed settings***	***Predicted effects***
Information Availability (categorical)	Amount of information in tools	[Low, High]	Less availability = lower confidence
Information Needed (categorical)	Amount of tools needed to review	[Low, High]	Less needed = higher confidence
Information Inconsistency (categorical)	Amount of conflicting information among tools	[Low, High]	Less inconsistency = higher confidence

Alert Difficulty was estimated based on the estimated information needed to make the correct decision, availability of information, and the consistency of available information. In order to create a consistent difficulty scale for the alerts, the three difficulty variables were setup to identify perceived difficulty from changing the proposed settings of low or high. Four levels of difficulty were created, in order to aid analysis.

The four types of difficulty include:

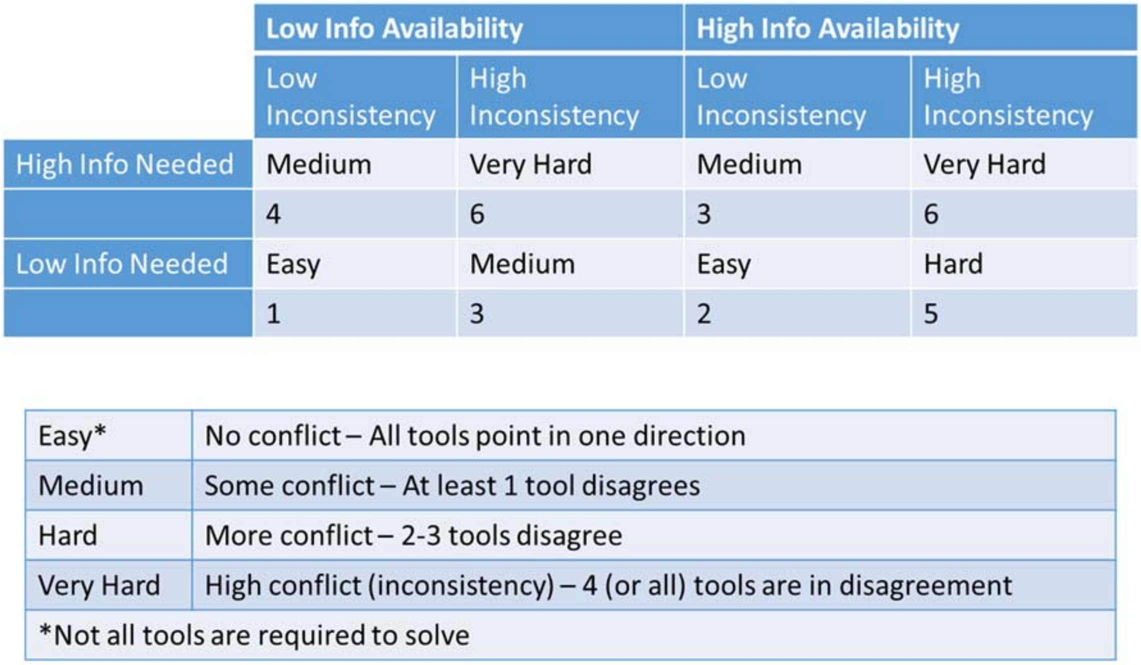
A. EASY

B. MEDIUM

C. HARD

D. VERY HARD

Eight possible alert situations were created using each combination of the three factors. Figure 1 illustrates the difficulties based on each possible setting of independent variables. The numerical values under each difficulty were determined by the subject matter expert (SME). Higher numerical values indicated increasing difficulty. The four difficulty levels were mapped to the six numerical scores.



	Low Info Availability		High Info Availability	
	Low Inconsistency	High Inconsistency	Low Inconsistency	High Inconsistency
High Info Needed	Medium	Very Hard	Medium	Very Hard
	4	6	3	6
Low Info Needed	Easy	Medium	Easy	Hard
	1	3	2	5

Easy*	No conflict – All tools point in one direction
Medium	Some conflict – At least 1 tool disagrees
Hard	More conflict – 2-3 tools disagree
Very Hard	High conflict (inconsistency) – 4 (or all) tools are in disagreement
*Not all tools are required to solve	

Figure 1: Alert Difficulty Breakdown

Using these six numerical scores, mapped to four difficulty levels, the independent variables could be modified in different ways in order to facilitate creating robust alerts. The proportion of the difficulties and number of alerts with false alarm or threat actions were not provided to the participants during the experiment, in order to avoid any counting or other related biases. Using the alert difficulty breakdown as a guideline for alert creation, a total of 10 Easy, 8 Medium, 6 Hard, and 6 Very Hard alerts were populated into the experiment database, and these correlated to 17 False Alarm and 13 Threat based actions.

It is hypothesized that a variance in these difficulties will roughly correlate to participant decision confidence – the more difficult an investigation, the less confidence the participant should experience in their decision-making.

Using the difficulty proportions, the 30 total alerts were randomly distributed into each of the 5 rounds, see Table 2.

Table 2: 30 Alerts with Associated Difficulties

AlertID	Difficulty	AlertID	Difficulty
1	Easy	**16**	Easy
2	Easy	**17**	Very Hard
3	Hard	**18**	Medium
4	Very Hard	**19**	Very Hard
5	Hard	**20**	Medium
6	Easy	**21**	Medium
7	Easy	**22**	Very Hard
8	Hard	**23**	Easy
9	Easy	**24**	Hard
10	Easy	**25**	Medium
11	Medium	**26**	Medium
12	Very Hard	**27**	Medium
13	Medium	**28**	Very Hard
14	Easy	**29**	Hard
15	Easy	**30**	Hard

3.3.1.2 Response Variables

Decision confidence is the primary response variable in the experiment. The self-reported comparisons, which measure decision confidence, are assumed to be dependent of the other choice the participants make, which is making a decision involving the selection of either “False Alarm” or “Threat” for an alert. Coordinating the psychometric data and the investigative process behavior will allow for each aspect of the experiment to be replayed and analyzed, as it will be logged and recorded.

Psychophysiological signals will be captured and analyzed in future studies, as the expertise of the experimenter does not support this analysis. The collection of psychophysiological signals is presumed to correlate to accurate subjective decision-making and decision confidence scoring. Alpha waves are associated with increases in memory load [43], [44]. Gamma waves are associated with memory load, stimulus novelty, attention, and reaction [45]–[48]. Theta waves are associated with decision certainty and error prediction [49], [50]. The response variables, which will be recorded and measured during the experiment, are listed in Table 3.

Table 3: Response Variable Summary

Response variable	***Normal operating level and range***	***Measurement precision and accuracy***	***Relationship of response variable to objective***
Decision choice (categorical)	["False Alarm", "Threat"]	Subjective	Correctness
Decision confidence (categorical)	["1", "2", "roughly the same"]	Subjective	Relative confidence
EEG (numerical)	0-131 Hz at 500 samples/sec 0-262 Hz at 1,000 samples/sec	0.7 µV RMS from 1-50 Hz	Alpha – (9-12 Hz) Gamma – (30-60 Hz) Theta – (4-8 Hz)
ECG (numerical)	60-100 beats per minute	Low noise	Stress/workload

EOG (numerical)	Depends on age/sex
	Mean = 17 blinks per minute
	Reading = 4.5 blinks per minute
		0.7 µV RMS from 1-50 Hz	Movement, vestibule-ocular reflex, blink rate, and saccade

Behavior	Ordering of tool use (categorical) Time per tool (categorical) Time to decision (categorical)	Subjective	The investigative process identifies exploration and/or techniques

3.3.1.3 Constant Factors

Table 4 shows the factors which will be constant for each run of the experiment. A total of 30 alerts were chosen to fit the 2 hour time window of the experiment, as this is the upper-bound generally assumed for electrophysiological experiments. A limit of 6 alerts per round was imposed limit in order to rely on the short-term memory of participants for the greatest subjective scoring efficiency. The number of alerts used for each difficulty will be as close to an even amount as possible, given 30 total alerts. The distribution of the alerts will be randomly distributed across the 5 rounds, and this distribution will then be used for all participants during the experiment. The information from each alert, including the ordering, will be identical for all participants.

Table 4: Constant Factors Summary

Factor	*Desired experimental level*	*How controlled?*	*Anticipated effects?*
30 alerts	Participant reliance on short-term memory	5 rounds of 6 alerts	Minimizes confusion/reliance on memory when comparing
Number of alerts (by difficulty)	Normal workflow	CIAT configuration	None
Alert ordering	Normal workflow	CIAT configuration	None

The **30 Alerts**, made up of five rounds of six alerts, were chosen to maximize the ability of the participants to quickly and reliably recall alerts, such that temporal ordering could be extrapolated from comparing small groups of alerts to each other.

The **Number of Alerts**, including the four types of alert difficulty, were created by a subject matter expert. The four levels are: Easy, Medium, Hard, And Very Hard. The four levels of alert difficulty allowed for flexibility in alert creation and tool information. Since the amount of alerts for each difficulty were withheld from the participant, they had no way of relying on counting alerts per round or overall when carrying out their investigation. Time to gain and analyze the information from the tools was hypothesized to be the single most important factor in determining an alerts difficulty. The amount of information available from each tool was modulated as part of the independent variables.

Alert Ordering is determined in pre-trial experimentation; the ordering was set to the same for all participants. The ordering of the alerts is anticipated to cause no effects.

3.3.1.4 Data Collection and Analysis

Data collection was performed using CIAT, the associated CIAT logging database, and the Cognionics system's physiological output file. Each measurement was stored during and after each participant's trial, but calculations on the data was only done post-experiment. The analysis involved looking for overall trends in the participant population, before analyzing the results from each participant individually.

3.3.1.5 Test Matrix

Table 5 shows the notional test matrix for the experiment. This matrix was performed on each of the 11 participants. It should be stated that the threat and false alarm distribution are not reflective of real-world alert distributions. The intent was to not cause the participant to select a blanket decision choice, knowing that the real-world threat amount is typically very low. Likewise, the alert difficulty distribution was intended to present a range of possible difficulties so that the participant was forced into states of low and high confidence, which can be used in mapping the behavioral data to the electrophysiological data in future work.

Table 5: Test Matrix

Round	Alert Difficulty	Truth Choice/Confidence	Expected Time | Confidence
1	EASY	THREAT	Short | High
1	EASY	THREAT	Medium | High
1	HARD	FALSE ALARM	Long | Medium
1	VERY HARD	FALSE ALARM	Long | Low
1	HARD	THREAT	Short | Low
1	EASY	THREAT	Medium | High
2	EASY	FALSE ALARM	Medium | Medium
2	HARD	FALSE ALARM	Medium | High
2	EASY	THREAT	Medium | High

2	EASY	FALSE ALARM	Medium | Low
2	MEDIUM	FALSE ALARM	Long | Medium
2	VERY HARD	FALSE ALARM	Long | Low
3	MEDIUM	THREAT	Medium | High
3	EASY	THREAT	Short | High
3	EASY	FALSE ALARM	Short | High
3	EASY	THREAT	Short | High
3	VERY HARD	FALSE ALARM	Long | Low
3	MEDIUM	FALSE ALARM	Short | Medium
4	VERY HARD	FALSE ALARM	Medium | Medium
4	MEDIUM	FALSE ALARM	Short | Low
4	MEDIUM	THREAT	Short | High
4	VERY HARD	FALSE ALARM	Medium | Low
4	EASY	THREAT	Medium | High
4	HARD	FALSE ALARM	Medium | Low
5	MEDIUM	THREAT	Long | High
5	MEDIUM	THREAT	Medium | Medium
5	MEDIUM	FALSE ALARM	Long | Medium
5	VERY HARD	FALSE ALARM	Short | Low
5	HARD	FALSE ALARM	Short | Medium
5	HARD	THREAT	Medium | Low

3.3.2 Participants

For this study 11 participants, all male, were recruited, see Appendix A and Appendix B. All participants in this study were voluntary military and government civilian personnel. Participants were not compensated for their participation. The participant's ages were between 22 to 34 years, with a mean age of 26, and a median age of 25 (one subject did not report demographic information). All participants had at a minimum a Bachelor's Degree, and used electronic devices in their job and on a daily basis in their lives. Exclusion criteria included inability to use a mouse and keyboard, visual impairment or inability to view information on a computer screen, and specific motor, perceptual, or cognitive conditions which precluded them from operating a computer. Additionally, because participant electrophysiological data was to be collected, they consented to the placement of electrodes on their head, face, and chest. Additionally, each participant's cyber security experience and whether they had earned any cybersecurity certifications, were recorded. Participant's consent was obtained prior to starting their participation in the study.

3.3.3 Materials

The synthetic task environment used in this study was a modified version of the Cyber Intruder Alert Testbed, also known as CIAT [23]. CIAT provided the underlying features and capabilities, which enabled this research to benefit from a stable interface and tested database system. CIAT, and the associated databases, were modified to reflect the addition of EEG equipment, and to allow for tailored cyber alerts more relevant to the experiments for this research.

For the experiment day, participants were asked to complete a pre-/post-experiment questionnaire. The pre-experiment questionnaire, see Appendix C, asked the participant to account for their most recent amount of sleep and caffeine intake for future correlation purposes. The post-experiment questionnaire, see Appendix D, asked the participant to rate the difficulty of the cyber investigations on a Likert Scale, from 1 to 5. Additionally, demographic information was requested, involving the participant's electronic device usage, electronic device usage in their job, whether they had cybersecurity experience, their age, gender, and highest education level.

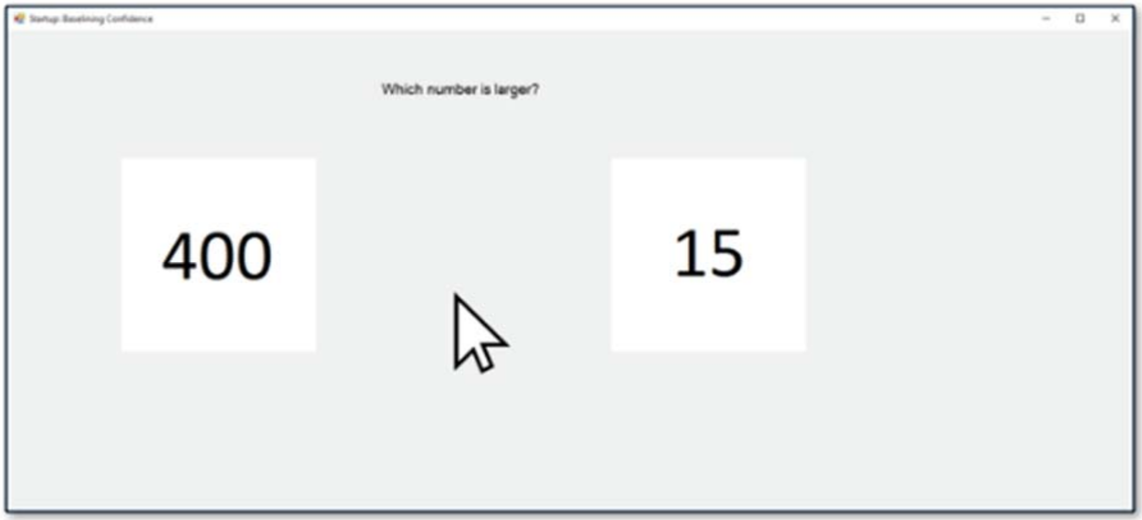
3.3.3.1 CIAT 2.0

For the purposes of this research a modified version of CIAT, named CIAT 2.0, was created. The changes are in three parts, one part focusing on the design of the interface, the next part on the database and the alerts created, and the last part on the program and databases interface with the EEG timing and sensor equipment. Henceforth CIAT 2.0 will just be referred to as CIAT.

The task for the study was a computer-based investigation activity. During each investigation, participants interact with the CIAT program through a computer interface using mouse and keyboard. The CIAT interface provides a method for recording the investigation steps the participant takes, and enables the participant to self-report decision confidence on each investigation. EEG, ECG, and EOG signals was triggered by CIAT and collected by the Cognionics Data Acquisition suite of tools, see 3.3.3.2 for more information on how the EEG data is collected [1], [51].

3.3.3.1.1 Interface

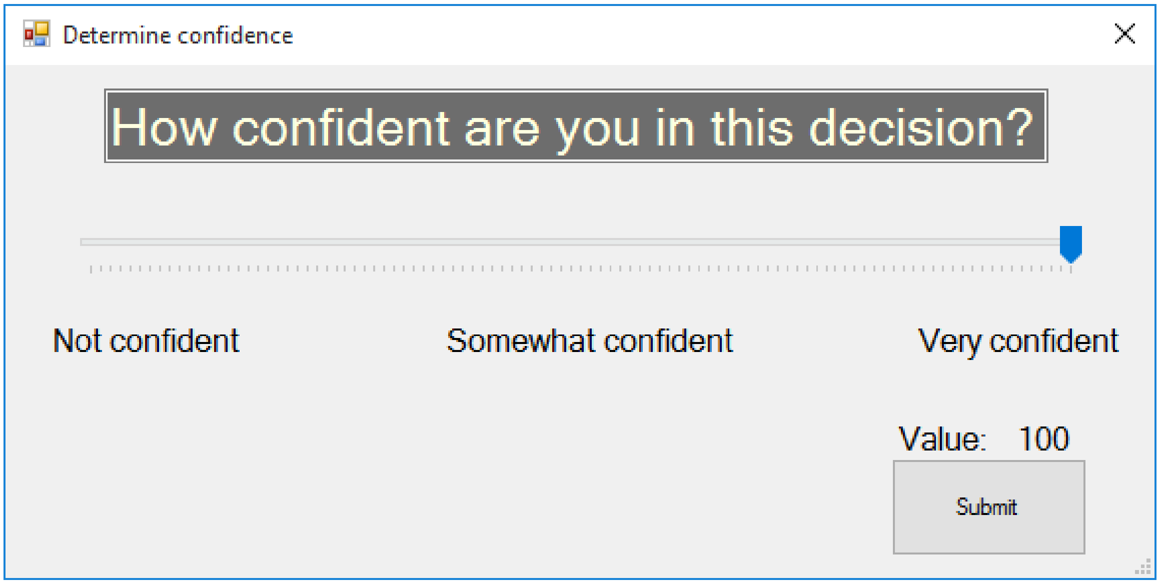
The interface in CIAT was split three main windows: Baselining Questions, Alert Screen, and Confidence Ranking. The Baselining Questions window was the first activity presented to the user when they opened the CIAT program.



Which number is larger?

Figure 2 : Baselining Questions Sample

Figure 2 shows an example baseline question, as it would appear for the user. The user would then have to use their mouse to click on their answer choice. Throughout the experiment, after a selection is made another window would appear, requesting the user measure their associated confidence for their answer choice.

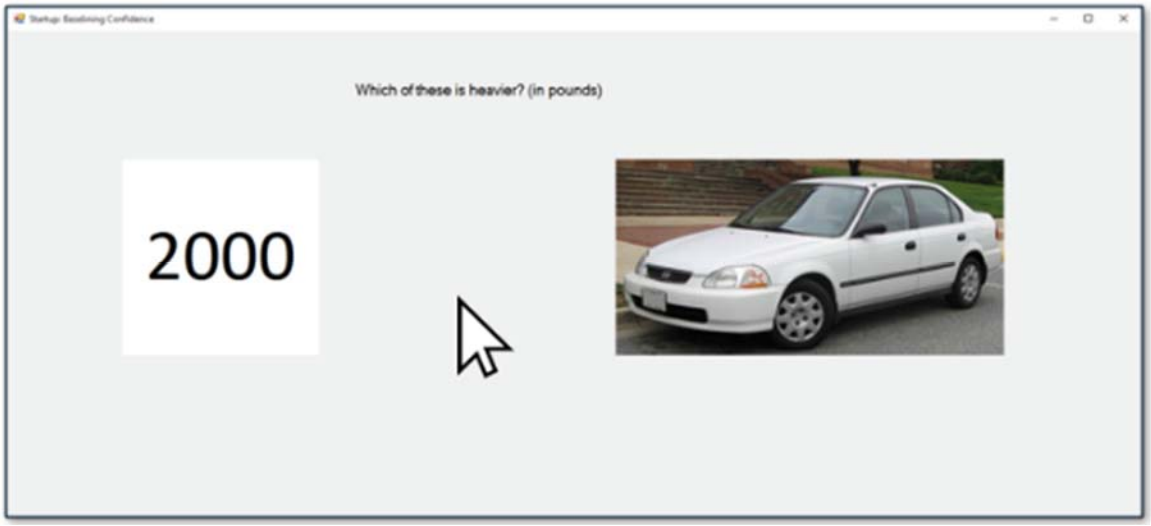

Determine confidence

How confident are you in this decision?

Value: 100

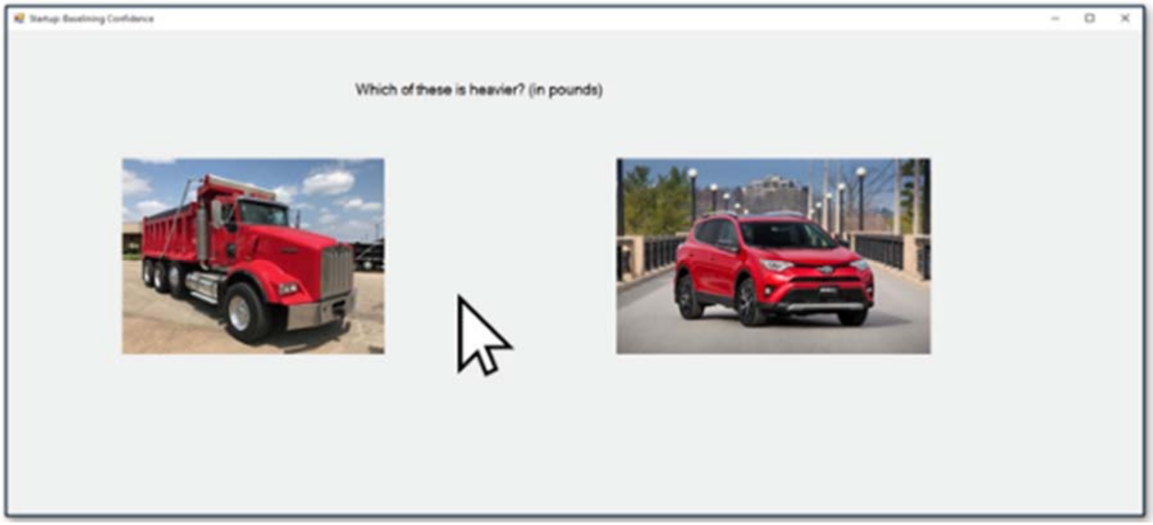
Figure 3 : Determine Confidence Sample

The user would use the slider, as shown in Figure 3, to rate their decision confidence on a scale of 0-100. Additionally, three subjective anchor words were used in order to provide further separation when reporting decision confidence. Both the verbiage and raw value, seen above the submit button, is visible for the user to rate their decision. The Baselining Questions consisted of three examples during the Practice round, and seven examples during the Experiment round. The Baselining Questions were the same for all users. In addition to the number comparison shown in Figure 2, two more Baselining Questions were asked.



Which of these is heavier? (in pounds)

Figure 4 : Number-to-Car Baselining Question Sample

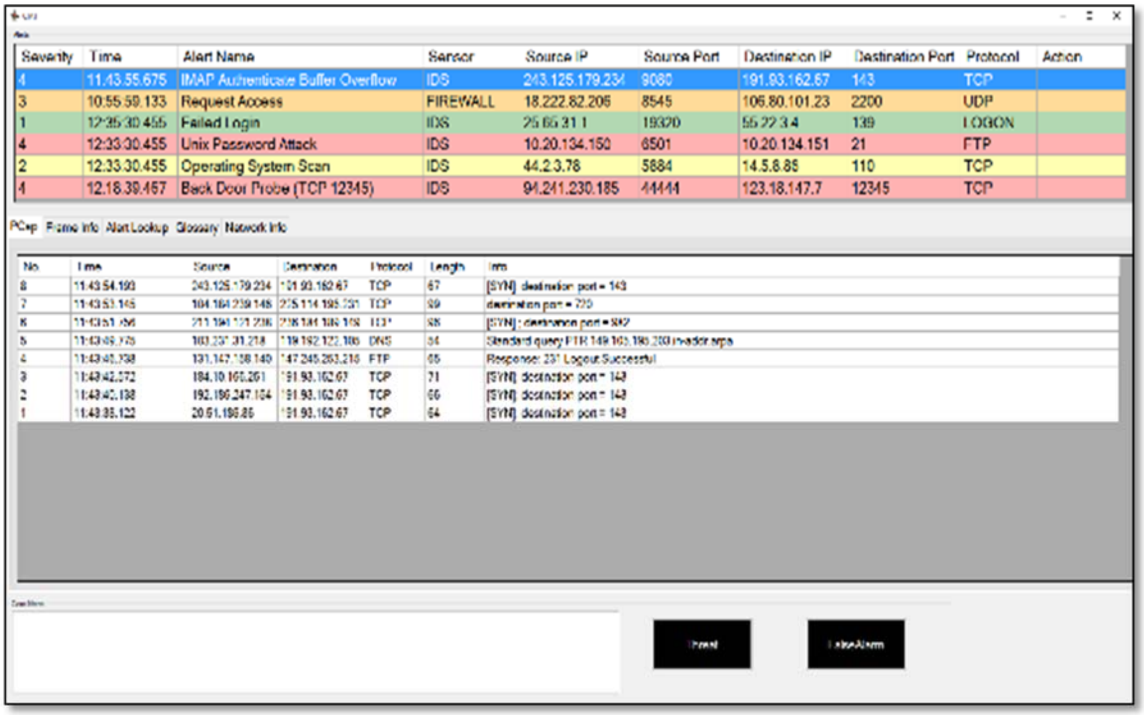


Which of these is heavier? (in pounds)

Figure 5 : Car-to-Car Baselining Question Sample

The second and third styles are shown, taken from the Practice round, in Figure 4 and Figure 5 respectively. Cars were selected in the baseline, as this required minimal background knowledge to answer, and was something that all participants could safely be assumed to see or interact with on a daily basis based on transportation norms in society. Other possible baselining questions, such as arithmetic problems or history-based questions involving United States Presidents, were ultimately decided against because of the possibility of involving other brain signals or memory recall which would have greatly varied based on participant's abilities outside of what was being measured.

The Alerts Screen was the main window the user would see, and where they would spend the majority of their time with CIAT. Figure 6 shows a sample of alerts taken from the Alert Screen during the Practice round. Appendix E labels the primary features of the Alert Screen.



Severity	Time	Alert Name	Sensor	Source IP	Source Port	Destination IP	Destination Port	Protocol	Action
4	11.43.55.675	IMAP Authenticate Buffer Overflow	IDS	243.125.179.234	9080	191.93.162.67	143	TCP	
3	10.55.59.133	Request Access	FIREWALL	18.222.82.206	8545	106.80.101.23	2200	UDP	
1	12:35:30.455	Failed Login	IDS	25.65.31.1	19320	55.22.3.4	139	LOGON	
4	12:33:30.455	Unix Password Attack	IDS	10.20.134.150	6501	10.20.134.151	21	FTP	
2	12:33:30.455	Operating System Scan	IDS	44.2.3.78	5884	14.5.8.85	110	TCP	
4	12.18.39.467	Back Door Probe (TCP 12345)	IDS	94.241.230.185	44444	123.18.147.7	12345	TCP	

No.	Time	Source	Destination	Protocol	Length	Info
8	11:43:54.193	243.125.179.234	191.93.162.67	TCP	67	[SYN] destination port = 143
7	11:43:53.145	104.184.239.148	225.114.195.231	TCP	99	destination port = 220
6	11:43:51.756	211.191.121.236	238.181.189.168	TCP	88	[SYN] ; destination port = 992
5	11:43:49.775	103.237.31.218	119.192.122.185	DNS	54	Standard query PTR 149.163.196.200.in-addr.arpa
4	11:43:46.738	131.147.138.140	147.245.263.216	FTP	65	Response: 231 Logout Successful
3	11:43:42.272	184.10.166.261	191.93.162.67	TCP	71	[SYN] destination port = 143
2	11:43:40.138	192.186.247.164	191.93.162.67	TCP	66	[SYN] destination port = 143
1	11:43:35.122	20.51.196.85	191.93.162.67	TCP	64	[SYN] destination port = 143

Figure 6 : Alerts Screen Sample

The Alerts Screen in CIAT displays the alerts at the top, in colors based on their relative severity level, the tool selection in the middle, and an area for note taking and the decision choices at the bottom. The severity levels are used to differentiate visually between the alerts, and do not associate with the alert difficulty. The participants were briefed during the training day that the severity level did not relate to the intended difficulty. With this setup, the user must directly select tools for results to be displayed. This allows for recording of every aspect of the investigative process, while the user researches and decides whether the alert is a Threat or False Alarm. Figure 7 demonstrate what is expected of a participant while they investigate each alert.

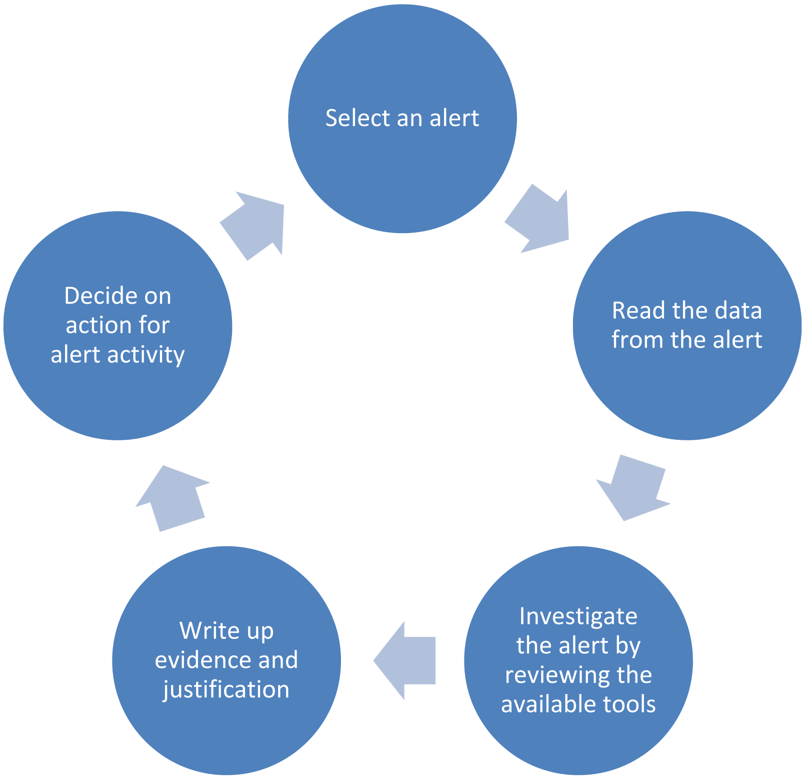
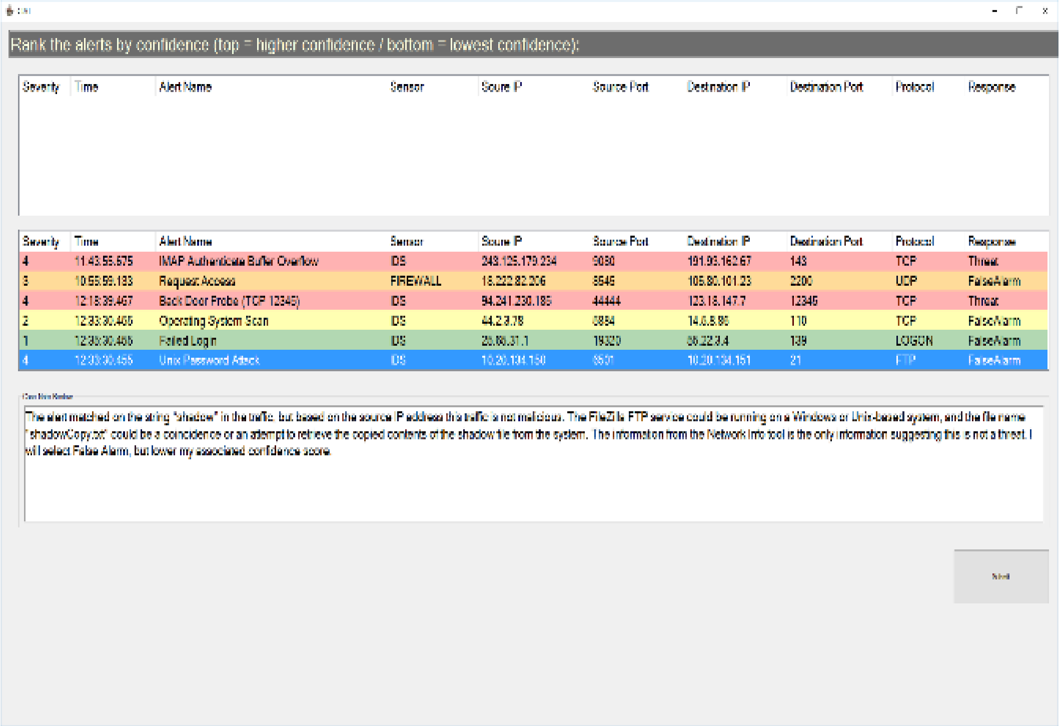

Select an alert
Read the data from the alert
Investigate the alert by reviewing the available tools
Write up evidence and justification
Decide on action for alert activity

Figure 7: Generalized Workflow

Once the user completed a round of six alerts, the user is presented with a new screen. The task for the alert confidence ranking screen, see Figure 8, is to move and sort the alerts from most confident (on the top) to least confident (on the bottom), by reviewing the notes the user submitted for each alert during their investigation. The alerts displayed in the top box must be dragged and dropped, and rearranged, in the bottom box before submitting. This task acts as a quality control mechanism for the analysis. By having the user verify the order of the alerts, without having the confidence scores they submitted from the previous screen, they must rely on their short term memory and feelings when ordering these alerts.



Rank the alerts by confidence (top = higher confidence / bottom = lowest confidence):

Severity	Time	Alert Name	Sensor	Soure IP	Source Port	Destination IP	Destination Port	Protocol	Response

Severity	Time	Alert Name	Sensor	Soure IP	Source Port	Destination IP	Destination Port	Protocol	Response
4	11:43:55.575	IMAP Authenticate Buffer Overflow	IDS	248.125.179.234	9080	191.93.162.67	143	TCP	Threat
3	10:55:59.183	Request Access	FIREWALL	18.222.82.206	8546	105.80.101.23	2200	UDP	FalseAlarm
4	12:18:39.467	Back Door Probe (TCP 12345)	IDS	94.241.230.185	44444	123.18.147.7	12345	TCP	Threat
2	12:33:30.465	Operating System Scan	IDS	44.2.3.78	3884	14.6.8.86	110	TCP	FalseAlarm
1	12:30:30.465	Failed Login	IDS	26.86.31.1	19320	86.22.3.4	139	LOGON	FalseAlarm
4	12:33:30.455	Unix Password Attack	IDS	10.20.134.150	8501	10.20.134.151	21	FTP	FalseAlarm

The alert matched on the string "shadow" in the traffic, but based on the source IP address this traffic is not malicious. The FileZilla FTP service could be running on a Windows or Unix-based system, and the file name "shadowCopy.txt" could be a coincidence or an attempt to retrieve the copied contents of the shadow file from the system. The information from the Network Info tool is the only information suggesting this is not a threat. I will select False Alarm, but lower my associated confidence score.

Figure 8: Alert Confidence Ranking

Once the user submits the alert confidence ranking, they will see the alert screen again, Figure 6, but with new alerts. This will repeat for 5 rounds of alerts, for a total of 30 alerts during the experiment. For the practice, the participants were given 2 rounds of alerts, for a total of 12 alerts.

3.3.3.1.2 Timing Database and Triggers

In order to track the behavior and investigative process of the users, mouse and keyboard input was logged. This allowed for a play-by-play reconstruction of each user’s exploration through the tool, including the notes they typed, and the choices they made. Coupled with each of these events were the triggers, which were time-synchronized with the data generated by the EEG sensors. In total, there were 19 different timing events generated by CIAT, which were logged in the timing database. These timing events included items such as when forms were displayed, when a button was pressed, and when decision choices were selected by the user. Coupled with this timing information were triggers sent to the EEG measurement equipment, which provided each timing event to be sliced and time synchronized, during post-processing analysis, with the database.

3.3.3.2 EEG / ECG / EOG Equipment

Participants interacted with the CIAT program running on a desktop machine in the lab, which was configured to send trigger time sync data to the researcher's laptop computer in order to synchronize the recording of the collected electrophysiological data. To collect EEG data, participants wore a dry electrode harness as shown in Figure 9. Purchased from Cognionics, Inc., the Cognionics Mobile Series Headset was made up of a harness capable of recording up to 72 channels. A total of 66 electrode channels were recorded on the EEG cap, including the ground electrode. One electrode, located near the neck behind the right ear, was used as a reference node. In addition, seven electrodes were added as three additional channels, for 69 total channels, in order to capture the EOG and ECG data. Six of these electrodes were set as pairs, one positive and one negative, and one electrode was a shared ground. Two pairs of electrodes, one pair per channel, were used for the EOG data, see Figure 10. One pair of electrodes was used for the ECG, which included the ground on its channel, see Figure 11. These electrodes measured brain activity and sent signals to the laptop computer using a wireless Bluetooth connection where the signals were recorded. The Cognionics Mobile Series Headset recorded at a rate of 1000 samples per second. The Cognionics Data Acquisition suite of tools was used to capture and process the EEG data into the Biosemi (.BDF) file format.

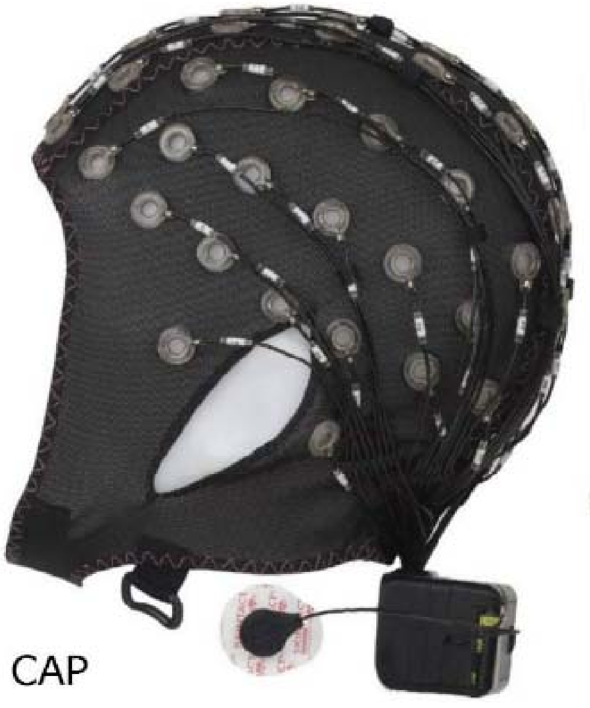


Figure 9: Cognionics Mobile Series Headset EEG Cap and Harness

The EOG electrodes were placed on four locations on the face, as shown in Figure 10, in order to measure the blink rate and direction of eye movement. A shared ground electrode was used between the EOG and ECG.

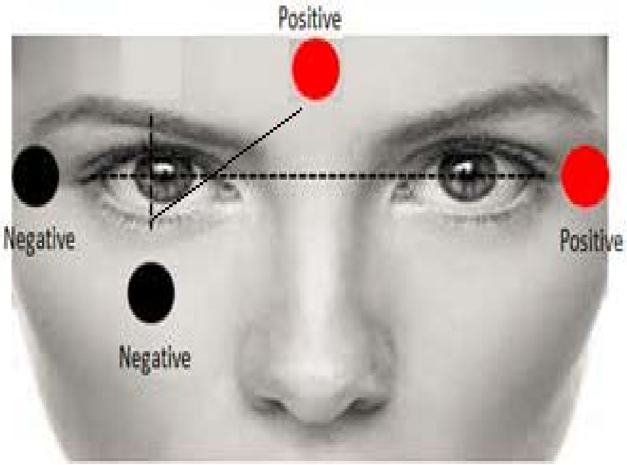


Figure 10: EOG Electrode Placement on Face

The ECG electrode placement, see Figure 11, shows where the two ECG electrodes would be placed on the participant's chest and also where the shared ground would be placed.

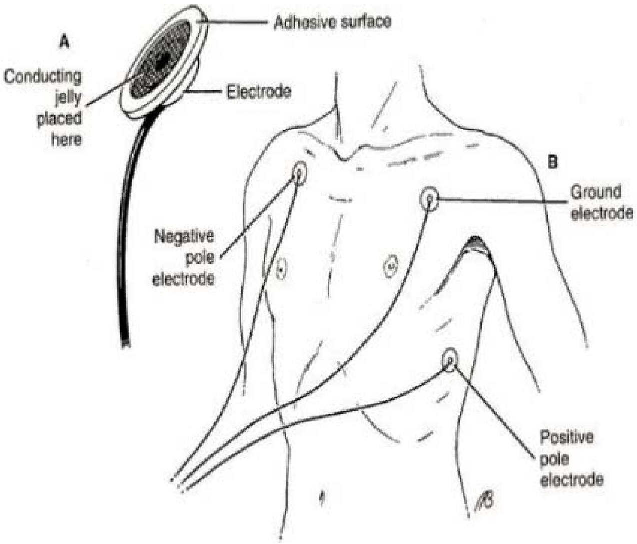


Figure 11: ECG Electrode Placement

Both the EOG and ECG data were collected as exterior (EXT) node measurements, as they were connected to the Cognionics EEG cap through a Universal Serial Bus 3.0 (USB 3.0) Data AcQuisition Module (DAQ), which fed the information to the researcher's laptop wirelessly.

3.3.3.3 EEG – Cognionics Mobile-72 Wireless EEG System

The Cognionics Mobile Series Headset collects all of the EEG data from the participant. The intent of collecting EEG measurements was to map the associated behaviors of the participant during the alert investigations. CIAT recorded the windows and tools that the participant used for each alert, as these were important for associating events that led to changes in EEG measurements. Due to time constraints, EEG will be left to future work.

3.3.3.4 ECG – Cognionics 1-channel + shared ground electrode

ECG measurements were associated with timestamps of the decision selections (e.g. False Alarm or Threat). Similar to EEG measurements, ECG were used to measure workload and stress as the participant conducts and validates their decisions [37]. ECG analysis will be left to future work.

3.3.3.5 EOG – Cognionics 2-channel + shared ground electrode

EOG measurement analysis recorded blinks, saccades and visual fixation, which are associated with levels of perception, concentration, awareness, and the learning and training progress of learners [2], [52]. The intent of measuring eye movement, and the associated dwell time, was to indicate levels of confusion or exploration by the participant. Additionally, rapid eye movements indicate other factors such as graphic user interface frustration, which might affect decision-making and decision-confidence. EOG recordings can be used to augment EEG-artifact cleaning process since eye muscle movements are a large source of these artifacts. EOG analysis will be left to future work.

3.4 Assumptions

The following assumptions are made for this experiment:

- The confidence of a decision is dependent on the decision-making process up to the choice of the decision. It is assumed that in this experiment structure, once a decision is made it cannot be changed.
- The participants are not withholding information, and are willing to honestly self-assess in their decision confidence, based on their decision-making.
- The participants have not been told of the experiment or prepped, by another participant, before participating in the experiment.

3.5 Procedures

The participant's activities were split between two days of up to two hours on each day. The first day included a familiarization lecture and hands-on training with the CIAT program, as well as cyber security fundamentals. Training involved multiple participants with one instructor, with class sizes between 2 and 4 participants. Four separate training days were used to train 11 participants. The training day activities were conducted in a classroom environment, with computers and a projector screen to present the training lecture and demonstrate the CIAT program. The first task the participants practiced was the decision confidence baseline, which involved three types of questions requiring the participant to pick the best answer from a set of two answer choices, see Section 3.3.3.1.1. The intent of the decision confidence baseline was to use familiar concepts and example questions to prepare the participant to understand how they must think about evaluating their decision confidence.

The second task involved interface familiarization and a workflow walkthrough for two alerts by the instructor. Each of the participants was given a general workflow process as a handout, which was also available to them during the experiment, see Appendix F. After these two alerts were completed, the participants were allowed to open the CIAT program and follow along with one example while the instructor guided all participants. After all of the participants had completed these three alerts as a class, three new alerts were provided. The participants were instructed to work at their own pace, and on their own, but they could seek help from the instructor. Once all the participants completed these three alerts, the instructor reviewed these alerts and provided their notes and confidence ratings as a comparison.

After the round of six alerts, a new screen was displayed in CIAT requiring the participant to rank each respective alert based on the relative decision confidence to each other alert. During each of these first six alerts, also referred to as the first round, the instructor provided their own decision selection, decision confidence score, and their associated case notes, which participants could read and ask questions about. After familiarizing the participants with the decision confidence ranking task, the participants were given the remaining time to complete six alerts at their own pace, but without any discussion about the decision, the decision confidence score, or the case notes from the instructor. In total, the training day consisted of two rounds of six alerts, for the participant to practice and understand their job and the task environment.

An individual 2-hour experiment block was scheduled for each participant. Only one participant was scheduled during a 2-hour block. Each participant had to first complete the training before being scheduled on a subsequent day for their experiment. All experiment days occurred within two weeks of the day the participant completed training.

In the experiment, the participant first completed a pre-experiment questionnaire, see Appendix C. The pre-experiment questionnaire asked the participant to quantify and qualify their sleep from the night before, and their level of alertness and ability to complete the task.

Next the participant was prepped and configured with the EEG, EOG, and ECG equipment before being asked to sit at a desk with the associated computer terminal loaded with the CIAT software. Once the systems were checked for accurate readings, the participant was allowed to begin the experiment by opening up the CIAT program. All three tasks were identical to what the participant had seen and practiced on the training day, albeit instead of 12 total alerts across 2 rounds, they were given 30 total alerts across 5 rounds. The partitioning of the alerts into 5 rounds of 6 alerts was intended to enable participants to recall the previous 6 decisions they made so they could reflect on those alerts during the decision-confidence ranking step. One by one, the participant would investigate each alert and determine whether it was a false alarm or threat. Additionally, the participant was required to input case notes justifying their reasoning for the decision before submitting a decision. This justification would also aid them in recalling the information during the decision confidence comparison stage since the tools were not available for review when they were tasked to perform the relative confidence comparison between rounds. Because each alert investigation was estimated to take 2-3 minutes to complete, a round of 6 alerts was expected to take up to 18 minutes to complete. Since the equipment necessary to conduct the electrophysiological measurements, and the posture of the participant, needed to be controlled during these sections, a pause between these rounds allowed for a short break to adjust before proceeding. Between each of the rounds, the participant was required to complete a decision confidence ranking.

Once the final alerts were investigated, and the final round was ordered by relative confidence, the participant was asked to complete a post-experiment questionnaire, see Appendix D. The post-experiment questionnaire asked the participant to rate how difficult the cyber investigations were overall. Additionally, computer usage experience and demographic information was surveyed in the questionnaire.

3.6 Analysis Strategy

All collected data was analyzed with python and statistics packages. Analysis focused on the results of decision choice and decision confidence. The decision confidence from participants was compared to the truth data, from an experienced analyst (the baseline), which was correlated with the control factors to determine which changes incurred the greatest effect on decision confidence.

First the baselining questions were reviewed, as they were important for EEG analysis. The baseline questions, if calibrated correctly, would establish known distinct difficulty levels which could be mapped to electrophysiological data. Since the difficulties are only ordinal, they would allow for a relative comparison between different states of physiology the participant might be in. Similarly, the patterns of behavior associated with baseline decision-making and alert decision-making could be reviewed for similarities, although the tasks are wildly different. The baseline questions do not require an investigation, and only rely on comparing numbers or car weights, therefore the electrophysiology may prove more relevant than the behavior patterns. A rank comparison will be done to validate that the difficulties were ordered as intended.

The expectations for the behavior pattern analysis involved reviewing and analyzing the recorded data from CIAT. Time to decision, for example, could be an indicator of confidence. Looking back at the Test Matrix, see Table 5, the expected averaged results for the time to decide and the confidence level for the alerts in each round based on the difficulty. The expected values acted as a hypothesis for the data analysis. The choices made by participants, and the correctness, indicated whether alert difficulty correctly aligned with our intended alert difficulty. Easy alerts were expected to have almost 100% accuracy, whereas very hard alerts were expected to be of much lower accuracy. A rank comparison will be done to validate the difficulties were ordered as intended. Notional results were illustrated in Figure 12 and Figure 13.

Figure 12 illustrates a notional representation of the participant times per alert difficulty. Data exploration, such as trend and correlation comparisons, enabled key decision-making behaviors to be identified. The questionnaire data, concerning the participant's computer skill or general degree of confidence, was analyzed in order to identify whether any correlation could be found with time to decision. It is hypothesized that participant's with past experience and skills in cyber will perform better than those participant's without these skills.

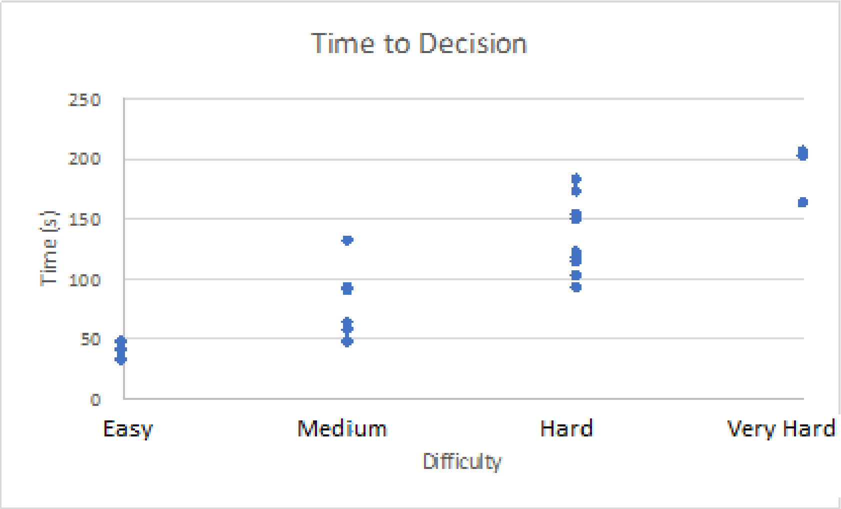


Figure 12: (Notional) Time to Decision

Figure 13 charts the relative confidence of each of the difficulty tiers of alerts. Grouping and clustering can be used to determine the general decision-making disposition of individuals, in order to see who and possibly analyze why individuals responded similarly.

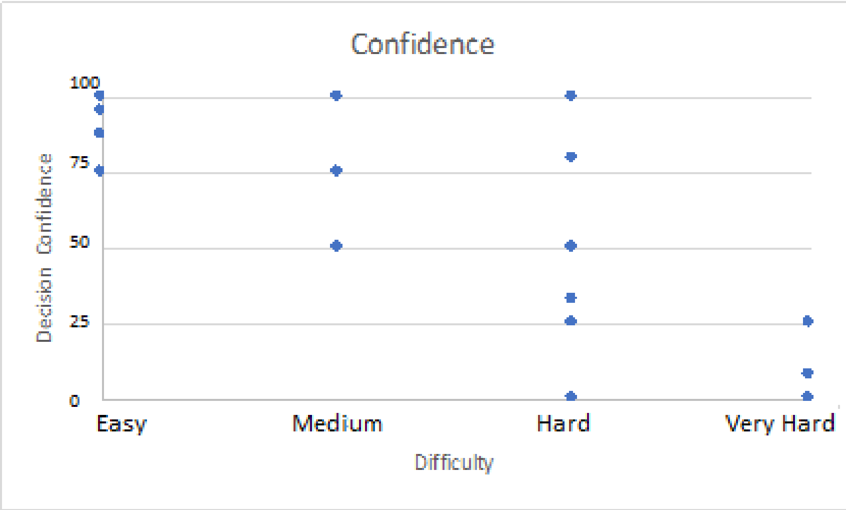


Figure 13: (Notional) Difficulty vs Decision Confidence

Other factors of analysis interest include tool usage trends, how long tools are used, and how frequently tools are transitioned between. Analyzing these behaviors may reveal exploration behavior, by the participant's, which may associate with lower confidence. For example, the participant may only need to consult one or a few tools in order to make a decision, in cases of easier alert difficulty, whereas they may have to spend more time and review tools countless times as the alert difficulty is increased.

Participant experience may play a part in understanding the behavioral differences, associated with how investigations may differ between alert difficulties. Therefore a general workflow guideline, for the participants to rely on, will be provided on both the training and experiment days. The training day will focus on teaching the workflow process in order to provide all participants a baseline level of knowledge for conducting cyber based alert investigations.

Analysis of the electrophysiological data will be future work. A recommended approach for the EEG data is applying the diffusion model. The diffusion model is a model of the cognitive processes made during one- or two-choice decisions [53]. The drift-diffusion model suggests participants will quickly decide upon an initial course of action based on the available information, and use future stimuli to either further fortify or contradict their decision [17], [53]. The correlation to the EEG measurements and the process by which the participant comes to a decision, would provide insight into how the participant's behaviors influence decision-making. Likely graphs to be presented include comparison-based and cluster-based overlay charts, to determine similarities among participants when conducting investigative behavior which will be cross-correlated with the tool and timing information collected from CIAT.

3.7 Summary

In summary, the methodology explained in this chapter establishes the foundation for how the experiment was created and set the expectations for data collection. By recording the behavioral data of participants, through the CIAT STE, this research allows for analyzing how confidence is affected by patterns of investigative behavior. This analysis strategy appropriately looks to review and calibrate the baseline questions and the investigative cyber alerts, prior to doing any behavior comparisons among the participants. After the difficulty is calibrated, data exploration will elaborate hypotheses which were tested in order to answer the research questions for this paper.

The next chapter describes the analysis conducted on the compiled data. It became evident in the early stages of the data analysis, that the initial difficulty classifications of some of the alerts needed to be fixed and recalibrated. Section 4.2.2 explains why this was needed, and how the alert difficulties were tuned after all of the participants had completed the experiment. Additionally, since the electrophysiological measurements were reliant on finding expertise to conduct the analysis, the primary focus was on identifying and creating a methodology which prioritized capturing behavioral metrics from the CIAT tool independent of the external EEG equipment.

IV. Analysis and Results

4.1 Chapter Overview

The purpose of this chapter is to explain the data exploration and analysis process which led to the results. The investigative patterns of behavior for each participant were explored. Each participant's results were compared to each other, and to the population of participants. Several identifiable behaviors were extracted from the data, and will be analyzed in the results section. The results will also be highlighted in the conclusions of Chapter V.

4.2 Behavioral and Subjective Analysis and Results

The initial analysis of the participant's investigation activity involved plotting both the accuracy and confidence scores against difficulty to determine whether the differences in alert difficulty had the intended effect of causing variations in the confidence scores when comparing alerts across the same participant or between participants.

4.2.1 Baseline Review

Reviewing the 7 baseline questions was done first in order to construct and validate a data analysis process which would be scaled to the 30 alerts from the experiment. These were made up of 4 Easy, 0 Medium, 2 Hard, and 1 Very Hard questions. Figure 14 shows the plotted confidence values of all participants for the baseline questions. The participant's reported different confidences for each of the alert types, but this needs to be validated by reviewing the rank correlation.

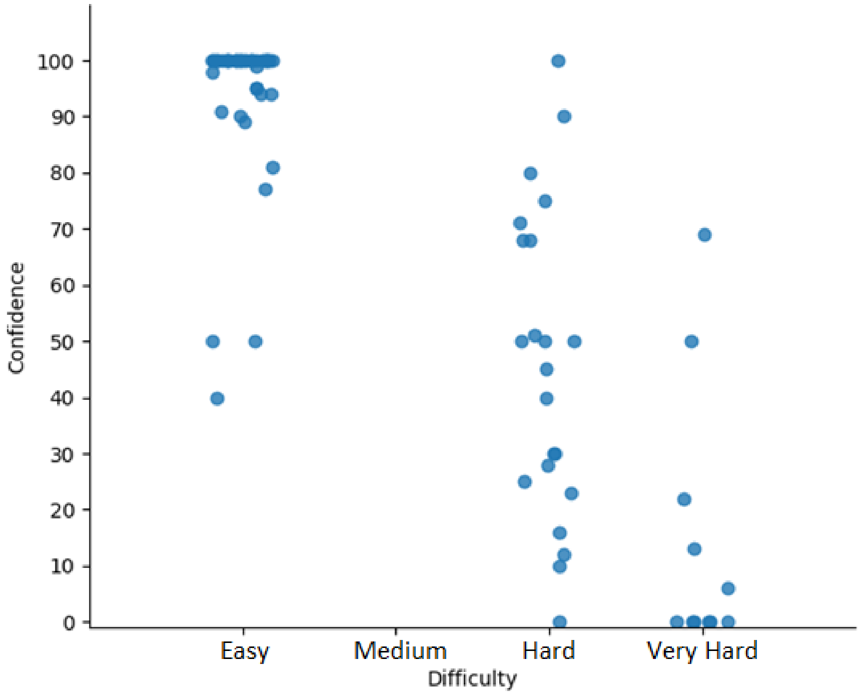


Figure 14: Baseline Comparison of Confidence versus Difficulty

A ranking correlation comparison was done using the Mann-Whitney U test. Comparing each of the questions' estimated difficulty with the confidence of the participants yielded statistically significant results, which correctly ordered the alerts by what was intended. The average confidence of each alert difficulty among each participant was input into the Mann-Whitney U test. The easy alerts were of a higher confidence relative to the hard alerts, which was statistically significant (U-stat(11) = 3.973, p = 7.105x10^-5), where the alpha value (significance level) = 0.05. The positive value of the U-stat means that the easy alerts were ordered higher than the hard alerts. This was repeated for each combination of alerts, in order to determine a rank ordering of the baseline alert difficulties with the reported confidence. For hard and very hard alerts, the U-stat was significant (U-stat(11) = 2.791, p = 0.005258). Likewise, the results for the easy and very hard alerts were also significant (U-stat(11) = 3.973, p = 7.105x10^-5). This confirms an ordering of the alert difficulties, from most to least confident, as easy > hard > very hard. Note that there were no medium alerts in the baseline. These results meant the calibration of the baseline alerts was correct.

A scatter plot comparing the difficulty of the baseline comparisons by difficulty versus accuracy was created, see Figure 15. The clusters of accuracy for each of the alerts was separated by 1 for correct, and 0 for not correct. A similar rank comparison was done with the alerts based on accuracy.

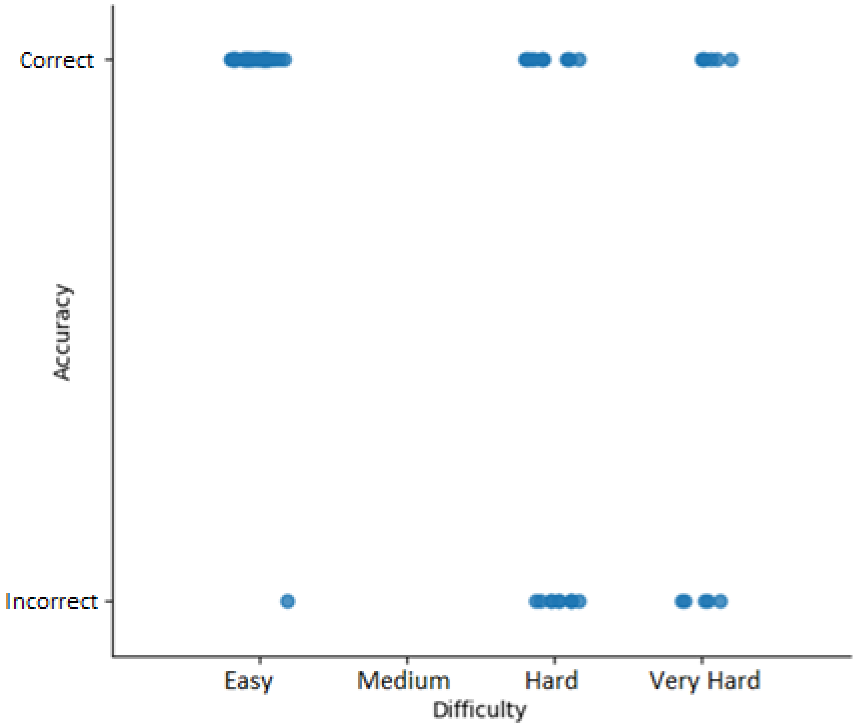


Figure 15: Baseline Comparison of Accuracy versus Difficulty

Using the Mann-Whitney U test again, the accuracy of each alert was ranked and compared for statistical significance. Using the same alpha value of 0.05, the only significant ordering was easy and hard alerts (U-stat(11) = 2.397, p = 0.01654). Therefore, the only ordering which could be correlated from the accuracy was that the easy alerts had a higher accuracy than the hard alerts.

The intent of the baseline was to familiarize the participant with how to answer questions and select a confidence. All the behavior and scores during this simple task were recorded, so that future calibration and analysis could be done when coupled with the electrophysiological measurements. With a known and calibrated baseline, the participant's EEG results could be compared from their performance on the alerts. Knowing that the difficulties were correctly ordered, by confidence, would also be useful for identifying and comparing behavioral trends.

4.2.2 Alerts Review

During the compilation of the results for the 30 alerts and the initial review of alerts, the SME raised concerns that alterations to the CIAT tools and database may have led to some alerts being incorrectly calibrated. The goal for creating 30 alerts was to make as close to an equal amount of alerts for each difficulty as possible. These 30 alerts were originally calibrated such that 10 alerts were easy, 8 medium, 6 hard, and 6 very hard. Table 6 shows the original breakdown of the 30 alerts by correct response and difficulty.

Table 6: 30 Cyber Alerts for Experiment (Original Calibration)

AlertID	CorrectResponse	Difficulty
1	Threat	Easy
2	Threat	Easy
3	FalseAlarm	Hard
4	FalseAlarm	Very Hard
5	Threat	Hard
6	Threat	Easy
7	FalseAlarm	Easy
8	FalseAlarm	Hard
9	Threat	Easy
10	FalseAlarm	Easy
11	FalseAlarm	Medium
12	FalseAlarm	Very Hard
13	Threat	Medium
14	Threat	Easy
15	FalseAlarm	Easy
16	Threat	Easy
17	FalseAlarm	Very Hard
18	FalseAlarm	Medium
19	FalseAlarm	Very Hard
20	FalseAlarm	Medium
21	Threat	Medium
22	FalseAlarm	Very Hard
23	Threat	Easy
24	FalseAlarm	Hard
25	Threat	Medium
26	Threat	Medium
27	FalseAlarm	Medium
28	FalseAlarm	Very Hard
29	FalseAlarm	Hard
30	Threat	Hard

During the construction of the CIAT tool and alert database, the self-imposed limitation of 2 hours for participant experimentation trials, led to changes being implemented for the tools. The information available in several tools and the alert metadata were changed to minimize confusion by participants without cyber experience and also to target a per alert investigation time of 2-3 minutes, so that the experiment could be conducted within the allotted time of 2 hours. All of the alerts were created by one SME with several weapon system certifications and three years of experience on an Air Force cyber defense weapon system. Since the computer experience and cyber skillset of the participants varied between those with cyber security experience and certificates and those without, it was important to calibrate the alerts in such a way as to allow anyone with minimal training to be able to identify signs of good and bad cyber based network activity. All participants were familiar with computer usage, and use computers on a daily basis for their jobs and with their daily life, but investigating cyber alerts was a task the majority of participants had not conducted prior to this study. Thus, confidence scores were scrutinized based on the intended thresholds set by the difficulty. For example, very low confidence scores on easy alerts and high confidence scores on very hard alerts were suspect and reviewed first to identify whether the alerts created the intended difficulty. The difficulty was set based on three main characteristics: information needed, information available, and consistency of the information. Validating the intended difficulty levels, such that the subjective confidence metrics were consistent across and between participants was important for identifying the influence of behavior during the investigations. Each difficulty category: easy, medium, hard, and very hard, was charted relative to each participant's accuracy, investigation time, and rated confidence score.

The initial review of these alerts suggested that the alerts were not calibrated to the difficulty level intended. Thus, all alerts were reviewed again by the SME to determine if any of the alerts had changed in difficulty due to changes to the initial quantity and verbosity of the information available in the tools in CIAT. After reviewing all the experiment alerts, it was determined that a total of 19 alerts needed alterations such as a correction to the difficulty score for 17 alerts and changes to the correct answer for 5 alerts. Any outliers identified in the scatter plots of the difficulty versus accuracy and difficulty versus confidence were now attributed to consistency per participants. The updated difficulty spread for these 30 alerts was updated to 11 easy, 6 medium, 9 hard, and 4 very hard. Table 7 shows the updated correct responses and recalibrated difficulties as the cells highlighted in yellow. Due to the recalibration, the amounts for the correct responses were also changed from 17 False Alarms and 13 Threats, to 14 False Alarms and 16 Threats.

Table 7: 30 Cyber Alerts for Experiment (Updated Calibration)

AlertID	Correct Response	Difficulty	AlertID	Correct Response	Difficulty
1	Threat	Easy	16	Threat	Very Hard
2	Threat	Easy	17	False Alarm	Hard
3	False Alarm	Hard	18	Threat	Easy
4	False Alarm	Very Hard	19	False Alarm	Very Hard
5	Threat	Hard	20	False Alarm	Medium
6	Threat	Easy	21	False Alarm	Easy
7	False Alarm	Medium	22	Threat	Medium
8	Threat	Hard	23	Threat	Medium
9	Threat	Hard	24	False Alarm	Hard
10	False Alarm	Easy	25	Threat	Hard
11	False Alarm	Medium	26	Threat	Easy
12	Threat	Very Hard	27	False Alarm	Hard
13	Threat	Easy	28	False Alarm	Easy
14	Threat	Medium	29	False Alarm	Easy
15	False Alarm	Hard	30	Threat	Easy

Using the newly recalibrated alerts, plots were created similar to the baseline comparisons in order to continue data exploration. Figure 16 displays the confidence versus difficulty of the original 30 alerts prior to recalibration. The same plots were generated for the 30 alerts as were generated for the baseline analysis. Figure 17 displays the confidence versus difficulty plot of all 30 alerts after the recalibration took place. Similar to the baseline comparison plots, the higher difficulties were illustrated by confidence scores which were more spread out. The Easy difficulty, represented as 1, showed the highest concentration in higher confidence scores.

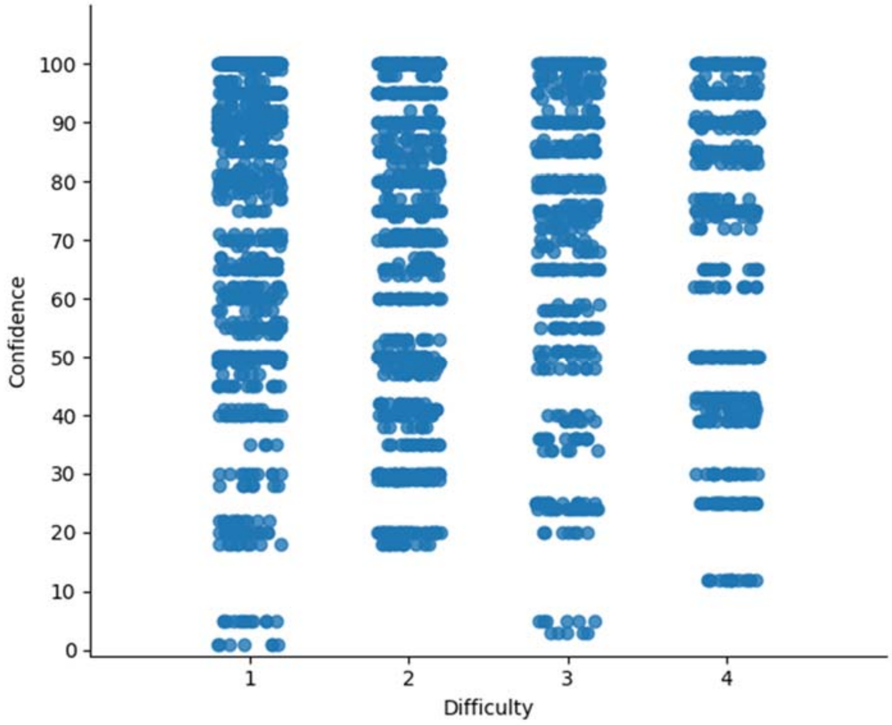


Figure 16: Cyber Alert Comparison of Confidence versus Difficulty (Original Calibration)

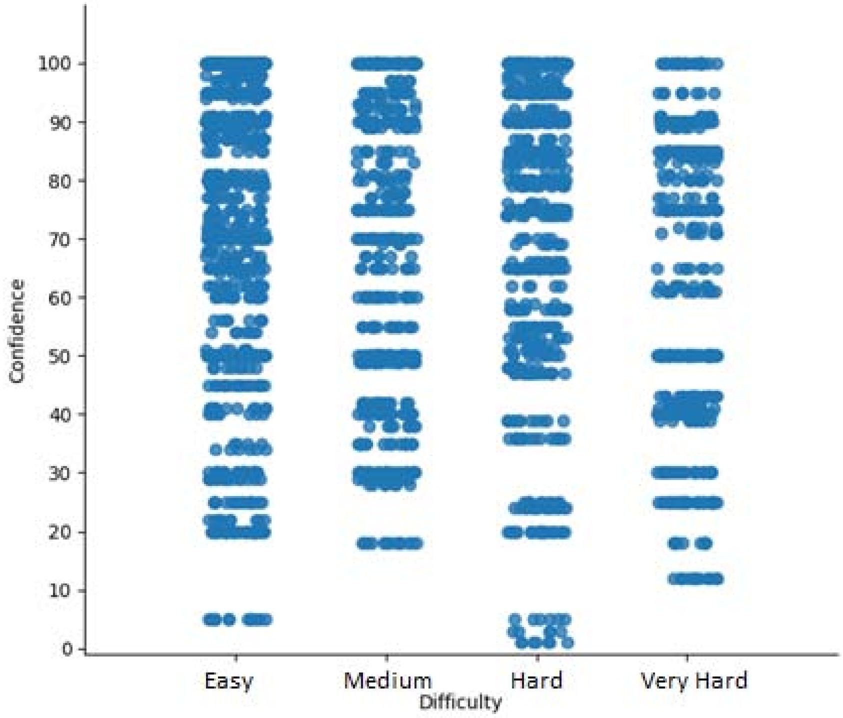


Figure 17: Cyber Alert Comparison of Confidence versus Difficulty (Updated Calibration)

The next task was to validate the intended difficulties of the alerts, as these were the main instrument for affecting the confidence level of participants. A large spread of confidence score is visible, see Figure 17, for each of the difficulties. Using the Mann-Whitney U test, the confidence of each alert was ranked and compared for statistical significance. Using an alpha value of 0.05, the only significant ordering was between easy and very hard alerts (U-stat(11,4) = 2.068, p = 0.03860). This means that only the easy and very hard difficulties have a statistical significance, allowing for rank ordering.

Going through the same process as was done for the baseline, a rank comparison test was completed. Again, the Mann-Whitney U test was used. Using an alpha value of 0.05, various orders were statistically significant. The easy and hard alerts (U-stat(11,9) = 2.594, p = 0.009493), easy and very hard alerts (U-stat(11,4) = 3.283, p = 0.001026), medium and very hard alerts (U-stat(6,4) = 2.791, p = 0.005258), and hard and very hard alerts (U-stat(9,4) = 2.856, p = 0.004284) were all statistically significant with p values greatly below 0.05. This means that the ordering of the alerts by accuracy, from most to least accurate, is easy > medium > hard > very hard.

4.2.3 Data Exploration

During the initial data exploration, it was hypothesized that the ordering of tool use, the time per tool, and the overall time to a decision would correlate to the differences when comparing the confidence of each alert decision, per participant.

Data exploration into the ordering of tool usage was cursory and did not provide for a sufficient way to readily compare within a participant or between participants. In order to allow for time to explore the other hypotheses, tool order was skipped in the hopes of being returned to later, when other trends had been identified which caused a need to review the ordering of tool usage. Thus, the focus of data exploration moved to reviewing the time per tool and frequency of tool use for alerts. The actual metric for time per tool was calculated by looking at the time in each tool, given various other factors.

4.2.3.1 Analyzing Time-in-Tool

Figure 18 illustrates the total time spent in a tool, combined per each alert, to showcase which tools the participant's spent the most time. This metric was called the time-in-tool. Each participant's time-in-tool performance was compared to the reported confidence scores, for identifying behavior trends.

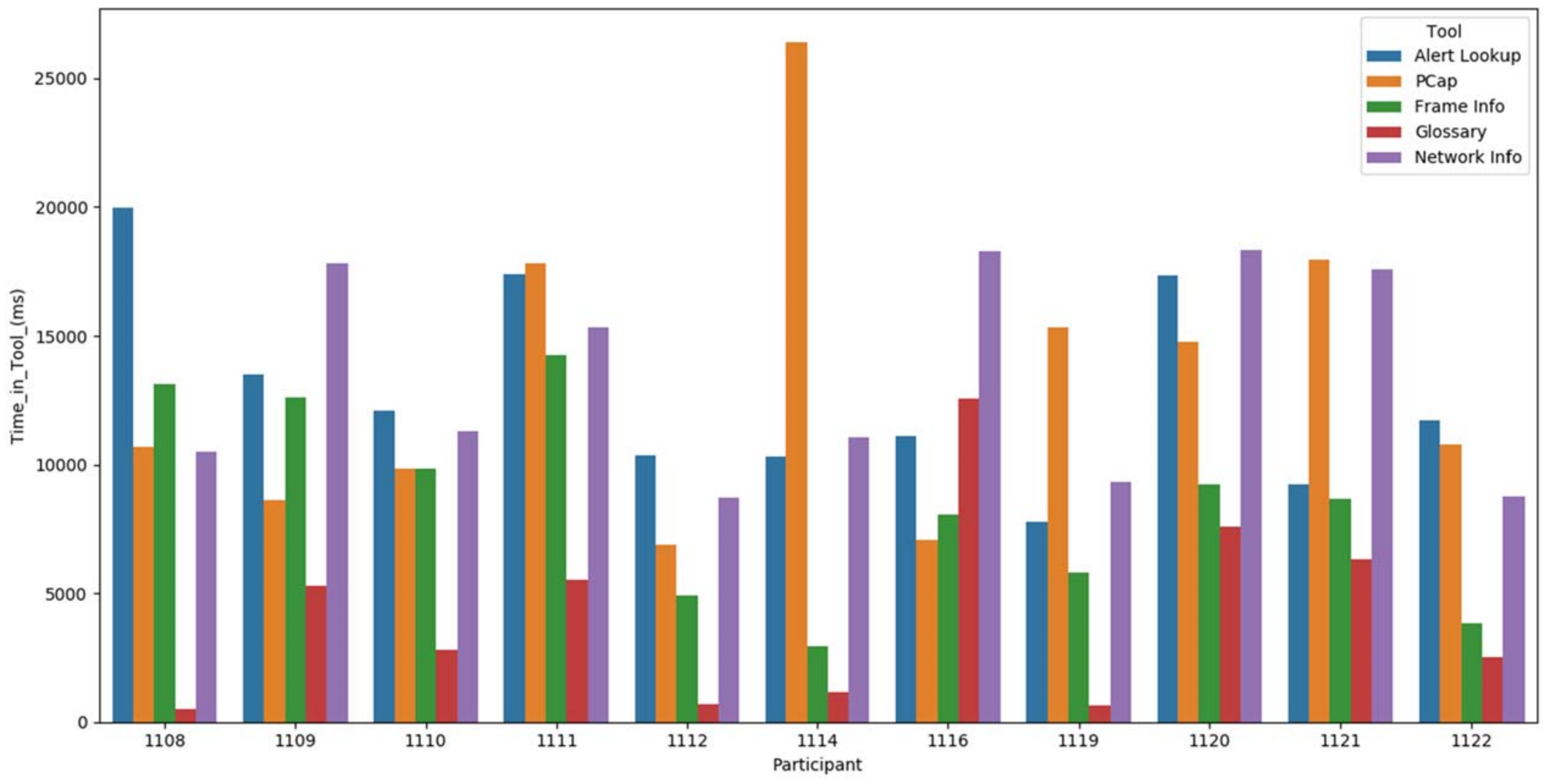


Figure 18: Time-in-Tool versus Participant (per Tool)

For two participants, the time-in-tool metric was significant. There was a significant effect for participant 1120: (F(1,28) = 4.869, p = 0.03571) and participant 1121: (F(1,28) = 4.692, p = 0.03896), with an alpha value of 0.05. For participant 1120 and 1121, the time spent in the tools was statistically significant for the confidence of the alert. A decreasing trend line is readily apparent in Figure 19, showing Confidence versus Time-in-Tool for Participant 1120. This trend line shows that tools are used for shorter periods of time, when the participant expresses higher confidence.

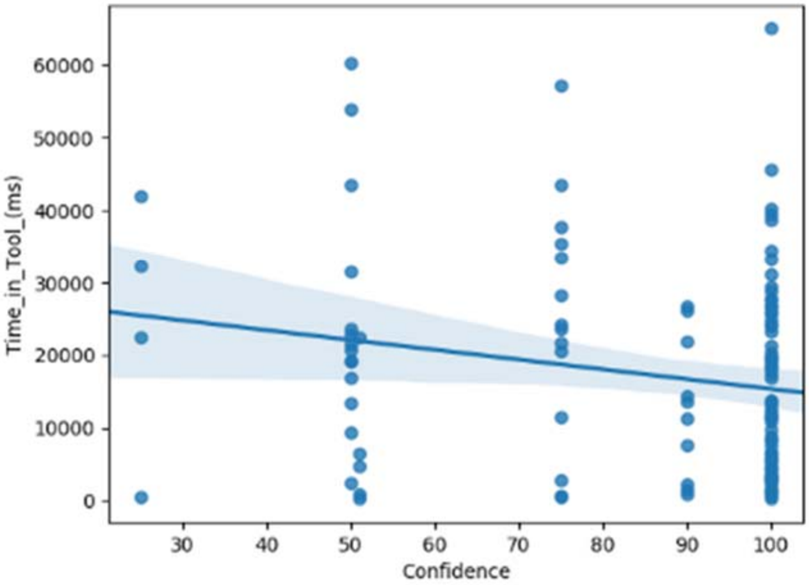


Figure 19: Confidence vs Time-in-Tool – Participant 1120

For participant 1121, a decreasing trend line is evident in Figure 20, although it is not as steep as Figure 19. The trend line shows that as confidence increases the time spent in tools decreases, which is the similar case for participant 1120.

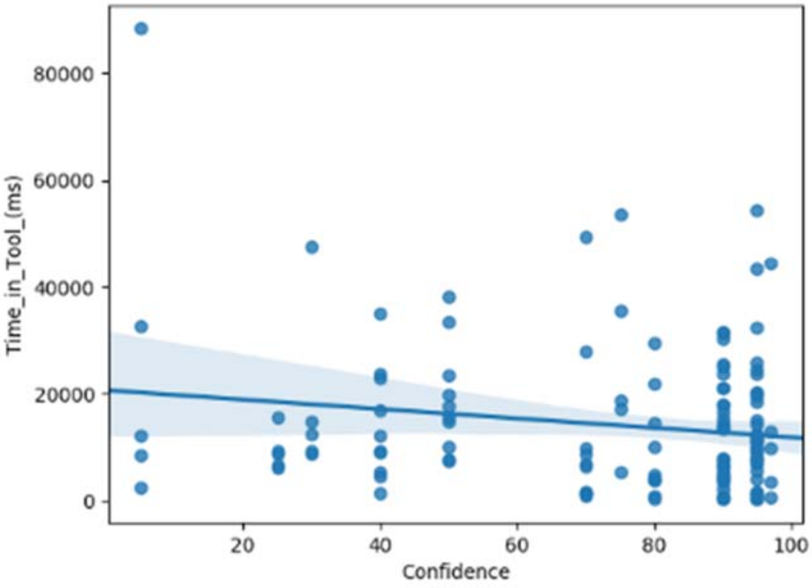


Figure 20: Confidence vs Time-in-Tool – Participant 1121

For the other participants, there was no statistical difference in their time-in-tool performance. Therefore further data exploration was necessary to find other patterns of behaviors which could be used to estimate confidence.

4.2.3.2 Analyzing Time-to-Decision

Figure 21 was plotted to illustrate the relationship of difficulty with the time to make a decision for each participant. Figure 21 seems to suggest that higher difficulty does not necessarily map to longer decision times. The initial speculation for this phenomenon was that Very Hard alerts may not have as much information readily available for the participant, thus causing the investigation to be shorter relative to the other difficulties.

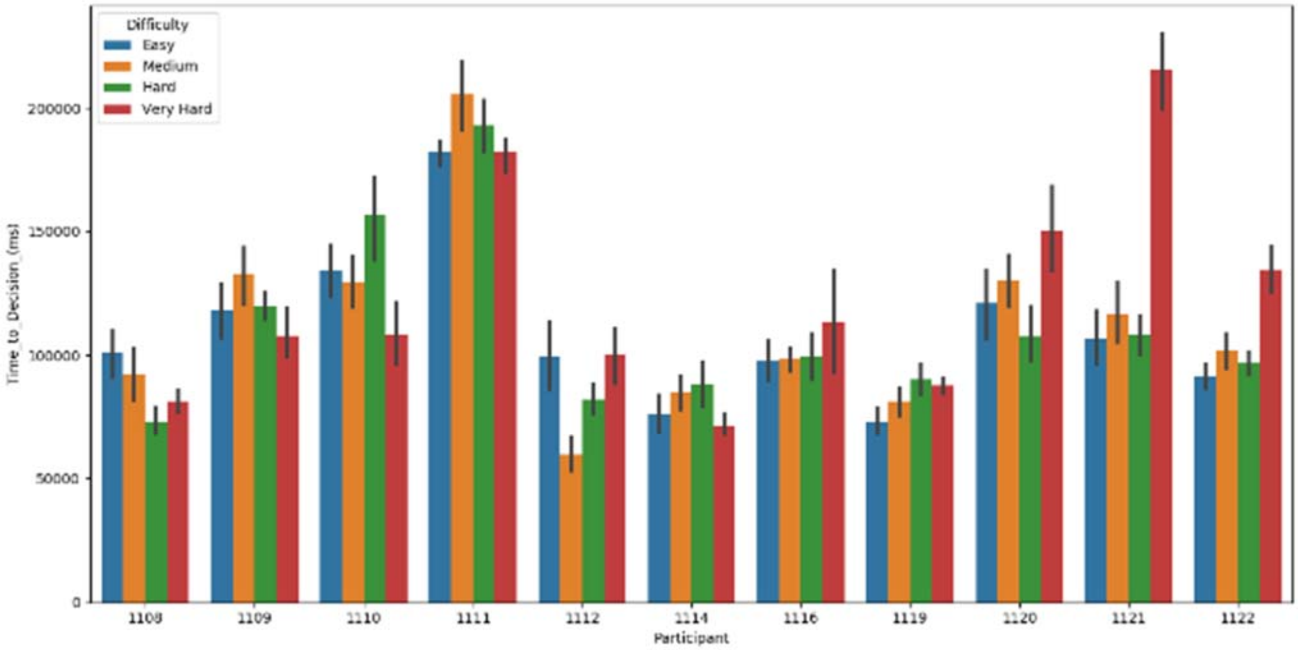


Figure 21: Time-to-Decision versus Participant (per Difficulty)

Alternatively, the relationship of time-to-decision and alert difficulty per participant could have something to do with confidence, which is another comparison that needing to be reviewed and interpreted. The trend lines in Figure 22 suggest that when participants are rating alerts with a lower confidence they tend to take a longer time to decide on their actions. The majority of the peaks in Figure 22 are Very Hard alerts, which actually ends up refuting our initial speculation on Very Hard alerts tending to take less time over all to decide on.

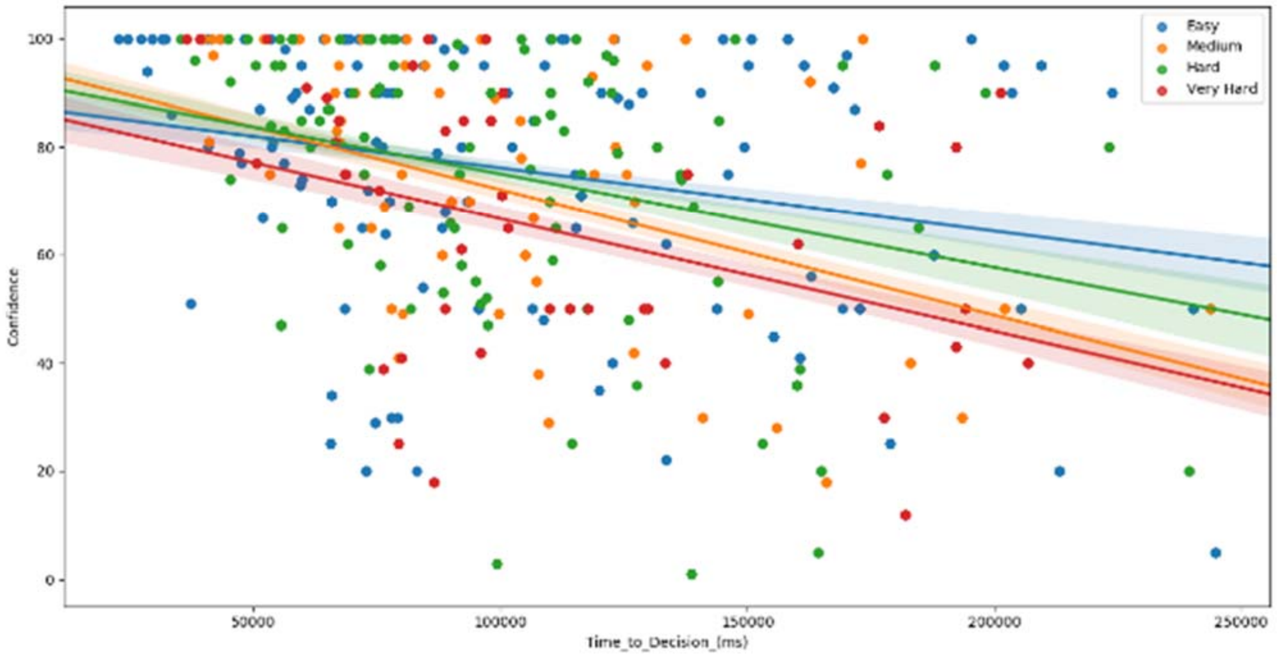


Figure 22: Time-to-Decision versus Confidence (per Difficulty)

A one-way ANOVA was used to confirm whether the time-to-decision had a statistically significant effect on reported confidence for participants. Six participants showed statistical significance, and the results are displayed in Table 8.

Table 8: One-way ANOVA for Time-to-Decision Based on Confidence Scores

	alpha = 0.05	Fcrit = 4.20
	df = 28	
	Time-to-Decision	
Participant #	**F-value**	**P-value**
1108	4.922	0.03480
1109	6.080	0.02006
1110	8.981	0.005661
1114	8.515	0.006871
1116	35.74	0.000002
1121	34.13	0.000003

Data exploration continued to compare and contrast the results across all of the participants, in order to identify if there was a generalization which could be made from the time-to-decision and confidence. Looking at all participants in the experiment, see Figure 23, there seems to be a downward trend overall in terms of time-to-decision regardless of difficulty. A stabilization of the time-to-decision did not seem to occur, also illustrating that the participants could become faster as they grow familiar with the task. Overall, training effects are acknowledged, and attempts were made to mitigate them, such as providing a training day and various alerts to practice investigating before the actual experiment. Confounding variables such as the participant fatigue with the length and rigor of the test, may need to be accounted for in future studies. The participant's cyber experience, which was expected to be a confounding variable, showed no effect on the time to decision.

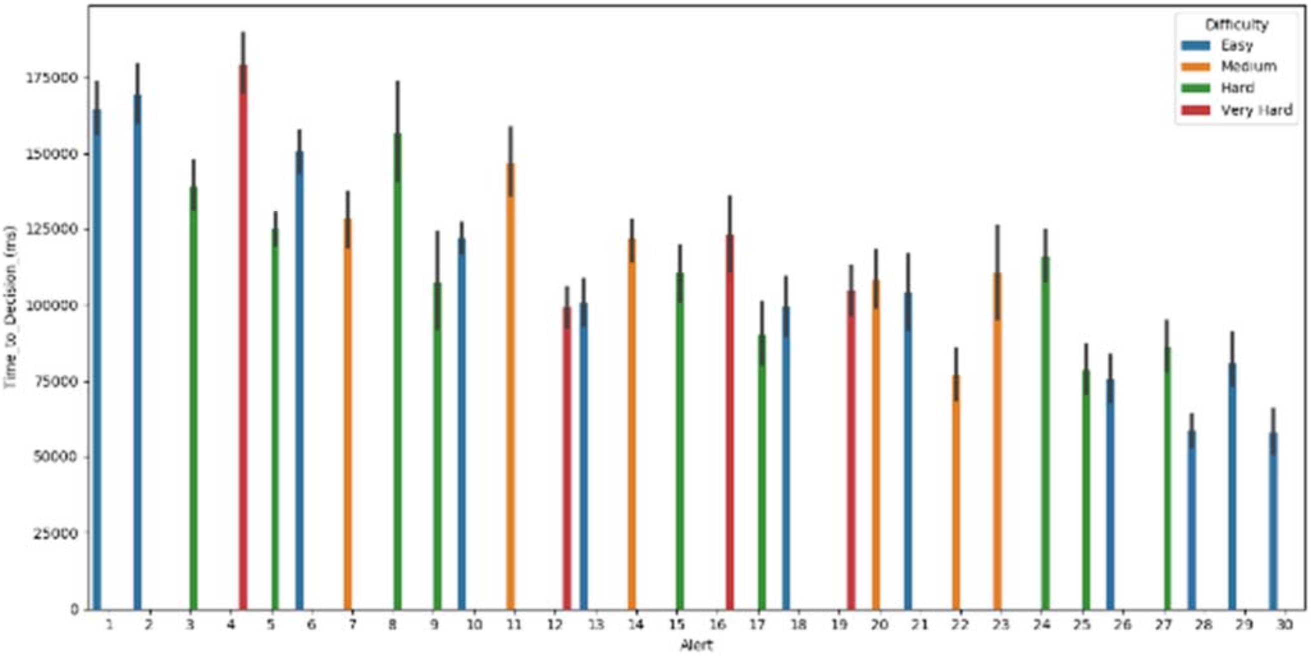


Figure 23: Time to Decision versus Alert (per Difficulty)

4.2.3.3 Analyzing Transitions

The next phase of analysis involved creating and evaluating transition probability matrices. Each transition probability matrix was constructed by summing the total amount of tool uses, while keeping track of the last used tool. These transition values were then graphically represented as heat maps. These heat maps visually illustrate the frequency of tool transitions. The heat maps show the quantity of transitions from the tool identified in the row to the tool identified in the column. The heat map does not identify which tool was the first or last used in the workflow.

A heat map was generated for each participant, by each alert. The alerts were first grouped by difficulties and reviewed. These heat maps were then reviewed across all participants, based on difficulty of the alerts. The intent behind reviewing the tool transitions was to determine whether the workflow process, printed on a sheet of paper and given to the participants during the experiment, was followed. The purpose of the workflow was to aid all participants, especially those without any cyber alert knowledge or experience.

Figure 24 shows how a strict adherence to the workflow would look like, assuming the participant was already familiar with all of the terms in the glossary and did not need to consult it. Figure 24 also assumes the participant would be starting from the Alert Lookup tool, as is specified in the workflow handout, see Appendix F. The glossary was removed from the strict heat map, as looking up a keyword or abbreviation could occur at any time and would make a generalized workflow impossible to construct.

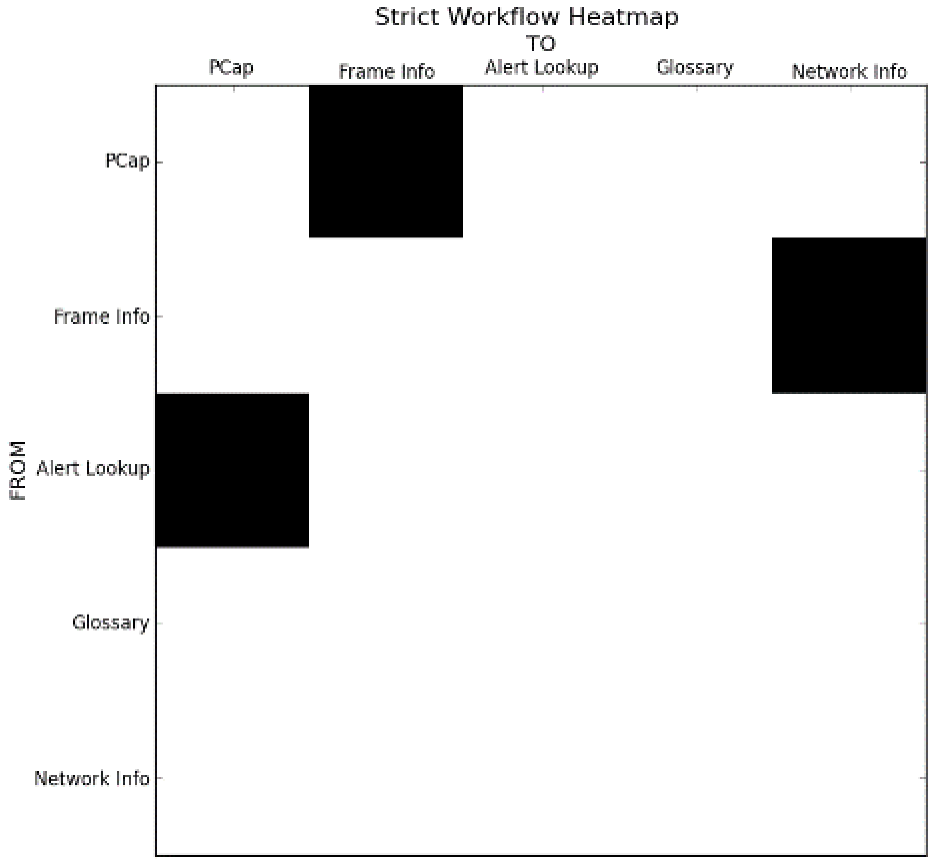


Figure 24: Strict Workflow Tool Transitions

The strict workflow heat map provided a baseline by which to compare the workflow process for all of the participants combined together, broken out by difficulty, or broken out per participant and each specific round. A heat map combining the workflow activity of all participants was created, see Figure 25. This provided a visible representation of the workflow process conducted by each participant across all 30 alerts. The darker colors represented heavier transitions from and to tools. The heaviest, and most frequent, tool transitions were from Alert Lookup to PCap, Frame Info to PCap, and PCap to Frame Info. Conversely, the lighter colors indicated less frequent tool transitions. Transitions from the Glossary to Frame Info, Glossary to PCap, and Frame Info to

Glossary were the least frequent transitions overall for participants, see Figure 25. This combined heat map provided interest for observing tool transitions by counting each of the tool transitions.

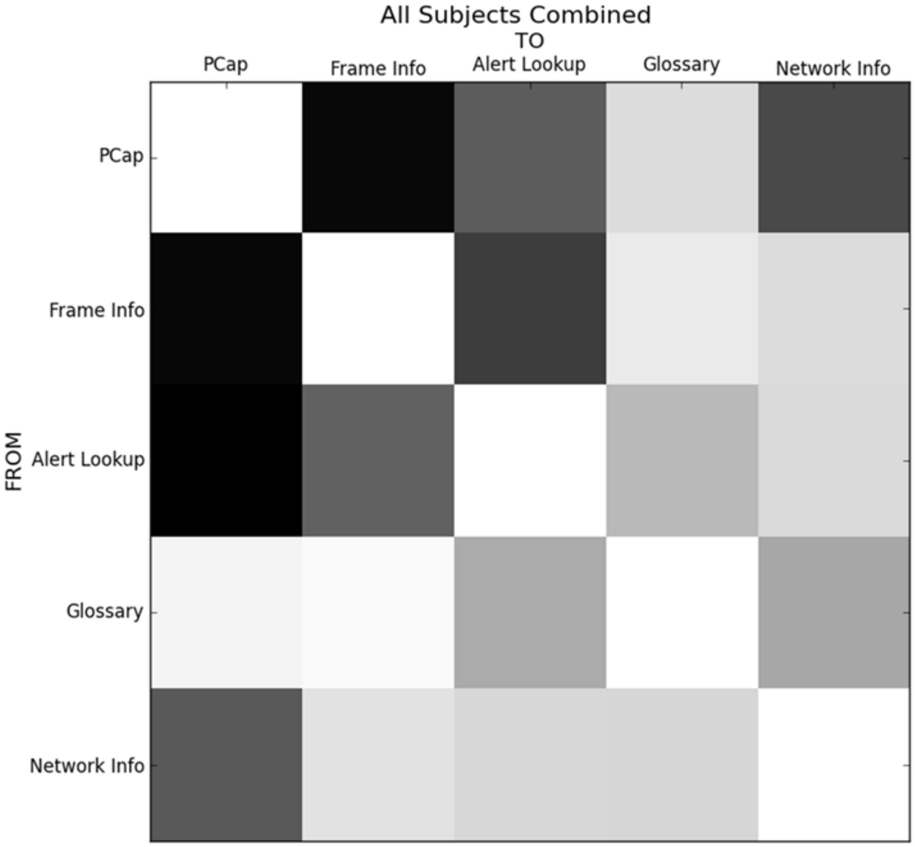


Figure 25: Combined Workflow of All Participants

It was determined that omitting the glossary uses made it easier to identify patterns of tool usage, as the infrequent use would be because of the learning effect by which participants are becoming more familiar with terms as they proceed in the experiment. Even by including the glossary tool, in some heat map samples, it shows up as only a few transitions. As the rounds progressed, and the participant completed more alerts, the usage of the glossary tool dropped, along with one other tool. Tool transitions into the Network

Info tool exhibited a noticeable drop, similar to the glossary, when looking at all of the participants across the rounds. This could mean that the information in both the Glossary and Network Info tools were becoming familiar to the participants. This explanation can be reinforced by understanding the information available from the tools.

Five tools which were available to participants, three of the tools were static information and two were dynamic. The two dynamic tools, i.e. changing the displayed information with every alert, were the PCap and Frame Info tools. The three static tools were Alert Lookup, Glossary, and Network Info. The Alert Lookup would be most likely be required to be reviewed on every alert, in order to explain the definition of the alert, whereas the Glossary and Network Info tools could be omitted in later alerts as the information did not change and was able to be imparted in short term memory of the participants, e.g. the learning effect.

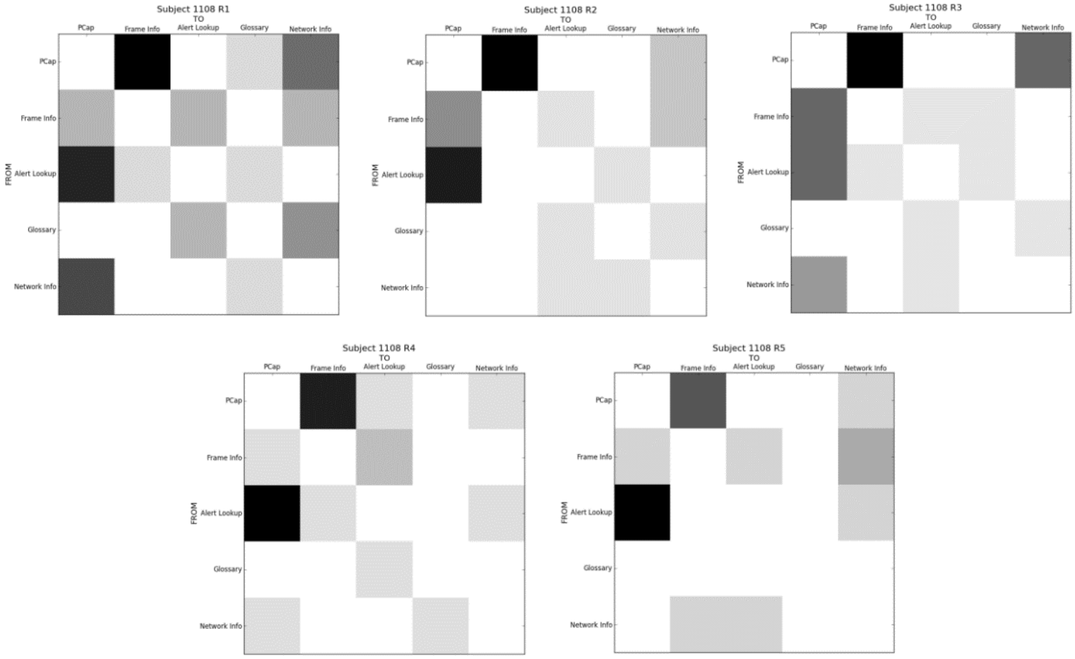


Figure 26: Participant 1108 Heat Maps by Round

The drop off of Glossary and Network Info usage was further confirmed when reviewing all of the participant's heat maps, based on the five rounds, although no statistical significance calculations were done to confirm this. Figure 26 showcases the first participant in the experiment, which highlights the drop in tool transitions to the Glossary and Network Info tools as the rounds proceed. This result seemed to carry over across all 11 participants, leading the future transition counts omitting transitions across Glossary and Network Info tools, in order to account for the learning effect.

Figure 27 shows the average tool use, across all participants, based on the difficulty of the alert. Tool usage seemed to be the highest, on average, across participants when looking at Very Hard alerts. Further analysis will be conducted on tool usage, in relation to the time spent in tools, later on in analysis.

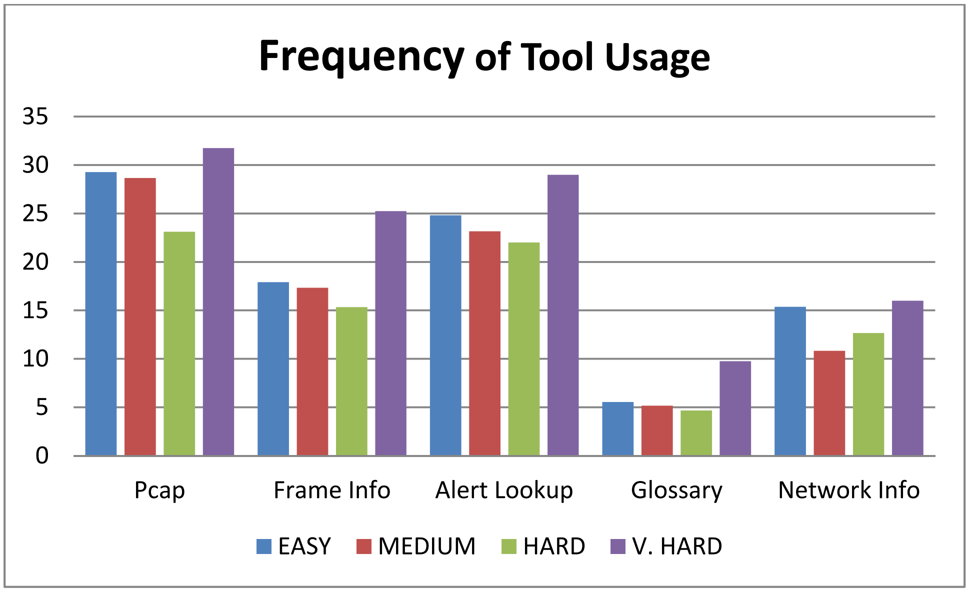


Figure 27: Frequency of Tool Usage by Amount of Alerts (per Difficulty)

4.2.3.4 Analyzing Tool Transition Counts

The next analysis effort was on tool transition counts and identifying whether they related to changing difficulties and confidence. It was hypothesized that the frequency of tool transition would correlate to lower or higher confidence levels in within-subject comparisons, while not necessarily being broad enough to relate to between subject comparisons. In Figure 28, transition counts trended upwards as difficulty increased for participants: 1111, 1112, 1114, 1119, 1120, 1121, and 1122.

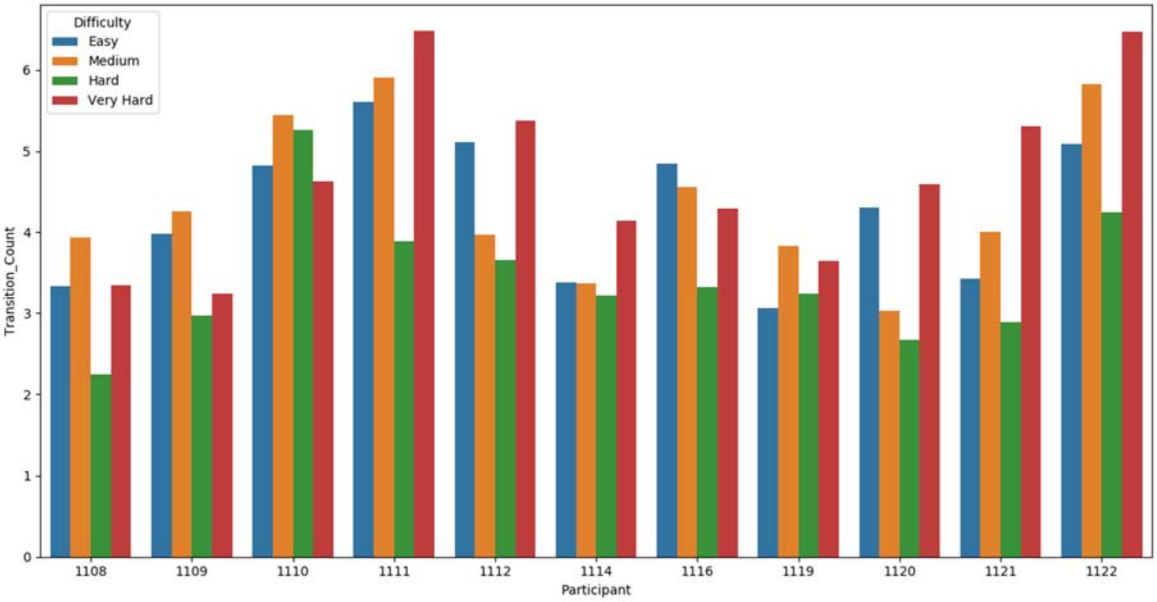


Figure 28: Participant versus Average Tool Transition Count (per Difficulty)

For all but two participants, the transition counts for Hard alerts were the lowest. Participant 1110's lowest average transition count occurred with Very Hard alerts, whereas Participant 1119's lowest average transition count occurred with Easy alerts. Both participant 1110 and 1119 had cyber security experience and cyber certifications, but the four other participant's with cyber security experience had the lowest transition counts with Hard alerts like the majority of the participant population. Other confounding factors may have played a role in the lower transition counts of Hard alerts, including the small amount of Very Hard alerts. Accounting for learning effects by counterbalancing the ordering of the alerts may explain this anomaly. Further analysis will be recommended in future work.

Table 9: Average Transition Count Based on Difficulty and Confidence Bins

		Average Transition Count			
Participant	Difficulty	CI (0,25]	CI (25,50]	CI (50,75]	CI (75,100]
1108	Easy	2	3.81	N/A	2.94
1108	Medium	N/A	N/A	2	4.13
1108	Hard	N/A	1.67	2.14	2.33
1108	Very Hard	3	N/A	3.45	N/A
1109	Easy	5.18	4.07	3.36	N/A
1109	Medium	6.08	3.78	3	3
1109	Hard	2.78	2.67	3.23	2.25
1109	Very Hard	4.22	3.91	2.25	1.6
1110	Easy	N/A	4.94	5.63	4.42
1110	Medium	N/A	6.4	N/A	3.84
1110	Hard	7.3	N/A	N/A	3.83
1110	Very Hard	N/A	N/A	N/A	4.63
1111	Easy	6.61	3.86	5.84	5.29
1111	Medium	N/A	6.4	3.57	5.68
1111	Hard	3.57	N/A	2.83	4.17
1111	Very Hard	N/A	6.78	N/A	5.45
1112	Easy	N/A	N/A	4.33	5.2
1112	Medium	N/A	N/A	4.75	3.38
1112	Hard	N/A	4.18	3.71	3.44
1112	Very Hard	N/A	6.09	N/A	3.27
1114	Easy	N/A	N/A	1.25	3.53
1114	Medium	N/A	5.36	N/A	1.69
1114	Hard	N/A	4.62	N/A	2.67
1114	Very Hard	N/A	3.57	5.45	3.1
1116	Easy	N/A	5.37	5.86	4
1116	Medium	N/A	6.23	N/A	3.15
1116	Hard	N/A	3.89	3.19	2.87
1116	Very Hard	5.47	N/A	N/A	3.37
1119	Easy	3.25	N/A	3.04	3.03
1119	Medium	N/A	3.57	4.59	3.05

1119	Hard	N/A	2	3.73	3.14
1119	Very Hard	N/A	3.47	3.8	N/A
1120	Easy	N/A	2.33	1.33	4.65
1120	Medium	N/A	N/A	2.55	3.25
1120	Hard	1.2	2.2	4.24	1.96
1120	Very Hard	N/A	2.33	N/A	5.07
1121	Easy	N/A	3.88	3.17	3.39
1121	Medium	N/A	3.7	4.89	3.35
1121	Hard	3	N/A	N/A	2.88
1121	Very Hard	5.12	5.46	N/A	N/A
1122	Easy	4.8	5.43	7.06	4.58
1122	Medium	N/A	4.33	7.39	2
1122	Hard	N/A	N/A	4.69	3.87
1122	Very Hard	N/A	7.3	6.46	5.79

Table 9 suggests that lower reported confidence mapped to higher transition counts, for several participants. The table represents the transition counts, broken out across bins of confidence scores, in order to illustrate the disparity in confidence scoring for certain alert difficulties.

For five participants: 1109, 1110, 1114, 1116, and 1121, the transition counts were statistically significant based on the confidence scores, see Table 10.

Table 10: One-way ANOVA for Transition Count Based on Confidence Scores

	alpha = 0.05	Fcrit = 4.20
	df = 28	
	Transition Count	
Participant #	F-value	P-value
1109	7.262	0.01177
1110	17.38	0.000266
1114	4.228	0.04918
1116	7.092	0.01269
1121	6.778	0.0146

The effect of transition counts on confidence, for each participant, respective of the alert difficulty is displayed as Table 11.

Table 11: Transition Count and Confidence per Difficulty (by Participant)

Participant	Difficulty	Tool Transition Count	Confidence
1108	Easy	3.333333	68.14667
1108	Medium	3.939394	85.06061
1108	Hard	2.243243	84.81081
1108	Very Hard	3.346154	61.69231
1109	Easy	3.979381	45
1109	Medium	4.26087	42.47826
1109	Hard	2.966102	44.25424
1109	Very Hard	3.242424	46.51515
1110	Easy	4.818966	71.89655
1110	Medium	5.447761	59.9403
1110	Hard	5.2625	63.5125
1110	Very Hard	4.630435	88.1087
1111	Easy	5.607692	68.91538
1111	Medium	5.902778	62.875
1111	Hard	3.888889	80.76389
1111	Very Hard	6.479167	53.95833

1112	Easy	5.108434	90.14458
1112	Medium	3.972973	85.32432
1112	Hard	3.662338	71.16883
1112	Very Hard	5.372093	62.7907
1114	Easy	3.380952	92.63492
1114	Medium	3.375	75.04167
1114	Hard	3.217391	78.17391
1114	Very Hard	4.142857	74.53571
1116	Easy	4.842105	77.42105
1116	Medium	4.5625	63.22917
1116	Hard	3.322034	65.11864
1116	Very Hard	4.294118	51.5
1119	Easy	3.056338	71.1831
1119	Medium	3.833333	71.64583
1119	Hard	3.246377	76.82609
1119	Very Hard	3.648649	52.18919
1120	Easy	4.303371	92.30337
1120	Medium	3.028571	92.14286
1120	Hard	2.666667	75.17647
1120	Very Hard	4.588235	86.17647
1121	Easy	3.424658	85.27397
1121	Medium	4	71.35417
1121	Hard	2.887097	81.67742
1121	Very Hard	5.311475	34.7541
1122	Easy	5.087302	70.83333
1122	Medium	5.823529	58.72549
1122	Hard	4.25	81.20238
1122	Very Hard	6.473684	63.42105

4.3 Electrophysiological Analysis and Results

Preliminary analysis of the recorded EEG data was conducted to ensure all of the channels were recorded through all of the participant's trials. EEGLAB, the primary MATLAB plugin used for analyzing EEG data, only showed a handful of channels upon inspection. Further constraints, such as unfamiliarity with electrophysiological analysis, led to the EEG results being compiled and cataloged but not analyzed. Therefore, this analysis will end up being exclusively left to future work.

4.4 Summary

It was hypothesized at the beginning of this chapter that that ordering of tool use, the time per tool, and the overall time to a decision would correlate to the differences when comparing the confidence of each alert decision, per participant. The analysis and results show, with statistical significance, that not only was time per tool, measured at time-in-tool, important, but the time-to-decision and tool transition count were all behaviors which affected reported confidence. Data exploration of the behavior data extracted from CIAT allowed for seven of the eleven participants to have behaviors mapped to their confidence. Four participants did not illicit a behavior pattern which could be identified with the analysis methods covered above. In addition, the ordering of tool use was not able to be validated with statistical significance, and is left to future work, see more in Section 5.3.4.

With only 30 alerts available to analyze, a larger data set of participants and alerts may lead to other factors becoming more relevant for identifying and mapping behavior to confidence.

The behaviors extrapolated from this research are specific to this synthetic task environment, and would need to be generalized and applied to other tools in order to expand the results across all cyber based investigations. Since this study only involved 11 participants it would be worthwhile to conduct an experiment with more participants, to determine if these behaviors can be generalized given a larger sample size.

In summary, the behavioral and subjective analysis led to the observation and statistical validation of three behavior factors which effect reported decision confidence. Dependent on a larger sample size and analysis of the EEG measurements, the findings of the behavioral analysis already allows for identifying behavior mechanics, specific to each participant, which map to reported decision confidence.

V. Conclusions and Recommendations

5.1 Conclusions of Research

The behavioral analysis of both the pool of participants and each participant specifically, allowed for the identification of key behavioral factors, which correlated with confidence. Table 12, displays a summary of the three behaviors analyzed in Chapter 4. The statistical significance of the ANOVA results are explained in the associated behavior analysis portions of Section 4.2.3. One participant's confidence, participant 1121, correlated with all three analyzed patterns of behavior. Additionally, four participants, participant 1111, 1112, 1119, and 1122, exhibited no behavioral effects on their confidence.

Table 12: Behaviors Which Effect Confidence

	Behavioral Correlation with Confidence (as confidence increases)		
Participant #	Time-in-Tool	Time-to-Decision	Transition Count
1108	~	↓	~
1109	~	↓	↓
1110	~	↓	↓
1111	~	~	~
1112	~	~	~
1114	~	↓	↓
1116	~	↓	↓
1119	~	~	~
1120	↓	~	~
1121	↓	↓	↓
1122	~	~	~

LEGEND	
↑	Increased
↓	Decreased
~	No effect

The time-to-decision behavior influenced the confidence in six out of eleven participants. Based on the results, no generalization of what effects of the tested population can be made at this time, based on the analyzed behavior patterns.

This conclusion assumes that the overall understanding of how confidence and behavior effect the formulation of a decision is correct. The alert difficulty was the factor which was varied in the experiment. This led to behavior and an associated confidence, which ultimately lead to a decision by the participant. The factors which effected behavior were attributed to time, both in tool usage and overall decision time, and the transition count among the available tools. Decision confidence was recorded after each decision was made. With future electrophysiological analysis, the goal would be to determine what specific behaviors correlate to increased or decreased confidence up to the point of a decision being made.

Answering RQ1, the results of this study indicate that three key behavioral factors correlated with participant confidence during the formulation of participant decisions. These patterns of behavior were Time-in-Tool, Time-to-Decision, and tool Transition Counts. Seven of the eleven participants in this study exhibited one or more of these patterns of behavior.

Answering RQ2, the results showed that even when participants were given time to practice and a workflow process to follow, they would deviate from the workflow regardless of confidence. Future analysis will need to be completed, to determine whether there is a statistical significance to workflow tool usage, as this could indicate reliance on experience or familiarity with the tools and investigation. The captured EEG data may reveal insight into behaviors associated with tool transitions and tool usage. Although five tools were available, the tools were not always reviewed in the same order or in the same amount, providing further behavioral differences when compared to the reported confidence levels during investigations. Statistical analysis will need to be completed in order to make a conclusion about tool usage ordering.

Answering RQ3, the results show that behavior patterns correlated with increases in decision confidence, but the converse needs be confirmed with further statistical analysis. Namely, the next logical question is: what behavior patterns are associated with a low confidence decision? Time-to-Decision decreased for participants who more confident in their decisions, possibly brought about by not having to spend arduous amounts of time repeatedly going over the same tools. Lower tool transition counts, and the associated time in these tools, also mapped to higher reported confidence in several participants. Participant 1121 exhibited all three behaviors with increased confidence.

Answering RQ4, the electrophysiological will be evaluated in future research, thus any quantifiable differences in EEG metrics are unable to be confirmed at this time.

Notwithstanding the EEG analysis, to reiterate the findings from RQ3, three distinct behaviors were observed to occur when participants were in lower confidence situations. Likewise, the inverse of these results showcases that higher confidence decisions tend to occur when decisions are made faster, relative to other alerts. The speed to a decision, should not be taken as the only behavior though, as this could lead to inaccuracy if purely looking at time, although this quantifiable measurements it the easiest to compare within and between participants.

5.2 Significance of Research

Thanks to the modularity of the CIAT STE, the additional features supported in CIAT 2.0, will allow for future studies in behavioral analysis to be completed with simple alterations to the dataset in the Microsoft Access database. Keeping with the modular design of CIAT, CIAT 2.0 allows for rounds to be of varying alert amounts, and for the addition or subtraction of tools for the users. Additionally, CIAT 2.0 enables EEG collection, as it supports timing and signal forwarding to the Cognionics Data Acquisitions suite of tools.

This study indicates that the investigative process in cyber defense, which ultimately leads to a decision based on varying degrees of confidence inferred from tool review and task understanding, requires further analysis to better understand how human behavior may be measured and analyzed. Some behavioral assertions can be extrapolated by only reviewing the workflow process or the self-reported metrics from simple questionnaires, but the underlying physiological activity may provide a keener insight into the degree to which data analysis and the investigative process effects the decision action and the associated confidence in this decision.

The primary significance of this study was the collection of human participant behavior, physiological data, and decision confidence from 11 participants, while they investigated and decided dispositions for cyber alerts. With this data, further behavioral analysis can be conducted, especially as it relates to physiological analysis, as no other cyber defense studies have looked into the physiological data of participants.

5.3 Recommendations for Future Research

5.3.1 Design Changes

Some design changes may be worth implementing in order to limit or eliminate confounding variables. During data exploration and analysis, it became apparent that several design decisions had created situations in which the participant's transition between the tools used during an investigation could not accurately account for their initial tool selection. This was because completing the investigation for a previous alert, and making a decision, did not reset the tool selection. The tool selection was left as the previous alerts information, thus if the tool was the PCap or Frame Info, the participant would already have new information displayed without having to manually request it by selecting the tool. This had the possibility of skewing the actual tool transition statistics toward less tool transitions. A remedy for this scenario would reset the entire tool selection area of CIAT to be empty, whenever a new alert is selected, thus forcing the participant to intentionally select their first tool each time they work on a new alert.

The structuring of six alerts per round in CIAT was intended to limit and balance the amount of information presented to the participant at one time, while also separating rounds by a consistent amount, given 30 total alerts and the experiment's 2-hour time limit. During the design of the synthetic task environment, the decision to display six alerts per round was justified in order to give a manageable amount of workload requiring the participant's focus and attention, prior to a short built-in break. The decision confidence ranking was structured as a validation of the previous decisions made during the investigative phase of the experiment, since the numerical confidence values were not displayed for the participant. This prevented the participant from simply ordering the numbers from largest to smallest, as well as forcing them to break any ties by arranging the alerts in order from most confident to least confident per round. One simple design change could be implemented in order to limit distractions. Since the alerts are all designed to be self-contained and independent, the interface could be structured to only display one alert at a time. This change, bundled with the reset of the tool selection would force the participant to make each investigative decision without any prior set tool. Additionally, this would prevent the participant from selecting an alert, reading some of the information, and changing to another alert.

Another benefit of only displaying one alert at a time to the participant, would allow for a restructuring of the alerts per participant. This would counter the learning effect currently observed in the data, as each participant could be configured to see a different preset ordering of the alerts. As mentioned previously, the investigation time did not plateau or trend to a specific bound for any of the participants during the 30 alerts. More alerts or a preset ordering of the alerts, could assist in determining the lower bound of investigative time required to make a decision, as well as countering the learning effect. Additionally, since the ordering of the alerts presented to the user would be controlled, this would eliminate differences in investigation timing, better accounting for possible confounding variables such as participant fatigue.

A recommended change to the ordering of the alerts, would help account for the learning effect. By counterbalancing the presentation of the alerts to the participants, it would become possible to review the alerts without having to look for a performance plateau. This would account for the learning effect or possible fatigue, which may explain why the time to decision continued to decrease in general throughout the experiment.

5.3.2 EEG Analysis

Since this study's analysis omitted EEG analysis, the data still needs to be reviewed. Everything from clicks and timing to decision and accuracy may end up providing additional behavioral factors which may validate whether the behaviors extrapolated from the subjective and behavioral analysis were valid.

5.3.3 Participant Selection for Future Trials

The participant selection pool was greatly limited for this research. Not only were all participants AFIT students or employees, but there were no female participants nor a sizeable amount of participants with cyber backgrounds in cyber defense. Future studies should extend the findings in this research by recording the behavior of those in the cyber defense community with this expertise. The pool of participants could be broken into groups based on time certified on cyber defense tools, and their level of computer or cyber security certifications. Participation from an Air Force cyber defense unit would garner additional insight into the types of tools and tasks which make up the investigative process, especially based on tool usage frequency and the general workflow or process dictated by the unit's job. Ultimately, there may not be much of a discrepancy between the efficiency or confidence of participants, when comparing those with cyber defense expertise versus those without, due to the provided training and self-contained structure of this experiment. This would validate the efficacy and usefulness of this synthetic task environment, allowing for continued modifications and modularity to pinpoint what factors affect the decision confidence of operators. For example, complementary tools may end up having a greater effect on investigation time, and therefore decision confidence, if it requires the operator to frequently switch between tools. Distinct tools, on the other hand, may make it easier to progress through an investigation in a workflow style, thus increasing decision confidence or accuracy.

5.3.4 Other Data Analysis

Continued analysis of the tool transitions could possibly explain the time-in-tool results. If statistical analysis is done for the tool transitions, it should also investigate the originating and final tool used for each investigation. Analysis of the originating and final tools used for an investigation was not investigated in this study.

Some tool pattern usage analysis was done, but no significant results were found, in part due to difficulties in establishing a method for comparing the order of tool usage per alert and per participants. One method for analyzing the transition probability matrices is to conduct distance measurements between each alert and round for each participant. Each tool transition count, made up of each source and destination tool pairing, can be compared to other alerts by calculating the Euclidean distance. This matrix comparison method was proposed, but due to time limitations it will be left as proposed future work.

5.3.5 Other Recommendations

Current systems exhibit either a machine-feeds-human or human-feeds-machine-feeds-human style of network defense, for most if not all general computer interaction. An example of machine-feeds-human can be illustrated by the relationship humans have with IDS devices. These host and network IDS devices rely on humans to make the decision on information recorded, organized, and displayed to human users. Human-feeds-machine-feeds-human situations are those in which a host and network intrusion prevention system (IPS) relies on the human feeding the machine rules and evaluation criteria, by which the system takes action, although the human is available at any moment to modify the criteria based on feedback. Knowing these limitations, the motivation for this research is to improve trust and confidence with the systems human operate, by allowing the machine to monitor the human and identify when the investigation process was compromised by poor analysis behavior. This would allow a machine to augment the human in any computer-focused task, as long as there is a sufficient baselining or patterns of behavior to extrapolate.

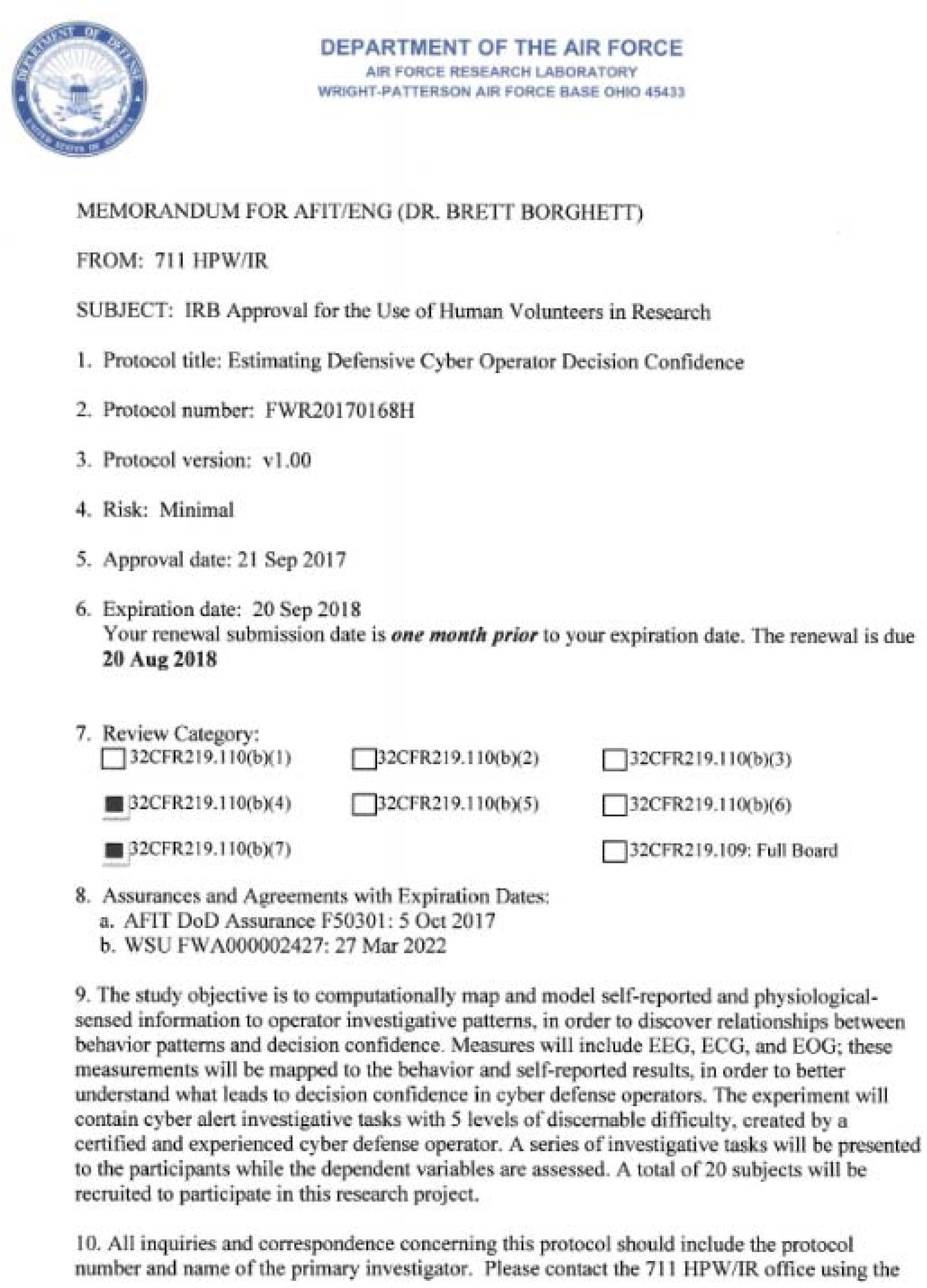
Monitoring and improving decision confidence enables consistent and expedited effectiveness in those people training on these cyber defense tools, as well as the ability to extend monitoring of decision confidence to quality assurance capability for those currently operating on the tools. Decision confidence can lead to improved quality assurance and work output, with minimal overhead, thanks to the computer agents that assist in determining confidence metrics from tool usage, timing, and even the humans' write-up. These three areas, quality assurance, training, and tool usage, can all be measured using a subset of the methodologies illustrated above. Additionally, it would make the most operational sense to task experienced human users, or team leads, as they would be able to review or share in making the final decision on ambiguous or alerts where low decision confidence is estimated. This research, due to its focus on human efficacy, should apply to any cyber defense tool or process, as long as human-in-the-loop decisions are required, as their decision-making and decision-confidence will always play a role in software.

5.4 Summary

In summary, this research fills an important gap in the literature regarding understanding the decision confidence of cyber defense analysts by looking at behavior patterns while they conduct investigations. The electrophysiological data may provide additional insight into how cyber analyst's behaviors may be influenced outside of what can be recorded from a computer interface. The decision-making process relies on confidence in the tools, but more heavily on the experience and understanding of the analysts who carry out reviewing the data. The identified behavior patterns allow for an estimation of decision confidence in regard to cyber based alert investigations. With an understanding of the behavior and estimated confidence level of analysts, assistive tools and techniques can be implemented to allow for quality assurance, tailored training, and other enhancements for the cyber warfighter.

Appendices

Appendix A: IRB Approval Letter



DEPARTMENT OF THE AIR FORCE
AIR FORCE RESEARCH LABORATORY
WRIGHT-PATTERSON AIR FORCE BASE OHIO 45433

MEMORANDUM FOR AFIT/ENG (DR. BRETT BORGHETT)

FROM: 711 HPW/IR

SUBJECT: IRB Approval for the Use of Human Volunteers in Research

1. Protocol title: Estimating Defensive Cyber Operator Decision Confidence

2. Protocol number: FWR20170168H

3. Protocol version: v1.00

4. Risk: Minimal

5. Approval date: 21 Sep 2017

6. Expiration date: 20 Sep 2018
   Your renewal submission date is ***one month prior*** to your expiration date. The renewal is due **20 Aug 2018**

7. Review Category:
   □32CFR219.110(b)(1) □32CFR219.110(b)(2) □32CFR219.110(b)(3)
   ■32CFR219.110(b)(4) □32CFR219.110(b)(5) □32CFR219.110(b)(6)
   ■32CFR219.110(b)(7) □32CFR219.109: Full Board

8. Assurances and Agreements with Expiration Dates:
   a. AFIT DoD Assurance F50301: 5 Oct 2017
   b. WSU FWA000002427: 27 Mar 2022

9. The study objective is to computationally map and model self-reported and physiological-sensed information to operator investigative patterns, in order to discover relationships between behavior patterns and decision confidence. Measures will include EEG, ECG, and EOG; these measurements will be mapped to the behavior and self-reported results, in order to better understand what leads to decision confidence in cyber defense operators. The experiment will contain cyber alert investigative tasks with 5 levels of discernable difficulty, created by a certified and experienced cyber defense operator. A series of investigative tasks will be presented to the participants while the dependent variables are assessed. A total of 20 subjects will be recruited to participate in this research project.

10. All inquiries and correspondence concerning this protocol should include the protocol number and name of the primary investigator. Please contact the 711 HPW/IR office using the

organizational mailbox at AFRL.IR.ProtocolManagement@us.af.mil or calling 937-904-8094 [DSN 674].

Rhonda C. allen For:

KIM E. LONDON, JD, MPH
Chair, AFRL IRB

1st Indorsement to AFIT/ENG (DR BRETT BORGHETTI), Approval for Use of Humans in Research, Expedited Review, FWR20170168H

MEMORANDUM FOR AFMSA/SGE-C

This protocol has been reviewed and approved by the AFRL IRB. I concur with the recommendation of the IRB and approve this research.

21 SEP 2017

TIMOTHY J. SAKULICH
Vice Director
711th Human Performance Wing

Appendix B: Blank Informed Consent Document (ICD)

Estimating Defensive Cyber Operator Decision Confidence
FWR20170168H v1.00

INFORMATION PROTECTED BY THE PRIVACY ACT OF 1974

Informed Consent Document
For
Estimating Defensive Cyber Operator Decision Confidence

Air Force Institute of Technology, AFIT/EN Human Systems Integration Laboratory
Building 640, room 340, Wright Patterson AFB, OH.

1. Principal Investigator

Dr. Brett Borghetti/AD-22/Associate Professor, Department of Electrical and Computer Engineering/AFIT/ENG, DSN 785-3636x4612, brett.borghetti@afit.edu

Nature and Purpose: You have been asked to volunteer to participate in the research project named above.

The purpose of the study is to study the relationships between decision confidence, brain signals and patterns of investigative behavior.

Duration: Up to 2 hours on two separate days within two weeks.

Experimental Procedures: The entire experiment will be completed on a PC-based computer where your responses will be recorded and used in future analyses. A mouse and keyboard will be used to interact with a computer while experimental materials are displayed on screen. The experiment is comprised of three phases: training, baseline measurement, and evaluation. You will train on, and make decisions on cyber security alerts. You will also answer a few questions before and after the evaluation phase. You will remain seated for the duration of the experimental session.

Day 1 is a training day where you will learn how to perform the tasks and make decisions. On Day 2, you will make decisions on a set of cyber alerts on your own. In addition, on Day 2 you will complete a pre-experiment questionnaire asking about your sleep, caffeine use, and readiness for the experiment. On Day 2 only, after completing the experiment, you will complete a post-experiment questionnaire asking about your computer experience and a few demographic questions.

During the Training Day (Day 1), you will learn about and practice the experimental tasks.

On Day 2, several sensors will be attached to your body. An Electroencephalograph (EEG) head cap will be applied to measure brain activity. Sensors placed near your eyes for Electrooculography (EOG) will measure eye movement and blink signals while sensors on your chest will record heart information using Electrocardiography (ECG). After applying and setting up the sensors, you will rest for a brief period so we can obtain a sensor baseline, then you will complete experiment activities during the decision-making phase. During the decision-making Phase, you will alternate between decision-making activities and assessing how confident you were about each of your decisions.

Please note that you are free to withdraw from this study at any time, for any reason without penalty. All data from this experiment are associated only with a participant number in the computer (not your name or other identifying information), so you can be assured of confidentiality. Your email contact information is stored separately from the data collected during the experiment. Following the Day 2 session you will be debriefed and given the opportunity to ask questions and provide comments.

Inclusion criteria. You must be able to use your right-hand to operate a mouse in order to participate in the study.

Exclusion Criteria: You are not eligible to participate in the study if you meet the following criteria:

- Unable to use a mouse in right hand
- Visual impairment or inability to view information on a computer screen
- Specific motor, perceptual, or cognitive conditions that preclude you from operating a computer, reading small characters on a computer monitor, or hearing and comprehending verbal commands presented by the experimenter or through computer speakers
- Use of certain hair products (e.g. hair gel) which will interfere with the EEG electrodes.
- Unusually thick hair which may prevent a proper fitting of the EEG cap.
- Head size which is not coverable by the available EEG caps (too large or too small).
- Unable or unwilling to attach electrodes to chest/abdomen for ECG collection

Benefits:
There is no direct benefit from the study for the participants. The study will allow you to take part in important research about decision-making activities and help the investigators detect the relationship between this variable, brain waves, and investigative patterns of behavior. If you are interested, researchers can refer you to published scientific articles available on the study topics.

Discomfort and risks:
This computerized portion of this study does not involve any more than minimal risk to you. In other words, there is no harm or discomfort beyond what is ordinarily encountered in daily life when using the computer or during the performance of routine physical or psychological tests.

There are other possible sources of risk & discomfort. We will attach sensors on your head, face and arms. Some participants may experience discomfort (due to limited movement during trials). Minor skin or eye irritation and/or discomfort may result when the electrodes are placed the head, face, chest and abdomen when you or the testers clean those locations to reduce electrical impedance in order to improve signal quality. Minimal, temporary hair loss is unlikely but may occur locally at the electrode sites. Because applying the electrical sensors to participants requires making contact with and, in some cases, scrubbing the skin with liquids and exfoliating materials, there is a theoretical risk of transmitting skin-borne pathogens during this process. All necessary equipment will be cleaned and disinfected before the procedure begins to minimize the risk of transmitting diseases. All equipment is standard and is used commercially.

If you choose to fill out the questionnaire, steps will be taken to protect your confidentiality as described below.

Compensation: Participation in the study is completely voluntary and there is no compensation.

Entitlements and confidentiality:

a. Records of your participation in this study may only be disclosed according to federal law, including the Federal Privacy Act, 5 U.S.C. 552a, and its implementing regulations and the Health Insurance Portability and Accountability Act (HIPAA), and its implementing regulations, when applicable, and the Freedom of Information Act, 5 U.S.C. Sec 522, and its implementing regulations when applicable. Your personal information will be stored in a locked cabinet in an office that is locked when not occupied. Electronic files containing your personal information will be password protected and stored only on a secure server. It is intended that the only people having access to your information will be the researchers named above and this study's Research Monitor, the AFRL Wright Site IRB, the Air Force Surgeon General's Research Compliance office, the Director of Defense Research and Engineering office or any other IRB involved in the review and approval of this protocol. When no longer needed for research purposes your information will be destroyed in a secure manner. Complete confidentiality cannot be promised, in particular for military personnel, whose health or fitness for duty information may be required to be reported to appropriate medical or command authorities. If such information is to be reported, you will be informed of what is being reported and the reason for the report.

   Your entitlements to medical and dental care and/or compensation in the event of injury are governed by federal laws and regulations, and that if you desire further information you may contact the base legal office (711HPW/JA, 937-656-5666 for Wright-Patterson AFB). In the event of a research related injury, you may contact the Deputy Director and Acting Chair of the AFRL IRB of this research study at (937) 904-8100 or AFRL.IR.ProtocolManagement@us.af.mil.

b. The decision to participate in this research is completely voluntary on your part. No one may coerce or intimidate you into participating in this program. You are participating because you want to. Dr. Brett Borghetti, or an associate investigator, has adequately answered any and all questions you have about this study, your participation, and the procedures involved. Dr. Brett Borghetti can be reached at (937) 255-3636x4612. Dr. Borghetti, or an associate investigator, will be available to answer any questions concerning procedures throughout this study. If significant new findings develop during the course of this research, which may relate to your decision to continue participation, you will be informed. You may discontinue participation at any time without penalty. Notify one of the investigators of this study to discontinue. The investigator or research monitor of this study may terminate your participation in this study if she or he feels this to be in your best interest. If you have any questions or concerns about your participation in this study or your rights as a research subject, please contact the IRB office via telephone at 937-904-8100 or email at AFRL.IR.ProtocolManagement@us.af.mil.

Taking part in this research study is completely voluntary. Your signature below shows that:

- *You agree to be in this study*
- *The researcher has explained the study to you and you have read and understand the information you have been given*
- *You were given the opportunity to ask questions about the study and all of your questions have been answered to your satisfaction*
- *You understand that signing this consent does not take away any of your legal rights*

You will be given a copy of this signed consent form for your records.

Volunteer Signature____**Date**______________

Volunteer Name (printed)____

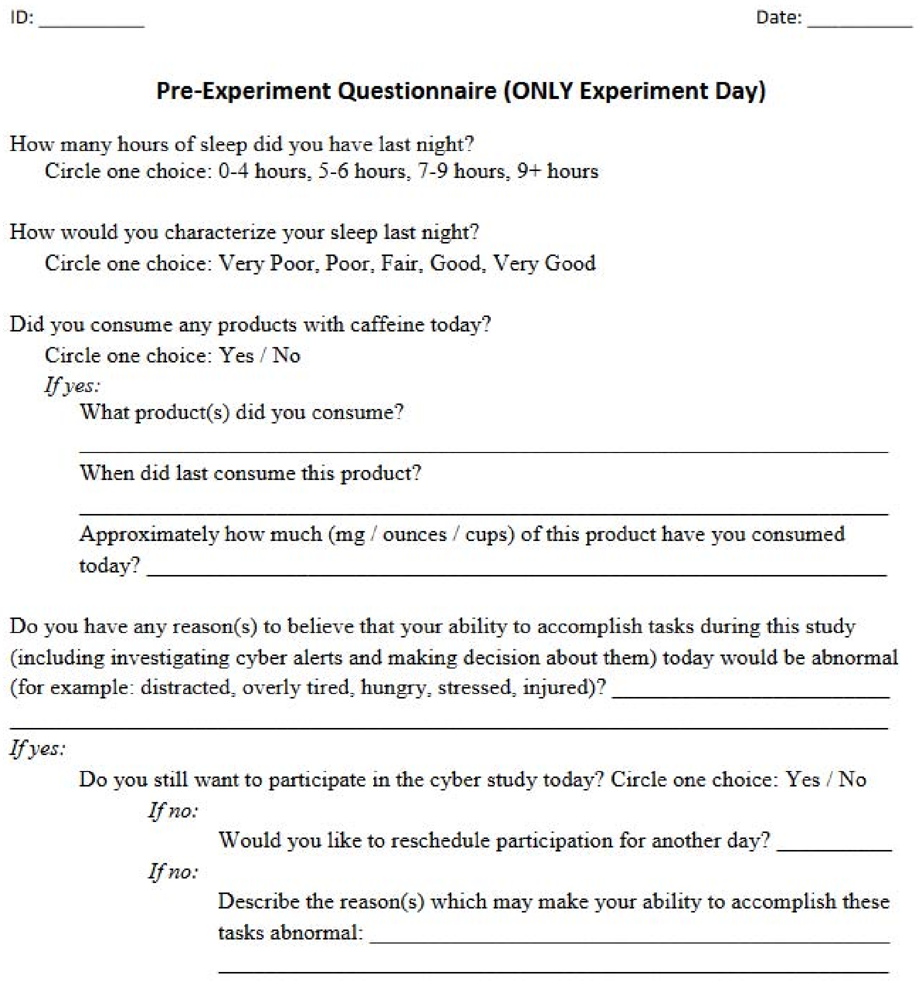
Advising Investigator Signature ____________________ **Date** ________________

Investigator Name (printed)____

Witness Signature _________________________________ **Date** ________________

Witness Name (printed)____

Appendix C: Pre-Experiment Questionnaire



ID: __________ Date: __________

Pre-Experiment Questionnaire (ONLY Experiment Day)

How many hours of sleep did you have last night?
Circle one choice: 0-4 hours, 5-6 hours, 7-9 hours, 9+ hours

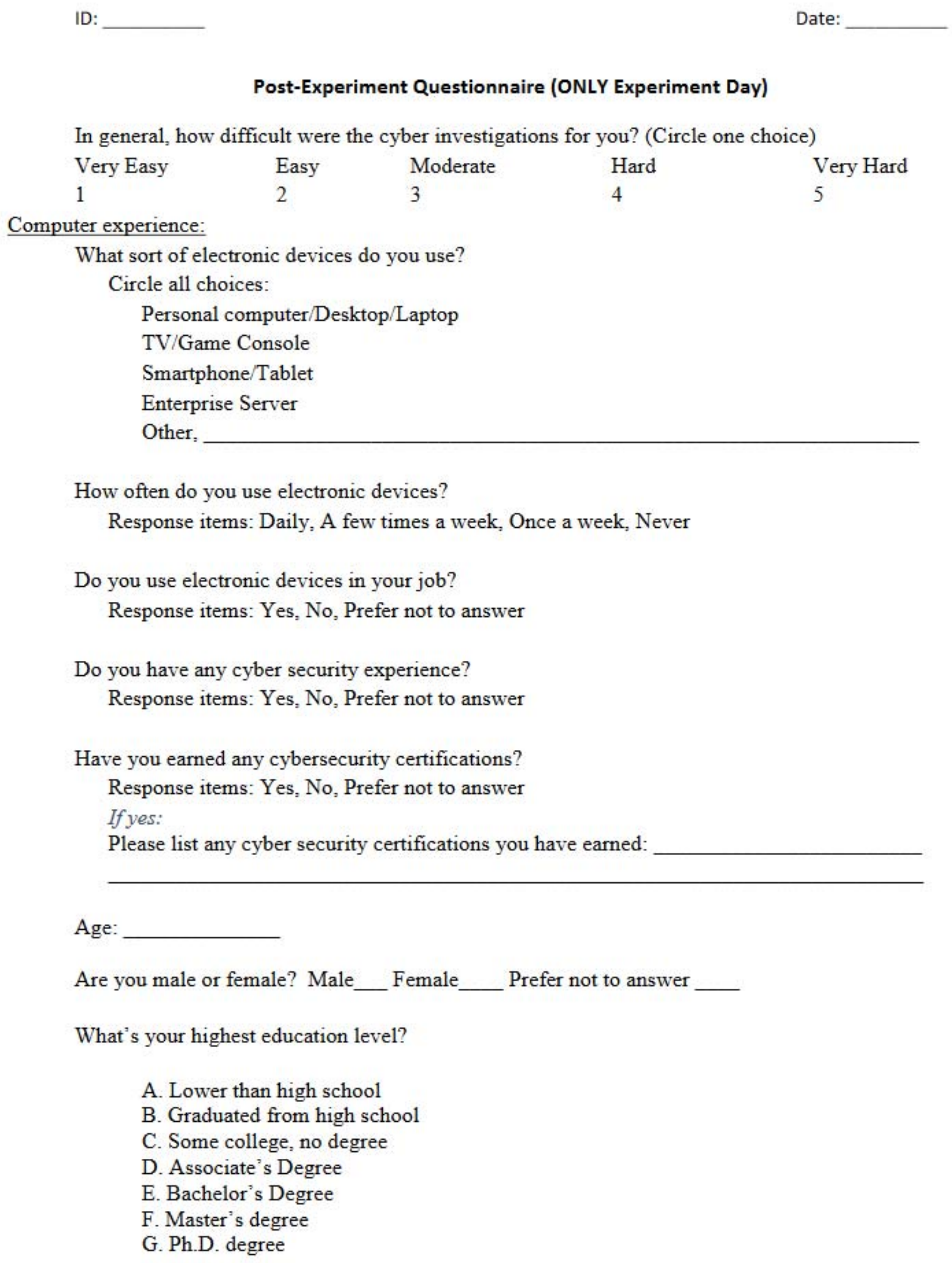
How would you characterize your sleep last night?
Circle one choice: Very Poor, Poor, Fair, Good, Very Good

Did you consume any products with caffeine today?
Circle one choice: Yes / No
If yes:
What product(s) did you consume?
____
When did last consume this product?
____
Approximately how much (mg / ounces / cups) of this product have you consumed today? ____

Do you have any reason(s) to believe that your ability to accomplish tasks during this study (including investigating cyber alerts and making decision about them) today would be abnormal (for example: distracted, overly tired, hungry, stressed, injured)? ________________________
____

If yes:
Do you still want to participate in the cyber study today? Circle one choice: Yes / No
If no:
Would you like to reschedule participation for another day? ___________
If no:
Describe the reason(s) which may make your ability to accomplish these tasks abnormal: ____
____

Appendix D: Post-Experiment Questionnaire



ID: __________ Date: __________

Post-Experiment Questionnaire (ONLY Experiment Day)

In general, how difficult were the cyber investigations for you? (Circle one choice)

Very Easy	Easy	Moderate	Hard	Very Hard
1	2	3	4	5

Computer experience:

What sort of electronic devices do you use?

Circle all choices:

Personal computer/Desktop/Laptop
TV/Game Console
Smartphone/Tablet
Enterprise Server
Other, ____

How often do you use electronic devices?

Response items: Daily, A few times a week, Once a week, Never

Do you use electronic devices in your job?

Response items: Yes, No, Prefer not to answer

Do you have any cyber security experience?

Response items: Yes, No, Prefer not to answer

Have you earned any cybersecurity certifications?

Response items: Yes, No, Prefer not to answer

If yes:

Please list any cyber security certifications you have earned: ________________________
____

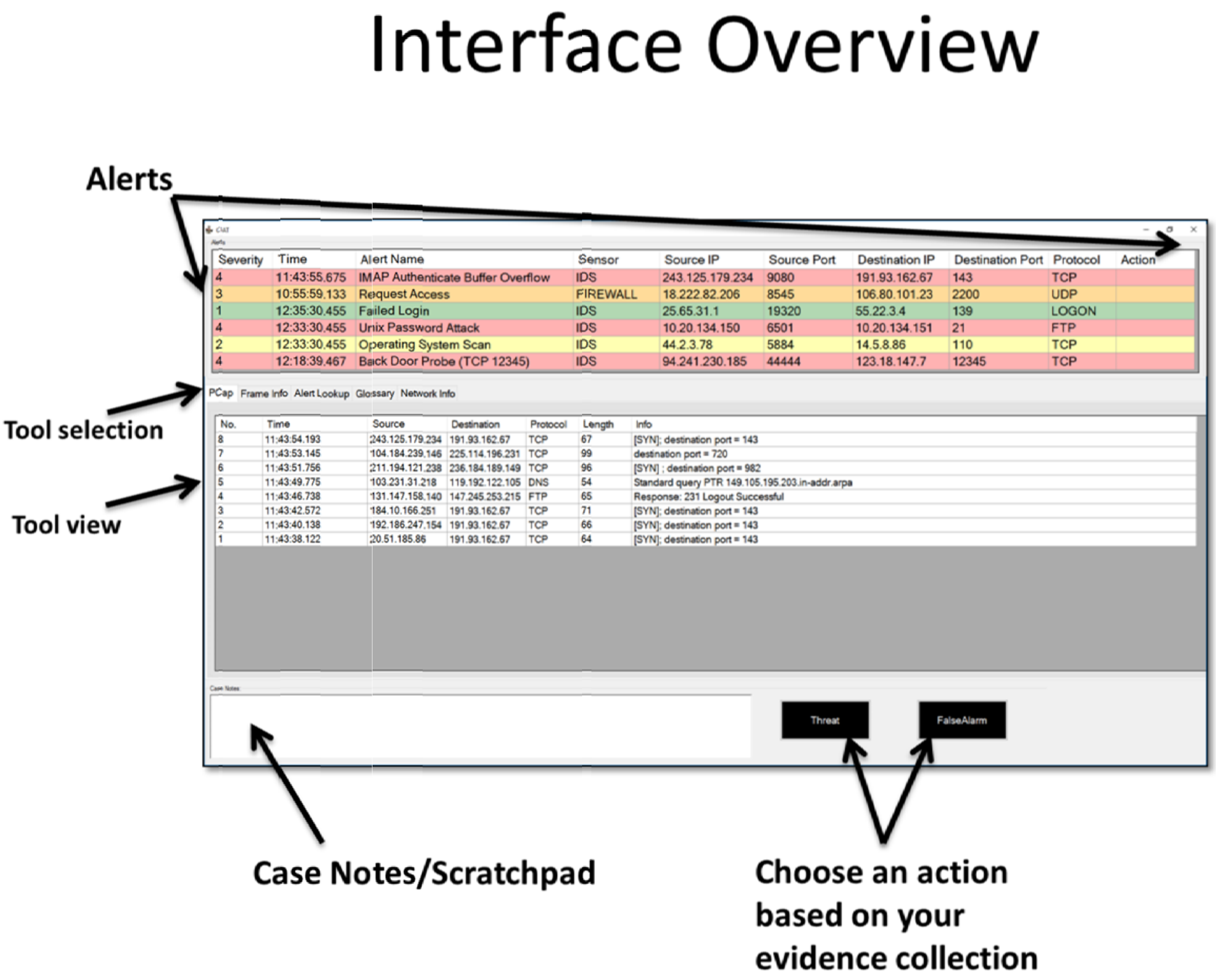
Age: ______________

Are you male or female? Male___ Female____ Prefer not to answer ____

What's your highest education level?

A. Lower than high school
B. Graduated from high school
C. Some college, no degree
D. Associate's Degree
E. Bachelor's Degree
F. Master's degree
G. Ph.D. degree

Appendix E: CIAT 2.0 Alert Interface Overview



Interface Overview

Alerts

Tool selection

Tool view

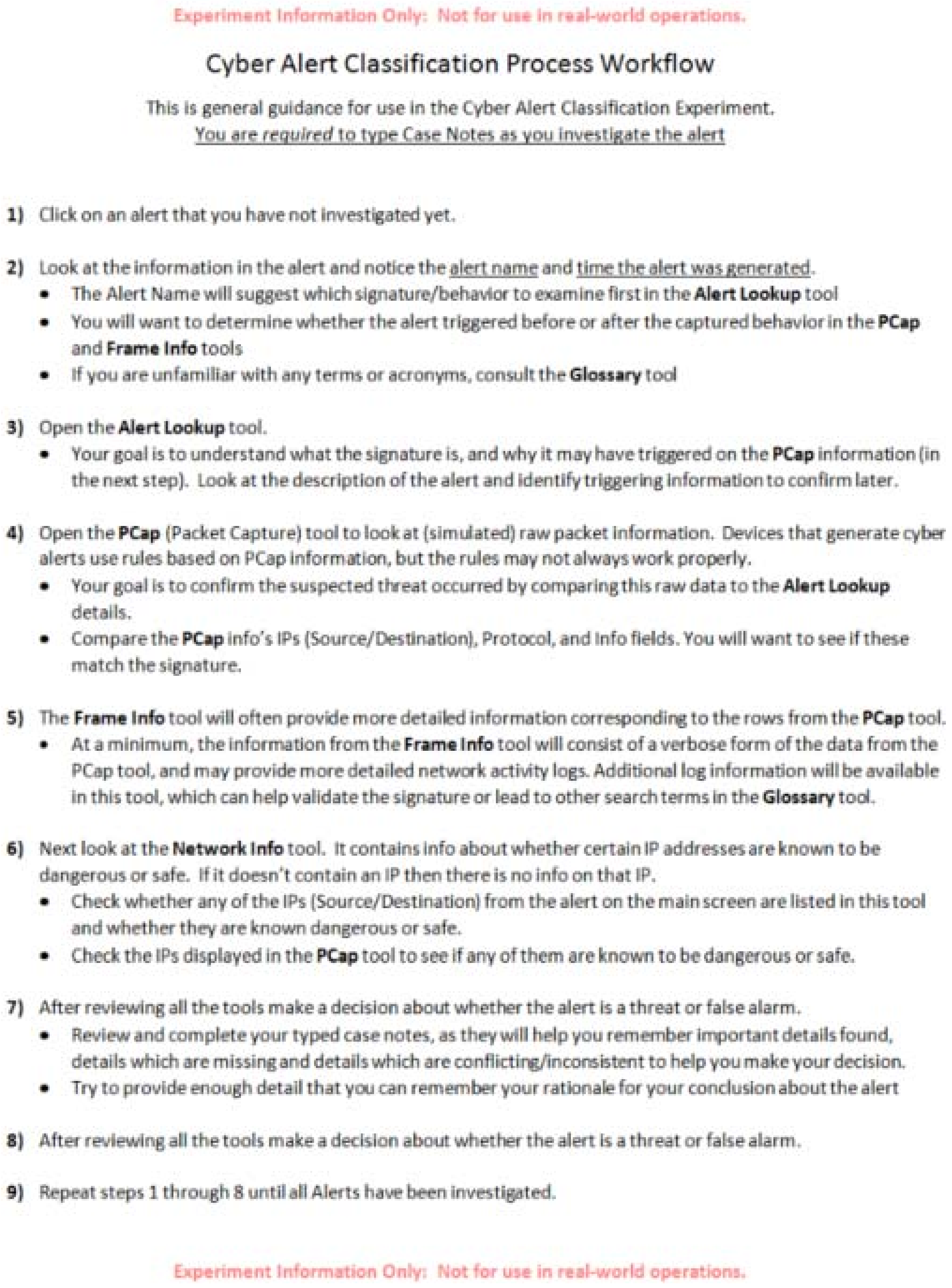
Case Notes/Scratchpad

Choose an action based on your evidence collection

Severity	Time	Alert Name	Sensor	Source IP	Source Port	Destination IP	Destination Port	Protocol	Action
4	11:43:55.675	IMAP Authenticate Buffer Overflow	IDS	243.125.179.234	9080	191.93.162.67	143	TCP	
3	10:55:59.133	Request Access	FIREWALL	18.222.82.206	8545	106.80.101.23	2200	UDP	
1	12:35:30.455	Failed Login	IDS	25.65.31.1	19320	55.22.3.4	139	LOGON	
4	12:33:30.455	Unix Password Attack	IDS	10.20.134.150	6501	10.20.134.151	21	FTP	
2	12:33:30.455	Operating System Scan	IDS	44.2.3.78	5884	14.5.8.86	110	TCP	
4	12:18:39.467	Back Door Probe (TCP 12345)	IDS	94.241.230.185	44444	123.18.147.7	12345	TCP	

No.	Time	Source	Destination	Protocol	Length	Info
8	11:43:54.193	243.125.179.234	191.93.162.67	TCP	67	[SYN]; destination port = 143
7	11:43:53.145	104.184.239.146	225.114.196.231	TCP	99	destination port = 720
6	11:43:51.756	211.194.121.238	236.184.189.149	TCP	96	[SYN] ; destination port = 982
5	11:43:49.775	103.231.31.218	119.192.122.105	DNS	54	Standard query PTR 149.105.195.203.in-addr.arpa
4	11:43:46.738	131.147.158.140	147.245.253.215	FTP	65	Response: 231 Logout Successful
3	11:43:42.572	184.10.166.251	191.93.162.67	TCP	71	[SYN]; destination port = 143
2	11:43:40.138	192.186.247.154	191.93.162.67	TCP	66	[SYN]; destination port = 143
1	11:43:38.122	20.51.185.86	191.93.162.67	TCP	64	[SYN]; destination port = 143

Appendix F: General Cyber Alert Investigation Workflow Handout



Experiment Information Only: Not for use in real-world operations.

Cyber Alert Classification Process Workflow

This is general guidance for use in the Cyber Alert Classification Experiment.
You are *required* to type Case Notes as you investigate the alert

1) Click on an alert that you have not investigated yet.

2) Look at the information in the alert and notice the alert name and time the alert was generated.
   - The Alert Name will suggest which signature/behavior to examine first in the **Alert Lookup** tool
   - You will want to determine whether the alert triggered before or after the captured behavior in the **PCap** and **Frame Info** tools
   - If you are unfamiliar with any terms or acronyms, consult the **Glossary** tool

3) Open the **Alert Lookup** tool.
   - Your goal is to understand what the signature is, and why it may have triggered on the **PCap** information (in the next step). Look at the description of the alert and identify triggering information to confirm later.

4) Open the **PCap** (Packet Capture) tool to look at (simulated) raw packet information. Devices that generate cyber alerts use rules based on PCap information, but the rules may not always work properly.
   - Your goal is to confirm the suspected threat occurred by comparing this raw data to the **Alert Lookup** details.
   - Compare the **PCap** info's IPs (Source/Destination), Protocol, and Info fields. You will want to see if these match the signature.

5) The **Frame Info** tool will often provide more detailed information corresponding to the rows from the **PCap** tool.
   - At a minimum, the information from the **Frame Info** tool will consist of a verbose form of the data from the PCap tool, and may provide more detailed network activity logs. Additional log information will be available in this tool, which can help validate the signature or lead to other search terms in the **Glossary** tool.

6) Next look at the **Network Info** tool. It contains info about whether certain IP addresses are known to be dangerous or safe. If it doesn't contain an IP then there is no info on that IP.
   - Check whether any of the IPs (Source/Destination) from the alert on the main screen are listed in this tool and whether they are known dangerous or safe.
   - Check the IPs displayed in the **PCap** tool to see if any of them are known to be dangerous or safe.

7) After reviewing all the tools make a decision about whether the alert is a threat or false alarm.
   - Review and complete your typed case notes, as they will help you remember important details found, details which are missing and details which are conflicting/inconsistent to help you make your decision.
   - Try to provide enough detail that you can remember your rationale for your conclusion about the alert

8) After reviewing all the tools make a decision about whether the alert is a threat or false alarm.

9) Repeat steps 1 through 8 until all Alerts have been investigated.

Experiment Information Only: Not for use in real-world operations.

Bibliography

[1] G. Funke *et al.*, “Development and Validation of the Air Force Cyber Intruder Alert Testbed (CIAT),” in *Advances in Human Factors in Cybersecurity*, 1st ed., vol. 501, Denise Nicholson, Ed. Switzerland: Springer, 2016, pp. 363–376.

[2] A. Insabato, M. Pannunzi, E. T. Rolls, and G. Deco, “Confidence-Related Decision Making,” *J. Neurophysiol.*, vol. 104, no. 1, pp. 539–547, 2010.

[3] G. M. Sullivan and A. R. Artino, “Analyzing and Interpreting Data From Likert-Type Scales,” *J. Grad. Med. Educ.*, vol. 5, no. 4, pp. 541–542, 2013.

[4] K. A. White, “Development and Validation of a Tool to Measure Self-Confidence and Anxiety in Nursing Students During Clinical Decision Making,” *J Nurs Educ*, vol. 53, no. 1, pp. 14–22, 2014.

[5] R. Likert, “A Technique for the Measurement of Attitudes,” *Archives of Psychology*, vol. 22 140. p. 55, 1932.

[6] Merriam-Webster, “Confidence,” *Merriam-Webster*. [Online]. Available: https://www.merriam-webster.com/dictionary/confidence. [Accessed: 22-May-2017].

[7] AM O’Connor, “User Manual - Decision Self-Efficacy Scale,” *User Manual - Decision Se lf-Efficacy Scale*, 2002. [Online]. Available: https://decisionaid.ohri.ca/eval_self.html. [Accessed: 01-Jan-2018].

[8] J. E. Ormrod, *Educational Psychology: Developing Learners*, 8th Editio. Pearson, 2014.

[9] S. L. Pfleeger, “Leveraging Behavioral Science to Mitigate Cyber Security Risk Leveraging Behavioral Science to Mitigate Cyber Security Risk © 2012 The

MITRE Corporation . ALL RIGHTS RESERVED .," pp. 1–44.

[10] D. A. Holland and J. E. Freeman, *A Ten-Year Overview of USAF F-16 Mishap Attributes from 1980-89*, vol. 39. 1995.

[11] R. Kiani, L. Corthell, and M. N. Shadlen, "Choice certainty is informed by both evidence and decision time," *Neuron*, vol. 84, no. 6, pp. 1329–1342, 2014.

[12] P. Dayan and N. D. Daw, "Decision theory, reinforcement learning, and the brain," *Cogn. Affect. Behav. Neurosci.*, vol. 8, no. 4, pp. 429–453, 2008.

[13] B. Y. Ng, A. Kankanhalli, and Y. (Calvin) Xu, "Studying users' computer security behavior: A health belief perspective," *Decis. Support Syst.*, vol. 46, no. 4, pp. 815–825, 2009.

[14] M. Tyworth, N. a Giacobe, V. F. Mancuso, M. D. Mcneese, and D. L. Hall, "A human-in-the-loop approach to understanding situation awareness in cyber defence analysis," vol. 13, no. June, pp. 1–10, 2013.

[15] M. R. Endsley, "Toward a Theory of Situation Awareness in Dynamic Systems," *Hum. Factors J. Hum. Factors Ergon. Soc.*, vol. 37, no. 1, pp. 32–64, 1995.

[16] P. K. Belling, J. Suss, and P. Ward, "Investigating Constraints on Decision Making Strategies," *Int. Conf. Nat. Decis. Mak.*, no. May, pp. 21–24, 2013.

[17] N. Yeung and C. Summerfield, "Metacognition in human decision-making: confidence and error monitoring," *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, vol. 367, no. 1594, pp. 1310–21, May 2012.

[18] P. Ward, J. Suss, D. W. Eccles, A. M. Williams, and K. R. Harris, "Skill-Based Differences in Option Generation in a Complex Task: A Verbal Protocol Analysis," *Cogn. Process.*, vol. 12, no. 3, pp. 289–300, 2011.

[19] W. Lu, S. Xu, and X. Yi, “Optimizing Active Cyber Defense,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, vol. 8252 LNCS, pp. 206–225.

[20] W. Chappelle *et al.*, “Sources of Occupational Stress and Prevalence of Burnout and Clinical Distress Among U.S. Air Force Cyber Warfare Operators,” 2013.

[21] R. Torkzadeh, K. Pflughoeft, and L. L. Hall, “Computer Self-Efficacy, Training Effectiveness and User Attitudes: An Empirical Study,” *Behav. Inf. Technol.*, vol. 18, no. 4, pp. 299–309, 1999.

[22] V. F. Mancuso, V. S. Finomore, K. M. Rahill, E. A. Blair, and G. J. Funke, “Effects of Cognitive Biases on Distributed Team Decision Making,” *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 58, no. 1, pp. 405–409, 2014.

[23] G. Funke *et al.*, “Development and Validation of the Air Force Cyber Intruder Alert Testbed (CIAT),” in *Advances in Human Factors in Cybersecurity*, 1st ed., Denise Nicholson, Ed. Switzerland: Springer, 2016, pp. 363–376.

[24] M. Champion, S. Jariwala, P. Ward, and N. J. Cooke, “Using Cognitive Task Analysis to Investigate the Contribution of Informal Education to Developing Cyber Security Expertise,” *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 58, no. 1, pp. 310–314, 2014.

[25] W. He, X. Yuan, and X. Tian, “The Self-Efficacy Variable in Behavioral Information Security Research,” *Proc. - 2nd Int. Conf. Enterp. Syst. ES 2014*, no. April, pp. 28–32, 2014.

[26] T. J. Cleary and J. Sandars, “Assessing Self-Regulatory Processes During Clinical Skill Performance: A Pilot Study,” *Med. Teach.*, vol. 33, no. 7, pp. e368–e374,

2011.

[27] J. McClain, A. Silva, G. E. Aviña, and C. Forsythe, “Measuring Human Performance within Computer Security Incident Response Teams,” no. September, 2015.

[28] G. L. Callan, “Self-Regulated Learning (SRL) Microanalysis for Mathematical Problem Solving: A Comparison of a SRL Event Measure, Questionnaires, and a Teacher Rating Scale,” *ProQuest Diss. Theses*, p. 196, 2014.

[29] A. C. Trujillo, “Evaluation of Electronic Formats of the NASA Task Load Index,” no. August, 2011.

[30] J. Kruger and D. Dunning, “Unskilled and Unaware of It: How Difficulties in Recognizing One’s Own Incompetence Lead to Inflated Self-Assessments,” *J. Pers. Soc. Psychol.*, vol. 77, no. 6, pp. 1121–1134, 1999.

[31] K. W. Eva and G. Regehr, “I’ll Never Play Professional Football and Other Fallacies of Self-Assessment,” *J. Contin. Educ. Health Prof.*, vol. 28, no. 1, pp. 14–19, 2008.

[32] M. C. Howard, “Creation of a Computer Self-Efficacy Measure: Analysis of Internal Consistency, Psychometric Properties, and Validity,” *Cyberpsychology, Behav. Soc. Netw.*, vol. 17, no. 10, pp. 677–681, 2014.

[33] D. R. Compeau and C. A. Higgins, “Computer Self-Efficacy: Development of a Measure and Initial Test,” *MIS Q.*, vol. 19, no. 2, p. 189, 1995.

[34] A. Bandura, “Guide for Constructing Self-Efficacy Scales,” *Self-efficacy beliefs Adolesc.*, pp. 307–337, 2006.

[35] A. D’Amico, K. Whitley, D. Tesone, B. O’Brien, and E. Roth, “Achieving Cyber

Defense Situational Awareness: A Cognitive Task Analysis of Information Assurance Analysts," *Proc. Hum. Factors Ergon. Soc. Annu. Meet. 2005*, vol. 49, pp. 229–233, 2005.

[36] R. Kiani and M. N. Shadlen, "Representation of Confidence Associated with a Decision by Neurons in the Parietal Cortex," *Science*, vol. 324, no. 2009, pp. 759–764, 2009.

[37] Z. Minchev, "Multiple Human Biometrics Fusion in Support of Cyberthreats Identification," *Cybern. Inf. Technol.*, vol. 15, no. 7, pp. 67–76, 2015.

[38] P. Brown, K. Christensen, and D. Schuster, "An Investigation of Trust in a Cyber Security Tool," *Proc. Hum. Factors Ergon. Soc. Annu. Meet. 2016*, pp. 1454–1458, 2016.

[39] J. McClain *et al.*, "Human Performance Factors in Cyber Security Forensic Analysis," *Procedia Manuf.*, vol. 3, no. Ahfe, pp. 5301–5307, 2015.

[40] N. Ben-Asher and Y. Paul, "Synergistic Architecture for Human-Machine Intrusion Detection," *J. Cyber Secur. Inf. Syst.*, vol. 5, no. 1, 2017.

[41] B. K. Phillips, V. R. Prybutok, and D. A. Peak, "Decision Confidence, Information Usefulness, and Information Seeking Intention in the Presence of Disconfirming Information," *Informing Sci.*, vol. 17, no. 1, pp. 1–24, 2014.

[42] M. W. Boyce, K. M. Duma, L. J. Hettinger, T. B. Malone, D. P. Wilson, and J. Lockett-Reynolds, "Human Performance in Cybersecurity: A Research Agenda," *Proc. Hum. Factors Ergon. Soc. Annu. Meet. 2011*, vol. 55, no. 1, pp. 1115–1119, 2011.

[43] O. Jensen and C. D. Tesche, "Frontal theta activity in humans increases with

memory load in a working memory task," *Eur. J. Neurosci.*, vol. 15, no. 8, pp. 1395–1399, 2002.

[44] A. Gevins, M. E. Smith, L. McEvoy, and D. Yu, "High-resolution EEG mapping of cortical activation related to working memory: Effects of task difficulty, type of processing, and practice," *Cereb. Cortex*, vol. 7, no. 4, pp. 374–385, 1997.

[45] M. W. Howard *et al.*, "Gamma Oscillations Correlate with Working Memory Load in Humans," *Cereb. Cortex*, vol. 13, no. 12, pp. 1369–1374, 2003.

[46] T. Gruber and M. M. Müller, "Oscillatory Brain Activity Dissociates Between Associative Stimulus Content in a Repetition Priming Task in the Human EEG," *Cereb. Cortex*, vol. 15, no. 1, pp. 109–116, 2005.

[47] C. Tallon-Baudry, O. Bertrand, M. A. Hénaff, J. Isnard, and C. Fischer, "Attention Modulates Gamma-Band Oscillations Differently in the Human Lateral Occipital Cortex and Fusiform Gyrus," *Cereb. Cortex*, vol. 15, no. 5, pp. 654–662, 2005.

[48] S. L. Gonzalez Andino, C. M. Michel, G. Thut, T. Landis, and R. G. De Peralta, "Prediction of Response Speed by Anticipatory High-Frequency (Gamma Band) Oscillations in the Human Brain," *Hum. Brain Mapp.*, vol. 24, no. 1, pp. 50–58, 2005.

[49] J. Jacobs, G. Hwang, T. Curran, and M. J. Kahana, "EEG Oscillations and Recognition Memory: Theta Correlates of Memory Retrieval and Decision Making," *Neuroimage*, vol. 32, no. 2, pp. 978–987, 2006.

[50] J. F. Cavanagh, M. J. Frank, T. J. Klein, and J. J. B. Allen, "Frontal Theta Links Prediction Errors to Behavioral Adaptation in Reinforcement Learning," *Neuroimage*, vol. 49, no. 4, pp. 3198–3209, 2010.

[51] A. Vieane *et al.*, “Coordinated Displays to Assist Cyber Defenders,” *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 60, no. 1, pp. 344–348, 2016.

[52] D. Druckman and R. a. Bjork, *Learning, Remembering, Believing: Enhancing Human Performance*. 1994.

[53] R. Ratcliff and G. McKoon, “The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks,” *Neural Comput.*, vol. 20, no. 4, pp. 873–922, 2008.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 074-0188*

1. REPORT DATE *(DD-MM-YYYY)* 22-03-2018	**2. REPORT TYPE** Master's Thesis	**3. DATES COVERED** *(From – To)* September 2016 – March 2018

TITLE AND SUBTITLE Estimating Defensive Cyber Operator Decision Confidence	**5a. CONTRACT NUMBER** F4FBGN7340J001
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Borneman, Markus M., Captain, USAF	**5d. PROJECT NUMBER** JON#18G147B
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S) Air Force Institute of Technology Graduate School of Engineering and Management (AFIT/EN) 2950 Hobson Way, Building 640 WPAFB OH 45433-8865	**8. PERFORMING ORGANIZATION REPORT NUMBER** AFIT-ENG-MS-18-M-013
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) 711th Human Performance Wing 2610 Seventh St Bldg 441 WPAFB, OH 45433 937-255-8222 rajesh.naik@us.af.mil ATTN: Dr. Rajesh Naik	**10. SPONSOR/MONITOR'S ACRONYM(S)** 711 HPW/CL
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRUBTION STATEMENT A. APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES
This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

14. ABSTRACT

As technology continues to advance the domain of cyber defense, signature and heuristic detection mechanisms continue to require human operators to make judgements about the correctness of machine decisions. Human cyber defense operators rely on their experience, expertise, and understanding of network security, when conducting cyber-based investigations, in order to detect and respond to cyber alerts. Ever growing quantities of cyber alerts and network traffic, coupled with systemic manpower issues, mean no one has the time to review or change decisions made by operators. Since these cyber alert decisions ultimately do not get reviewed again, an inaccurate decision could cause grave damage to the network and host systems. The Cyber Intruder Alert Testbed (CIAT), a synthetic task environment (STE), was expanded to include investigative pattern of behavior monitoring and confidence reporting capabilities. By analyzing the behavior and confidence of participants while they conducted cyber-based investigations, this research was able to identify a mapping between investigative patterns of behavior and decision confidence. The total time spent on a decision, the time spent using different investigative tools, and total number of tool transitions, were all factors which influenced the reported confidence of participants when conducting cyber-based investigations.

15. SUBJECT TERMS
cyber alert investigation, decision confidence, Cyber Intruder Alert Testbed (CIAT), synthetic task environment, behavior patterns

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Dr. Brett J. Borghetti, AFIT/ENG	
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(Include area code)*	
U	U	U	UU	131	(937) 255-6565, ext 4612 brett.borghetti@afit.edu	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18