SAND2015-1796C

Measuring Expert and Novice Performance within Computer Security Incident Response Teams

Austin Silva[1], Glory Emmanuel, Jonathan T. McClain[2], Laura Matzen, Chris Forsythe

[1]Corresponding Author, Sandia National Laboratories, Albuquerque, New Mexico
aussilv@sandia.gov
[2]Principle Investigator, Sandia National Laboratories, Albuquerque, New Mexico
jtmccl@sandia.gov

Abstract. There is a great need for creating cohesive, expert cybersecurity incident response teams and training them effectively. This paper discusses new methodologies for measuring and understanding expert and novice differences within a cybersecurity environment to bolster training, selection, and teaming. This methodology for baselining and characterizing individuals and teams relies on relating eye tracking gaze patterns to psychological assessments, human-machine transaction monitoring, and electroencephalography data that are collected during participation in the game-based training platform Tracer FIRE. We discuss preliminary findings from two pilot studies using novice and professional teams.

Keywords. Cybersecurity, training, teams, visual search, eye tracking, EEG, in-situ testing, measuring individual differences, psychological measures

1 Introduction

As the trend towards highly sophisticated cyber-attacks continues to rise in the U.S., the need for more highly skilled professionals to counter these attacks has become critical. There is an increased demand for qualified individuals across the nation in both private and government domains. Currently, cyber incident responders (IRs) are generally selected through an interview-based hiring process and trained through tool-centric workshops. There is little understanding of the progression from novice to competent to expert to elite, within cyber security, nor is there a clear understanding of the ideal structure and allocation of duties within cyber IR teams. This lack of basic understanding is compounded by the fact that the number of individuals training for positions in cyber security is insufficient to meet the demands projected over the next decade. In order to defend against dynamic and intelligent cyber-attacks, there is a need for a rigorous, quantitative approach to selecting and training IRs that will produce more world-class experts, at the individual and team level.

An interdisciplinary team of researchers at Sandia National Laboratories ("Sandia") has developed an empirical methodology to quantify novice and expert differences. Performance data was collected through the instrumentation of a training environment known as Tracer FIRE. Cognitive and behavioral processes were quantified through the use of eye tracking, electroencephalography (EEG), self-report measures, and computer instrumentation. The purpose of the current project is to test the hypothesis that we can differentiate between effective and ineffective individuals and teams by utilizing cognitive measures such as EEG, eye tracking, self-report measures, and human-machine transactions during domain-specific and domain-general tasks. The integration of technical capabilities to develop a human performance measurement system for IR teams will offer a cornerstone for cyber security training programs.

This work is challenging for two reasons. At the psychological level, identifying causal relationships in an unstructured environment is very challenging. At the instrumentation level, capturing and synchronizing the data from a network of machines and other sensors (e.g., EEG and eye tracking) at the millisecond level in a way that supports causal analysis at the psychological level is a significant data fusion problem. In this way, this work is unique in its close approximation to an operational environment while allowing controlled quantitative measurement.

In our paper, we separately discuss the relationship of Tracer FIRE, eye tracking, human-machine transaction tracking capabilities, EEG, and self-report measures to cyber security. We specifically present how each of these technologies can be used to collect data and quantify attributes related to cyber expertise. From there we highlight the methodology we used to collect data from expert and novice IRs. This methodology is unique due to the combination of these various technologies in a single cyber training environment. We are aware of both the strengths and weaknesses of our methodology and present how this experimental design can be used on a larger scale to understand the human dimension in cyber security environments.

While cybersecurity relies heavily on software tools that are used to protect and analyze a network and its traffic, the human in the system is sometimes understated. The human-centric approach to understanding cybersecurity personnel and their aptitude was developed and piloted. This approach leveraged metrics that range from psychological, behavioral, and neuronal measurements. There is a heavy emphasis on visual search since the tools used by IRs tend to include an abundance of text information that requires downselection through search. Overall, our methodology provides a mechanism to acquire a deeper understanding of the characteristics of high performing individuals in cyber security, as well as their impact on incident outcomes. Such understanding could lead to a broad range of capabilities such as:

1. The ability to identify individuals with a high aptitude to excel in the cyber security domain in order to inform recruiting and enhance training.
2. The ability to build a better cyber security workforce that will directly contribute to the crucial task of protecting organizations' information and infrastructure.

2 Design and Instrumentation

2.1 Tracer FIRE

Tracer FIRE is a multi-day training method that uses lectures regarding cybersecurity techniques and tools that culminates in a two-day competitive team event. During the lectures, the participants learn in a classroom setting with hands-on examples of the tools and topics of study. The final days allow students to team up together in groups of up to six students and compete in a challenge-based puzzle game in which the teams must analyze and solve a forensic narrative. During the main event, teams are allowed to allocate challenges to one another or team up to collectively solve aspects of the narrative. Teams are then given points for each of the challenges they solve, in addition to unlocking new content (i.e., challenges.

The Tracer FIRE environment has been used before to study the dynamics of teams including communication and task delegation [1]. But there were problems in comparing performance based on point awards, with questions arising concerning the degree to which points were meaningful and comparable across teams and challenges [2]. The new scenario presented a linear story line, which allowed for teams to be compared against each other. Using this test bed, a number of data collection methods were used: psychological measures for understanding individual decision-making styles as well as team composition; human-machine transaction monitoring of the activity of each team member; eye tracking on a domain-specific task to compare metrics to experts in the field; and EEG assessment of the relationship between memory processes and performance in the cyber exercise.

2.2 Participants

The data collected in the pilot consisted of participants from two separate events. The first Tracer FIRE was for experienced cyber IR professionals and had 9 participants out of 20 that were willing to complete the psychological measures. It should be noted that the EEG and eye tracking tasks were not available at this time. The second population consisted of student interns in a separate Tracer FIRE exercise. The novice teams were self-selected and ranged from students with limited exposure to cyber security to students that were graduate students specializing in cybersecurity. Some participants volunteered for data collection that occurred within an instrumented training exercise environment, as well as a test battery outside of the exercise setting. For the student exercise, 31 of 40 participants consented to data collection. Of this group, twelve participants agreed to undergo the EEG and eye tracking tasks. Participation in the any of the measures was completely voluntary.

2.3 Eye Tracking

Eye tracking was used to examine if there would be differences between visual search patterns of novice IR personnel compared to experts. The tools used by IRs are mostly text-centric and require the users to scan through vast data in the form of logs files and other visual output. The process of how experts downselect this information tends to be user specific and changes through experience. More experienced users are able to ignore extraneous data, but novice tool users may spend more time inspecting every visual element. Metrics to examine differences included response accuracy, reaction time, and time until first fixation within the region of interest (the area on the screen in which the answer was contained). While the data collected during the Tracer FIRE exercise were only novice information, a subject matter expert provided an "optimal search path" for a notional expert data reference.

Materials. The equipment used in the eye tracking experiments was the Smart Eye Pro 6.0 eye tracking system. This system consisted of a three-camera configuration, which is key to ensuring a large head box since often the text on the screen in many of the cyber tools was small and required the user to lean in towards the screen. This is often a problem with a two-camera system where users can freely move in and out of the head box and camera frame. The SmartEye system recorded raw gaze patterns and sampled at 60Hz. The raw gaze data was also time synced with the on-screen video recording through the use of EyesDX's Record Manager and Video Streamer software packages. The combined data was then processed in the MAPPS data analysis software for determination of fixations and gaze length.

Procedure. The first task was domain-specific and consisted of ten questions regarding screenshots from common cyber tools used in the field. These screenshots used static images instead of letting the user dynamically operate the software to provide comparable measures of reaction times and eye tracking, making it unnecessary to adjust for on-screen differences. The software tools that were used in this experiment include Network Miner, Wireshark, ENCASE Enterprise, AccessData Registry Viewer, PDF Dissector, and Hex Workshop. Subjects were displayed a question regarding a static screenshot prior to being shown the screenshot. This allowed the subject to read the question and get a preview of what tool was about to be shown. After the subject determined they understood the question, they pressed the space bar, and a one-second fixation cross appeared in the center of the screen. Then, they were presented with a full-screen static image. The image would stay on the screen until the participant pressed a key that corresponding to their answer. All answers were single character answers such as the last number of an IP address or the first letter of a password. This allowed the subject to respond using a single button.

After completing the cyber tools task, subjects were asked to complete a domain-general visual cognitive battery. This battery was developed by Sandia National Laboratories and has been used for research in domains involving visual search such as studies of synthetic aperture radar analysts [11] and TSA baggage screeners. The battery includes a mental rotation task, an attentional beam task, a rotation span task, an O/Q visual search task, a T/L discrimination task, and the Sandia Progressive Matrices task.

Preliminary Findings. The preliminary findings suggest that the there are differences between the expert and novice population. The expert reference was based on a subject matter expert that was able to provide an expected gaze path for each of the trials. The novice users often took longer to find the region of interest and were often distracted by extraneous text that drew their attention. Thus, reaction time and time to the first fixation were longer (approximately 3 times on average) in our sample.

2.4 Psychological Measures

Self-report measures allowed quantification of attributes associated with individuals' personality, cognition, and social processes. Using these measures, the gaze search patterns identified by the eye tracking system were compared to the attributes identified by the measures. Three specific attributes were measured, with the general hypothesis that these attributes would differentiate across individual skill level. The following was expected:

- High performing individuals that responded quickly or entered the regions of interest fast would have a higher level of need for cognition, indicating a strong desire to pursue and solve hard problems.
- High performing individuals that answered soon after entering the region of interest would be rational, intuitive decision makers while low performing with individuals eye tracking demonstrating participants entered the region of interest, but failed to answer without first investigating the screen in its entirety, would be avoidant or spontaneous. High performing teams would have higher levels of dependent decision-making style.
- High performing individuals would demonstrate higher levels of conscientious and openness to experience. Roughly half of high performers would be extroverted, indicating that extroversion is not a predictor of performance.

The three assertions were assessed using the following scales. For personality, the Big Five was used. The Big Five Personality Inventory consists of items used to measure neuroticism, agreeableness, conscientiousness, and openness to experience as well as characteristics of extraversion and introversion [3]. Respondents are asked on a Likert scale of 1 to 5 to state how strongly they disagree (1) or agree (5) with a statement about themselves. Example statements are: "Is outgoing, sociable," "Is talkative," and "Is sometimes shy, inhibited." The desire to pursue and solve problems was measured using the Need for Cognition Scale. This is a self-report assessment instrument that quantitatively measures "the tendency for an individual to engage in and enjoy thinking" [4]. Cacioppo and Petty created the Need for Cognition Scale in 1982. The original scale included 34 questions. Two years later, Cacioppo and Petty collaborated with Chuan Feng Kao to shorten the scale to the 18-item format, which is used in the Wabash National Study of Liberal Arts Education. Based on previous research, the Need for Cognition Scale appears to be a valid and reliable measure of individuals' tendencies to pursue and enjoy the process of thinking—that is, of their "need for cognition" [4,5]. Need for Cognition scores are not influenced by whether an individual is male or female, or by differences in the individual's level of test-taking anxiety or cognitive style (the particular way that an individual accumulates and merges information during the thinking process). In general, scores on the Need for Cognition

Scale are not impacted by whether or not the individuals are trying to paint a favorable picture of themselves [4]. Finally, decision-making style was quantified using the General Decision-Making Style inventory (GDMS) [6]. The GDMS measures five different decision-making styles: rational, intuitive, dependent, avoidant and spontaneous. The instrument has 25 questions (5 items for each dimension) rated on a 5-point Likert-type scale ranging from "strongly disagree" to "strongly agree". The GDMS has been shown to be a reliable and valid scale for assessing decision-making. Reliability (Cronbach's alphas) for the different dimensions vary between 0.62 and 0.87 and patterns of correlations with values, measure of social relations, work conditions and other variables provide convergent validity for the GDMS [7,8,9].

2.5 Cyber Environment Instrumentation

Instrumenting a highly distributed competitive cyber training exercise, such as Tracer FIRE is fraught with difficulties. First, the distributed nature of the exercise, meaning the fact that much of the target behavior is occurring across networks of multiple machines, means any data collection solution must also be distributed. Second, because it is a competitive environment, data collection solutions must not interfere with the task at hand, nor be presented as a target for attack within the training environment.

Within the Tracer FIRE environment, we chose three distinct targets for behavioral data collection; the individual participant's workstations, the game server from which challenges are obtained, and the news server from which task-relevant contextual information is presented. We outline each instrumentation approach below.

Hyperion. One of the key sources of behavioral data within the cyber environment is the individual actions taken by participants at the machine level. This includes information such as when the user is active, what tools they are using, and how they are using them. To capture this information, we developed a low-level tool called Hyperion. The Hyperion agent is installed on each of the workstations within the Tracer FIRE environment. Hyperion is capable of logging individual keystrokes, mouse clicks, and window switches. In addition, Hyperion has a window-pathing collection method, which generally allows us to understand exactly which part of an application received an event.

Game Server Instrumentation. We couple the low-level behavioral data collected by Hyperion with high-level Tracer FIRE-specific game information by instrumenting the Tracer FIRE game server. This instrumentation provides information such as a team's progress within the training scenario, challenge receipt, and correct and incorrect solutions. This can be linked with Hyperion data to understand what tools and methodologies were being used during specific aspects of the game scenario.

News Server Instrumentation. Another source of data is the news server. The news server provides non-technical information contextual injects throughout the Tracer FIRE scenario in the form of a CNN-like news site (e.g., media claims by a hacktivist organization). This information can be used within the exercise to facilitate the solution of challenges. The news site also contains “red-herring” type information, which can lead teams astray. Our instrumentation collects when a news article becomes available within the scenario and when a participant actually accesses the information. This information can be correlated with game server and Hyperion data to better understand the impact of outside information on team performance.

Preliminary Findings. Analysis of the cyberspace specific behavioral data is still ongoing. However, initial analysis of the first Tracer FIRE data indicates that there may be a relationship between an increased use of general-purpose tools, such as command-line and scripting tools, and performance in the Tracer FIRE exercise [10]. We hypothesize that one aspect of expertise within cyber security may be knowing the limitations of existing cyber security-specific tools, and knowing when to use general purpose tools to “do it yourself.” These findings will then be compared to see if a relationship between tool use, eye tracking gaze patterns, and personality style exists.

2.6 Electroencephalography (EEG)

The EEG task used in this experiment was a recognition memory task that incorporated repeated and quizzed items. Prior research has shown that the amplitude of event-related potential (ERP) repetition effects elicited by repeated words can indicate whether or not an individual is using an effective learning strategy and can be predictive of future memory performance [12]. Participants with larger repetition effects, indicative of self-quizzing, outperform participants with smaller repetition effects, indicative of more passive study strategies. We hypothesized that participants who used effective learning strategies on a simple memory test would also be high performers on the cyber security training tasks.

Materials. The materials for the memory task consisted of 255 common English nouns. They were divided into 16 counterbalanced study/test lists such that every word appeared in every condition. During the study block, each participant studied 30 words that were presented once, 15 words that were presented and then repeated after a short lag (one intervening item), 30 words that were repeated after a long lag (nine intervening items), 15 words that were studied once and then quizzed after a short lag, 30 words that were quizzed after a long lag, and 45 words that were quizzed but had not been studied previously in the list. After the study block, participants took a short break before beginning the test block. On the test block, participants were tested on all of the studied words (120 total), plus 90 new, unstudied words.

Procedure. A fixation cross appeared in the center of the screen throughout the experiment. Participants were asked to fixate on the cross and to avoid blinking or moving their eyes during the presentation of a word. Prior to the presentation of a study word, a yellow dot appeared above the fixation cross for one second. This was a cue to the participant, indicating that he or she should prepare to memorize a word. Prior to the presentation of a quiz word (during the study block) or a test word (during the test block), a red dot appeared. This meant that participants would be asked to respond to the next word, indicating whether or not they believed they had studied that word previously. After the dot disappeared, the word appeared above the fixation cross. The interval between the disappearance of the dot and the appearance of the word varied randomly between 600 and 800 milliseconds. The word remained on the screen for one second. On quiz and test trials, the word was followed by a question mark. The question mark remained on the screen until the participants pressed a button on the keyboard to indicate their response (yes or no). The next trial began after 250 milliseconds. EEG was recorded using a 16-channel Emotiv headset, a widely-available consumer-grade EEG headset with electrodes soaked in saline solution.

Preliminary Findings. The behavioral results of the memory test were consistent with prior studies. Participants had the lowest memory for words that were only studied once. They had improved memory for the words that were repeated during study, and the best memory performance for words that were quizzed during study. However, the EEG signal from the Emotiv headset proved to be too noisy to extract useable ERPs. The repetition effect is relatively small and it typically centered over the top of the scalp. The position of the Emotiv electrodes and the signal-to-noise ratio were insufficient for calculating the magnitude of the repetition effects. In future research, a lab-grade EEG system with electrodes on the top of the scalp would be needed to test our ERP hypotheses.

3 Discussion

The preliminary findings and data collection proved that it is feasible to perform in-situ testing on cybersecurity individuals while performing domain-specific tasks. These strategies can be leveraged to then compare and contrast difference between high- and low-performers to understand if there are different attributes associated with each group. While some testing, such as EEG, may be suited better for isolated testing outside of the main operations location, others metrics can be collected in real-time to assess behavior, strategy, and teaming dynamics.

While the data collection was mostly focused on the individual, team dynamics can be understood using the answer submissions of each team and its members. However, it was noted that attribution of action is essential to understand if the behavior that is collected under user screen name is what had actually occurred in the scenario or if the other users were operating the keyboard under someone else's name. If the correct data is collected, the team composition and cohesion could be assessed such that personnel with similar decision-making styles are paired together to ensure positive social outcomes. However, with a larger sample of experts, there may be a need to disperse and intermingle the different psychological styles assessed in the self-report measures for optimal success.

For future data collection, the team seeks to acquire more physiological and behavioral data from experts to further compare the differences in the novice data set. For instance, using the SME-generated gaze patterns was enough to provide interesting anecdotal evidence of differences, with a large population of experienced cyber defenders and incident responders, the appropriate statistical measures could be calculated. The combination of all the measurements during the same domain-specific task would also enable more robust comparison of how visual search styles may be related to various neural and behavioral activity in the task.

Acknowledgments

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94AL85000.

4 References

1. Stevens-Adams, S., Carbajal, A., Silva, A., Nauer, K., Anderson, Reed, T., & Forsythe, C. Enhanced Training for Cyber Situational Awareness. Foundations of Augmented Cognition, 8027, 90-99 (2013)
2. Reed, T., Silva, A., & Nauer, K. Instrumenting competition-based exercises to evaluate cyber defender situation awareness. Foundations of Augmented Cognition, 8027, 80-89 (2013)
3. Benet-Martínez, V., & John, O. P. Los Cinco Grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the Big Five in Spanish and English. Journal of personality and social psychology, 75(3), 729-750 (2013)
4. Cacioppo, J. T., Petty, R. E., & Feng Kao, C. The efficient assessment of need for cognition. Journal of personality assessment, 48(3), 306-307 (1984)
5. Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. Psychological bulletin, 119(2), 197-253 (1996)
6. Scott, S. G., & Bruce, R. A. Decision-making style: The development and assessment of a new measure. Educational and psychological measurement, 55(5), 818-831 (1995)
7. Loo, R. A psychometric evaluation of the general decision-making style inventory. Personality and Individual Differences, 29(5), 895-905 (2000)
8. Spicer, D. P., & Sadler-Smith, E. An examination of the general decision making style questionnaire in two UK samples. Journal of Managerial Psychology, 20(2), 137-149 (2005)
9. Thunholm, P. Decision-making styles and physiological correlates of negative stress: Is there a relation?. Scandinavian Journal of Psychology, 49(3), 213-219 (2008)
10. Silva, A., McClain, J., Reed, T., Anderson, B., Nauer, K., Abbott, R., and Forsythe, C. Factors Impacting Performance in Competitive Cyber Exercises. Proceedings of the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC2014) (2014)
11. Matzen, L. E. Effects of Professional Visual Search Experience on Domain-General and Domain-Specific Visual Paper presented at the HCI International, Los Angeles, CA (2015).
12. Matzen, L. E., Haass, M. J. Using Computational Modeling to Assess Use of Cognitive Strategies. Foundations of Augmented Cognition, 6780, 77-86 (2011)