Enhanced Training for Cyber Situational Awareness

Susan Stevens-Adams, Armida Carbajal, Austin Silva, Kevin Nauer,
Benjamin Anderson, Theodore Reed, and Chris Forsythe

Sandia National Laboratories, Albuquerque, NM, USA
{smsteve,ajcarba,aussilv,ksnauer,brander,
tmreed,jcforsy}@sandia.gov

Abstract. A study was conducted in which participants received either tool-based or narrative-based training and then completed challenges associated with network security threats. Three teams were formed: (1) Tool-Based, for which each participant received tool-based training; (2) Narrative-Based, for which each participant received narrative-based training and (3) Combined, for which three participants received tool-based training and two received narrative-based training. Results showed that the Narrative-Based team recognized the spatial-temporal relationship between events and constructed a timeline that was a reasonable approximation of ground truth. In contrast, the Combined team produced a linear sequence of events that did not encompass the relationships between different adversaries. Finally, the Tool-Based team demonstrated little appreciation of either the spatial or temporal relationships between events. These findings suggest that participants receiving Narrative-Based training were able to use the software tools in a way that allowed them to gain a greater level of situation awareness.

Keywords: cyber security, training, situational awareness.

1 Introduction

Situation awareness is essential to effective cyber security analysis and incident response team performance. However, cyber situation awareness has not been well studied (Tadda, 2008). This research sought to help clarify the cyber situation awareness problem, while providing insights that will improve training effectiveness for cyber defenders.

An explosion of new vendor and open source tools has occurred in the past few years to address the growing cyber problem, with U.S. Government enterprise networks and their incident response teams being a primary market. However, these new tools have not always improved the situation awareness of cyber security analysts. Consequently the return on investment has been questionable given the costs of purchase, development and integration of the new technologies.

Nonetheless, cyber security analysts need tools to assist them in fathoming the vast quantities of data and deciphering ever-more sophisticated network attacks. There is

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 90–99, 2013.

need for research to understand why tools that ought to increase the productivity of cyber security analysts often fail to realize this objective. We believe that this failure may be partially attributable to insufficient training and, particularly, the fact that intended users often lack fundamental knowledge essential to effectively use the tools being provided to them. Today, there is no scientific basis for asserting that one mode of training cyber defenders to use software tools is superior to any other mode of training. Likewise, there has been no openly published empirical assessment of students receiving alternative modes of training. The objective of this project was not to compare alternative software tools and no data was collected that reflected on the relative performance or utility of alternative software tools. Instead, through laboratory research employing human performance measurement, the current project scientifically addressed the question of what type of training is needed to maximize the effectiveness of new tools being introduced to improve the situation awareness of cyber security analysts.

1.1 Purpose of Study

The current project employed a suite of network analysis tools comparable to those commonly used in operational cyber settings. Two modes of training were considered. The baseline training condition (Tool-Based training) was based on current practices where classroom instruction focused on reviewing the software functionality with various exercises in which students apply those functions. In the second training condition (Narrative-Based training), classroom instruction addressed software functions, but in the context of adversary tactics and techniques. Upon completion of training, participants were evaluated during a Tracer FIRE (Forensic and Incident Response Exercise) simulated blue team exercise. It was hypothesized that students receiving Narrative-Based training would gain a deeper conceptual understanding of the software tools and that this would be reflected in better performance during the Tracer FIRE exercise.

Three hypotheses were tested. Hypothesis 1: The narrative-based training is different from the tool-based training and will result in better performance in an assessment of students' abilities to use software tools to interpret events associated with a cyber-attack. Hypothesis 2: Personality has an effect on team success and dynamics. Certain personality attributes will result in lower team scores. Hypothesis 3: Cognitive aptitude has an effect on team success. Certain cognitive aptitudes will result in superior team scores.

While research of this nature is commonplace in other high consequence domains (e.g., military operations), there exists little precedent within the cyber security domain. Accordingly, the cyber domain introduces unique challenges. For instance, scenarios must be presented that are unique and somewhat realistic, yet offer equivalent outcome measures of performance. Process measures must be identified and implemented that allow data to be collected in a non-obtrusive manner such that measurement does not interfere with participants exercising the skills and knowledge being measured. Furthermore, outcome and process measures must be identified that are generalizable to and predictable of performance within operational settings. By beginning to address these issues, the proposed project advances the domain of cyber science through development of unique experimental methodologies, while providing a deeper understanding of situation awareness within the cyber domain. Furthermore, the current study offered an opportunity to collect data regarding secondary research questions concerning the effectiveness of cyber operations. Cyber security is a major challenge for DOE and other government agencies and there has been little scientific study of the human dimension of cyber operations.

Through the current study, data was collected that addressed group processes, the relationship between certain cognitive and personality attributes and the behavior and performance of cyber defenders, and the use of narrative in constructing stories to understand, explain and remember events in the cyber domain.

2 Methods

2.1 Participants

Thirteen employees from Sandia National Laboratories volunteered to participate in the experiment. All participants met the following requirements: (1). be 18 years or older, (2). have a background in computer science, (3). have an interest in cyber security/cyber incident response, (4). have not participated in any prior Tracer FIRE events and (5). be available on the designated dates for five full days of training and three full days to participate in the Tracer FIRE evaluation exercise.

2.2 Materials

The suite of network analysis tools used in the experiment included Encase Enterprise, Wireshark, IDA Pro, Volatility, Hex Workshop and PDF Dissector. Teams were additionally provided IRC chat as a means for intra-team communication and Plotweaver as an aide in creating a record of events.

2.3 Procedure

Participants were first asked to fill out a consent form and then complete a prescreening questionnaire. Next, the participants were asked to fill out a demographic questionnaire and a detailed questionnaire assessing general computer security and cyber incident response skills. This information was later used to assign individuals to the two training conditions and subsequently to place the participants into teams for the Tracer FIRE exercise. The objective was to assure that the three teams competing in the exercise were relatively balanced with respect to the knowledge and experience of team members.

Training. Participants were assigned to either the Tool-Based (7 participants) or the Narrative-Based (6 participants) training conditions. The two training groups received 3 days of training appropriate for their condition. The two training groups were then combined for 2 additional days of training which addressed details concerning the use of the selected tools. This training was not as extensive as that provided in the Tool-Based training and emphasized the knowledge participants would need to solve the challenges in the Tracer FIRE exercise.

Tool-Based Training. Participants assigned to the Tool-Based training condition received 3 days of training focused on the functions incorporated into the tools and the mechanics of using the tools. This training involved relatively little information concerning adversary tactics and techniques and was comparable to training commonly provided by software vendors and included canned examples showing how the tools work, with relatively little emphasis on the application of the tools to real-world problems.

Narrative-Based Training. Participants assigned to the Narrative-Based training condition received 3 days of training emphasizing the theory of adversary tactics, application of tools and a detailed understanding of the role as a cyber incident responder. This training involved little consideration of the functionality of tools used for conducting network analysis. The training was structured in a manner that sought to help students comprehend the complex ideas and information in a form that was personal and formed relationships between their prior knowledge and personal experiences.

Tracer FIRE exercise. Following the 5 days of training, the participants were placed in one of three teams for the Tracer FIRE exercise. The Tool-Based team comprised of four participants whom had all received Tool-Based training. The Narrative-Based team comprised of four participants whom had all received the Narrative-Based training. The third group, the Combined team, composed of five participants; three of whom had received the Tool-Based training and two of whom had received the Narrative-Based training.

Each team was asked to solve multiple challenges to receive points with the score for each team continuously displayed and teams encouraged to compete against each other. The challenges were built around a coordinated series of events involving the same multi-level attack upon a host network of each team. The challenges required the teams to use the software tools addressed during training to analyze network traffic. This provided the basis for their interpreting events and establishing overall situational awareness. Points were awarded on the basis of successfully answering challenge questions concerning specific aspects of the attack, as well as their ability to form an accurate picture of the overall pattern of events (i.e., situational awareness).

Secondary Measures. Subjects were asked to complete a personality assessment consisting of the Big Five Inventory (BFI) from the website www.similarminds.com. Participants were also asked to perform three cognitive tasks: syllogism, comprehension span and mental rotation. These tasks have been used in previous studies and address different cognitive aptitudes associated with adaptive thinking and decision making. The object was to assess whether these same aptitudes correlated with performance for cyber defender tasks.

Syllogism. This task is a measure of reasoning. Participants were given a logical argument in which a proposition is inferred from a set of premises and were asked to indicate whether the proposition was true given the premises.

Comprehension span. This task is a measure of verbal comprehension and associated memory recall. The participants saw a sentence and had to indicate whether the sentence made sense or not. After a series of sentences, the participant was asked to recall the last work of every sentence in order.

Mental rotation. This task is a measure of visual-spatial ability and mental flexibility. The participant was presented with a series of 20 pairs of figures. The task was to indicate whether the two figures, one of which was often rotated a specific amount of degrees, corresponded to the same object. The number correct that were classified in 60 seconds was taken as a measure of mental rotation ability.

Finally, at the beginning of the Tracer FIRE exercise, participants were told that there was a story embedded within the upcoming series of challenges. Furthermore, it was their task to discover this story as they solved the various challenges. It was encouraged that teams pay attention to cues associated with the stories and take notes to help them later piece together these cues. Then, at the end of the exercise, teams were given 30 minutes to construct an illustration depicting their interpretation of events and the underlying story.

3 Results

3.1 Descriptive Statistics

Participants were assigned to teams in a manner that provided a relative balance in the skills and experience of the individual team members. With respect to the questionnaire assessing general computer security and cyber incident response skills, the sums of the test scores for each team were Tool-Based training (Team 1) = 354, Combined training (Team 2) = 374, and Narrative-Based training (Team 3) = 347.

3.2 Training Type and Team Differences

The Narrative-Based team received the most points (11,182) followed by the Tool-Based team (10,480) and Combination team (9,811), respectively. This was also reflected in the average number of points received by team members; members of the Narrative-Based team individually scored more points on average than members of the other two teams.

A general linear model ANOVA with two factors was conducted to determine if there was a “training type” or “team” effect on the number of points obtained by teams. There were two levels in the training type: Narrative-Based or Tool-Based training. There were three levels in the team factor: Team 1 (Tool-Based), Team 2 (Combined), and Team 3 (Narrative-Based) training. Neither training type nor team factor was significant and there was no statistical difference between team scores (i.e., "success") based on the training type or team.

3.3 Team Narratives

The teams were asked to prepare an illustration describing the story underlying the various events encompassed in the Tracer FIRE exercise. Figure 1 shows the ground truth depicted by the Plotweaver tool. As can be seen, there were multiple actors who intersected one another at key point in time. This was a multi-layered scenario that unfolded over time and it was not expected that any of the teams would be able to fully deduce all of the relationships that occurred across time and space.

Figure 2 shows the illustration prepared by the Narrative-Based team. It is apparent that this team failed to deduce many of the relationships between actors and events. However, this team did recognize five separate plot lines that loosely corresponded to those depicted in the ground truth storyline. Likewise, they recognized seven of the thirteen points at which the plotlines intersected one another. These two measures are believed to be indicative of the team's overall situation awareness which, as discussed below, was superior to that of the other two teams.

Figure 3 shows the illustration by the Combined team. This team deduced plotlines that loosely corresponded to four of the five plotlines within the ground truth depiction. Likewise, this team recognized the sequential development of events across time. However, it is striking that this team did not recognize any of the points where the individual plotlines intersected with one another. In fact, in both their hand-drawn illustration and their verbal account, the Combined team presented a linear sequence of events that did not involve any interactions between events, or individual actors. This team pieced together a story involving four separate actors that, for the most part, operated independently, when, in fact, the actors operated in concert with one another and this was a key element to interpreting the overall sequence of events. While this team clearly grasped the temporal structure of events, as well as the importance of individual actors, they were unable to deduce the relationships between different actors that were evidenced through their interactions as the scenario unfolded.

Figure 4 shows the illustration produced by the Tool-Based team. The Tool-Based team produced an even more impoverished illustration than either the Narrative-Based or Combined teams. They recognized three of the five plotlines. Yet, they recognized none of the relationships between the separate plotlines and two of the three plotlines that they did recognize consisted of a single event. Furthermore, their depiction captured none of the relationships between events or the relationships between different actors. Each member of this team seemed to have deduced one or more elements of the story independently; however, as a team, they were unable to put these elements together and did not seem to recognize that there was a coordinated action being taken by the adversaries.

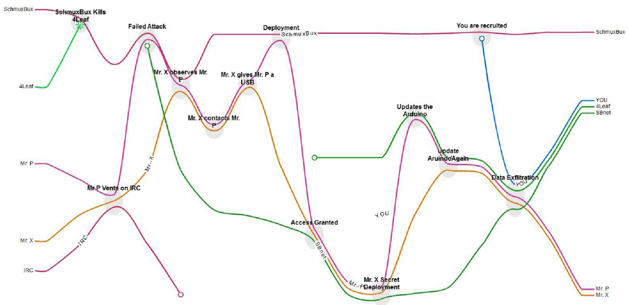

Fig. 1. Plotweaver Depiction of Ground Truth for Tracer FIRE Scenario

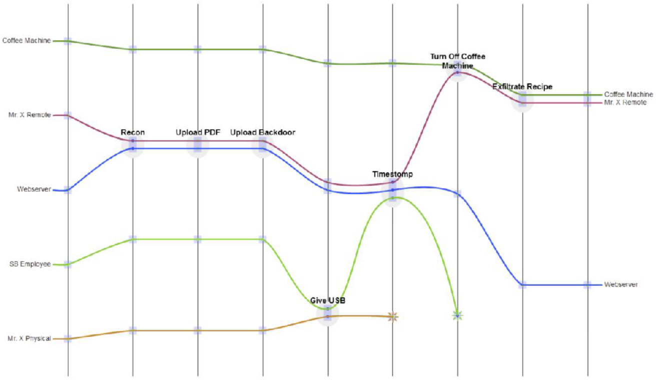

Fig. 2. Plotweaver Illustration Prepared by the Narrative-Based team

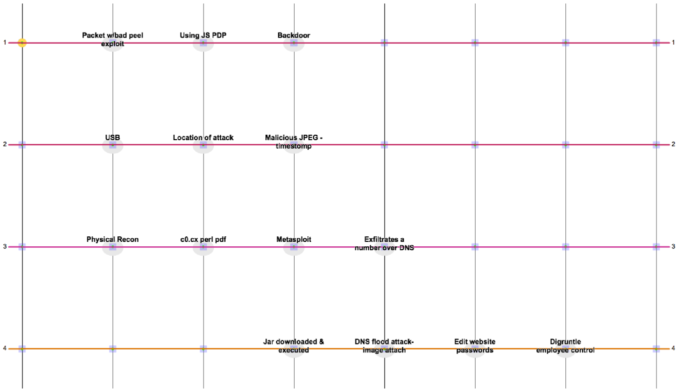
1
2
3
4
Packet w/bad peel exploit
Using JS PDP
Backdoor
USB
Location of attack
Malicious JPEG - timestomp
Physical Recon
c0.cx perl pdf
Metasploit
Exfiltrates a number over DNS
Jar downloaded & executed
DNS flood attack-image attach
Edit website passwords
Digruntle employee control
1
2
3
4

Fig. 3. Plotweaver Illustration Prepared by the Combined team

Interestingly, it was noted that all three teams deduced about the same number of story elements. During the Tracer FIRE exercise, there were specific challenges that, if successfully completed, teams learned a key element of the storyline. While the Narrative-Based team earned the most points in these challenges, there was not a huge difference between the points earned by the Narrative-Based and the other two teams. This indicates that all three teams had many of the key story elements available to them but only the Narrative-Based team was able to put those story elements together in a way that corresponded to the actual relationships between events.

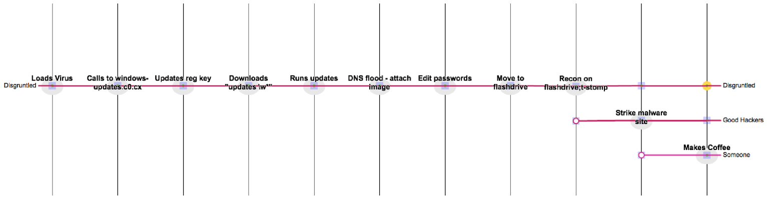

Fig. 4. Plotweaver Illustration Prepared by the Tool-Based team

3.4 Personality Factors

The BFI data was analyzed to determine if personality measures were associated with team success. Two subjects opted out of the personality assessment portion of the study. Therefore, only 11 participant's data was analyzed. A correlation matrix was calculated to determine if there was multicollinearity (Penney et al., 2011). The variables showed random scatter and no significant correlation.

A stepwise regression was conducted to determine if any of the variables were significant (alpha = 0.15, was set for selection in the stepwise regression). The final model included TimeTotal (total amount of time spent working on challenges), Inquisitiveness, and Emotional Stability. There were no departures from normality or outliers, and the residuals displayed constant error variance, with the error terms normally distributed. These data indicate that participants fell within the range that would be considered normal within the overall population and, therefore, results cannot be attributed to individual subjects with extreme scores on the Inquisitiveness or Emotional Stability personality dimensions. TimeTotal was not significant, but was included in the model as β1. Inquisitiveness and Emotional Stability were significant (R^2 = 58.87 and R^2-*adjusted* = 41.25, β0 = -7491, β1 =1.148e-12, *t*-value = 1.83, *p*-value < .1093). Inquisitiveness was marginally significant (β2 = 63, *t*-value = 2.02, *p*-value= .083) as was Emotional Stability (β3 = 88, *t*-value = 2.98, *p*-value= .021). Only Emotional Stability was included in the final model.

3.5 Cognitive Factors

The three cognitive tasks, Mental Rotation (MRScore), Comprehension Span (CompS) and Syllogism (Syllo), were analyzed to determine if they were associated with team success. Four subjects opted out of the cognitive task portion of the study. Therefore, only 9 participant's data was included.

The final model included CompS (R^2 = 47.25, R^2-*adjusted* = 39.71, β0 = 1459, β1 = 32 with a *t*-value =2.50, *p*-value <0.041). There were no departures from normality, no outliers, the residuals displayed constant error variance, and the error terms were normally distributed. Thus, the results could not be attributed to individual subjects who exhibited extreme scores, as all subjects were within the range that would be considered normal for the population.

4 Conclusion

The results from this study provide insights concerning alternative methods for delivering training for cyber defenders, as well as a better understanding of factors contributing to team situation awareness and individual and team performance of cyber defenders. Most notably, this study highlights the importance of the narrative, or the capacity to interpret events and put them into the context of a story, to the effective use of software tools by cyber defenders. Furthermore, the study also illustrates the importance of individual characteristics to the ability of individuals to effectively work together within a cyber incident response team.

With only three teams, it was not possible to demonstrate a statistically significant difference in the performance of the teams receiving alternative modes of training, although the team receiving Narrative-Based training did earn more points than their counterparts. Likewise, on average, the members of the Narrative-Based team individually earned more points than their counterparts on the other teams. While not statistically significant, these results are in the expected direction and are consistent with detailed analysis of overall situation awareness exhibited by the three teams.

Assessments of personality and cognitive factors revealed two variables that were significantly correlated with individual performance during the cyber exercise. With respect to personality, those who exhibited higher scores on the Emotional Stability dimension performed better. Those scoring high on this dimension tend to be more secure and confident, whereas those scoring low exhibit a greater tendency to show unpleasant emotions such as anger, anxiety, depression and vulnerability. It should be noted that while the participants in the current study exhibited a range of scores on this dimension, their scores fell within the range considered normal for the overall population.

There are two important ramifications for the finding that individual performance correlated with Emotional Stability. First, during training, the Emotional Stability of individual students may be expected to affect both the benefit derived from the training experience, as well as the performance during training exercises, such as Tracer FIRE. Thus, it is proposed that mechanisms be employed that allow individual and team performance to be more closely monitored in real-time so that instructors may effectively intervene when students have become non-productive and are struggling. Likewise, in composing teams, it may be beneficial to combine individuals with varying experience and maturity to provide some degree of scaffolding for weaker team members who may become easily discouraged.

Second, and perhaps more importantly, within operational settings, it may be expected that personnel will exhibit varying levels of Emotional Stability and this will have an indirect, and perhaps direct, effect on their performance. This may be manifested in their capacity to effectively function within teams, as well as their capacity to cope with ongoing stressors. It is uncertain what countermeasures may be most appropriate; however this represents an important consideration given the nature of the Cyber domain where technically qualified personnel are in high demand and many organizations find it difficult to retain their best talent.

A second individual factor that correlated significantly with performance was Comprehension Span. In this task, subjects were presented a series of sentences and after each sentence, they were required to indicate if the sentence made sense. Then, their memory span was tested by requested that they recall the last word in each sentence. To perform well, an individual must have both proficient at interpreting verbal content and possess good short-term memory. Previous studies have shown that individuals who perform well on this measure also perform well in tasks requiring adaptive decision making. Here, adaptive decision making is defined as the capacity to recognize that a strategy is ineffective and thus, there is need to either alter an existing strategy or abandon an existing strategy for an alternative strategy (Abbott et al., 2011). It is proposed that the challenges presented through the Tracer FIRE exercise place similar demands for adaptive decision making upon the participants and that Comprehension Span represents a fundamental cognitive attribute underlying effective performance.

References

1. Abbott, R., Haass, M., Trumbo, M., Stevens-Adams, S., Hendrickson, S., Forsythe, C.: Robust Automated Knowledge Capture, SAND 2011-8448, Sandia National Laboratories (October 2011)
2. Penney, L.M., David, E., Witt, L.A.: A review of personality and performance: Identifying boundaries, contingencies, and future research directions. Human Resource Management Review 21, 297–310 (2011)
3. Tadda, G.P.: Measuring the Performance of Cyber Situational Awareness Systems. In: Proceedings of the 11th International Conference on Information Fusion, Cologne GE, June 30-July 3 (2008)