Automated Performance Assessment in Cyber Training Exercises

Robert G. Abbott, Jonathan McClain, Benjamin Anderson, Kevin Nauer, Austin Silva & Chris Forsythe

Sandia National Laboratories

Albuquerque, NM USA

rgabbot@sandia.gov, jtmccla@sandia.gov, brander@sandia.gov, ksnauer@sandia.gov, aussilv@sandia.gov, jcforsy@sandia.gov

ABSTRACT

Cyber threats have become ubiquitous as criminals extend their reach and cyber becomes a front in conflicts between different peoples and a major source of revenue for criminal organizations. Personnel responsible for cyber defense are becoming increasing critical. However, there is a shortfall between the number of individuals training to enter cyber security and the projected demand for these skills. Consequently, methods and technologies are needed to enhance and accelerate the training of cyber security personnel.

Previous research has demonstrated the benefits of automated performance assessments as a means to target training to the specific needs of individual students. The current paper describes an extension of these capabilities to cyber security training exercises. In these exercises, students are placed in teams and must work together, using appropriate software tools and online resources, to conduct forensic analysis for cyber crimes. Individual and team performance is assessed on the basis of successfully solving individual challenges and applying information from individual challenges to correctly ascertain an overall picture of the who, what and why of the crimes.

The current paper describes a framework for conducting cyber security training exercises with an emphasis on instrumentation to enable automated performance assessment. Instrumentation captures students' computer-based transactions in a log that is time-synched with the game-server used to deliver challenges and register student responses. Analyses were conducted to better understand the factors that distinguish more or less effective student performance and techniques developed to automatically parse logs of student activities into meaningful blocks of task-oriented activity. These capabilities are a prerequisite for the development of real-time automated assessment of student performance within the context of cyber security exercises.

ABOUT THE AUTHORS

Dr. Robert G. Abbott is a Principal Member of the Technical Staff in the Cognitive Systems group of the Cyber Engineering Research Institute where he leads research in cognitive and behavior modeling.

Jonathan T. McClain is a member of the Cognitive Systems organization. His research interests include instrumentation for automated knowledge capture and measuring human performance.

Benjamin Anderson is a member of Sandia National Laboratories' Information Design Assurance Red Team (IDART) and a member of the RECOIL Lab where he conducts research and training in cyber forensics.

Kevin Nauer has over ten years experience conducting forensic analysis and leading a team of analysts to conduct incident response operations. He leads a development effort to create a framework to support collaborative cyber security incident response operations, including training cyber analysts through competition-based exercises.

Austin Silva is a cognitive scientist trained in Educational Neuroscience. His expertise and research interest lies at the intersection of human performance measurement, novel technologies, and applied learning.

Dr. Chris Forsythe is a Distinguished Member of the Technical Staff in the Human Factors organization. He has over 25 years of experience conducting research regarding human performance and the use of technology to enhance performance in training and operational environments.

Automated Performance Assessment in Cyber Training Exercises

Robert G. Abbott, Jonathan McClain, Benjamin Anderson, Kevin Nauer, Austin Silva & Chris Forsythe

Sandia National Laboratories

Albuquerque, NM USA

rgabbot@sandia.gov, jtmccla@sandia.gov, brander@sandia.gov, ksnauer@sandia.gov, aussilv@sandia.gov, jcforsy@sandia.gov

INTRODUCTION

Cyber security has become a priority as cyber warfare emerges as a new threat in conflicts between nations and groups, and criminal organizations and nations turn to cybercrime as a source of revenue (Bernik, 2014; Lee, 2014; Pederson, 2014). Consequently, within government, military and private organizations, personnel responsible for cyber defense are becoming increasingly critical. At the same time, a "human capital crisis" has been described wherein the number of individuals receiving the essential cyber security training does not meet the current and projected demands of the cyber security workforce (Fourie, et al., 2014; Rob, et al., 2014)). Various innovative approaches have been proposed for delivering cyber security training to students ranging from secondary schools to college to cyber security professionals (Danforth & Lam, 2014; Martini & Choo, 2014).

The technical, practical and pedagogical challenges of delivering effective training in cyber security for students and professionals have been documented (Boesen, et al., 2014; Goodman, 2014). A common element within most proposals has been the use of exercises that simulate the experiences of a professional cyber defender (Gavas, Memon & Britton, 2012). These exercises may be designed to teach basic concepts, provide exposure to adversary tactics and techniques, exercise the use of cyber security software tools, practice forensic analysis and other skills, and promote the development of teamwork (Reed, Nauer & Silva, 2013). These training platforms additionally provide a basis for conducting research to better understand factors that contribute to effective performance, and distinguish novice, competent, expert and elite cyber professionals (McClain, et al, 2015; Silva, et al, 2013).

However, cyber security exercises, and particularly, large competitive exercises, present challenges for instructors in providing the opportunity to monitor student progress and effectively intervene to address gaps in the knowledge and skills of individual students. This issue has previously been described in other complex settings, such as military simulation-based training, in which teams of students must apply what they have learned to address dynamic situations and there is a low ratio of instructors to students (Stevens, et al., 2009a, 2009b). It has been proposed that automated performance assessment might provide a remedy that enables instructors to cope in situations that do not allow the desired level of attention to individual students. Specifically, through automated techniques, student performance may be tracked on a moment-to-moment basis and evaluated relative to expert performance or other criteria. Instructors may periodically check in with individual students to identify specific needs and target interventions to those needs. Furthermore, as the automated system monitors performance for low to medium level knowledge and skills, instructors may devote their time to assessing higher level performance objectives (e.g., situation awareness), which are often difficult to assess through automated techniques. Implemented within the context of naval aviation, this concept has been demonstrated to produce superior training outcomes (Stevens-Adams, et al., 2010a, 2010b).

Automated performance assessment requires development of models that capture the essential features that distinguish performance within the targeted domain. The current paper describes research to develop models of performance within cyber security training exercises that provide a foundation for development of automated performance assessment. The context is one in which students participate in a cyber security exercise using computer systems that have been instrumented to capture a log of their ongoing activities with software applications and accessing Internet resources. A previous paper described techniques that allow behavioral activity logs to be automatically parsed into blocks of activity corresponding to specific high level goals (e.g., solving a specific challenge within the training exercise) (Abbott, et al., 2015). It was found that most blocks of activity extended for

a relatively brief duration (17-18 min, on average). During this time, on average, there were 45 distinct human-machine transactions involving 4-5 different software tools, with an average of 22 transitions between software tools. This suggests a pattern where participants engaged in relatively brief bouts of activity in which they employ, transition between and must integrate across several different software tools.

Further analysis has compared participants with different levels of self-reported knowledge and experience (McClain, et al., 2015). It was found that at a behavioral level, the patterns of activity for students with differing levels of experience were similar within blocks of activity for measures such as duration, the number of software tools used and the number of transitions between software tools. Thus, at a basic level, those with varying levels of experience behave similarly. However, as had been previously reported (Silva, et al., 2014), those with more experience performed better and made more use of general purpose software tools (e.g., Windows command line, Windows Task Manager, Windows Explorer file directory). Additionally, it was fund that those who performed best, utilized a wider range of software tools, than those who performed less well.

The current paper describes further research to understand behavior associated with problem solving within the context of cyber security training exercises, and applies the resulting insights as a foundation for techniques enabling automated assessments of student performance. Whereas previous research addressed relationships between the use of cyber security software tools, and experience and knowledge of cyber security topics and software tools, the current paper considers the behavioral factors that contribute to superior performance within a cyber security exercise. These analyses offer a basic understanding of the patterns of behavior demonstrated within exercises and are the basis for development of techniques to automatically assess student performance. Refinements are described to techniques previously reported for automatically parsing logs of student computer transactions into task-oriented blocks of activity (Abbott, et al., 2015), including a validation study comparing automatically parsed and human manually parsed records. The results should benefit instructors currently conducting exercise-based cyber security training in suggesting clues that certain students or teams may be struggling, and offering a basis for cursory appraisals of the general level of capability of individual students. These findings should be of particular interest to those involved in development of programs to deliver cyber security training through online mechanisms where students have minimal contact with human instructors, yet instructors have a detailed record of student computer system transactions (Son, Irrechukwu & Fitzgibbons, 2014).

METHODS

Data utilized for the current study were collected during cyber security exercises conducted by Sandia National Laboratories known as Tracer FIRE (Forensic Incidence Response Exercise). The Tracer FIRE platform was developed for the U.S. Department of Energy for training of cyber security incident response teams. It consists of a multi-day program combining classroom instruction with a competition-based cyber security exercise. In the competitive exercise, participants work as teams of three to six individuals. Each individual is provided a laptop equipped with software tools commonly used to conduct cyber security forensic analysis (e.g., Encase Enterprise, Wireshark, PDF Dissector), as well as utilities available with Microsoft Windows (e.g., Windows Explorer, Windows Task Manager, Notepad) and Microsoft Office (e.g., Excel, Powerpoint). Throughout the exercise, students are provided with Internet access and are free to download and install other software tools on the assigned laptop computers.

The exercise involves a multi-level challenge. Participants are presented a series of individual challenges that require them to apply forensic analysis skills and utilize the cyber security software tools provided to them. For example, students may be provided with a pdf file that is infected with malware and to solve the challenge, they must extract the malware and determine its function. A game server provides the interface used by students to access challenges, submit answers and receive feedback regarding their submissions. A scoreboard lists each team and their current point totals.

The objective is to solve individual challenges to gain points, but at the same time, students are asked to piece together clues presented through the individual challenges to make inferences concerning a larger scenario, which involves determining the who, what and why of a collection of cybercrimes. In addition to these clues, there is a news server that periodically posts news stories related to the overall context in which the cybercrimes occurred (e.g., public announcement by hactivist group). News stories can be accessed at the discretion of students and

provide clues supplementing those obtained by solving the individual challenges. At the end of the exercise, teams present their conclusions regarding the overall crime(s) and explain how they reached these conclusions.

Subjects

The subjects providing data for the current study were 26 individuals who consented to data collection during two separate Tracer FIRE cyber security training exercises. There were 11 subjects from an event that occurred during the spring of 2014 and 15 subjects from a second event that occurred in the summer of 2014.

Procedure

At the beginning of the competitive exercise, an announcement was made offering students the opportunity to participate in a study to understand behavior associated with cyber security forensic analysis. Those volunteering to participate were asked to complete the informed consent process.

The Tracer FIRE platform is instrumented to log records of student human-machine transactions. These logs capture data regarding events that includes:

- Participant UserID
- Timestamp
- Interval since previous transaction (i.e., duration)
- Challenge ID, for transactions involving the game server
- Event Type, for transactions involving game server
- Submission, answer submitted for transactions involving submitting answer to game server
- Points Awarded, for transactions involving submitting answers to the game server
- Software Tool, for transactions involving software tools
- Class of Event (Windows, Game Server or News Server)
- Article ID, for transactions involving the News Server

Within the course of the Tracer FIRE event, data was collected for all of the participants. After the event, data for participants consenting to participate in the research study was transferred to another system, and logs were wiped clean, erasing the data from subjects that did not consent to serve as research participants. Other than the informed consent, data collection occurred non-intrusively, with the experience of research participants being no different than that of students that did not consent to participate.

RESULTS

Data Parsing

A separate publication has described the techniques utilized to parse data logs into blocks of activity associated with periods of time in which a student worked on a single high-level objective (Abbott, et al., 2015). In Abbott et al. (2015), descriptive statistics are reported that provide a characterization of the behavior observed within blocks of activity, as well as the utilization of software tools over the course of the exercise. For the current study, several refinements were made to the automated parsing techniques described by Abbott, et al. (2015). These refinements primarily concerned a clearer delineation of activities associated with the initial set-up and configuration of individual computers for the exercise and improved handling of the log entries generated by the game server.

A key question for the current study concerned the validity of the automated parsing technique and whether the blocks of activity identified by the automated techniques would correspond to a human scoring the same activities. To address this question, a sample 2,364 log entries from the total of 48,481 log entries were manually scored. Each log entry consisted of a discrete action taken by the student that could involve opening/exiting a software tool, opening/closing a file, opening/closing a window, switching between windows, issuing an Internet request, etc. Within the sampled block of activity, the human scorer identified 54 blocks of activity, compared to the automated technique which identified 56 blocks. The scoring of blocks by the human and automated technique was compared

by considering each boundary between the end of one block and the beginning of the next block identified by the human scorer and determining whether the automated technique identified a boundary on the same row, or within 1, 2, 3, 4 or 5 rows. The results are shown in Figure 1. It can be observed that while the automated technique identified a boundary on the same row as the human in 43% of the cases, the automated technique was within one row of the human in 70% of the cases and within two rows in 94%.

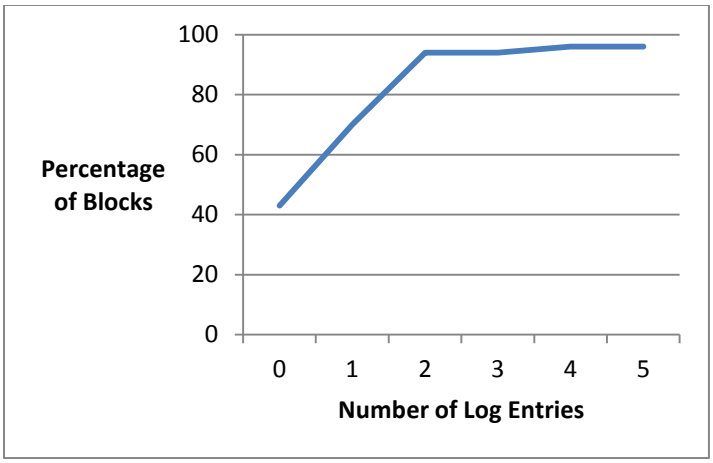

Figure 1. Comparison of scoring of blocks of activity by human scorer and automated technique with percentages reflecting the proportion of blocks identified by the human for which the automated technique identified a boundary between the end and beginning of a block within a given number of log entries.

Performance Analysis

McClain, et al. (2015) discussed an initial analysis of performance with regard to the self-reported level of experience of participants. This analysis found that there was no relationship between self-reported experience and several measures of behavioral activity (e.g., duration of blocks of activity, number of tools used within blocks of activity). However, self-reported experience was related to the use of specific software tools. For example, there were more instances using the Chrome Internet browser for more experienced participants. The current analysis extends these findings by considering performance on the individual challenges.

The Tracer FIRE game server is configured so that each member of a team can open a given challenge. However, once a team member has opened a challenge, no other member of the team can open that challenge. Furthermore, each individual can only have one challenge open at any given time. While individuals on a team will often collaborate on individual challenges, the configuration of the game server implicitly encourages students to open a challenge and focus their efforts on that particular challenge. The current analysis considered what factors were related to successfully solving challenges.

For this analysis, blocks of activity were separated into those for which participants submitted an answer and those for which no answer was submitted. There were a total of 379 blocks of activity, with submissions occurring within 155 blocks. The initial analysis considered what factors indicated that a student was likely to make a submission within a given block of activity. A stepwise regression was calculated in which Submission (Y/N) served as the response and the Total Time, Actions per Tool Usage, Average Time per Tool Usage, Number of Different Tools Used, Number of Transitions between Tools and Number of Returns to Tools for blocks of activity served as predictors. The resulting model accounted for very little variance (R-Sq=4.82 (Adjusted R-Sq=4.30)), with two factors reaching statistical significance. When subjects frequently returned to a previously utilized tool within a block of activity, they were more likely to make a submission ($t=2.94$; $p<0.003$) and when subjects devoted more actions to specific tools (i.e., fewer transitions between tools), they were more likely to make a submission ($t=-2.68$; $p<0.008$). It is worth noting, as illustrated in Figure 2, the duration and the number of tools used in blocks of activity in which a submission was made was remarkably similar to blocks for which there was no submission.

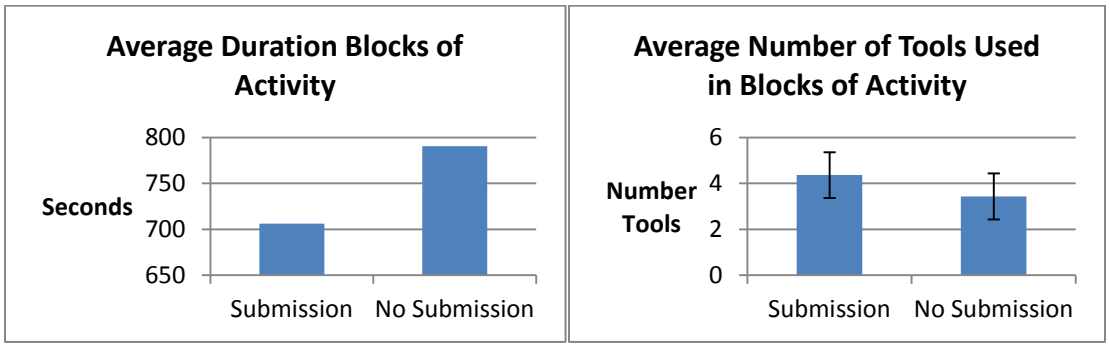

Figure 2. These figures illustrate the similarity of blocks of activity in which there was an answer submission and those in which there was not. Standard deviations for duration are not shown in the left figure. For Submission, standard deviation was 942.8 sec and for No Submission 980.1 sec.

Subsequent analysis assessed specific software tools as predictors of whether students made a submission. For this analysis, the number of actions with each software tool within a block of activity was considered. A stepwise regression produced a model with three factors (R-Sq=8.97 (Adjusted R-Sq=7.96)). In order of predictive value, blocks of activity with a submission were predicted on the basis of the number of actions with Wireshark ($t=4.65$; $p<0.0001$), and Java Decompiler ($t=2.30$; $p<0.05$), with the number of actions with the Windows Task Manager being a negative predictor ($t=-2.03$; $p<0.05$).

The second question considered what factors predict that a student will make an incorrect submission. As described previously for submissions, characteristics of the blocks of activity (e.g., duration) accounted for very little variability (R-Sq=2.82 (Adjusted R-Sq=2.17)). There was one predictor in this model: the number of different tools, with students using more different tools being more likely to submit an incorrect answer ($t=2.09$; $p<0.05$). Stepwise regression analysis of specific software tools generated a model with three predictors (R-Sq=11.64 (Adjusted R-Sq=9.85)). The one positive predictor of an incorrect submission was the number of actions with Windows Explorer ($t=3.05$; $p<0.003$), with the number of actions with the command line interface ($t=-2.98$; $p<0.003$) and WinRAR ($t=-2.15$; $p<0.05$) having a negative relationship.

The final analysis asked what factors are predictive of students making a correct submission. None of the characteristics of the blocks of activity reached the level of statistical significance criteria set for inclusion in the model (i.e., $p<0.05$). A consideration of the number of actions using specific tools yielded a model with three predictors (R-Sq=9.14 (Adjusted R-Sq=7.30). In the order of their predictive weight, the three software tools were: java ($t=2.52$; $p<0.05$); Chrome ($t=2.38$; $p<0.05$) and java decompiler ($t=2.03$; $p<0.05$).

DISCUSSION

Within the current research paradigm, a basic capability essential to development of automated performance assessment for cyber security exercises is an ability to automatically parse logs of human-machine transactions into meaningful blocks of activity. Here, a meaningful block of activity is defined as a series of actions directed to achieving a high-level goal. Once data logs have been parsed into blocks of activity, specific activities can be assessed to determine the most likely objective and patterns of activity evaluated with respect to their appropriateness in achieving the identified goal. Techniques for accomplishing the latter capabilities have previously been demonstrated within other contexts and should fairly readily transfer to cyber security forensic analysis (Stevens-Adams, et al., 2010a; 2010b).

The current study found that the automated techniques placed the boundary between the end of one activity and the beginning of the next within two rows of a human scorer in 94% of the cases. To place this in perspective, on average, a block of activity extends for 17-18 min, encompassing 45 actions, with each action being 17-18 seconds in duration. Thus, the worst case using automated techniques is that in 6% of the blocks of activity, there would be

1-2 actions, out of a set of 40-50 actions, inadvertently included/excluded. This level of performance should be sufficient as a basis for further development that evaluates actions taken with respect to the goal(s) attributed to a given block of activity.

Within the course of an exercise, students frequently undertake an extensive series of activities that ends with their having failed to obtain the information necessary to formulate an answer to a challenge. A common cause for failing to make a submission is that the student either does not understand the question or does not know the initial steps needed to make progress toward an answer. Additionally, students may misunderstand the challenge and undertake an unproductive series of activities. From an instructional perspective, it would be helpful to recognize when these situations exist, particularly given that students have a limited period of time to participate in exercises and often, there is significant cost associated with their participation. Results from the current study suggest two distinguishable patterns of activity. In one, the student understands the challenge and the software tool(s) needed to solve the challenge, and while they may utilize other tools in the course of solving the problem, they return to one or more tools. With the second pattern, a student either does not understand the challenge or does not recognize the software tool(s) essential to solve the challenge, and jumps around from tool-to-tool looking for some insight into how to proceed. This does not mean that successful students use more tools overall, but instead, within a block, they frequently transition from tool-to-tool. In fact, when subjects used more tools within a block of activity, they were more likely to submit an incorrect answer. However, it is important that students understand that successful cyber forensic analysis within the context of training exercises (McClain, et al., 2015; Silva, et al., 2014) and in actual operations (Reed, et al., 2014) involves the integrated use of multiple software tools.

The nature of the challenges incorporated into a competitive exercise will dictate the software tools that are most appropriate to successfully solve the associated problems. However, software tools may be distinguished on the basis of their being either specialized for use in cyber security domains (e.g., Wireshark, IDA Pro) or being general purpose tools that may be used in a wide range of applications. Previous studies have reported that the more successful students made broader use of general purpose tools than less successful students (Silva, et al., 2014) and that the more experienced cyber security professionals made more use of general purpose tools than less experienced participants (McClain, et al., 2015).

The current study found support for the contention that successful performance in cyber security forensics is associated with less exclusive use of specialized software tools. Use of the command line interface was associated with a lower likelihood of incorrect submissions and use of the Chrome web browser was predictive of correct submissions. However, in the current analysis, use of specialized software tools was predictive of success. Those making greater use of Wireshark and java decompiler were more likely to make submissions and those making more extensive use of java and java decompiler were more likely to make correct submissions. These findings do not refute previous assertions that successful cyber forensic analysts more thoroughly integrate the use of specialized and general purpose software tools, however the findings do highlight the importance of specialized tools. Given that use of the Windows Task Manager was negatively predictive of successful submissions and use of Windows Explorer was predictive of incorrect submissions, the less successful students may have spent more time searching for files (Windows Explorer) and more time searching for processes (Windows Task Manager). Depending on the training objective, both activities could reflect unproductive time searching through file structures and ongoing system processes, which given their negative effect on performance, could be eliminated through efficient design of the exercise and supporting materials.

ACKNOWLEDGEMENTS

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. (SAND2014-2123 C)

REFERENCES

Abbott, R.G., McClain, J.T., Anderson, B., Nauer, K., Silva, A. & Forsythe, C. (2015). Log analysis of cyber security training exercises. In *Proceedings of the Applied Human Factors and Ergonomics Conference*, Las Vegas, NV.

Bernik, I. (2014). *Cybercrime and cyber warfare*. John Wiley & Sons.

Boesen, S., Weiss, R., Sullivan, J., Locasto, M. E., Mache, J., & Nilsen, E. (2014, August). EDURange: meeting the pedagogical challenges of student participation in cybertraining environments. In *Proceedings of the 7th USENIX conference on Cyber Security Experimentation and Test* (pp. 9-9). USENIX Association.

Danforth, M., & Lam, C. (2014, August). Four-Week Summer Program in Cyber Security for High School Students: Practice and Experience Report. In *Proceedings of the 7th USENIX conference on Cyber Security Experimentation and Test* (pp. 10-10). USENIX Association.

Fourie, L., Pang, S., Kingston, T., Hettema, H., Watters, P., & Sarrafzadeh, H. (2014). The global cyber security workforce: an ongoing human capital crisis.

Gavas, E., Memon, N., & Britton, D. (2012). Winning cybersecurity one challenge at a time. *IEEE Security & Privacy*, 10(4), 0075-79.

Goodman, S. E. (2014). Building the Nation's Cyber Security Workforce: Contributions from the CAE Colleges and Universities. *ACM Transactions on Management Information Systems (TMIS)*, *5*(2), 6.

Lee, J. A. (2014). The Red Storm in Uncharted Waters: China and International Cyber Security. *University of Missouri-Kansas City Law Review*, *82*(4).

Martini, B., & Choo, K. K. R. (2014). Building the Next Generation of Cyber Security Professionals. *Martini B and Choo KK R*.

McClain, J.T., Silva, A., Emanuel, G., Anderson, B., Nauer, K. & Forsythe, C. (2015). Human performance factors in cyber security forensic analysis. In *Proceedings of the Applied Human Factors and Ergonomics Conference*, Las Vegas, NV.

Pedersen, C. (2014). Much Ado about Cyber-space: Cyber-terrorism and the Reformation of the Cyber-security. *Pepperdine Policy Review*, *7*(1), 3.

Reed, T., Nauer, K., & Silva, A. (2013). Instrumenting Competition-Based Exercises to Evaluate Cyber Defender Situation Awareness. In *Foundations of Augmented Cognition* (pp. 80-89). Springer Berlin Heidelberg.

Rob, R., Tural, T., McLorn, G. W., Sheikh, A., & Hassan, A. (2014, September). Addressing cyber security for the oil, gas and energy sector. In *North American Power Symposium (NAPS), 2014* (pp. 1-8). IEEE.

Silva, A., McClain, J., Reed, T., Anderson, B., Nauer, K., Abbott, R. & Forsythe, C. (2014). Factors impacting performance in competitive cyber exercises. *Proceedings of the Interservice/Interagency Training, Simulation and Education Conference*, Orlando FL.

Son, J., Irrechukwu, C., & Fitzgibbons, P. (2014). Virtual Lab for Online Cyber Security Education. *Communications of the IIMA*, *12*(4), 5.

Stevens, S., Forsythe, C., Abbott, R. & Giesesler, C. (2009). Experimental assessment of accuracy of automated knowledge capture. *Proceedings of the International Conference on Human-Computer Interaction*, San Diego, CA.

Stevens, S., Forsythe, C., Abbott, R. & Giesesler, C. (2009). Experimental assessment of automated knowledge capture. *Proceedings of the Interservice/Interagency, Training, Simulation and Education Conference*, Orlando, FL.

Stevens-Adams, S., Basilico, J., Abbott, R.A., Gieseler, C. & Forsythe, C. (2010). Using after-action review based on automated performance assessment to enhance training effectiveness. Proceedings of the Human Factors and Ergonomics Society, San Francisco, CA.

Stevens-Adams, S., Basilico, J., Abbott, R.A., Gieseler, C. & Forsythe, C. (2010). Performance assessment to enhance training effectiveness. Proceedings of the I/ITSEC Conference, Orlando, FL.