Instrumenting Competition-Based Exercises to Evaluate Cyber Defender Situation Awareness

Theodore Reed, Kevin Nauer, and Austin Silva

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

Abstract. Cyber defense exercises create simulated attack and defense scenarios used to train and evaluate incident responders. The most pervasive form of competition-based exercise is comprised of jeopardy-style challenges, which compliment a fictional cyber-security event. Multiple competitions were instrumented to collect usage statistics on a per-challenge basis. The competitions use researcher-developed challenges containing over twenty attack techniques, which generate forensic evidence and observable second-order effects. The following observations were made: (1) a group of defenders performs better than an individual; (2) situation awareness of the fictional event may be measured; (3) challenge complexity does not imply difficulty. This research introduces a novel application of system instrumentation on competition-based exercises and describes an exercise development methodology for effective challenge and competition creation. Effective challenges correctly represent difficulty and reward competitors with objective points and optional forensic clues. Effective competitions compliment training goals and appropriately improve the knowledge and skill of a competitor.

1 Introduction

Information (cyber) security exercises have become powerful tools for simulating and planning for emergency scenarios, training, and competition. This paper focuses on the latter examples of training and competition. These exercises create simulated attack and defense scenarios where participants organize into groups and interact hands-on with operating systems, hardware, and software.

The exercise format varies, including modes with a sizable red (or attack) team versus many blue (or defending) teams, all red versus red teams, or all blue versus blue [CA1]. The red versus red is considered an attack and defense exercise where each team functions as both blue and red; they must maintain their security posture while decreasing their opponent's. The red versus blue is an interactive defense where each blue team is evaluated by their security posture after a complex and distributed set of red team attacks. A blue versus blue exercise uses point-valued challenges; the team that correctly solves the most challenges is the exercise victor [DE1].

The blue versus blue, or challenge-based exercises, are well-suited for training. The instructor develops interactive-challenges (i.e., a capture of forensic

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC/HCII 2013, LNAI 8027, pp. 80–89, 2013.

data containing a reportable sliver of evidence) which requires comprehension of course material to solve. Students may be motived to learn the material such that they can demonstrate competitive mastery (we do not make this assertion).

Challenge-based exercises are also the most flexible. Participants typically use their own hardware and tools, and may compete remotely and asynchronously (e.g., an exercise may not be bounded by time). Unfortunately this flexibility creates a difficulty for instrumentation; it is difficult to observe behavior and interaction. In this paper we describe a methodology for competition-based exercise development that yields measurable usage data and allows competition-designers introspection into player-challenge interaction.

1.1 Purpose of Study

Competition-based, continuous [GG1], exercises have proven successful for multiple applications and have become a pervasive [CB1] method of comprehension verification and community entertainment. Similar formatted exercises have been commonplace in high consequence domains (e.g., military) [MT1]. However, there have been few studies on the development and operation of these exercises and the human interaction in the cyber-security domain.

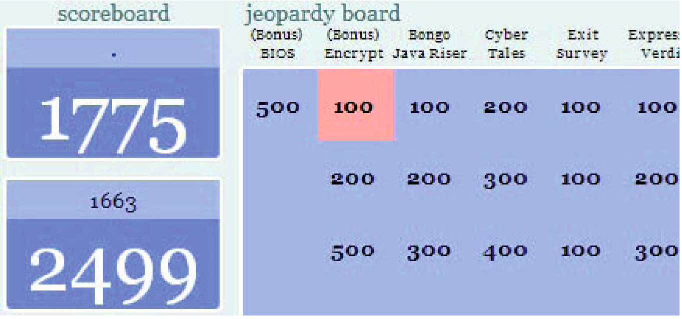
This research introduces an exercise platform and challenge development methodology that allows study of player-exercise, and player-player interaction. Example studies include: (1) a comparison of training modes; (2) player and tool adaptability; (3) situation awareness comprehension variability [T1]; (4) defensive solution-path discovery [SH1]; and (5) challenge playability tolerance. The last example uses the exercise to collect interaction statistics and create an arbitrary game mechanic called tolerance [GD1]. This demonstrates the exercise platforms ability to verify the challenge development, and is part of the development methodology. The methodology defines four categories of tolerance: simple, difficult, confusing, and unsolvable. A well-defined challenge should both be simple or difficult, and generate measurable feedback effects.

2 Approach

2.1 Exercise Platform

This research used a jeopardy–style interface containing categories of increasing-value challenges to represent the exercise. This game–board uses username and password account (or user) authentication and associates each user to a team. If any user correctly solves a challenge the team will receive the point-value; a team score is the aggregate of its users. The interface presents a robust configuration to the competition-designer.

The designer chooses from an XML-defined repository of challenges, with the ability to set time-thresholds and custom point values for each. A challenge is defined as a block of instruction, suggested time to complete, suggested point value, and solution. A solution may be an input string, a review process, or a trigger event. These challenges are organized into categories and categories are organized into boards. The designer configures the board availability (i.e., start and stop time) as well as trigger events (i.e., stop conditions) and submission rules. Fig. 1 shows an example participant view of the game–board. Note that one 100-point challenge has been solved by the user.



scoreboard

.

1775

1663

2499

jeopardy board

(Bonus) BIOS	(Bonus) Encrypt	Bongo Java Riser	Cyber Tales	Exit Survey	Expres Verdi
500	100	100	200	100	100
	200	200	300	100	200
	500	300	400	100	300

Fig. 1. Game–board from a participant's view

2.2 Methodology

Developing exercise challenges is non-trivial. Challenges should test a participant's critical thinking and knowledge application abilities. Challenges should implement a 1:1:2 ratio of effort required for a solution. This ratio represents 1-part discovery, 1-part understanding, and 2-parts solution development. The participant should spend the discovery phase analyzing the challenge to find a starting point. The understanding phase should be spent researching what skills, tools, and techniques are required for a solution. The solution development should stress the participants technical and critical-thinking prowess.

The challenge developer must maintain the highest level of fidelity for their challenge. Environment and data anomalies jeopardize the tolerability of a challenge and degrade any potential experiment or assessment. Example anomalies may include: (1) improper use of IP-space when creating a synthetic environment for forensic data generation, (2) unmatched operating system version artifacts left in physical memory, (3) poorly synchronized timing seen in network data, file systems, and descriptions, and (4) typographic fixes or incorrect checksums.

Dependent Challenge. A challenge (*c1*) may include artifact data needed to solve a separate challenge (*c2*). Challenge *c2* is called a dependent challenge. Dependent challenge development is particularly difficult; the development must be conscience of the playability implied by lack of depended knowledge.

2.3 Participants and Data

This research used five exercises. Each spanned at least two working-hour days, comprised of the same challenge set and over 220 combined participants. The participants represent a combination of high school students, undergraduate and graduate college students, and industry professionals. There were a total of 95 teams with a majority of 1-player teams with an assumed[1] maximum of 7-player teams. For this research no identifiable information was collected. When each exercise is completed usage data is exported with teams and users represented as arbitrary integer placeholders.

The exercises used 97 challenges per-event. Challenges were worth 100-500 points each, and most were solvable independent of the others. In all of the exercises recorded, wrong answers had no penalty and awarded 0 points; challenges were attempted until a successful submission (if any). The exercises attempted to measure participant situation awareness about a fictional cyber-security event. The challenges contained forensics data which required little interaction with the exercise platform. Thus it was very important that the challenges generate second-order effects such as (1) red-herring[2] submissions, (2) fictional names, services, or IP-addresses, or (3) additional forensics data.

The analysis uses an example assessment of challenge playability tolerance. Submissions and incorrect actions are compared to create a tolerance. Over six thousand submissions were recorded with just fewer than one thousand correct submissions. Over one million actions were recorded with a ratio of 4:1 incorrect to correct actions per challenge.

3 Results

3.1 Data Sanity

The data from all five exercises is combined and visualized in the following sections. In Fig. 2 the total score for each team is plotted in ascending order. The score distribution follows the exponential trend-line very closely. This is expected as better-performing teams solve higher-valued challenges across all categories. Problems with challenge confusion, which require participants to guess, may create a deviation. An imbalance in scores is highlighted indicating a potential guessing situation.

In Fig. 3 the number of correct and incorrect submissions per-challenge are plotted with a logarithmic trend-line. This describes a global interaction for every challenge. Challenge developers should expect a global logarithmic distribution, indicating a well-formed exercise with increasingly difficult challenges.

[1] One of the exercises was played virtually, thus any team may contain an unknown number of human players whom share user accounts. However, it is unlikely that accounts are shared as the game platform does not allow simultaneous challenge solving (i.e., only one challenge can be viewed at a time).

[2] A known-wrong submission that is easy or obvious but indicates progress.

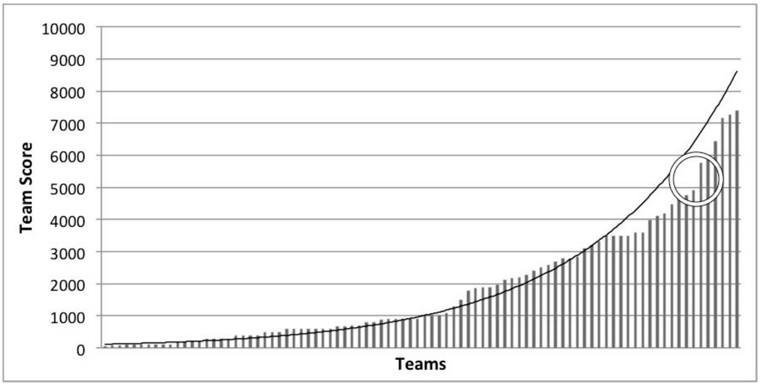


Fig. 2. Total score distribution with exponential trend-line

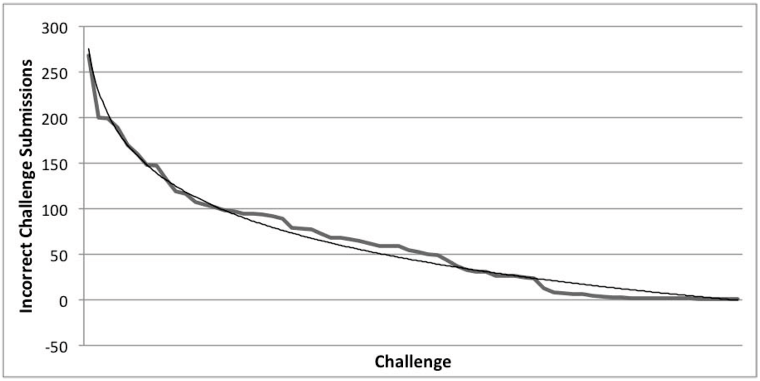


Fig. 3. Submission count per-challenge with logarithmic trend-line

Any abnormalities or deviations in submission counts may indicate confusion. The ratio of incorrect to correct submissions in Fig. 9 is used to enhance this visualization and help identify poorly-defined challenges. The two highlighted challenges have enormous incorrect to correct submission ratios.

Using a linear-trend with a bisection creates four quadrants of challenge ratios. Challenges with high submissions and low correct submissions (Q1) are candidates for review. Balanced ratios with high submissions are also candidates if not defined as difficult challenges. The same applies to the inverse if not defined as simple challenges. Finally, challenges without correct submissions are flagged as potentially unsolvable[3].

[3] Occasionally a 'solvable-but-near-impossible' challenge is useful for attracting curiosity.

3.2 Activity

In Fig. 4 the average momentum for the top three teams overall is shown as the dark line. The average momentum for the top three teams from one standard deviation (*sd*) away is shown as the light line. The momentum is seemingly linear for both groups. In this representation where momentum is a function of score versus time the reason for a dramatic (30%) point spread is unknown.

The point spread is more obvious when comparing Fig. 5 and 6. These figures show a normalized delay between submissions for each set of three teams. The longer each team plays, the more frequently they experience delayed submissions. Note, this does not represent periods of non-play. Delay normalization is a function of incorrect submissions. These plots may suggest teams are encountering more difficult challenges. The plots corroborate a similar momentum in Fig. 4 with a similar delay from point 15.

3.3 Tolerance

In Fig. 7 and 8 participant tolerance is show as the average for the top three teams and the average for the top three teams from one *sd*. To assess tolerance the exercise platform measures a combination of player frustration (f) and promotion (p). A promotion p, is defined as any positive feedback provided by the exercise platform to the player. A frustration f is a continually increasing value assigned to each player; f is reset to an initial state upon p. The exercise platform measures team frustration using a gain calculation based on incorrect actions and time.

$$f_t = \sum\nolimits_{i=p}^{n} n\,(t_i - t_{i-1}) \qquad \text{where } p \text{ is the last promotion event }. \tag{1}$$

Fig. 8 shows a significant amount of frustration toward the end of the measurement which most likely leads the disparity in points. Within the five exercises a p is a correct submission or a positive action taken by a participant (i.e., acquiring an additional piece of forensics data, gaining access to a services, or disabling an attacker).

3.4 Situation Awareness

Situation awareness is assessed by comparing the average performance of event related challenges to non-event related challenges. The event related challenges implicitly include artifacts and relations to other event challenges. These relations are not dependent challenges; the related challenges are solvable independently. However, knowledge of additional event related challenges builds context around possible attack vectors, techniques, and tools. If the participant has situation awareness and can build this context, the assertion is they will solve event related challenges more efficiently.

Out of the 97 challenges, 13 tightly related challenges were compared against an unrelated 13. These pairs were assessed by the challenge developers as having

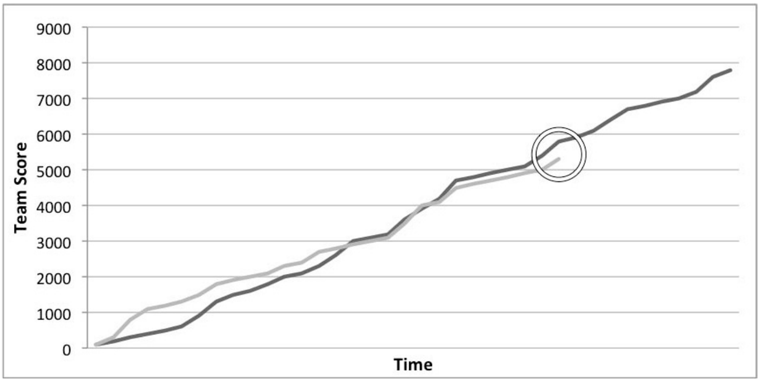


Fig. 4. Average momentum of top three teams (dark) and top three teams from one standard deviation (*sd*)

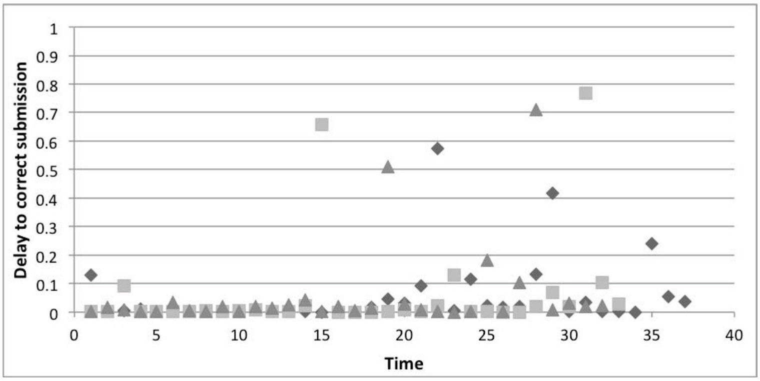


Fig. 5. Normalized time delay between submissions for top three teams

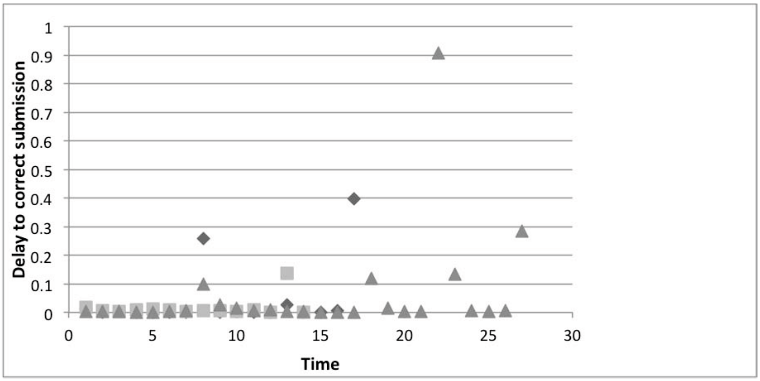


Fig. 6. Normalized time delay between submission for top three teams from one *sd*

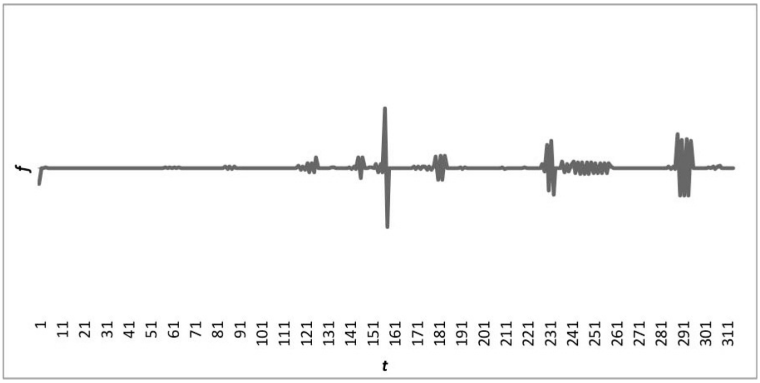


Fig. 7. Average frustration for top three teams

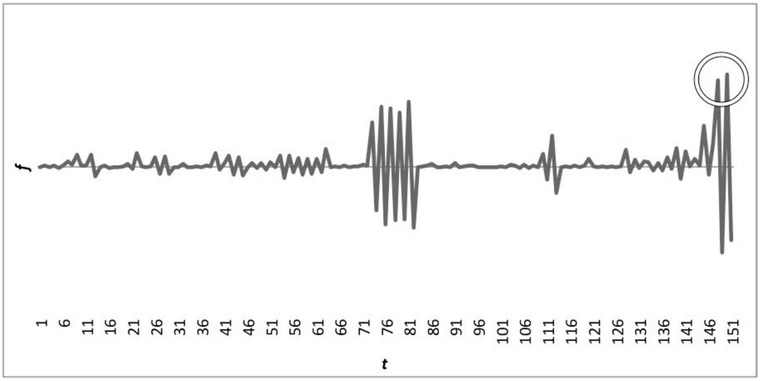


Fig. 8. Average frustration for top three teams from one *sd*

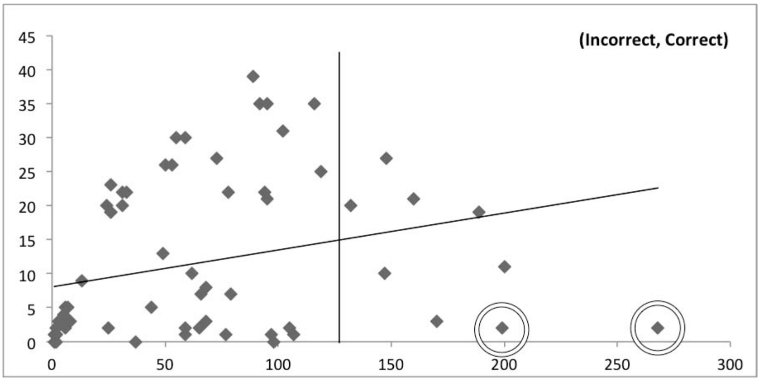


Fig. 9. Challenge submission ratios plotted (incorrect versus correct)

similar difficulty with little knowledge overlap. It is possible to solve the pairs in either order (non-event then event, or event then non-event) without confounding performance. Over 96% of participants demonstrate a better average performance on the event related challenges. A degree of situation comprehension may be measured within the exercise using a tool called *Plotweaver*[O1].

4 Conclusion

This exercise platform successfully demonstrates an example evaluation of cyber defender situation awareness. Participants are evaluated by their comprehension of a fictional cyber event through narration and plot description. Participants are also evaluated based on event related challenge performance versus non-event related challenges. If both a related and non-related challenge exists with similar difficulty and no overlap in knowledge requirement or other confounds: then the solution path can evaluated based on insight. The platform generates these statistics by comparing measurements generated through instrumented challenges.

The platform successfully validates challenge tolerance through usage statistics. This feedback is given to challenge developers and functions to remove unwanted difficulty confounds. Challenges that move from unsolvable or confusing to difficult make the exercise more enjoyable, reduce potentially harmful frustration, and generate more statistically-relevant usage data.

Instrumentation of challenges to provide measurable second-order effects created observations on player activity fallout based on frustration thresholds. This activity was not apparent in objective interaction data such as game-board activity and score momentum.

5 Future Work

Additional objective and subjective usage measures will continue to enhance the community's ability to use cyber defense exercises to improve domain knowledge and event response. The existing measures can be engineered into the exercise platform to provide real-time feedback to the designer. If player frustration and interaction threshold classes can be defined, a designer can provide in-line challenge and exercise augmentations. These augmentations can reduce frustration and experiment confounds to generate better data and a more enjoyable exercise experience.

Finally, challenge solution paths should be more closely monitored. Additional rewards can be granted to players demonstrating unique solutions. This encouragement may potentially enhance situation awareness, generate richer usage data, and reduce future frustration thresholds.

References

[T1] Tadda, G.P.: Measuring performance of Cyber situation awareness systems. In: Proceedings of the 11th International Conference on Information Fusion. Rome Res. Site, Air Force Res. Lab., Rome, NY, pp. 1–8 (2008)

[GG1] Glicksberg, I., Gross, O.: Notes on Games over the Square. In: Kuhn, H.W., Tucker, A.W. (eds.) Contributions to the Theory of Games. Annals of Mathematics Studies 28, vol. II, pp. 173–183. Princeton University Press (1950)

[GD1] Gilleade, K., Dix, A.: Using frustration in the design of adaptive videogames. In: Proceedings of the 2004 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology (ACE 2004), pp. 228–232. ACM, New York (2004)

[O1] Ogievetsky, V.: PlotWeaver (2013), https://graphics.stanford.edu/wikis/cs448b-09-fall/FP-OgievetskyVadim

[MT1] Mullins, B., Lacey, T., Mills, R., Trechter, J., Bass, S.: How the Cyber Defense Exercise Shaped an Information-Assurance Curriculum. In: IEEE Symposium on Security and Privacy, pp. 40–49 (2007)

[CB1] Childers, N., Boe, B., Cavallaro, L., Cavedon, L., Cova, M., Egele, M., Vigna, G.: Organizing large scale hacking competitions. In: Kreibich, C., Jahnke, M. (eds.) DIMVA 2010. LNCS, vol. 6201, pp. 132–152. Springer, Heidelberg (2010)

[DE1] Doup, A., Egele, M., Caillat, B., Stringhini, G., Yakin, G., Zand, A., Cavedon, L., Vigna, G.: Hit 'em where it hurts: a live security exercise on cyber situational awareness. In: Proceedings of the 27th Annual Computer Security Applications Conference (ACSAC 2011), pp. 51–61. ACM, New York (2011)

[CA1] Cowan, C., Arnold, S., Beattie, S., Wright, C., Viega, J.: Defcon Capture the Flag: defending vulnerable code from intense attack. In: Proceedings of the DARPA Information Survivability Conference and Exposition (2003)

[SH1] Sommestad, T., Hallberg, J.: Cyber Security Exercises and Competitions as a Platform for Cyber Security Experiments. In: Jøsang, A., Carlsson, B. (eds.) NordSec 2012. LNCS, vol. 7617, pp. 47–60. Springer, Heidelberg (2012)