Cyber security exercises and competitions as a platform for cyber security experiments

Teodor Sommestad, Jonas Hallberg

Swedish Defence Research Agency, Linköping, Sweden
{Teodor.Sommestad, Jonas.Hallberg}@foi.se

Abstract. This paper discusses the use of cyber security exercises and competitions to produce data valuable for security research. Cyber security exercises and competitions are primarily arranged to train participants and/or to offer competence contests for those with a profound interest in security. This paper discusses how exercises and competitions can be used as a basis for experimentation in the security field. The conjecture is that (1) they make it possible to control a number of variables of relevance to security and (2) the results can be used to study several topics in the security field in a meaningful way. Among other things, they can be used to validate security metrics and to assess the impact of different protective measures on the security of a system.

Keywords: research method, data collection, security competitions, security exercises

1 Introduction

Cyber security is an important topic and a lively research area. A plethora of frameworks, methods, tools, and principles can be found in literature. However, few of these have been formally or empirically validated. Verenedel's [1] review of security metrics illustrates the lack of empirical basis underlying theories used in the security field. Of the 90 papers reviewed, a minority attempted to empirically validate the metrics or measurement method proposed, and these tests are often limited to a comparison to the beliefs of a group of experts. One reason for the scarcity of empirical testing is the difficulty to acquire relevant data.

There are several reasons for the lack of empirical data. Data related to security (e.g., incidents and security mechanisms used) are sensitive and often treated as confidential by organizations [2, 3]. This naturally limits the availability of relevant data for research in the field. Several analyses have been made of the information sharing issue [4, 5] and actors such as national Computer Emergency Response Teams have taken on a role as information mediators. However, information and data on security from operational environments is still unavailable to the public and the research community as a whole. Also, if actual incident data would become publicly available for a representative sample of organizations, it could be difficult to draw general

adfa, p. 1, 2011.

conclusions from it. For example, if organizations would report their security solutions and the incidents they have experienced, this data would probably be biased because of the incidents they have failed to detect (e.g., cases when confidential data is read by unauthorized persons).

Although the lack of empirical data makes it difficult to study the security of operational systems, there are examples of studies that explore the strength of security measures using experiments or simulations. These include studies of protection against memory corruption exploits (e.g. [6, 7]), studies of denial-of-service attacks (e.g. [8]) and studies of the detection capabilities of intrusion detection systems (e.g. [9]). While these studies (and others not mentioned here) are able to quantify aspects related to security, the results lack properties necessary to make them a good support for decision makers in an operational scenario. In particular, they do not state how well different solutions work in practice when they are exposed to representative attackers. With few exceptions, quantitative security studies are made in settings with predefined attacks whose validity, considering operational scenarios, is unclear. For instance, the tests made in [6, 7] show how well different measures respond to a number of different attacks, but they do not show which of these attacks that will occur in practice when the system is attacked by intelligent adversaries.

This paper discusses the use of security competitions and security exercises (henceforth collectively referred to as security competitions) as a platform to obtain knowledge in the security field. Security competitions involve real attackers and, possibly, defenders drawn from a population chosen by the arranger of the competition and they are executed in an environment designed and implemented by the arranger. Additionally, competitions make it possible to specify goals for the attackers and the defenders. Their potential as a tool for education has been described in [10, 11] and lessons learned when using competitions as a tool in education can be found in [12]. This paper argues that competitions are usable as a tool for studying several attributes related to realistic attackers, realistic defenders, and realistic security processes.

The contribution of this paper is an analysis of the possibility to use security competitions as a platform for experimentation in the cyber security field. The paper describes typical arrangements in security competitions, suggests fields which could be further explored through competitions and presents examples of experiments that have been performed in similar arrangements. Its outline is as follows. Section 2 goes through the typical setup of security competitions and how this relates to experiments. Section 3 presents a non-exhaustive list of topics that would be possible to study through cyber security competitions, and relates previous work to these. In section 4, the overall idea is discussed. In section 5, the paper is concluded.

2 Security competitions

Security competitions can take many forms. Examples include: iCTF [13], CSAW's Cyber Security Competition [14], Cyber Security Challenge [15], and the National Collegiate Cyber Defense Competition [16]. Variation exists between them when it comes to the actors involved, the environments used, and the incentives created for the actors. This section provides an overview of common components in competition design and how the design of a competition relates to the design of an experiment.

2.1 The actors and teams

Security competitions will involve a number of actors that participate in and manage the competition. The designer of a competition must determine to what extent the competition should be defense-oriented and to what extent it should be offense-oriented [17]. In a defense-oriented competition the focus is to practice methods that can be used to defend a system against cyber attacks; in an offense-oriented competition the aim is to carry out attacks. A defense-oriented setup will involve one or several teams that defend systems against attacks; an offence-oriented setup will involve one or several teams set out to carry out attacks. Defensive teams are often called blue teams and offensive teams are often called red teams. Mixed approaches involving both active blue teams and active red teams are also possible, i.e., where the red teams attack the blue teams' systems or all teams attack each other. The members of the teams involved in the competition are referred to as the *participants* of the competition.

When a competition is designed the orientation of the competition can be controlled and participants of a certain *competence* can be selected. To some extent, the arranger of a competition can also influence the *organization* of the teams involved, for example, by assigning roles and responsibilities to the participants.

The participants of the red and blue teams are the active actors of the competition. Two other types of actors are frequently involved in competitions: members of the green and white teams. The green team manages the environment and ensures that the systems used in the competition operate as intended, e.g., that all actors have proper access to the environment. The white team referees the competition and manages the incentives for the red and blue teams, e.g., creates the competition scenario.

2.2 The competition environment

The competition environment, i.e. the technical infrastructure, is managed by the green team. Among other things, it includes: network topology, operating systems, application software, configurations and user accounts. How these elements are composed will depend on the scenario and the purpose of the competition. The arranger will thus determine what the *targeted system* shall be, the *security mechanisms* included, and the known *vulnerabilities* it contains. Moreover, the arranger can determine what information is available to the participants concerning the competition environment. In other words, the arranger can control the *intelligence* the participants have.

The goal in most competitions is to represent some real-life situation or hypothetical security problem staged in a realistic manner, i.e., to make the competition environment realistic. However, creating a realistic environment can require a significant number of person-hours, hardware, software, and expertise. As a consequence, the realism is sometimes traded against costs and the competition environment is rarely a full-blown enterprise network, but rather a small and controlled environment. Another difference to typical enterprise networks is that computer networks used in competitions often are designed to include a large number of known *vulnerabilities*. A computer network with known *vulnerabilities* often fit the purpose of the competition better than well-hardened networks since a well-hardened network might require attackers to spend a significant amount of time searching for new novel software vulnerabilities (an aspect which few competitions focus on).

2.3 Goals and rules for the actors

Competitions may have several different goals, for example to offer a challenge, train, test the competence or increase the awareness of the participants. There are several articles describing setups with the goal to train the participants, e.g. [10, 18–22]. It is natural to dictate the rules and award systems for events where the competition-aspect is central, but it is also recommended for competitions that focus on the training and learning experience [23]. In general, some types of objectives are suitable for blue teams and some are suitable for red teams [23].

Rules and rewards are in place to make the competition develop as intended and would for example describe allowed (and disallowed) practices during the competition, how participants are scored in the competition, and non-disclosure clauses. A competition could for example disallow execution of distributed denial-of-service attacks since the infrastructure is not built for it, or forbid blue teams to change the software on certain machines because it is business critical in the stipulated scenario. The arranger could also limit the use of *tools* to a predefined set.

Scoring systems could for example reward a red team if they manage to read some file, gain certain privileges on a machine, or to cause denial-of-service on a machine, whereas blue teams could be rewarded for maintaining systems operational. As performance measurements are a natural part of a competition, they provide means to control the *goals of the participants*. It is also natural to record the *participants' time /success* when they work towards these goals.

2.4 Competitions versus experiments

In an experiment, data collection and analysis follow a carefully worked-out plan. The basic requirement of an experiment is that different treatments are administered to different subject-groups or repeatedly to the same subject, and measurements are recorded after the treatment. A variable is either seen as a *dependent* variable, an *independent* variable, or a *nuisance* variable. A *dependent* variable is a measurement of the particular aspect one wish to observe; an *independent variable* is a treatment one wish to examine the effect of; *nuisance variables* are the factors which may have an effect on the *dependent variable*, but are outside of scope for the particular study. For the experiment to produce a reliable result, *nuisance variables* and *independent* *variables* needs to be controlled. Such control can be imposed by design or by randomization. [24]

As described above there are several variables that are possible to control and measure in a competition. The extent of this control is in some cases limited or costly (e.g., it is difficult to control the presence of software vulnerabilities which have not been made public yet). However, this paper argues that competitions can be designed so that the influence of independent variables on a dependent variable is measurable and the nuisance variables are handled in an acceptable manner. In this paper, the variables of concern to a security experimenter will be coarsely grouped into the following categories (introduced in italic in section 2.1-2.3): *Targeted system*, *Security mechanisms*, *Goals of participants*, *Vulnerabilities*, *Competence*, *Tools*, *Participants' time /success*, *Intelligence*, and *Organization*. These categories are referred to in the text of section 3.

3 Topics that can be investigated in security competitions

The first five subsections below describe security-related topics that are possible to perform experiments on in conjunction to security competitions. The last subsection summarizes previous experimental studies similar to those proposed here. It should be noted that the enumeration below does not aim to be exhaustive. It should rather be seen as a set of suggestions made to inspire.

3.1 The process model of an attack

There are several models over the steps an attacker takes during an attack. McQueen et al. [25] present a model over attacks with three "attacker subprocesses"; Olovsson and Jonsson [26] divided attacks into three phases; Schudel et al. [27] present another process model; Branlat and Morison [28] describe actions and interplay between attackers and defenders. However, there are few quantitative results showing that these qualitative models match the processes of actual attacks.

The student experiments performed at Chalmers [26, 29] and the experiments performed within DARPA's Information Assurance program [27, 30] are exceptions which offer some quantitative data on the different phases of an attack. In both these experiments observations are made of the time that attackers spend on different phases and if they succeed or not. More precisely, they investigate how *participants' time* is distributed over different phases in a process model and assess how different factors (e.g., *competence*) influence this.

A number of different factors influence how attackers spend their time and the activities they chose to perform. For example, the type of *intelligence* they have about the targeted system (e.g., network diagrams), the type of system that the *targeted system* is (e.g., web server or web browser) and the *security mechanism* used (e.g., dynamic defense [31]) can all be expected to influence the attack process in different ways. Experiments in conjunction to competitions could investigate how these factors influence the performance of the participants.

3.2 The attributes of successful attackers and defenders

There is no established theory concerning the *competences* an attacker or defender should have to be successful in its role. The *competence* of participants in a competitions (e.g. the factors enumerated in [32]) could be compared to the *participants' time/success* data in order to identify important determinants of success. Success for attackers could for example be measured as the time spent to achieve the goal; success for defenders could for example be measured as the portion of attacks that were prevented respectively detected by the defenders. In case the attackers or defenders operate as teams, the teams' mixes of *competence* can be assessed, perhaps in combination with the teams' *organization*.

Experimentation has been made on this topic as well. For example, the relationship between defensive success and the number of years at university (freshmen, sophomores etc.) has been assessed in [33]. The result suggests that seniors are tougher targets than freshmen. In [34] interviews of members in a red team are used to investigate what the key factors are for their effectiveness. The result of these interviews suggests, among other things, that it is difficult to identify attributes that reflect the effectiveness of a red team as well as the relationships between the effectiveness of individuals and teams. In other words, additional research is needed on this topic.

3.3 The impact of security mechanisms on success

Defenders can choose between a considerable number of *security mechanisms* to increase the security of their systems. By running competitions with and without a *security mechanism*, data on its practical effectiveness can be obtained. *Security mechanism* for a wide range of attacks can be tested in conjunction to a competition. This includes security mechanisms that protect against: buffer overflow attacks, denial-of-service attacks, SQL-injection, password-cracking, and network reconnaissance.

Two examples of tests involving security mechanisms are described in [35] and [36]. In [35], the impact of two *security mechanisms* against distributed denial-of-service attacks (and the combination of these two) is assessed during a competition. In [36] the targeted system is designed with four different sets of *security mechanisms* in order to assess the effectiveness of the defense-in-depth concept.

Moreover, in competitions that involve an active defense (i.e., a blue team) the impact of defender's management procedures and incident handling capabilities can be investigated. For instance, an experimenter can assess the effectiveness of administrators actively responding by different tactics. Effectiveness can for example be measured as the portion of attack goals that they prevented the attackers from accomplishing or the time it took the attackers to compromise the targeted system. In other words, *security mechanisms* that are active and adaptive so that they disturb the attack process can be evaluated in exercises involving live attackers. The "dynamic defense experiment" described in [31] is an example of this.

3.4 The accuracy of detection and incident analysis methods

The targeted system can be equipped with systems that log events and states during the competition. Such logs are used together with other data in investigations that attempt to reconstruct the chain-of-events that occurred during an incident. This capability is relevant both in incident management [37, 38] and forensic investigations [39].

If the additional data collected during the competition includes detailed descriptions of the steps taken by the attackers, they offer a straightforward method to evaluate the accuracy of methods for reconstructing the chain of events from logs. For example, in the experiment described in [40], the capabilities of a system administrator to detect and analyze intrusions made during a competition is analyzed. The red team provided detailed logs on their actions and these were compared to the assessments made by the system administrator.

In general, a number of factors can be expected to influence the difficulty of identifying and analyzing incidents. For example, *tools* used by attackers, their *attack-goals* and their *success* can be compared to the assessment made by the incident analysis team and/or their incident analysis tools. An intuitive hypothesis is that attackers with *competence,* special *tools,* and good *intelligence* will be more difficult to analyze and detect than other attackers.

3.5 The accuracy of security assessment methods

Critique has been directed towards methods used in the security community to assess the security of systems. In particular, few methods that quantitatively predict or assess operational security has been validated empirically [1]. Security competitions can be used to test assessment methods that are classified as "system" or "vulnerability" in [1]. In other words, methods that aim to describe how components and their structure in the system relate to security and methods that aim to describe the existence or appearance of system vulnerabilities.

There are several possibilities to evaluate the usefulness of assessment methods in conjunction to offensive competitions. As long as the security competition can be constructed to represent scenarios that are covered by the model/metric, the "correct" value can be compared to the calculated value. In [41] assessments made by security experts on *participants' success* for remote code execution attacks made under certain conditions are compared to empirical observations. In [42] 16 security metrics based on vulnerability ratings are compared to the *participants' time* spent and *participants' success* when attacking a set of strategically designed servers. In [43] the relationship between two metrics is assessed.

3.6 Past experimentation in conjunction to competitions and exercises

To perform security experiments involving human actors is not a new proposal. Several lessons have been learned and documented from experimentation with real attackers: in [26, 29] issues in experiments involving live attackers are discussed, in [44, 45] experiences from several red team experiments performed at DARPA are summarized, and in [35] lessons learned from a long multi-team experiment are described.

Table 1 categorizes a number of experiments involving human attackers. It should be noted that all studies in Table 1 are not referred to as experiments in the original article, e.g., in [33] the data collection method is described as ”observations of actual attacks carried out during the cyber defense exercise”. However, they all hold all the features required (e.g., with respect to control) to be experiments. The table denotes if a variable within the type is treated as a dependent variable (D) or an independent variable (I) that is varied.

Table 1. Past experimental and observational studies performed with human attackers.

Ref.	Topic investigated	Targeted system	Security mechanisms	Attack-goals	Vulnerabilities	Competence	Tools	Participants’ time /success	Intelligence	Organization
[26, 29]	3.1							D[a]		
[27, 44]	3.1							D[a]		
[36]	3.1 & 3.3		I	I				D		
[31]	3.1 & 3.3		I	I				D		
[33]	3.2 & 3.3			I		I		D		
[46]	3.3	I	I	I				D		
[45]	3.3		I	I				D		
[35]	3.3		I	I				D		
[41]	3.5		I					D		
[42]	3.5				I			D		
[43]	3.5							D[b]		

[a] *Time spent on different attack-activities is assessed.*

[b] *Two* metrics *on success are compared.*

As can be seen from the table, not all setups include independent variables that are varied. In particular, [27, 44] explores the process of attacks and the behavior of a certain type of attacker and [43] assess the relationship between a theoretical metric for attack effectiveness and success in an offensive competition. In these cases a single experimental setup is repeated multiple times. It should also be noted that strict experimental control is not always imposed on all known nuisance variables before the observations are made in the studies in Table 1. For instance, the attackers in [33] and [43] are convenience samples rather than the result of a controlled selection process. Such limitations are a natural effect of the conditions for these experiments and do not necessarily affect the validity and relevance of the data for the problem at hand.

4 Potential issues

This section comprises three subsections. The first subsection discusses the response variable(s) used in previous research. The second subsection discusses the cost of arranging competitions. The third subsection discusses some other issues with designing cyber security competitions as experiments and exploiting the produced data in research.

4.1 The response variable(s) – participant's performance

As can be seen from Table 1, the success of and/or time spent by participants is a popular response variable in experimental cyber security research based on competitions. Thus, the response variable is the success or failure of these actors when they attempt to perform well-defined tasks, or alternatively, the time it takes for them to complete the tasks. The time attackers need to accomplish a task is often referred to as the "time-to-compromise" [47–49] or the "adversary work factor" [44]. It is an established metric in the cyber security field to express the security level of a system. It is also commonly used when the security of physical systems is expressed. For instance, "net working time" is used to rate the security offered by safes [48]. Likewise, the probability that an adversary can succeed with an attack given specified conditions is commonly used in security assessments, e.g., in probabilistic risk analysis.

Time-to-compromise and success rate have an apparent value in decision making processes and research with this focus would certainly produce results with direct practical value. Few objections have been made against these metrics in literature. In fact, no direct arguments against them were found. The closest to an objection that has been found is given in [50], where it is argued that the process of discovering vulnerabilities in a software product is "believed to be chaotic" and that this would impede measurements of the time needed to complete a discovery. Still, even though the movement of a thrown dice is chaotic, a fair dice will on average yield six every sixth throw. If the chaoticness of the process of discovering vulnerabilities should be tested, competitions where the goal is to discovery new vulnerabilities seem to be a good place to make observations. For instance, it could be tested how variables related to *Competence* influence the process.

4.2 The cost of arranging a competition

Arranging a competition is costly [22]. As discussed above it will be costly to simulate a full-scaled enterprise environment in a competition, however, even a more limited target system can introduce high costs. For example, realistic software environments will often involve licensed products imposing direct monetary costs. Moreover, competitions are often built around a fictive scenario to make them more intriguing. To build a target system that fits this scenario could require that changes are made to applications design, naming conventions used, etc. These costs will limit the freedom a researcher has to re-design a competition to produce a better experiment.

Furthermore, for some aspects of security it will be inherently costly to arrange competitions as a meaningful experiment. For instance, experimenters who wish to investigate the effort required to find new vulnerabilities in a software product can do so in a security competition. Different software products can be provided, and these may be developed according to different standards, red teams can be challenged to find previously unknown security vulnerabilities in the products, and their performance can be recorded. While such setups may be arranged, discovering vulnerabilities may in some cases require months or years of effort for a security professional [51]. Hence, attackers involved in this kind of competition must be prepared to spend a significant amount of time to generate useful data if relatively secure products should be studied.

4.3 Other practical issues

In addition to cost considerations and the competition setup there are matters that are difficult to study in a competition. For instance, it will be difficult to test attributes of social engineering attacks since the participants will be in a different context than during normal operations. For the same reason it will be difficult to test a blue team’s active defense capability when attacks are sudden and unexpected (as they are for many organizations). Since the blue team will expect attacks, they will probably act in a different (more alert) manner. Other topics which appear difficult to assess via competitions include, but are not limited to, the monetary losses caused by adversarial activities, the incentives that real attackers or defenders act upon, and properties of the cyber security market.

Other practical considerations are of a more administrative nature. An experiment will produce data that is used to test or formulate hypotheses. The scientific process requires that the data is made available to other researchers and that experiments are possible to repeat by others. Some competitions might be impossible to perform under such circumstances. For instance, if a participating organization wishes to test their secret tools for their attacks, then the competition’s data may be labeled as confidential and it will be difficult for others to repeat the experiment or even review the data produced. Moreover, researchers might need to change the competition in order to perform a well-designed experiment on the matter they are interested in. Such changes may conflict with other objectives of the competition, e.g., to offer an interesting, educating, and lively competition. The possible conflict between scientific research and other objectives of the competition is worth considering. It is the only important difference between an experiment made in conjunction to a competition and a traditional (standalone) experiment on the same topic.

The basic proposal of this paper is to perform experiments in conjunction to cyber security competitions. Either by formulating experiments as competitions to attract study-subjects or by getting involved in existing competitions in order to get access to an arrangement already paid for. We have been in contact with three established arrangers of competitions to gauge the possibility to use them for experimentation, given the issues discussed above. More precisely, they were asked if (1) it would be possible to collect data from the exercise and use it for research and (2) if it would be possible for researchers to influence the design of the exercise in order to produce a more interesting experiment. We believe that their answers points towards the potential of security competitions as a basis for experimentation.

Officials of Security Challenge [15] see no problem with the idea as such. However, it was also judged as likely that the privacy regulations in the Great Britain would make it difficult to use collected data for research purposes.

The director (Dwayne Williams) of the National Collegiate Cyber Defense Competition [16] replied that data recordings are unproblematic and have been made during past competitions, e.g. in the form of network packet recordings. When it comes to the design of the competition, it is likely that smaller changes to the targeted systems are acceptable; however, it is unlikely that alterations of the scoring mechanisms or putting requirements on the participants to perform additional tasks would be accepted.

The iCTF [13] is arranged by professor Giovanni Vigna at University of California Santa Barbara and already includes the collection of data for research. In fact, the last two competitions have been designed with this particular objective in mind. Data from the competition is published online [13] and empirical research results (see [43]) have been produced from the competition. Professor Vigna is also open for input on the design of the competition.

5 Conjecture

Cyber security competitions are popular. They are primarily held to offer the participants an interesting challenge and to educate and train the participants for real situations. The arranger of a competition controls a number of variables that are important in the cyber security field. This control makes it possible to design them as experiments where the subjects under investigation are the participants, their interaction with an external environment, and the actual system. To arrange competitions in order to conduct experiments can be costly. However, a large number of competitions are already carried out today. This includes high-profile events with hundreds of participants, national events involving multiple organizations, and small events executed by a couple of organizations or individuals. Experimenters who can influence the setup or take advantage of their existing setup can limit their experimentation cost significantly, while producing empirical data to test and formulate hypotheses.

References

1. Verendel, V.: Quantified security is a weak hypothesis: a critical survey of results and assumptions. New Security Paradigms Workshop. 37-50 (2009).

2. Geer Jr, D., Hoo, K.S., Jaquith, A.: Information security: why the future belongs to the quants. Security & Privacy, IEEE. 1, 24–32 (2003).

3. Kotulic, A., Clark, J.G.: Why there aren't more information security research studies. Information & Management. 41, 597–607 (2004).

4. Gal-Or, E., Ghose, A.: The Economic Incentives for Sharing Security Information. Information Systems Research. 16, 186–208 (2005).

5. Gordon, L.: Sharing information on computer systems security: An economic analysis. Journal of Accounting and Public Policy. 22, 461–485 (2003).

6. Wilander, J., Nikiforakis, N., Younan, Y., Kamkar, M., Joosen, W.: RIPE: Runtime Intrusion Prevention Evaluator. In Proceedings of the 27th Annual Computer Security Applications Conference, ACSAC, 41-50 (2011).

7. Shacham, H., Page, M., Pfaff, B., Goh, E.: On the effectiveness of address-space randomization. ACM conference on. 298 (2004).

8. Khattab, S.M., Sangpachatanaruk, C., Melhem, R., Znati, T.: Proactive server roaming for mitigating denial-of-service attacks. Information Technology: Research and Education, 2003. Proceedings. ITRE2003. International Conference on. pp. 286–290. IEEE (2003).

9. Ktata, F.B., Kadhi, N.E., Ghédira, K.: Agent IDS based on Misuse Approach. Journal of Software. 4, 495–507 (2009).

10. Conti, G., Babbitt, T., Nelson, J.: Hacking Competitions and Their Untapped Potential for Security Education. IEEE Security & Privacy. 56–59 (2011).

11. Fanelli, R.L., O'Connor, T.J.: Experiences with practice-focused undergraduate security education. Proceedings of the 3rd Workshop on Cyber Security. , Washington, DC, United states (2010).

12. Werther, J., Zhivich, M., Leek, T.: Experiences in cyber security education: The mit lincoln laboratory capture-the-flag exercise. the 4th Workshop on Cyber Secuirty Experimentation and Test. , San Francisco, CA, United states (2011).

13. Vigna, G.: The UCSB iCTF, http://ictf.cs.ucsb.edu/.

14. Polytechnic Institute of NYU: CSAW - CyberSecurity Competition, http://www.poly.edu/csaw2011.

15. Cyber Security Challenge: Cyber Security Challange, https://cybersecuritychallenge.org.uk/.

16. National Collegiate Cyber Defense Competition: Welcom to the National Collegiate Cyber Defense Competition, http://www.nationalccdc.org/.

17. Patriciu, V.V., Furtuna, A.C.: Guide for designing cyber security exercises. Proceedings of the 8th WSEAS International Conference on E-Activities and information security and privacy. pp. 172–177. World Scientific and Engineering Academy and Society (WSEAS) (2009).

18. Wagner, P.J., Wudi, J.M.: Designing and implementing a cyberwar laboratory exercise for a computer security course. Proceedings of the 35th SIGCSE technical symposium on Computer science education - SIGCSE ’04. 402 (2004).

19. Schepens, W.J., Ragsdale, D.J., Surdu, J.R., Schafer, J.: The Cyber Defense Exercise: An evaluation of the effectiveness of information assurance education. The Journal of Information Security. 1, (2002).

20. Conklin, A.: Cyber Defense Competitions and Information Security Education: An Active Learning Solution for a Capstone Course. Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS’06). p. 220b–220b. IEEE (2006).

21. Hoffman, L.J., Rosenberg, T., Dodge, R., Ragsdale, D.: Exploring a National Cybersecurity Exercise for Universities. IEEE Security and Privacy Magazine. pp. 27–33. IEEE (2005).

22. Childers, N., Boe, B., Cavallaro, L., Cavedon, L.: Organizing large scale hacking competitions. Proceedings of the 7th international conference on Detection of intrusions and malware, and vulnerability assessment. pp. 132–152. Springer Verlag, Bonn, Germany (2010).

23. Schepens, W.J., James, J.R.: Architecture of a cyber defense competition. Systems, Man and Cybernetics, 2003. IEEE International Conference on. pp. 4300–4305. IEEE (2003).

24. Keppel, G., Wickens, T.D.: Design and analysis: a researcher’s handbook. Pearson Education, Upper Saddle River, NJ, USA (2004).

25. McQueen, M.A., Boyer, W.F., Flynn, M.A., Beitel, G.A.: Time-to-Compromise Model for Cyber Risk Reduction Estimation. In: Gollmann, D., Massacci, F., and Yautsiukhin, A. (eds.) Quality of Protection. pp. 49–64. Springer US, Boston, MA (2006).

26. Jonsson, E., Olovsson, T.: A quantitative model of the security intrusion process based on attacker behavior. IEEE Transactions on Software Engineering. 23, 235–245 (1997).

27. Schudel, G., Wood, B., Parks, R.: Modeling behavior of the cyber-terrorist. RAND National Security Research Division, proceeding of workshop. pp. 45–59 (2000).

28. Branlat, M., Morison, A.: Challenges in managing uncertainty during cyber events: Lessons from the staged-world study of a large-scale adversarial cyber security exercise. Human Systems Integration Symposium (2011).

29. Olovsson, T., Jonsson, E., Brocklehurst, S., Littlewood, B.: Data collection for security fault forecasting: Pilot experiment, Dept. of Computer Eng., Chalmers Univ. of Technology, and ESPRIT/BRA Project no. 6362 (PDCS2), Toulouse (1993).

30. Levin, D.: Lessons learned in using live red teams in IA experiments. DARPA Information Survivability Conference and Exposition, 2003. Proceedings. pp. 110–119. IEEE (2003).

31. Kewley, D.L., Bouchard, J.F.: DARPA Information Assurance Program dynamic defense experiment summary. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans. 31, 331–336 (2001).

32. Guard, L., Crossland, M., Paprzycki, M., Thomas, J.: Developing an empirical study of how qualified subjects might be selected for IT system security penetration testing. Citeseer. 2, 413–424 (2004).

33. Dodge, R.C., Carver, C., Ferguson, A.J.: Phishing for user security awareness. Computers & Security. 26, 73–80 (2007).

34. Kraemer, S., Carayon, P., Duggan, R.: Red team performance for improved computer security. Human Factors and Ergonomics Society Annual Meeting Proceedings. pp. 1605–1609. Human Factors and Ergonomics Society (2004).

35. Mirkovic, J., Reiher, P., Papadopoulos, C., Hussain, A., Shepard, M., Berg, M., Jung, R.: Testing a Collaborative DDoS Defense In a Red Team/Blue Team Exercise. IEEE Transactions on Computers. 57, 1098–1112 (2008).

36. Kewley, D.L., Lowry, J.: Observations on the effects of defense in depth on adversary behavior in cyber warfare. Proceedings of the IEEE SMC Information Assurance Workshop. pp. 1–8 (2001).

37. Mitropoulos, S., Patsos, D., Douligeris, C.: On Incident Handling and Response: A state-of-the-art approach. Computers & Security. 25, 351–370 (2006).

38. Werlinger, R., Muldner, K., Hawkey, K., Beznosov, K.: Preparation, detection, and analysis: the diagnostic work of IT security incident response. Information Management & Computer Security. 18, 26–42 (2010).

39. Meyers, M.: Computer forensics: the need for standardization and certification. International Journal of Digital Evidence. 3, 1–11 (2004).

40. Sommestad, T., Hunstad, A.: Intrusion detection and the role of the system administrator. Proceedings of International Symposium on Human Aspects of Information Security & Assurance. , Crete, Greece (2012).

41. Holm, H., Sommestad, T., Franke, U., Ekstedt, M.: Success rate of remote code execution attacks – expert assessments and observations. Journal of Universal Computer Science. 18, 732–749 (2012).

42. Holm, H., Ekstedt, M., Andersson, D.: Empirical analysis of system-level vulnerability metrics through actual attacks. IEEE Transactions on Dependable and Secure Computing, Accepted (2012).

43. Egele, M., Caillat, B., Stringhini, G.: Hit'em where it hurts: a live security exercise on cyber situational awareness. Computer Security. (2011).

44. Schudel, G., Wood, B.: Adversary work factor as a metric for information assurance. Proceedings of the 2000 workshop on New security paradigms. pp. 23–30. ACM (2001).

45. Levin, D.: Lessons learned in using live red teams in IA experiments. Proceedings DARPA Information Survivability Conference and Exposition. 110–119 (2003).

46. Ryder, D., Levin, D., Lowry, J.: Defense in depth: A focus on protecting the endpoint clients from network attack. Proceedings of the IEEE SMC Information Assurance Workshop (2002).

47. Paulauskas, N., Garsva, E.: Attacker skill level distribution estimation in the system mean time-to-compromise. Information Technology, 2008. IT 2008. 1st International Conference on. pp. 1–4. IEEE (2008).

48. Leversage, D., Byres, E.: Comparing Electronic Battlefields: Using Mean Time-To-Compromise as a Comparative Security Metric. Computer Network Security. 1, 213–227 (2007).

49. McQueen, M., Boyer, W., Flynn, M., Beitel, G.: Time-to-compromise model for cyber risk reduction estimation. Quality of Protection. (2006).

50. McHugh, J.: Quality of protection: Measuring the unmeasurable? Proceedings of the 2nd ACM Workshop on Quality of Protection, QoP’06. Co-located with the 13th ACM Conference on Computer and Communications Security, CCS'06. pp. 1–2. , Alexandria, VA (2006).

51. Sommestad, T., Holm, H., Ekstedt, M.: Effort estimates for vulnerability discovery projects. HICSS’12: Proceedings of the 45th Hawaii International Conference on System Sciences. , Maui, HI, USA (2012).