# Measuring team effectiveness in cyber-defense exercises: A cross-disciplinary case study

Magdalena Granåsen and Dennis Andersson


The original publication is available at www.springerlink.com:
Granåsen, M., Andersson, D., (2016), Measuring team effectiveness in cyber-defense exercises: A cross-disciplinary case study, *Cognition, Technology & Work*, 18(1), 121-143. https://doi.org/10.1007/s10111-015-0350-2

Original publication available at:
https://doi.org/10.1007/s10111-015-0350-2

# Measuring team effectiveness in cyber-defense exercises: A cross-disciplinary case study

Magdalena Granåsen, Dennis Andersson
magdalena.granasen@foi.se, dennis.andersson@foi.se
Division for Information- and Aeronautical Systems
Swedish Defense Research Agency
Box 1165, SE-581 11 Linköping, Sweden

## Abstract

In 2010, IT-security experts from northern European governments and organizations gathered to conduct the first of a series of NATO-led cyber-defense exercises in a pilot attempt of training cyber defense. To gain knowledge on how to assess team effectiveness in cyber-defense exercises, this case study investigates the role of behavioral assessment techniques as a complement to task-based performance measurement. The collected data resulted in a massive data set including system logs, observer reports, and surveys. Six different methods were compared for feasibility in assessing the teams' performance, including automated availability check, exploratory sequential data analysis, and network intrusion detection system attack analysis. In addition, observer reports and surveys were used to collect aspects relating to team structures and processes, aiming to discover whether these aspects can explain differences in effectiveness. The cross-disciplinary approach and multiple metrics create possibilities to study not only the performance-related outcome of the exercise, but also why this result is obtained. The main conclusions found are (1) a combination of technical performance measurements and behavioral assessment techniques are needed to assess team effectiveness, and (2) cyber situation awareness is required not only for the defending teams, but also for the observers and the game control.

**Keywords:** cyber defense exercise, cyber SA, performance assessment, team cognition, team effectiveness

Preprint. Published in Cognition, Technology & Work (2016) 18:121-143, DOI 10.1007/s10111-015-0350-2

# 1. Introduction

Modern organizations are often voluminously dependent on reliable and secure information systems, making them vulnerable to cyber crime, warfare, and terrorism. Incidents such as the cyber-attacks on Estonia in 2007 and the attacks on U.K., U.S., German, and French resources in 2005 prove that this threat is real (Greenemeier 2007). However, the amount of publicly available data from such incidents is low, which makes it difficult to study cyber defense in depth from real events. Cyber defense exercises (CDX) offer an opportunity for researchers to study how organizations respond to cyber-related threats, subject to the limitations of simulation.

In conjunction with a multinational CDX in 2010, a study was performed aiming to improve knowledge on how to assess team effectiveness in a CDX. The study was conducted as a cross-disciplinary case study of a scenario where hastily deployed cyber-defense teams protected safety-critical industrial networks against antagonistic hacker groups. Such a case study is motivated by the fact that SCADA (Supervisory Control and Data Acquisition) networks are prone to hacking and becoming a prime concern for society at large (Igure, Laughter, & Williams 2006). In reality, it is unknown how likely it is that external cyber-defense teams with little prior knowledge of the company network would be deployed in an actual emergency. On the other hand, it is not uncommon that corporations lack the ability to handle such extreme events without external assistance.

To address the objective of assessing team effectiveness, a dataset was compiled of log data, observations, and subjective self-ratings of e.g. performance, team cognition, within-team interactions, team organization, team composition, and strategies. Thus, the main focus of this study is not to give a valid evaluation how the teams in the studied CDX performed, but rather to evaluate methods for assessing team effectiveness. Analysis results on performance and cognition are included for a discussion of validity and applicability of the analytic tools and metrics used.

## 1.1. Description of the Baltic Cyber Shield case

Baltic Cyber Shield 2010 (BCS), a multi-national civil-military CDX, aimed to improve the capability of conducting future CDXs and increase knowledge on how study attacks on, and defense of, critical information infrastructure in such settings (NATO 2010).

### 1.1.1. Scenario and setup

BCS was co-hosted by the Cooperative Cyber Defence Centre of Excellence (CCDCoE) and the Swedish National Defence College (SweNDC). The scenario used during the BCS described a volatile geopolitical environment in which a rapid response team of network security personnel was deployed to defend critical infrastructure from cyber-attacks sponsored by a non-state terrorist group (NATO 2010).

Six defending (blue) teams each assumed control over a simulated power generation company and was tasked to protect their respective corporation against cyber threats. The actual terrorists (red team) were role-played by professional hackers and coordinated by game control (white team). The white team (distributed between SweNDC and CCDCoE) monitored the exercise and acted as judges of the blue teams' performance. The technical platform was designed and implemented by a *green team* from the Swedish Defence Research Agency, who also monitored the technical infrastructure and the data collection during the exercise (NATO 2010).

The defending (blue) teams were provided identical, pre-configured computer network composed of 20 physical PC servers running a total of 28 virtual machines in a computer cluster. These networks were divided into four VLAN segments labeled DMZ, INTERNAL, HMI, and PLC (NATO, 2010). The blue team networks were further connected to several in-game servers that provided additional business functionality to fictitious users within their corporations. Network connections were established through Virtual Private Networks (VPNs)

Preprint. Published in Cognition, Technology & Work (2016) 18:121-143, DOI 10.1007/s10111-015-0350-2

enabling the teams to be physically distributed. Real Programmable Logic Controllers (PLCs) simulated a power generation infrastructure, including steam engines, solar panels, a virtual distribution network, and miniature factories with actual butane flames that could be detonated by the red team by breaching into the PLC segment This mixed reality simulation added to the realism of the exercise, with the purpose of increasing the participants' motivation (Hammervik, Andersson, & Hallberg 2010).

The computer network was preconfigured with vulnerabilities to be exploited by the red team. Each blue team's main task was to defend their network against intrusion attempts by the red team. As a motivating factor for the blue teams, the exercise was setup as a competition where the groups were awarded points for preventing attacks and reporting suspicious activity (NATO 2010). The scenario also imposed rules, restricting the blue teams' solution space and penalized them when breaking rules. For example, some services were required to be operational and open to external communication over time and some systems were not allowed to be patched at all.

The exercise lasted two full days with scheduled breaks for organizational purposes. The exercise was divided into four phases, each with different objectives and rules of engagement for the red team. During the exercise, within-team communication was primarily conducted in native language, while between-team communication was conducted mostly in English. In Geers (2010), a more thorough presentation of the exercise setup and scenario is presented, while the most immediate lessons learned and a brief presentation of participants, exercise activities and performance is given in CCDCoE after action report (NATO 2010). A deeper analysis of the technical metrics used is discussed by Holm, Ekstedt, and Andersson (2012). The analysis approach has been briefly described before by Andersson, Granåsen, Sundmark, Holm, and Hallberg (2011).

### 1.1.2. Exercise participants

Four different kind of teams, with different functions, were involved in the exercise: Blue (defending), red (attacking), white (controlling, scoring) and green (technical). The presented case study analysis had an explicit focus on the blue teams.

Six blue teams participated in the exercise, formed from northern European governmental, military, and academic institutions. The teams were located in four different northern European countries (NATO, 2010). The team leaders' responsibilities included manning the team with appropriate competencies based on the exercise information given. Team sizes varied from four to ten participants. Each blue team remained intact for the duration of the exercise, with some minor exceptions. One of the six teams opted out of any data collection; hence the case study concerned five of the blue teams. A total of 43 persons participated in these five blue teams.

Team A was composed of five IT-security professionals working mainly at the same department at a governmental institution.

Team B was temporarily formed by nine IT-security professionals from various governmental institutions.

Team C was composed of ten IT-security professionals from three different governmental institutions in the security sector.

Team D was composed of nine reputable IT-security experts, all members of a professional network within IT security. One of the members was located in different facility from the rest of the team during the exercise, making the team partly distributed.

Team E was a student team, composed of ten graduate students attending an advanced-level IT security course.

Thus, some teams were temporarily formed (ad hoc), while other teams were composed of people normally working together (functional, departmental). Team E chose to use preconfigured workstations handed to them by the green team; the other teams chose to use their own computers.

The *red team* was composed of 16 skilled experts and technicians within the IT-security domain, which were distributed between two sites during the exercise and collaborating using tools for computer-supported cooperative work. The red team was part of game control, and performed attacks on the blue teams' systems. The *white team* consisted of a multinational mix of decision-makers, with the objectives of coordinating the activities of the red team and acting as judges of the blue teams' performance. The *green team* consisted primarily of technicians, consultants, and researchers, and was responsible for designing and implementing the technical infrastructure and data collection. During the exercise, the green team monitored system status and data collection processes and remotely coordinated the dispatched observers.

A background survey of nine questions was distributed to the blue-team participants before game start on the first day in order to obtain an overview of how the teams differed with respect to age, IT-security skills and familiarity with each other (professionally and personally). Response rate for the background survey was 72% (31 out of 43). Unfortunately, only three of the nine team members of team D responded to the background survey. Response rate for the other teams ranged between 60% and 100%. Of all participants responding to the background survey, 30 (97%) were males and 1 (3%) was female, ages ranging from 23 to 53 with a mean age of 33 years. 21 (68%) reported professional experience of working with IT-security issues and only 9 (29%) reported that they had a formal education within the IT-security domain. It should be noted that the numbers on the last two questions are likely to be skewed since some participants refrained from answering those two questions for anonymity reasons. Mean values on age, self-rated expertise in IT security and familiarity with other team members are displayed team-wise in Table 1. The questions on expertise and familiarity with other team members were rated on a 5-point Likert scale ranging from 1 (very low/not familiar) to 5 (very high/very familiar).

**Table 1** Descriptive statistics of defending teams based on survey responses of the background survey

| Team | Responses (N) | Mean age (years) | IT security expertise (1-5) | Personal familiarity (1-5) | Professional familiarity (1-5) |
|---|---|---|---|---|---|
| A | 5 | 32.40 | 3.00 | 3.25 | 3.00 |
| B | 9 | 34.89 | 3.00 | 2.11 | 1.56 |
| C | 9 | 37.11 | 3.89 | 3.22 | 2.67 |
| D | 3 | 41.00 | 4.67 | 2.67 | 2.33 |
| E | 9 | 27.11 | 2.83 | 2.67 | 2.17 |
| Overall mean | | 33.41 | 3.44 | 2.74 | 2.26 |

Team A's high scores of personal and professional familiarity between team members, fits well with the already known fact that this team was composed of personnel working at the same department in a governmental institution. In all teams, including the ad hoc ones, at least some of the team members had a prior relationship, either personal or professional. For most teams, the highest familiarity scores were reported by the team leaders, who had been responsible for recruiting the team members. Team D, composed of highly ranked IT experts from different organizations, stand out in their self-assessment of individual expertise within the IT security domain. This team also had the highest mean age, and can be assumed to have the most experience, although this cannot be verified as information on experience was not collected in the survey. The lower mean age of Team E compared to the other teams can be explained by that it consisted mainly of graduate or PhD students. Their assumed lack of work experience may explain why they rated themselves lower on expertise compared to the other teams.

## 1.2. Assessing team effectiveness in the cyber defense domain

*Team performance* may be defined as "the outcomes of the team's actions regardless of how the team may have accomplished the task" (Salas, Sims, & Burke 2005, p. 557), while *team effectiveness* "takes a more holistic perspective in considering not only whether the team performed (e.g., completed the team task) but also how the team interacted (i.e., team processes and teamwork) to achieve the team outcome" (Salas et al., 2005 p. 557). There is thus an interrelationship between the two concepts, and they are both relevant in their own right. *Team cognition* refers to the cognitive structures and processes within teams (Cooke, Salas, Kiekel, & Bell, 2004 p. 85), which are inherently important to teams' functioning and thus may correlate to their performance and effectiveness. Champion et al. (2012) identifies team structure, team communication and information overload as three additional factors affecting team performance. Decision making and other team processes may in turn be affected by aspects such as trust, risk behavior and various cognitive biases (Pfleeger & Caputo, 2012). Analyzing team effectiveness in general is thus a complex task that calls for system boundaries and delimitations to become feasible in practice.

Cyber security teams (both attacking and defending) operate in a highly uncertain and complex environment which is characterized by low visibility of what is occurring on the other side, making it difficult to make sense of the situation. Furthermore, the cyber security environment is highly collaborative, involving a number of different roles (Branlat 2011). During analysis of a CDX or a real incident, technical data such as system and event logs may give an answer to *what, where, who* and *when*. However, to understand *why* and *how,* more thorough analysis is needed, including not only actual performance from event logs, but also what motivated the actions taken, such as team decisions, strategies and team organization.

In this study, team performance refers to blue teams' accomplishment of mission objectives and successful reporting, while team effectiveness takes into account both performance and team cognition. Focus was put on studying team effectiveness, which may be judged partly by studying their performance, according to the above definitions.

Incorporating social and behavioral research methods into the cyber security field can give new possibilities of understanding causes to a given effect. If the causes to an outcome can be identified, the needs for training and technological, methodological and organizational development can be pinpointed. Why did the team not manage to eliminate the threat? Was it not discovered, was it discovered but not considered a threat, or was it considered but unknown by the team how to manage? Was it a matter of information overload, technological failure, ignorance, lack of cues or knowledge, insufficient procedures or simply an unfortunate series of misunderstandings? There are numerous methods to assess different aspects of team cognition, and when choosing an appropriate method, trade-offs have to be made between researcher time, cost, amount of data obtained, level of obtrusiveness/interruption of participants, reliability, and validity (Wildman, Salas, & Scott 2013). Studying cyber-security teams involves a number of challenges, including capturing participants' activities simultaneously (between teams, within teams), the representation of what occurs so that it is intelligible by an audience (including the analyst) and how to reduce the complex situations to make them tangible (Branlat 2011).

Endsley (1995) identifies three levels of *situation awareness* (SA; Endsley 1995): perception, comprehension, and projection. Barford et al. (2010) describe the three levels of SA in the cyber field as:

1. *Perception* involves identifying the type and source of an attack, awareness of the quality of information, and understanding capabilities, vulnerabilities and intents on both sides.
2. *Comprehension* is obtained through impact assessment and causality analysis of why and how events happened.
3. *Projection* is the ability to detect how the situation evolves.

SA in a CDX environment has been assessed in a simulation environment by measuring actual performance of how well teams defended networks and comparing it to teams' self-assessment of how well they completed different tasks in the scenario (Champion et al. 2012). However, SA in terms of the completion of specific tasks differs from awareness of the situation as a whole. Due to the complexity of the cyber security sector as well as lack of information sharing between functional domains, no individual is thought to have the full picture. Instead cyber SA is distributed across individuals and technological agents operating in different functional domains (Tyworth, Giacobe, Mancuso, & Dancy 2012). Little experimental research has been conducted on cyber SA, which is why there is only little data, and few validated assessment methods, available on the actual impact of cyber SA on performance (Franke & Brynielsson 2014).

Well-designed experiments, with sufficiently described sampling procedures help reducing biases and confounding variables and make hypotheses and research questions clear and understandable (Pfleeger & Caputo 2012). However, a perfectly controlled experimental study does not guarantee that cognitive aspects can be studied. Motivation, risk behavior and other drivers are challenging to reproduce in an experimental setting. Riegelsberger, Sasse, and McCarthy (2003) particularly note the challenges researchers face when studying trust in computer-mediated communication. They stress the importance of ecological validity, i.e., realistic tasks that are solved in a realistic setting. The perfectly controlled experiment may therefore not be sufficient for studying real-world phenomena, whilst the uncontrolled field study may contain too many disturbances to make valid conclusions.

In exercises, the main objectives often concern evaluation or validation of skills and processes; in the cyber field it has become increasingly common that exercises adopt a competition format (Sommestad & Hallberg 2012). These competitions can be considered somewhere between controlled experiments and field studies, and consequently yield both interesting possibilities and challenges when used as a platform for data collection and analysis. While cyber security competitions are common (e.g. Conklin 2006; Doupé et al. 2011; Geers 2010; Hoffman, Rosenberg, Dodge, & Ragsdale 2005), the resulting data logs are not as commonly used for scientific purposes. For researchers, collecting data from a competition is typically less costly than to set up a specific experiment, and may be beneficial to ecological validity as competitions can attract both professional and hobbyist hackers. The drawbacks of using this approach include lack of experimental control of design, scoring system, and software; as well as confidentiality issues with the collected dataset (Sommestad & Hallberg 2012). Further problems to studies in these settings include professionals' reluctance to share their tricks of the trade or even to be evaluated at all as they may feel it violates their privacy (Andersson 2011).

## 1.3. Reconstruction and Exploration of massive datasets

Collecting and combining data from a multitude of sources allow revelation of new insights and may increase validity of the study, but it is often time consuming. Methods and technology for displaying heterogeneous data sources simultaneously can enhance efficiency in the analysis phase. The Reconstruction and Exploration (R&E) approach enables exploratory analysis of the scenario, which can give the researcher a much needed ability to post-hoc visualize events at multiple locations to get a birds-eye view of the chain of events before digging into the details (Pilemalm, Andersson, & Hallberg 2008). The externalized causal model of the event flow can provide a ground truth for post-hoc analysis of decisions and actions, i.e. to get a deeper understanding the dynamics of the controlling teams and their interactions with the environment. If focused on team level interaction with the environment, such models can also be useful for introspection of the team's sensemaking process as they allow analysts to study the link between decisions and cues, context, and constraints (Andersson, 2014). Further, during the exercise, such models can assist game controllers, judges and observers in maintaining situational awareness; a task that may otherwise become overwhelming when dealing with highly specialized teams deeply engaged in tasks that are not easily observed.

R&E in itself does not include methods for how to analyze the data as shown in Fig. 1. This decoupled research approach has the benefit of providing flexibility for the researcher, at the expense of analysis guidance.

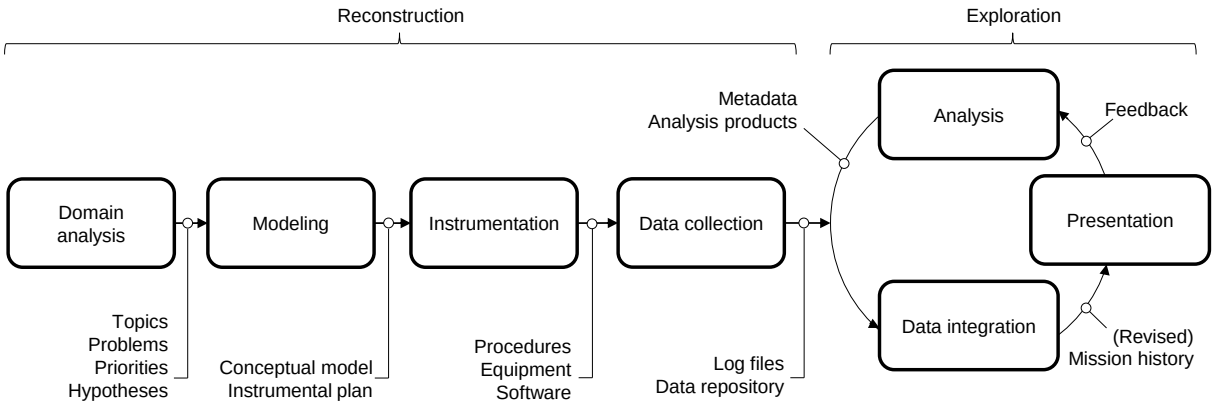


**Fig. 1** The seven phases of Reconstruction & Exploration (from Andersson 2009)

Andersson (2009) states that R&E could be descried as consisting of seven phases: (1) Domain analysis, (2) Modeling, (3) Instrumentation, (4) Data collection, (5) Data integration, (6) Presentation and (7) Analysis. In R&E, exploration is cyclic, with analysis results and presentation comments being fed back into the model to create revised mission histories, although the analysis method itself is not restricted by R&E. The R&E phases are quite similar to a case study research process, which Yin (2009, p. 24) describes in the phases plan, design, prepare, collect, analyze, and share (Table 2). The primary difference between the two models is the explicit focus on intermediate products that the R&E model provides with the purpose of generating analyzable mission histories, while the case study process comes with a higher degree of freedom of how to conduct data collection.

**Table 2** Comparison between R&E and the case study research process

| **Concept** | **Phases** | | | | | |
|---|---|---|---|---|---|---|
| **R&E** (Andersson, 2009) | Domain analysis | Modeling | Instrumentation | Data collection | Analysis, data integration | Presentation |
| **Case study** (Yin, 2009) | Plan | Design | Prepare | Collect | Analyze | Share |

Analysis in R&E can be performed using e.g. exploratory sequential data analysis (ESDA), an empirical approach for quantification of qualitative data (Sanderson & Fisher 1994). ESDA is based on eight steps (8C's) which guide the researcher through the analysis: chunks, comments, codes, connections, comparisons, constraints, conversions and computations; all of which can be performed independently -- but if worked sequentially they effectively help the analyst breaking down large datasets into smaller pieces enabling quantification, comparison, and ultimately explanation of interesting phenomena (Andersson et al. 2011). Because ESDA assumes temporal ordering of events, it is an absolute requirement that sequential integrity of the data is preserved (Sanderson & Fisher 1994), and as such it fits well with the R&E approach of post-hoc assembly of data into mission histories.

# 2. Method

This study uses the BCS to draw conclusions on how to assess team effectiveness in future exercises and competitions, with no ambitions of generalizing the performance results of the exercise to real-life situations. The study was conducted as a cross-disciplinary case study. A case study may give insight into specific situation of intrinsic interest, but the case can also give a more general understanding for certain phenomenas (Stake

1995, p. 3). The case study approach promotes the use of multiple methods and multiple data sources for validation (Yin 2009, p. 18) and the combination of qualitative and quantitative methods warrants both depth and breadth in the analysis (Flyvbjerg 2011). It was assumed that the massive, explorative data collection and in-depth, multidimensional analysis of the exercise would reveal insights regarding assessment methods for CDX teams in the full spectrum of team effectiveness; that is including team performance as well as team cognition aspects, and the interactions between those.

It was early recognized that a multitude of data sources would be needed to validate and approach BCS's research objectives 5-7, i.e. to train IT-security students and professionals, to improve the capability of conducting technical exercises in the cyber domain, and to study IT attacks and defense in critical infrastructure and SCADA (NATO 2010). Data collection included system logs, video and audio recordings, observer reports, and surveys. R&E was selected as the main approach to coordinate data collection and prepare for analysis since thorough planning and structuring of data collection was identified as a success factor. ESDA was the preferred analysis method to analyze qualitative data and extract certain performance measures. Complementing interviews and statistical analysis have been performed outside the main R&E process to attain additional insights, e.g. regarding team behavior. The analyses presented in this section are relative in nature, i.e. they are based on between-group comparisons of the blue teams. Although absolute metrics are powerful to provide a baseline for development, they require a more thorough analysis of optimal performance and validated performance metrics. Therefore, relative metrics are preferable in settings where performance is not scripted on beforehand and therefore unknown to exercise managers and analysts during execution.

## 2.1. Instrumentation and data collection

System logs and reports from teams and observers provided results on the teams' activity in the exercise network. The collected data includes human factors and team aspects, such as team organization, strategy, and teamwork, largely gathered through surveys and observer reports. The full data collection plan thus contained a set of quantitative and qualitative, subjective and objective parameters as visualized in Fig. 2.

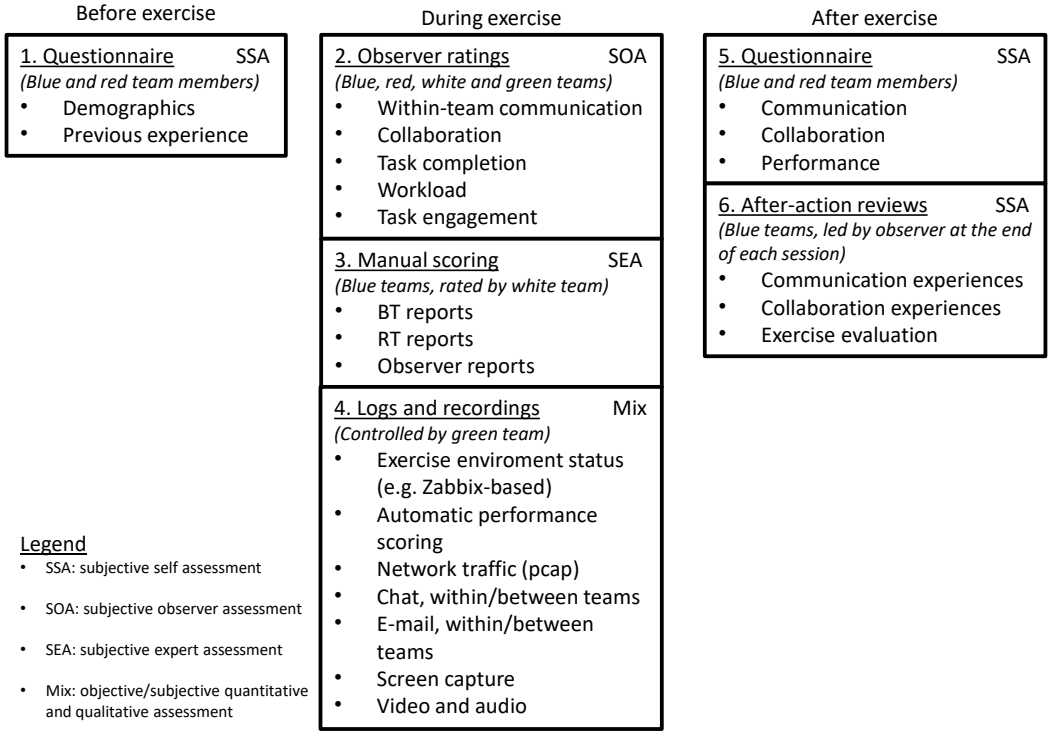


**Fig. 2** Data collection model employed during Baltic Cyber Shield

Data collection was conducted during the full two days of the exercise. In total 3 TB data was collected. Due to technical and security issues, the quality and amount of data captured differed between teams. These circumstances make it difficult to generate fully comparable data sets for each, and therefore negatively affects the validity of the comparisons. With respect to the objective as testing the methods rather than generating actually valid team performance assessments, the data set holds great value despite such validity issues.

Team members of blue team E received workstations with preinstalled screen capture and keyboard tools, which allowed detailed tracking of the actions of each team member in that team. Blue teams A and B agreed to install the software for screen capture on their own workstations, whereas teams C and D opted out. Also within teams A and B, there are individual differences as some members chose not to, or failed to, deliver screen capture videos. Table 3 and Fig. 3 list the data sources per team, note that data was also captured from the red, green, and white teams, for analysis also of how these organizer teams can be supported by R&E, however this data has not investigated further in this case study.

**Table 3** Data capture during BCS CDX

| | White team | Green team | Red team | Blue team A | Blue team B | Blue team C | Blue team D | Blue team E |
|---|---|---|---|---|---|---|---|---|
| Observer | X | X | X | X | X | X | X | X |
| Audio recorder | X | X | X | X | X | | X | X |
| Video cameras | X | X | X | X | X | | X | X |
| Screen capture tools | | | | X | X | | | X |
| Keyboard logging | | | | | X | | | X |
| Surveys | X | X | X | X | X | X | X | |
| Network traffic (pcap)[1] | X | X | X | X | X | X | X | X |
| Network traffic (netflow)[1] | X | X | X | X | X | X | X | X |
| Computer status (zabbix)[1] | X | X | X | X | X | X | X | X |
| E-mail logs | X | X | X | X | X | X | X | X |
| Chat logs | X | X | X | X | X | X | X | X |
| Collaboration wiki capture | X | X | X | X | X | X | X | X |

[1] Network traffic and computer status in the exercise network. All teams' activities in the cluster were captured. However, the status on individual computers was not captured.

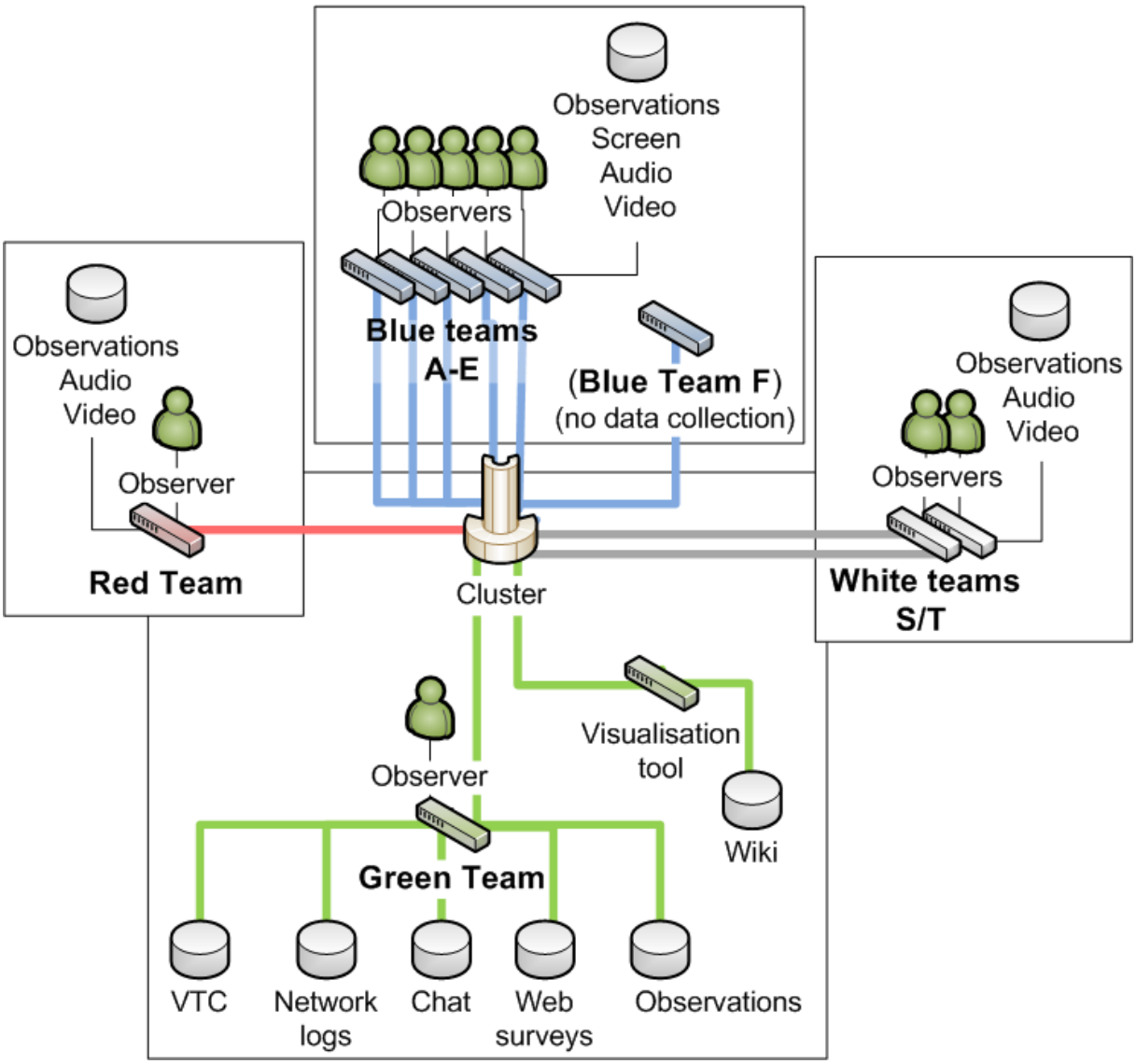


**Fig. 3** Logical distribution of teams, observers and data collection nodes. The white team members were split between Stockholm and Tallinn

The data logs captured from technical systems included e-mails, chat sessions, keyboard interactions, network traffic and utilization of memory, processors and hard disk space on each node in the virtual network. In order to capture screen video and keyboard interactions, custom made scripts were installed on most computers that were connected to the network. Since some of the teams used their own computers, the participants' willingness to cooperate and install these scripts on their respective computers was crucial. For blue team E, that was supplied workstations by the exercise organizers, it was easier to setup and control this logging. For the supplied Windows computers a custom-developed screen-capture program was used, while on Linux the participants were recommended to use the open source software XVidCap (http://xvidcap.sourceforge.net), but any other appropriate application was allowed. Some users operated from Mac OS X platforms, and consequently were unable to use the two provided solutions for screen capture. To capture terminal I/O a platform-independent custom script was supplied as part of the team packages.

Each team was accompanied by an observer who reported events using a pre-defined coding scheme including reporting categories (codes) relating to accomplishment of tasks as well as team interactions. Incoming observer reports were monitored in real-time by the green and white teams. The observers were native-speaking in their respective team's language, but reported in English, which made it possible for the white and green teams during game, as well as for analysts in post-exercise work sessions, to get insight into the teams

internal processes, and of what actions they took, even though not necessarily being able to understand respective teams' intra-team communication.

To enable a structured reporting of events, the observers were equipped with a handheld device running custom software for reporting *Network-Based Observation Tool* (NBOT) (Thorstensson 2012). Events reported by the observers were immediately visualized in chronological order at the green and white teams. On two occasions during the exercise, the pre-defined coding schema given to the observers was deemed unsuccessful since the observers found it difficult to apply the coding categories to their reports. To remedy these premature commitments, the analyst team was forced to create and distribute new schemas for the NBOT reports. The premature commitment did not compromise the data set as such, since the reports were still being generated - only not accurately coded which demands time for the analysts as they needed to recode the reports after the exercise.

On two occasions per day, on the green team signal, observers were instructed to collect information on the team workload, the team members' current priorities, and engagement in the task. These ratings were given on a 5-point Likert scale ranging from very low (1) to very high (5). The observers were instructed to complete the task with minimal interruption of participants. Additionally, at the end of each day, observers facilitated a discussion with their team, asking about the complexity, difficulty, clarity of the task, what they would need in order to receive a better overview of the situation. On these occasions the observers also collected general comments on the scenario and exercise. During an after-action review (AAR) approximately one week after the exercise, the observers gave their personal views on performance, communication issues and team strategy. The day after the exercise, most team leaders from blue, red, white and green teams, as well as the rest of the personnel involved in planning and executing the exercise, participated in a virtual AAR to summarize their experiences.

In addition to observer reports, survey data was collected from the blue team members through one background survey and two additional surveys, distributed to the blue teams at the end of each day's activities. The purpose of the surveys was to capture the participants' understanding of the teamwork and tasks they received. The two post-action surveys were identical to each other (36 questions). The introductory part contained three questions of team affiliation and age. 16 teamwork-related questions (11 Likert, 5 open-ended) dealt with comfort of working with other team members, team composition of competencies, amount of collaboration, team organization, team strategy, team priority and team performance. Nine individually oriented questions (6 Likert, 3 open-ended) asked for each participant's specific tasks, priorities and struggles, individual skills, situation overview of the network, information exchange, individual performance and workload. The last eight questions (3 Likert, 5 open-ended) concerned the exercise in terms of realism of the scenario and game network and needs for exercise and data collection improvement. The Likert questions were graded from 1-5, typically anchored at 1=to a very low extent, 3=neither low or high and 5=to a very high extent. All participants decided for themselves if they wanted to answer any questions at all due to privacy. The surveys were web-based and all answers treated anonymously.

## 2.2. Performance assessment

A semi-objective performance measure was implemented by the exercise management team as a motivator for the teams to do their best to defend their team. The measure was a score composed of an automatic and a manual part. The automatic part calculated an availability score by interrogating selected business services on the defending teams' company networks. This additive score was updated in real time and available to the participants at all time. The manual scoring was designed to encourage the team to report their activities. This score was based on e-mail reports from the teams, and assigned by judges in the white team who subjectively rated all incoming reports on content accuracy, timing and level of detail. Deduction of points was given for the failure of detecting or reporting incidents. All details of the scoring system were known by the teams during the

exercise. Additionally, the reports helped the white team maintain situational awareness during execution, and researchers during post-exercise analysis.

An additional set of team performance metrics was defined after the exercise by the analysts, including attack success rate, mean time to compromise, attack discovery, vulnerability removal, and vulnerability discovery (Andersson et al. 2011; Holm et al. 2012). These measures were constructed post hoc strictly for academic purposes and, thereby, not included in the after action report (NATO 2010).

1. *Exercise performance measure: Service availability* simulated the availability of the companies' business services as was specified in the exercise task, i.e., web services that the defending teams were instructed to keep online at all times. The services were automatically interrogated every five seconds, and points were assigned for each successful interrogation. Service availability was displayed to the teams during the game in order to enhance motivation by increasing the competitive component of the exercise.

2. *Exercise performance measure: Manual scoring* was assigned by the white team to the defending teams for reporting incidents, both proactive and reactive. Each report was assessed by the same judge and rated for content accuracy, timing and level of detail. The sum of service availability and manual scoring formed the total exercise performance measure and was revealed to the teams immediately after the exercise.

3. *Attack discovery* was measured as the ratio between reported attacks from the attacking team and the defending team. This measure relied on the accuracy of the reporting process, and is subject to both false positives and false negatives since the defending team may lack understanding of what is happening in the network at a given point in time.

4. *Vulnerability removal* represented the number of removed vulnerabilities, while *vulnerability discovery* represents the number of discovered vulnerabilities. Since the defending teams' networks were identical, they initially had the same number of vulnerabilities to discover.

5. *DMZ Attack success rate* was measured as the ratio between successful and attempted attacks on the network. Attack success rate was based on the DMZ portion of the network only, since the rest of the network was encrypted.

6. *DMZ Mean time to compromise (MTTC)* was a measure based only on successful attacks, and rated the average time from an initiated attack until the attacked subsystem was compromised. As with attack success rate, MTTC was based on the DMZ portion of the network only. In the attack success rate metric, a lower score indicates better performance, while MTTC has the opposite relationship.

All the above metrics were measured for blue teams A-E during the exercise, and the relative ordering of the teams based on the different metrics was compared to initiate a discussion on team performance metrics in cyber defense exercises.

## 2.3. Data integration

With vast amounts of media-rich qualitative data to analyze, a structured approach is needed to reduce the risk of information overload for the researchers analyzing the dataset. Time-synchronized mission histories can greatly reduce the efforts of understanding cause and effect relationships (Andersson 2013). Creating a mission history model suitable for ESDA requires thorough synchronization of data sources to maintain sequential integrity. Synchronization was achieved using internal network time protocol (NTP) servers connected to all devices in the virtual environment, guaranteeing sufficient accuracy on all computer clocks. Past experiences has taught that long sequences of video and audio can still become skewed when digitalized and compressed using standard encoders on common-of-the-shelf (COTS) hardware, and that the resulting logs can be off by several seconds at the end of the recording even though they are synchronized at start. To remedy the skew problem, all video and audio recordings (including screen capture videos) were cut and time stamped by the

NTP enabled clock at regular intervals. In this way it could be ensured that even if the reported sample rate differed from the actual, it can easily be compensated for when a new recording started. A more precise way of solving the problem would be to measure the actual sample rate and adjust the recordings afterwards, however since such a solution would generate non-standard sample rates in file headers, the resulting files would risk not being playable by several standard media players. A third approach would be to use hardware with enough precision and a codec with enough accuracy to limit the skew, although such a solution comes with the drawback of higher associated costs. The first approach was selected as it was rated good enough for the projected needs.

Non-networked devices, such as surveillance cameras and audio recorders were synchronized using virtual synchronization points (VSP). For the cameras, these VSPs were generated by filming a clock at several occasions. For audio, the VSPs were generated by observers who spoke the current time into the microphone. Using these VSPs, accurate offsets between logged timestamps and actual time could be calculated and compensated for, and time-related errors, such as skew and offset, in the collected data could be corrected afterwards.

## 2.4. Analysis

During analysis of system logs and observer reports for the performance metrics, emphasis was put on three parts of ESDA: codes, comparisons, and computations. The analysis used a combination of tools such as Microsoft Excel, IBM SPSS[1], F-REX[2] and Snort[3].

The first reconstruction cycle used only chat logs, e-mail communication and observer reports, laid out in F-REX as in Fig. 4 with observer reports in the top left corner, chat room logs in the mid left section and e-mail logs in the bottom left corner. The right hand side of the figure displays a timeline of the included events in the mission history, used for navigation during presentation and analysis. This limited dataset could not reveal much about the teamwork processes; but the lean mission history maintained high analyzability and allowed quantitative analyses of performance. Based on findings from the first cycle, portions of network traffic and selected video screens were selected for inclusion into the second cycle to allow a deeper analysis with more accurate coding and connections. With this data set certain attacks, or attempted attacks, in the DMZ could automatically be identified by Snort due to well-known signatures in the network traffic data. In the next analysis cycle, network traffic was analyzed using snort, an open source network intrusion detection system (NIDS). The NIDS analysis generated a high number of alarms which had to be clustered, using time and target team, since a normal attack consists of more than one exploit.

[1] IBM SPSS, Commercial statistical analysis software, http://www.ibm.com/software/analytics/spss
[2] F-REX, Tools for Reconstruction & Exploration of heterogeneous datasets (Andersson, 2009)
[3] Snort, Open source network intrusion detection software, http://www.snort.org

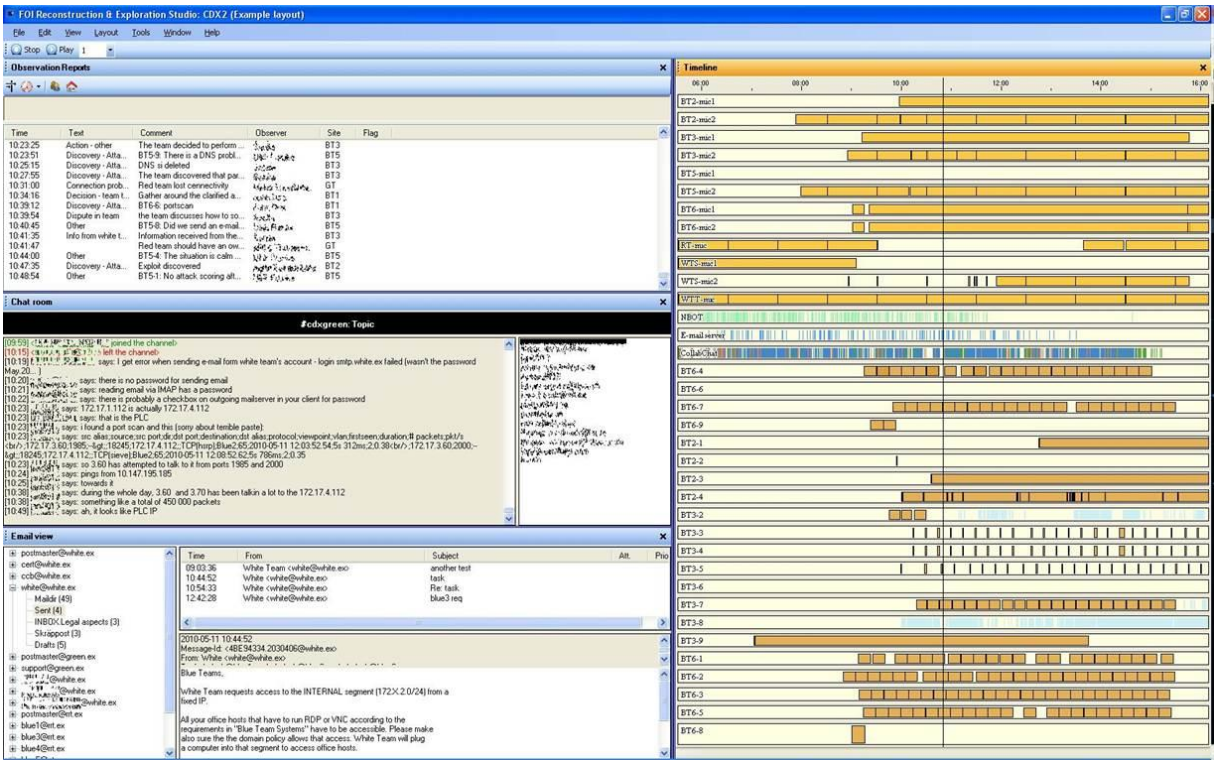


**Fig. 4** F-REX screenshot showing integrated data for analysis in the first cycle. Note that some information has been scrambled for anonymity reasons

Retrospective analysis of teamwork in an open-ended scenario without predefined metrics advocates a mission history model with high media richness, which is less effective and efficient than lean media models in presenting analyzable tasks (Lim & Benbasat 2000; Otondo, van Scotter, Allen, & Palvia 2008). Coded observation reports and event logs were used to navigate and prioritize information in the massive dataset and selectively reduce the presentation to chunks that could be analyzed.

Statistical analyses were performed on survey data, investigating differences between teams and correlating them with performance measures. It should be noted that there has been a scientific debate on whether significance testing is relevant when a full population is studied, as can be considered the case here as we compare the teams as separate entities and do not consider them as part of a larger population. Cowger (1984 1985) claims that a total population contains no sampling error, and therefore there is no motive for significance testing. That is, any difference detected between subgroups should be considered a significant difference. He is opposed by Rubin (1985), who claims that significance testing is valid to increase credibility of identified differences between subpopulations. In this analysis, significance tests have been performed, to compensate for the fact that the survey response rate was less than 100%, meaning that the entire population did not respond. The risk of using significance testing in this case is that some differences between teams are neglected (type II errors), however this risk is considered as less severe than the risk of type I errors (coincidental differences are treated as significant) which increase if significance testing is not performed. Significance level was set to 0.05; however results within 0.05-0.1 are also displayed and discussed in those cases when they are in line with other significant results.

In order to understand and explain the performance and survey results, reports from observer-held discussions, assessment of team workload, and AAR reports were used for triangulation.

# 3. Results

The results section includes analysis of logged system-system and human-system interactions, reports from observers, after-action reviews and survey results.

## 3.1. Surveys

36 (84%) of the 43 blue team participants responded to one or both surveys delivered after game stop each day. As for the background survey, response rate was lowest for team D, in which only four of the nine team members responded to either or both of the surveys. In total, 33 responded to the first survey and 30 responded to the second survey. This means that there are gaps in the data for those cases where participants only responded to one of the surveys or chose to not answer a specific question. In order to be able to perform analyses, the gaps were filled with the individual's value of the corresponding question the other day ± the mean difference between the two surveys of the other team members' responses for that particular question. By this manipulation, the complete data set could be used, albeit with a small error introduced. An alternative approach is to use the responses of only those 27 participants who completed both surveys, however this would also introduce errors since several actual respondents would then be completely ignored. As the objective of this study is to evaluate methods for performance assessment rather than to actually conduct the assessment, the introduced errors do not affect the validity of this study.

A repeated-measures analysis with team as between-groups factor and survey responses for each day as within-groups factor was conducted in order to discover any significant differences between the two days on similar questions. Three questions differed significantly between days. During day 2, system vulnerabilities were discovered to a larger extent (Day 1: M= 2.93, Day 2: M=3.23, $p<0.05$), network overview improved (D1=3.29, D2=3.49, $p<0.05$) and individual performance increased (D1=3.06, D2=3.31, $p<0.05$). However, more interesting is how different teams responded on the two days, as this may reveal some insights in how team dynamics and strategies changed in the teams during the game. All identified differences, significant at $p<0.05$, are presented in Table 4 below. Each cell number pair corresponds to day 1 and day 2 mean response value from each team, the higher marked in bold face. Note that only the significant differences are listed, and that identified differences are manifested both as increasing and decreasing from day 1 to day 2.

**Table 4** Significant intra-team mean value differences between day 1 and day 2 responses

| | Blue team | | | | |
|---|---|---|---|---|---|
| | A | B | C | D | E |
| Within-team collaboration | **3.92** / 3.26 | | | **4.75** / 3.75 | |
| Information exchange | **5.00** / 3.50 | | | | |
| Follow team strategy | **3.20** / 2.20 | | 3.05 / **3.62** | | |
| Team performance | **4.20** / 3.20 | | 3.11 / **3.67** | | |
| Individual performance | **3.20** / 2.70 | | 3.11 / **3.67** | | |
| Individual skills | **3.80** / 3.30 | | | 3.25 / **4.25** | |
| Team cohesion | 3.40 / **4.40** | | | | |
| Network overview | | 2.69 / **3.19** | | | |
| Decide on team organization | | | | 4.00 / **5.00** | **4.00** / 3.75 |
| Discovery of system vulnerabilities | | | | 3.25 / **4.25** | |

Table 4 reveals that teams B and E are fairly consistent in their ratings between the days, whereas team A in general rates themselves lower on several of the questions on day 2, and vice versa for teams C and D. Since the performance results were based on the accumulated results from both days, the rest of the quantitative survey analyses were performed on the mean values between the surveys for days 1 and 2.

Overall, the participants perceived the scenario as sufficiently realistic (M=3.46) and were highly motivated throughout the exercise (M=4.02). Teamwork was experienced as smooth (M=4.22), team members were

confident in other team members skills (M=3.91) and the teams were composed of mainly sufficient competencies (M=3.72). On within-team collaboration, the participants reported that they, to a high extent, collaborated (M=3.56) and exchanged information (M=3.86) with other team members during the exercise. Workload was considered as relatively high throughout the exercise (M=3.54).

A correlation analysis comparing the background survey questions on previous personal knowledge and familiarity between team members with the teamwork questions of the daily surveys reveals no significant correlations. Thus, for this type of task, we have found no evidence to support that previous knowledge and familiarity with other team members correlate with the ability to collaborate during the exercise.

A MANOVA was conducted with survey responses for the seven survey questions dealing with team composition, individual and team skills, organization, and collaboration as dependent variables and team as independent variable. Using Pillai's trace, there is a significant team effect on the these survey responses ($V=1.406$, $F(28, 128)=2.477$, $p<0.001$). Between-team effects show significant differences between teams regarding their assessment of whether the team was composed of sufficient competencies, individual skills, to which extent they decided upon a team organization and change of organization. Table 5 reports the mean values for ratings of team composition, skills and organization.

**Table 5** Team effect and mean values for ratings of team composition, skills and organization

| | | Composition of competencies* | Individual sufficient skills* | Decide team organization* | Change organization* | Confidence others' skills | Comfort working with team* | Amount of collab. |
|---|---|---|---|---|---|---|---|---|
| *F(4, 35)* | | *3.69* | *5.36* | *2.88* | *8.34* | *No sig. effect* | *2.78* | *No sig. effect* |
| *p* | | *0.013* | *0.002* | *0.037* | *<0.001* | | *0.042* | |
| Team | N | M | M | M | M | M | M | M |
| A | 6 | 2.92 | 2.98 | 2.92 | 1.56 | 3.23 | 3.23 | 3.02 |
| B | 9 | 3.65 | 3.39 | 3.40 | 2.93 | 3.57 | 3.79 | 3.11 |
| C | 11 | 3.67 | 3.80 | 3.64 | 2.15 | 4.15 | 4.41 | 3.95 |
| D | 4 | 5.00 | 5.00 | 4.50 | 1.00 | 5.00 | 4.83 | 4.29 |
| E | 10 | 3.46 | 3.49 | 3.81 | 2.51 | 3.91 | 4.24 | 3.53 |
| Tot. | 40 | 3.63 | 3.63 | 3.60 | 2.21 | 3.91 | 4.09 | 3.56 |

*Significant effect of team ($p<0.05$).

On the survey question of whether the team had an adequate set of competencies (Table 5, Composition of competencies), all teams except team A rated themselves as above average, that is, the teams believed that they had the essential competencies needed for solving the task. Pairwise comparisons reveal that team D rated team composition higher than teams A and E (A-D: $p=0.006$, D-E: $p=0.045$). On the open-ended follow-up question asking for which competencies were lacking in the team, members of team A reported lack of Unix and Linux skills, team C lacked system administrator and firewall configuring skills, and Team E lacked Unix and Windows administration skills.

On the question of whether the respondent as an individual had sufficient skills (Table 5, Individual sufficient skills), team D members reported higher individual skills than members of team A, B and E (Teams A-D: $p=.001$, B-D: $p=.007$, D-E: $p=0.011$).

On the question of whether the team had decided upon a team organization (Table 5, Decide team organization), the only significant difference is seen between team D and team A ($p=.031$). On the open-ended question on initial team organization, respondents of team C, D and E described that tasks and responsibilities were assigned to different team members. Team D changed their organization least, and differed significantly from team B ($p<.001$) and C ($p=.043$). For the questions of confidence in other team members' skill, how they perceived working with each other and to which extent they collaborated, there are no significant differences between specific teams, although a team effect is detected for how they perceived working with other team members.

In summary, the between-teams analyses of skills and organization show that teams A, B, C, and E responded relatively homogenous, most significant differences between teams concern team D in terms of higher assessment on team composition, individual skills, and less changes of team organization than other teams.

A MANOVA was conducted with the responses of the seven survey questions assessing individual and team performance, network overview, information exchange, and team strategy as dependent variables and team as independent variable. Using Pillai's Trace, there was a significant effect of team, V=1.589, F(28, 128)=3.012. Table 6 reports the mean values for ratings of performance, information exchange, and strategy. Between-team effects show significant differences between teams regarding their assessment of individual and team performance, network overview, information exchange and following the intended strategy.

**Table 6** Mean values for ratings of performance, network overview, information exchange and strategy

| | | Individual performance* | Team performance* | Network overview* | Information exchange* | Decide strategy | Change strategy | Follow strategy* |
|---|---|---|---|---|---|---|---|---|
| *F(4, 35)* | | 4.22 | *10.16* | *4.01* | *3.90* | *No sig.* | *No sig.* | *10.25* |
| *P* | | 0.007 | *<0.001* | *0.009* | *0.010* | *effect* | *effect* | *<0.001* |
| Team | N | M | M | M | M | M | M | M |
| A | 6 | 2.53 | 1.98 | 2.60 | 3.51 | 2.60 | 1.56 | 2.19 |
| B | 9 | 3.07 | 3.21 | 2.74 | 3.38 | 3.21 | 2.29 | 3.48 |
| C | 11 | 3.40 | 3.43 | 3.23 | 4.17 | 3.31 | 2.10 | 3.56 |
| D | 4 | 4.00 | 3.92 | 4.42 | 4.83 | 4.25 | 1.00 | 4.75 |
| E | 10 | 2.98 | 3.19 | 3.44 | 3.39 | 2.93 | 1.84 | 3.52 |
| Total | 40 | 3.15 | 3.15 | 3.20 | 3.77 | 3.18 | 1.89 | 3.44 |

*Significant effect of team (p<0.05).

Team A responded significantly lower than team D on rating of individual performance (Table 6, Individual performance (p=.007). Team A members rated their team performance lower than all the other teams (A-B: p=.001, A-C: p<.001, A-D: p<.001, A-E: p=.001).

On the question "To what extent did you have an overview of what was happening in your team's network?" team D assessed that they had a better overview than team A (p=.014) and team B (p=.015). On the question "to which extent did you exchange information with other team members" team D reported more information exchange compared to teams B (p=.038) and E (p=.036).

There is no team effect on to which extent the teams decided upon or changed their strategies. There is, however, a strong team effect of to which extent the teams assessed that they followed their strategy. Team D reported that they followed their strategy to a higher extent compared to all other teams (A-D: p<.001, B-D: p=.020, C-D: p=.028, D-E: p=0.24), and Team A followed their strategy less than all other teams (A-B: p=.005, Team A-C: p=.001, A-D: p<.001, A-E: p=.002).

To summarize, the results emanating from Table 6 show that the ratings of team D stand out in terms of higher performance assessment, a better network overview, more information exchange and maintaining their strategy. Team A was least prone to follow a strategy.

## 3.2. Observer reports and after action reviews

Observers' NBOT reports, delivered instantly during the exercise, were useful in obtaining situational awareness for the white and green teams during the exercise. The blue team observer report template highlighted team actions and teamwork; however the actual focus of the reporting differed between the observers, which can be attributed to individual factors such as domain knowledge and motivation. For instance, the observer in team B was not a cyber-security expert, and therefore found it difficult to report on the team´s actions and detection of attacks and vulnerabilities. While a few reports from this observer did capture these aspects, this observer instead reported more extensively on teamwork factors. The rest of the

observers were PhD students within IT Security and most focused almost entirely on the actions executed and instead produced very few reports on teamwork aspects. One team member from team D noted that only sporadically did they feed data to their observer, and thus the observer reports correspond only to a small sample of the events that should have been reported. The reason for this, the team member elaborated, was that the team had no self-interest in filing reports, even though the scoring system rewarded this behavior. Thus, based on observer reports alone, no conclusions can be drawn regarding differences between teams on teamwork aspects, however they were useful for performance assessment.

The blue teams' workload and engagement in the task was rated as high by all observers (Workload: M=3.95, Mdn=4; Engagement: M=4.1, Mdn=4) throughout the exercise, all ratings between 3 and 5.

During an AAR with all team leaders on the day after the exercise, participants confirmed that they had been highly motivated during the game and that the scenario and game pace had been sufficient. AARs with the observers were conducted separately. The observers reported that it was difficult to make sense of the situation and that they had needed heads-up information on the red team's actions in order to be better prepared on what, when and where to observe in the blue teams. It took a lot of effort for observers to achieve adequate situation awareness to comprehend and report on the team task achievements. They only rarely had time to focus on assessing team strategy, organization and other team work issues. Furthermore, as the observer did not have access to the chat tools, they were unable to follow all team interactions, as teams primarily chose to use the chat-tool for within-team communication although they were located in the same room and could have used voice communication.

## 3.3. Performance measures

The exercise performance score, designed by the exercise management, contained both automatic availability check and manually assigned scores based on red and blue team reports. The post-exercise log analyses were conducted by analyzing observer reports, chat room logs and e-mail communication using ESDA as well as NIDS analysis.

### 3.3.1. Exercise performance score

The exercise performance scores are displayed in Table 7, broken down into the components that made up the total score presented to each team after the exercise. The automatic availability check was displayed to the teams throughout the game. It should be noted that the automatic scoring system was inactive during the fourth and final phase, which severely reduced motivation for some teams (NATO 2010, p. 10).

**Table 7** Performance scores from the exercise

| Team | Auto. Availability check (1p/5s) | Manual, tack accomplishment | Manual, analysis & reporting | Manual, security incident | Manual, summary | Total score |
|---|---|---|---|---|---|---|
| A | 849.2 | 0 | 50 | -1595 | -1545 | -695.8 |
| B | 1346.3 | 0 | -35 | -885 | -920 | 920.3 |
| C | 1147.1 | 130 | 210 | -885 | -545 | 602.1 |
| D | 1332.7 | 120 | 255 | -395 | -20 | 1312.7 |
| E | 1202.4 | 155 | 125 | -495 | -215 | 987.4 |

As seen in Table 7, Team B was most successful in keeping the system operational (automatic availability check), tightly followed by Team D. Team D was most successful in the manual (reporting). It was noted that teams A and B suffered temporarily from weaker communication lines, and consequently were sometimes unable to report on accomplishment of tasks. This may explain the low scores from those two teams in the analysis and reporting column. The white team based their manual scores on security incidents mainly on the red team's reports of how they managed to compromise the blue teams' systems. Based on the sum of automatic and manual scoring, team D was proclaimed the winner of the competition at the end of the

exercise. It is worth noting that this team got high scores on all components, but not always the highest. The same result has been found later at the *Locked Shields* sequel exercises (NATO 2012; 2013).

### 3.3.2. Post-exercise log analyses

A mission history recreated from observer reports, chat room logs, and e-mail communication, collected after the exercise, was instrumental in quantification of the reported attacks. The first version of the mission history enabled finding an initial classification of the targets for all discovered compromises, as reported by the red and blue teams respectively (Table 8). According to the red team reports, the most frequently attacked services during BCS were the historian, the public web server and the customer portal. The defending teams seem to have reported most of the incidents on the public web servers and the customer portals, while the attacks on the historian would be more likely to have passed undetected. Also there are reports from the blue teams having discovered attacks that were never reported by the red team. Whether this depends on false positives from the blue team or false negatives from the red cannot be determined based on only this analysis. More data is available, however the amount of time needed for conducting such an investigation is not motivated since number of samples is too low to make statistical analyses.

**Table 8** Compromised services as reported by attacking and defending teams

| Service | # reports by red team ($s_a$) | # reports by blue teams ($s_d$) | Discovery ratio (%) ($s_d/s_a$) |
|---|---|---|---|
| Customer portal | 6 | 7 | 116.7 |
| Database | 3 | 3 | 100.0 |
| DNS/NTP | 1 | 3 | 300.0 |
| External firewall | 4 | 3 | 75.0 |
| Fileserver | 5 | 1 | 20.0 |
| Historian | 8 | 3 | 37.5 |
| Intranet | 3 | 2 | 67.0 |
| Mail server | 6 | 9 | 150.0 |
| News server | 4 | 5 | 125.0 |
| Operator | 2 | 1 | 50.0 |
| Other | 7 | 13 | 185.7 |
| Public web server | 11 | 12 | 109.1 |

In Table 9, all reports related to discovered vulnerabilities are listed. It seems the most frequent reporters ($a_d$+$v_d$) are teams D and E, are also the teams against whom the attackers were least successful ($a_s$). The attacking team was instructed to distribute their efforts equally against each defending team, which implies that the number of attempted attacks should be evenly distributed among them. Unfortunately the red team reports do not include failed attack attempts, which make it difficult to verify that each team actually received the same number of attacks delivered upon them. It should be noted that some attack vectors where never deployed against team D, since the red team had decided it was pointless due to their proactive defense. Further, the reported communication problems that some teams experienced during the exercise made them unable to report. As it is unknown what attack vectors these lost reports adhere to, it remains unknown how this affects the above statistics. To be able to get the number of attacks that were actually attempted, another round of ESDA analysis would be needed with a revised version of the mission history incorporating network traffic logs and selected screen dumps to search for signature attacks and cross-reference them against these reports. The amount of time and effort to conduct such analysis is only motivated to follow up on the red

team's performance in relation to the pre-game instructions, an objective which has not been considered relevant for this case study. Consequently, this analysis was not performed.

**Table 9** Total number of attack and vulnerability reports per defending team

| Team | # successful attacks ($a_s$) | # discovered attacks ($a_d$) | Discovery of attack (%) ($a_d$ / $a_s$) | # discovered vulnerabilities ($v_d$) | # removed vulnerabilities ($v_r$) | Removal of vulnerability (%) ($v_r$ / $v_d$) |
|---|---|---|---|---|---|---|
| A | 19 | 7 | 36.8 | 16 | 4 | 25.0 |
| B | 15 | 5 | 33.3 | 18 | 3 | 16.7 |
| C | 12 | 2 | 16.7 | 13 | 12 | 92.3 |
| D | 6 | 5 | 83.3 | 25 | 25 | 100.0 |
| E | 9 | 4 | 44.4 | 26 | 18 | 69.2 |

As can be seen from Table 9, teams D and E reported a larger number of discovered and removed vulnerabilities compared to the others ($v_d$ and $v_r$). The number of reported successful attacks on these two teams was also lower than on teams A, B and C. This may indicate that teams D and E had a more proactive strategy, while teams A and B were more reactive. The lower number of successful attacks against teams D and E suggests that their strategy of identifying vulnerabilities was the most successful in preventing attacks, while the $v_r/v_d$ ratio shows that teams C and D may have been exceptionally good at removing the vulnerabilities they discovered. Cross-referencing with Table 7 showing that teams C and D received high manual scores by the judges seems to confirm that this was actually the case and not the alternative explanation that teams C and D chose to not report vulnerabilities they failed to remove.

Table 10 shows the number of attempted attacks found through NIDS analysis and the number of successful ones, together with the calculated probability of success and the mean time to compromise (MTTC) for each team. The MTTC was calculated as the time from the NIDS generated alarm until the actual compromise. It should be noted that the NIDS analysis was conducted on three first phases of the four in the exercise, and could only be conducted on data in the DMZ since the rest of the data was encrypted when captured and as such will not yield any alarms in the NIDS analysis. The data presented in Table 10 therefore only concerns the DMZ portion of the networks, as opposed to Table 9 that shows reported attacks in all parts of the network.

**Table 10** Attempted and successful attacks on each team's DMZ, calculated by NIDS analysis

| Team | # attempted DMZ attacks ($a_a$) | # successful DMZ attacks ($a_s$) | Successful DMZ attacks (%) p($a_s$/$a_a$) | DMZ MTTC |
|---|---|---|---|---|
| A | 89 | 8 | 8.99 | 02:09:41 |
| B | 80 | 14 | 17.50 | 02:50:26 |
| C | 71 | 6 | 8.45 | 04:01:34 |
| D | 43 | 7 | 16.28 | 01:11:56 |
| E | 66 | 9 | 13.64 | 05:18:03 |
| Overall mean | | | | 03:06:20 |

Table 10 shows a slightly different rating as compared to Table 9. According to Table 10, the teams A and C defended well to keep the probability of successful attacks below 10% and teams C and E had a very high MTTC compared to the others. Team D on the other hand, which seemed successful on the reporting table, shows the lowest MTTC and the second highest probability of success for the attackers.

## 3.4. Summary of results

Table 11 displays a summary of the different performance measures. For each measure, the teams are ranked from 1 - 5 where 1 = first and 5 = last. The final column shows the aggregated rankings, calculated by adding each performance measurement rank. As the analyses displayed in Table 10 were only conducted on the DMZ part of the network, it can be disputed whether the DMZ measure is a good indicator of overall performance. Therefore the aggregated scores excluding DMZ rankings are shown in parentheses of the aggregate column. It should be noted that not all performance measures are mutually exclusive, and neither can they be considered

comprehensive as there are too many confounding variables not being compensated for. This means that the aggregate ranking presented in Table 11 may serve as food for discussion, but can hardly be regarded a valid performance measure by itself.

**Table 11** Ranking of teams based on performance measures

| Team | Auto. availability check | Manual scoring | Discovery of attack | Removal of vulnerability | DMZ success rate | DMZ MTTC | Aggregate (DMZ excluded) |
|---|---|---|---|---|---|---|---|
| A | 5 | 5 | 3 | 4 | 2 | 4 | 23 (17) |
| B | 1 | 4 | 4 | 5 | 5 | 3 | 22 (14) |
| C | 4 | 3 | 5 | 3 | 1 | 2 | 18 (15) |
| D | 2 | 1 | 1 | 1 | 4 | 5 | 14 (5) |
| E | 3 | 2 | 2 | 2 | 3 | 1 | 13 (9) |

According to the aggregated rank in Table 11, Team E received the best overall performance score, while team D performed best when DMZ scores are excluded. Data also hints that the different teams used different strategies and priorities, which has been confirmed in the after action report (NATO 2010) and by post-hoc interviews with team members. Team D, which was considered most successful regarding exercise performance (automatic and manual scoring) as well as in discovery of attacks and removal of vulnerabilities received a low score on the DMZ measures, indicating that they made a deliberate strategy choice, which has also been confirmed through qualitative inquiries. Team C on the other hand, defended the DMZ best of all teams according to the table, but was not as successful in discovering and preventing attacks in the rest of the network. Had the MTTC and success rate analyses been conducted on the whole network instead of on the DMZ only, the rankings in the corresponding columns are likely to differ.

A summary of the significant survey results are shown in Table 12, aggregated per team. The presented results represent rank order in the same way as Table 11. It must be noted that rank order in this case does not necessarily indicate better or worse performance, just to what extent the teams differ. For instance, it is not necessarily the case that deciding on a team organization beforehand or to be familiar with the other team members will enhance team performance. Ranking is based on significant differences and teams not differing significantly (at $p<0.05$) are given the same rank.

**Table 12** Ranking of teams based on significant differences found in survey results

| | Performance and SA | | Team composition | | | Teamwork | | |
|---|---|---|---|---|---|---|---|---|
| Team | Team performance[2] | Network overview[2] | Expertise[1] | Team composition[2] | Prof. familiarity[1] | Decide team org[2] | Follow strategy[2] | Exchange info[2] |
| A | 2 | 2 | 3 | 2 | 1 | 2 | 5 | 2 |
| B | 2 | 4 | 3 | 2 | 5 | 5 | 2 | 4 |
| C | 2 | 4 | 2 | 2 | 2 | 2 | 2 | 2 |
| D | 1 | 1 | 1 | 1 | 3 | 1 | 1 | 1 |
| E | 2 | 2 | 5 | 2 | 4 | 2 | 2 | 4 |

[1] Rated in background survey before the game started. No between-team significance testing of responses.

[2] Rated in surveys after each day. Rating based on significant differences.

As seen in Table 12, team D's responses stand out as ranked as no 1 on all aspects except for the question on whether team members were familiar of working with each other prior to the exercise. The other team responses were more diverse, teams responding high on some aspects and low on other aspects.

## 3.5. A note on team D's strategy

Post-hoc interviews have revealed that team D chose a different strategy than the other teams, largely based on reconfiguration of the networks and proactive defense. This strategy made it difficult for the red team to breach and map their new networks, and harder to launch the preplanned attacks. Since the attackers did not

even bother some of the attacks that they played on the other teams, the results from the NIDS analysis are somewhat misleading.

Further, a member of the same team reported to have talked very little within the team, i.e. sparse within-team communication, allegedly due to a high level of trust in each other's competencies. Therefore, team D did not employ a coordinated decision making strategy, but rather each team member acted as a separate cell taking their own decisions and coordinating as a team only on individual initiatives, effectively creating a self-synchronizing edge command structure. According to the team member's own perception this led to high SA and effectiveness, a statement that has been confirmed by red team members. However it is noted that this lack of delegated responsibilities may lead to unknown unknowns and thus false SA. Consequently it is not known whether the team was actually successful in removing the vulnerabilities per se, just that they were able to hide them well enough for the red team to give up their attempts at exploiting some of them. In a way this can be seen as a valid success criterion, captured by the low number of reported successful attacks from the red team.

# 4. Discussion

The presented CDX analysis is an example of the added value of using multiple data sources. It early became clear that different types of data were useful for analyzing different aspects of team effectiveness. In the BCS case study, the observer reports and the teams' action reports were the most valuable resources for understanding team performance. For assessment of team effectiveness, surveys served the purpose of improving understanding on how the teams perceived the team composition, team organization, team strategies, information sharing, and performance. The divergent results from the different team performance measures make it difficult to compare performance with team cognition aspects to draw conclusions regarding team effectiveness. However, the survey results highlight interesting aspects of team effectiveness in the CDX domain as well as methodological implications of assessing team effectiveness in the cyber defense field.

## 4.1. Assessment of team performance measures

*Real-time availability check* seems to be a good measure when it comes to differentiating between teams in real-time. The exercise management chose to display the automatic availability score to the teams during the game as a motivator, enhancing the competitive nature of the exercise. However, it is important to note that displaying such a performance score for the teams also directs their strategy, i.e., there is a risk that teams adapt their strategy to maximize their scores instead of protecting the system. Also, basing such systems only on availability is not entirely realistic as in a real situation it may not be the wisest strategy to keep the system operational at any cost. Thus, this measure must be used with care when a CDX is set up for training of IT-security personnel.

*The manual scoring* seems to have been a successful method to force the blue teams to report on their actions. The received reports increased the white team's SA during the game and were critical for the post-exercise analyses. The manual scoring based on these reports was more subjective than the post-exercise measures since different types of reports were given different value and subjective aspects such as the creativity in solutions were given credit. Because some blue teams suffered from technical communication problems which prevented them to report, the manual scoring was not fair between the teams in BCS, especially as timeliness in reporting was given extra credit. Furthermore, the blue and red teams' ability to correctly assess whether they were successful in attacking/defending depends on their situational awareness, and it was difficult to verify the reports during the game. Both the red and blue teams sent reports that were identified as incorrect during the post-exercise ESDA analysis. However, manual in-game scoring is a low-cost measure since scores are assigned in real-time during the exercise, as compared to the time-consuming post-exercise analyses (Wildman et al. 2013), and the result was available to participants immediately after the exercise. The possibly higher accuracy of post-exercise analyses must be valued against the extra cost.

For a total exercise performance score, the exercise judges simply added the manual and automatic scores. A weighted score may be more useful to give the desired impact of each sub-score and compensate for the lack of mutual exclusiveness in each measure. Optimizing these weights however, is a challenging task and assumes that is possible to predict roughly in what range the sub-scores will end up for each team.

In the BCS setup, the measures of *attack discovery* and *removal of vulnerabilities* relied on the accuracy of the reporting process. During the ESDA analysis it was possible to identify and correct cases of false reporting which had not been detected during the game. Another advantage compared to the manual in-game scoring was that the post-exercise measures were not as dependent on the blue teams' ability to communicate with the white team during game since also observer reports were taken into account. Thus, the attack discovery and removal of vulnerability metrics were more accurate and more objective than the manual scoring, but accuracy was still affected by false reporting. Since the defending teams' networks were identical, they had the same number of vulnerabilities, while, as will be discussed in the next part of the discussion, the blue teams were not exposed to the same number of attacks. Thus, removal of vulnerabilities or attack prevention is in some sense more fair metrics than attack discovery. However, it should be noted that teams can successfully take measures to render vulnerabilities impossible to exploit without even being aware of the particular exploit, as demonstrated by team D. Therefore vulnerability removal reports are not always a reliable metric.

As with the previously discussed metrics, the *NIDS analysis* to some extent relied on reporting, however only from the red team. Analysis showed that there is an almost complete overlap between red team observer's and red team-members' reports, suggesting high data quality in red team reports (Holm et al. 2012). Unfortunately the NIDS analysis could not be performed on the full network in BCS since a large portion of the collected data was encrypted. It would have been interesting to investigate whether the contradicting results of the NIDS analyses compared to the other analyses had remained if the analysis had been performed on the complete network. However, for the purpose of this study, the differing NIDS analysis results gave some insights on team strategies' effect on performance.

Analysis of the reports showed that teams D and E were most successful in detecting vulnerabilities and removing them. The lower number of successful attacks against teams D and E suggests that their strategies of preventing attacks were successful, while the $v_r$ / $v_d$ ratio shows that teams C and D were exceptionally good at removing the vulnerabilities they discovered. This is not consistent with results from the partial NIDS analysis which proclaimed teams A and C as the most resilient defenders. The MTTC measure shows greater differences, and indicates that attacks on team D were least time consuming for the attackers. Alternative explanations include that the teams were not evenly attacked (as confirmed through post-hoc analysis).

In addition to these performance measures, self-assessed individual and team performance were collected through surveys. The self-assessed performance correlated fairly well with the automatic availability score. However, as the automatic availability score was visible to the teams in real-time, it is probable that the teams' performance assessments were influenced by the display of the availability score. The validity of the self-assessed performance is thus difficult to assess. It is worth noting that team C assessed their team performance as good, although they did not perform as good on the automatic available check. Team C did, however, perform better according to the NIDS analysis. Thus, it can be concluded that team C based their performance assessment on other factors than system uptime, such as the ability to keep the DMZ intact.

## 4.2. Assessment of team effectiveness

The diverging results of the different performance metrics have implications for assessment of team effectiveness. As performance results diverged, it is difficult to use performance score as a baseline, against which survey results are compared. For instance, the teams' assessment of whether they decided upon a team organization is in line with performance according to the exercise scoring measures, discovery of attack and removal of vulnerabilities, while not in line with to the DMZ performance measures. Survey responses showed that team D to a higher extent than other teams decided upon a team organization, followed their intended

strategy, and exchanged information. Team D only briefly explained their strategy in the survey; however the ESDA as well as the NIDS analyses gave some information on the strategies of the teams. The analyzed data gives reason to believe that teams D and E had a more proactive strategy, while teams A and B were the most reactive. AAR sessions confirmed that team D chose a proactive strategy and needed to spend less time focusing on identifying and preventing ongoing attacks, compared to the other teams.

Self-assessment surveys and observer reports are resource-effective methods to collect data on how within-team processes affect performance. Surveys can capture the team members' perception and intentions, essentially giving a picture of their shared mental models. Self-assessment is a subjective metric, however for some types of questions, participants' self-assessments are still more accurate than any observer or other currently known method can detect, e.g., only the participants can give the answer to how well they know the other team members. For reasons discussed later in this section, observers found it difficult to observe team interactions. Surveys then aided in grasping aspects that were difficult to collect using other methods. Comparing survey results to performance also revealed one clearly unreliable example of self-assessment, i.e., team E had the lowest expertise ranking of all teams (2.83) but performed second best of all teams on most performance measures. Thus, self-assessed expertise is an inadequate predictor of performance. A possible explanation to the team E underestimation of their expertise is found in the structural differences between teams E and D. Team E was composed of younger individuals with a high degree of motivation but less professional experience and confidence. Team D, on the other hand, was composed of reputed IT-security experts. Even if they did not know each other personally, or had even worked together professionally, it is likely that they had more trust in each other's skills and professional profiles compared to the other teams, not to mention more experience.

It should be noted that some participants did not see the added value of the surveys and reporting mechanisms, and consequently neglected them. To make analysis easier for both game control and after-action analysis it is recommended that this kind of behavior is discouraged through explicitly stated rules or through in-game motivators such as well-defined scoring bonuses for completing surveys in the same way as teams were encouraged to submit reports during the exercise. However, there is always a risk that participants opt out from data collection, e.g. due to privacy issues (Malek 2005).

The observers, who were fluent in the teams' native tongues, were given the additional task of assessing and reporting on teamwork aspects during the game according to predefined categories in the observer reporting template. In addition, some teamwork aspects were collected during the observer led discussions after each day and during after action reviews. Using observers is a straight-forward way of getting insight into team cognition, such as communication, behavior, strategies and to which extent the teams actually follow their intended strategies and maintain the planned organization (Yin 2009). What had not been foreseen was that although the team members were located in the same room, a lot of the within-team interaction, coordination and information sharing in some teams occurred through the chat tool. The observers did not have access to the team chat, which meant that they could not observe all interactions between team members. This had a negative effect on the observers' abilities to monitor the information flows between team members, e.g. related to decisions and strategies.

Cyber-security experts with in-depth knowledge of the field were initially perceived as a requirement for them to be able to comprehend the situation (SA level 2), and compare this to the team's selected strategy and organization. On the other hand, the observers also need expertise in team cognition for them to understand the inner teamwork processes. Observing teamwork in this high-technological context thus put high demands on observers to assess the situation and relate it to team actions, and it can be difficult to identify such individuals. To complement observer reports, qualitative analysis methods such as ESDA (Sanderson & Fisher 1994) can be used to dig deeper into what strategies were actually employed by the team.

SA or Cyber SA in a CDX environment was not specifically investigated in this case study. The self-assessment of performance measure was flawed as the automatic availability check was displayed to the teams during the game. This severely impaired the possibility to assess situation awareness by comparing performance with self-assessed performance (Champion et al. 2012). Self-assessment of the teams' network overview resulted in only small differences between the teams, and there is not enough observer data to make an assessment of their actual overview. However, what is clear from this study is that SA is a matter not only for the participants of the exercise. During the CDX, observers and exercise management were critically dependent on understanding what happened in the game in order to make valid observations, monitor progress, direct the red team and conduct the manual scoring. SA was required also for the observers to allow them to direct their attention to important interactions and progress in the observed team.

Team composition elements such as team size, skill composition and equipment were given to the team leaders to decide, with only minimal restrictions and guidelines. The many components that were not controlled by the exercise management lead to uncertainties that make any results from comparison between the teams hard to generalize to IT professionals, and the effect of team cognition aspects on team performance hard to assess. A more thorough sampling procedure and investigation of teams' individual differences may violate the anonymity of participants already reluctant to share any data that may reveal their identities and skills.

It is utterly difficult to analyze to what extent the different teams received equivalent treatment by the red team. The schedule for attacks was roughly set beforehand, but the white teams coordinated the red team's actions to fine-tune against each blue team dynamically to keep an appropriate tempo. The white team may have been influenced by the reports they received from the blue teams. Also, the red team acted goal-oriented and kept attacking a team until they reached their target, with the notable exception of team D. Therefore, the number of attacks received by a team could partly be associated with how well they defended, making the related performance scores unjust.

## 4.3. Guidelines for assessment of cyber teams

Pros and cons were found for all tested performance metrics and effectiveness analysis methods, therefore desired quality and objectivity in the results must be weighed against the amount of time and resources needed to reach conclusions. This study has applied several different performance measurement methods to study assessment of defending cyber teams. By far, the cheapest and quickest of the tried metrics is the automatic availability score, which constantly provided an indication of the teams' level of performance by accumulating uptime of critical services. However, such metrics are only as valid as the objective to keep the system running at all costs, and they yield very little insight into team effectiveness.

The report assessment and the combined scoring algorithms where designed to motivate participants to cooperate and to provide exercise control with updated reports on the teams' situational awareness. For this particular purpose, the reports served their purpose. However, when ground truth is not available for the judges in real time, it becomes very challenging to assign consistent scores to each report matching their accuracy. For fairness the scoring system should therefore allow post-hoc revision of report scores.

The quality of the incident and action reports from the teams depends on their situation awareness, which was difficult to verify during the game. While methods of assessing the SA exist, e.g. SAGAT (Endsley 2000), these methods were deemed too invasive to use during BCS. Instead the validity of reports could be verified after the exercise, making them more accurate, but also more resource-demanding, for use in performance assessment. There is also the problem of teams simply ignoring the instructions to report incidents and counter-measures which severely affects their performance scores, although not necessarily their ability to prevent attacks and protect the system. Still, it can also be argued that reporting should be standard procedure and consequently a team that refuses to do so is violating regulations and therefore not performing well. The reports are also

valuable for post-exercise reconstruction and sense-making of the exercise, as they should capture key incidents and events.

Self-assessment is normally not seen as a reliable performance measure in itself but it is useful when studying how aware teams are of their performance and which aspects they consider when assessing themselves. Such insights may give the analyst insight into team and personal strategies and objectives and may be used to explain performance.

The other quantitative measurements applied, i.e. Discovery of attack, Removal of vulnerability, Attack success rate, and MTTC, they all seem to capture some element of performance which corresponds to specific defense mechanisms. Therefore each such performance metric favors a certain strategy, and consequently it is reasonable to assume that the combination of such instruments may be instrumental in analyzing team strategy.

As it stands, none of the methods applied for analysis of BCS are comprehensive enough to merit being described as a single objective performance measurement of cyber team performance. In the end it boils down to what definition of performance is being used. The suggested trick to succeed with this grand task is to limit the scope of performance as much as possible and select appropriate methods fit to the task based on availability of resources.

Understanding cyber team performance is a challenging task, requiring the analyst to obtain sufficient situational awareness, not only in the physical and cognitive domains, but also in the virtual. To make the task tangible it is advised to decide early on which aspects of effectiveness to study, and setup plans for how to assess it. Since most of the action in a CDX occurs in the virtual domain, data capture related to performance is largely a matter of running software that log system-system and human-system interactions. Capturing such data is normally a matter of getting access to system logs and in the case of customized software, implementing log tools that capture the users' interactions. Depending on the level of analysis and the format of the logs, post-processing may be required, and followed by analysis using methods such as ESDA, particularly chunking, coding and commenting to make the dataset quantifiable and comprehensible for the analysts.

However, this study also shows that it is not always easy to capture these interactions in a format suitable for quantification, as many programs, protocols and data formats being used are proprietary and lack logging capability. The chosen fallback solution to secure data for post-hoc analysis was to capture screen videos, keyboard interactions and network traffic for all systems where no better solution could be found. This data capture is very crude and hard to interpret as needed to do chunks, connections, and codes. Despite being crude and resource intensive, it is important to note that the effort proved doable. The process became especially cumbersome since some of the blue teams were allowed to use their private laptops, and the data capturing was therefore dependent on their willingness to install and run special software and scripts to capture these interactions. Getting their consent to do so was difficult, and thus the outcome is that the data set is missing some interactions that could potentially be instrumental when making detailed analysis of the teams' interactions and progress. Whether the unwillingness depends on privacy concerns or simply on lack of interest has not been further investigated.

## 4.4. Limitations and opportunities for developed methodology

The data collection plan for cyber team performance assessment during BCS 2010 started out as a tabula rasa, with no clearly defined definition of performance, and no well-documented prior cyber team performance assessment attempts of this magnitude to lean against. The selected methods and tools were therefore adopted from the analysts own experience from kinetic tactical and operational exercises in dynamic environments, combined with input from cyber experts with less experience in team performance analysis and macrocognition. This combination of social and technological science resulted in novel techniques for

performance and effectiveness assessment of cyber teams, but as with many novel ideas they proved prone to teething problems.

Ideally, automated availability scores are a great way of providing a quick and dirty indicator of team performance, however they should be designed in such a way that they always reflect current objectives and reward good solutions. Designing such algorithms may be difficult in two-sided exercises with a high degree of freedom, and it requires deep knowledge of how the scenario may develop over the course of time. There is also a trade-off between making the scores available to the teams as in BCS, which motivates the teams to perform – but also reveals information on their current progress which may otherwise be dependent on their current SA.

To complement technical scores, live judges were used in BCS that assigned scores to each team based on reports from the red and blue teams and observers. These scores are only as accurate as the reports themselves and the judges' ability to assess the accuracy. Therefore, there is a need for judges (and game control) to maintain SA during the exercise. There is also an opportunity to use inter-rater reliability techniques to measure the quality of the judges' assigned scores, i.e. to let multiple judge assess each report independently.

As for the observers, they too can be multiplied to increase reliability. During BCS there were a few adjustments to the coding schemas to fit better with the tasks as observed during exercise. While it is natural for a newly designed coding schema to iterate a number of turns before becoming stable, it is ultimately advisable to do so before the exercise starts, as opposed to during. The fact that they were altered during BCS results in the codes of observer reports not being completely comparable between the different episodes of the exercise, and thus making quantification of observed behaviors more difficult.

The self-assessment surveys were used to get an understanding of the team's own perception of their behavior. The results have primarily been used for comparing the teams' answers against each other on a question by question level. With other techniques, such as factor analysis and correlation analysis, there are a lot of results that can be extracted from this set of data.

Finally, and arguably most importantly, with a mixed and comprehensive dataset such as the BCS, there are opportunities for integrated measures, e.g. for combining statistical communication data with content analysis of e-mail, chat logs and voice recordings. Video analysis of team interaction can be correlated with performance, human-machine interaction correlated with individual performance, etc. One of the major strengths of the R&E approach is the diversity of research opportunities, but it also forces analysts to shut many ajar doors to avoid getting trapped in the dataset.

# 5. Conclusion

Task work during a CDX typically occurs mostly in the virtual realm, while team interactions occur primarily in the physical ditto. It follows that to assess team effectiveness observers must (1) maintain awareness of the team's activities in both dimensions and (2) be proficient in both cyber security and team cognition. Furthermore, methods and tools to improve the observers' ability to conduct their work and obtain sufficient situation awareness are needed.

Several different methods were used to assess the teams' performance including self-assessments, automated availability check, mean time to compromise, ESDA analysis of reports, and NIDS analysis, but also some more novel constructs based on the dynamics between attacking and defending teams: e.g. attack discovery ratio and vulnerability removal ratio. The tested metrics measured different aspects of team performance and consequently yielded different results. By adopting a holistic approach on team effectiveness, taking into account the survey results and observer reports on team interactions, some of the differences in performance

results could be explained. For instance, it was found that the team performing best according to most metrics (except the NIDS analysis) chose a proactive strategy by redefining their network and in this manner prevent the red team from establishing a foothold in the team's network in the same way as they did in the other teams. The fact that this effect was not reflected by the NIDS analysis shows that different performance assessment methods can indeed complement each other.

To sum up, team effectiveness assessment is colored by the team performance metrics used, which is why performance measures need to be designed carefully. There are several issues to consider when choosing performance metric(s) and the current study shows the value of triangulation in order to detect anomalies and unsuspected aspects of the metric, as well as fostering a discussion on which aspects should be valued as constituting good performance.

The main lesson learned from BCS is that both technical performance measures and observer reports of team activity are invaluable both during the game for the judges to maintain control over the events; and for post-exercise analysis of team effectiveness. To measure cyber team SA, performance and effectiveness are grand challenges that require scoping and careful planning.

# References


Andersson D (2009) F-REX: Event Driven Synchronized Multimedia Model Visualization. Proceedings of the 15th International Conference on Distributed Multimedia Systems, pp 140–145. Redwood City, CA: Knowledge Systems Institute

Andersson D (2011) Privacy and Distributed Tactical Operations Evaluation. Proceedings of the 4th International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services. Barcelona, Spain

Andersson D (2013) A Knowledge Base for Capturing Comprehensive Mission Experience. P Ann HICCS 46. IEEE, Wailea, HI. doi:10.1109/HICSS.2013.40

Andersson D (2014) An Externalizable Model of Tactical Mission Control for Knowledge Transfer. I J ISCRAM 6(3), pp 16-37. IGI Global. doi: 10.4018/IJISCRAM.2014070102

Andersson D, Granåsen M, Sundmark T, Holm H, Hallberg J (2011) Analysis of a Cyber Defense Exercise using Exploratory Sequential Data Analysis. Proceedings of the 16th International Command and Control Research and Technology Symposium. DoD CCRP, Québec City, Canada.

Barford P, Dacier M, Dietterich TG et al (2010) Cyber SA: Situational Awareness for Cyber Defense. In Jajodia S, Liu P, Swarup V, Wang C (eds.) Cyber Situational Awareness: Advances in Information Security 46, pp 3–13. doi: 10.1007/978-1-4419-0140-8_1

Branlat M (2011) Challenges to Adversarial Interplay Under High Uncertainty: Staged-World Study of a Cyber Security Event. Dissertation, Ohio State University

Champion MA, Rajivan P, Cooke NJ, Jariwala S (2012) Team-Based Cyber Defense Analysis. P CogSIMA 2. IEEE, New Orleans, LA, pp. 218–221. doi:10.1109/CogSIMA.2012.6188386

Conklin A (2006) Cyber Defense Competitions and Information Security Education: An Active Learning Solution for a Capstone Course. P Ann HICCS 39. Kauai, HI. doi:10.1109/HICSS.2006.110

Cooke NJ, Salas E, Kiekel PA, Bell B (2004) Advances in Measuring Team Cognition. In: Salas E, Fiore SM (Eds.) Team Cognition: Understanding the Factors that Drive Process and Performance, American Psychological Association, Washington, DC, pp 83–106

Cowger CD (1984) Statistical Significance Tests: Scientific Ritualism or Scientific Method? Soc Serv Rev 58:358–372

Cowger CD (1985) Author's reply. Soc Serv Rev 59:520–522

Doupé A, Egele M, Caillat B et al (2011) Hit 'em Where it Hurts: A Live Security Exercise on Cyber Situational Awareness. In P ACSAC 27:51-61. ACM, Orlando, FL

Endsley MR (1995) Toward a Theory of Situation Awareness in Dynamic Systems. Hum Factors 37:32–64. doi:10.1518/001872095779049543

Endsley MR (2000) Direct Measurement of Situation Awareness: Validity and Use of SAGAT. In: Endsley MR, Garland DJ (eds.) Situation Awareness Analysis and Measurement. Mahwah, NJ: Lawrence Erlbaum.

Flyvbjerg B (2011) Case Study. In: Denzin NK, Lincoln YS (eds.) The Sage Handbook of Qualitative Research, 4th edn. Sage, Thousand Oaks, CA, pp 301-316.

Franke U, Brynielsson J (2014) Cyber situational awareness – a systematic review of the literature. Comput Secur 46:18–31. doi:10.1016/j.cose.2014.06.008

Geers K (2010) Live Fire Exercise: Preparing for Cyber War. J Homel Secur Emerg 7. doi:10.2202/1547-7355.1780

Greenemeier L (2007) China's Cyber Attacks Signal New Battlefield Is Online. Sci Am: Sep 18. Scientific American, New York, NY

Hammervik M, Andersson D, Hallberg J (2010) Capturing a Cyber Defence Exercise. In Proceedings of The first national symposium on Technology and Methodology for Security and Crisis Management, Linköping, Sweden

Hoffman LJ, Rosenberg T, Dodge R, Ragsdale D (2005) Exploring a National Cybersecurity Exercise for Universities. IEEE Secur Priv 3:27–33. doi:10.1109/MSP.2005.120

Holm H, Ekstedt M, & Andersson D (2012). Empirical Analysis of System-Level Vulnerability Metrics through Actual Attacks. IEEE T Depend Secure 9:825–837. doi:10.1109/TDSC.2012.66

Igure VM, Laughter SA, Williams RD (2006) Security issues in SCADA networks. Comput Secur 25: 498–506. doi:10.1016/j.cose.2006.03.001

Lim KH, Benbasat I (2000) The Effect of Multimedia on Perceived Equivocality and Perceived Usefulness of Information Systems. MIS Quart 24:449–471. doi:10.2307/3250969

Malek, J. (2005). Informed Consent. In C. Mitcham (Ed.), Encyclopedia of Science, Technology and Ethics, vol. 2 (pp. 1016–1019). Detroit, MI: Macmillan.

NATO (2010). Cyber Defence Exercise Baltic Cyber Shield 2010: After Action Report. CCDCoE, Tallinn, Estonia

NATO (2012). Cyber Defence Exercise Locked Shields 2012: After Action Report. CCDCoE, Tallinn, Estonia

NATO. (2013). Cyber Defence Exercise Locked Shields 2013: After Action Report. CCDCoE, Tallinn, Estonia

Otondo RF, van Scotter JR, Allen DG, Palvia P (2008) The complexity of richness: Media, message, and communication outcomes. Inform Manage 40:21–30. doi:10.1016/j.im.2007.09.003

Pfleeger SL, Caputo DD (2012) Leveraging behavioral science to mitigate cyber security risk. Comput Secur 31:597–611. doi:10.1016/j.cose.2011.12.010

Pilemalm S, Andersson D, Hallberg N (2008) Reconstruction and Exploration of Large-scale Distributed Operations: Multimedia tools for Evaluation of Emergency Management Response. Journal of Emergency Management 6:31–47

Riegelsberger J, Sasse MA, McCarthy J (2003) The Researcher's Dilemma: Evaluating Trust in Computer-Mediated Communication. Int J Hum-Comput St 58:759–81. doi:10.1016/S1071-5819(03)00042-9

Rubin A (1985) Significance Testing with Population Data. Soc Serv Rev 59:518–520

Salas E, Sims DE, Burke CS (2005) Is there a "Big Five" in Teamwork? Small Gr Res 36:555–599. doi:10.1177/1046496405277134

Sanderson PM, Fisher C (1994) Exploratory Sequential Data Analysis: Foundations. Hum-Comput Interact 9:251–317. doi:10.1207/s15327051hci0903&4_2

Sommestad T, Hallberg J (2012) Cyber Security Exercises and Competitions as a Platform for Cyber Security Experiments. In: Jøsang A, Carlsson B (eds.) P NordSec 17. Springer, Karlskrona, Sweden, pp 44-60. doi:10.1007/978-3-642-34210-3_4

Stake RE (1995) The Art of Case Study Research. Sage, Thousand Oaks, CA

Thorstensson M (2012) Supporting Observers in the Field to Perform Model Based Data Collection. In: Rothkrantz L, Ristvej J, Franco Z (eds.) P ISCRAM 9. Simon Fraser University, Vancouver, Canada

Tyworth M, Giacobe NA, Mancuso V, Dancy C (2012) The Distributed Nature of Cyber Situation Awareness. In: P CogSIMA 2. IEEE, New Orleans, LA, pp. 174–178. doi:10.1109/CogSIMA.2012.6188375

Wildman JL, Salas E, Scott CPR (2013) Measuring Cognition in Teams: A Cross-Domain Review. Hum Factors 56:911–941. doi:10.1177/0018720813515907

Yin RK (2009) Case Study Research: Design and Methods, 4th edn. Sage, Thousand Oaks, CA