UNIVERSITY OF TARTU
Institute of Computer Science
Cyber Security Curriculum

Kaie Maennel

Improving and Measuring Learning Effectiveness at Cyber Defence Exercises

Master's Thesis (30 ECTS)

Supervisors: Rain Ottis, PhD
Liina Randmann, PhD
Raimundas Matulevičius, PhD

Tartu 2017

Õpitulemuste parendamine ja mõõtmine küberkaitseõppustel

Lühikokkuvõte: Küberõppusi peetakse üheks efektiivseimaks meetodiks erinevate sihtgruppide koolitamisel, see sobib nii (sõjaväelistele) professionaalsetele meeskondadele kui individuaalsetele õpilastele. Samas põhinevad teadmised õppustel saavutatud õpitulemustest peamiselt suulisel infol ja metoodika efektiivsust pole tõestatud. Käesolev töö käsitleb õppimist küberkaitseõppustel ning keskendub õpitulemuste hindamisele. Erinevate õppuste formaatide seast on antud töö aluseks valitud tehnilised küberkaitseõppused, milles on esindatud punaste ja siniste meeskonnad. Töös analüüsitakse kübekaitseõppusi lähtuvalt täiskasvanu õpiteooriatest ja õpitulemuste mõõtmise hetkeolukorda küberkaitseõppuste raamistikus. Õpitulemusi mõõdeti kahel küberkaitseõppusel, Locked Shields ja Crossed Swords. Neist esimene on suurim avalik küberkaitseõppus maailmas peaaegu 900 osalejaga ning peamiseks koolitusgrupiks on siniste meeskonnad. Teine õppus on väiksemahuline punaste meeskonna õppus. Locked Shields ja Crossed Swords on korraldatud NATO küberkaitsekeskuse poolt. Sellised õppused on tehniliselt väga kompleksed ning nii korraldajatele kui osalejetele keerukad. Seetõttu vajavad nii õppuse disain kui õpitulemuste mõõtmine suuremat tähelepanu. Käesolev töö pakub välja uudse ja skaleeritava õpitulemuste mõõtmise metoodika, nn. "5-ajatempli metoodika". Metoodika hõlmab nii efektiivset tagasisidet (s.h. võrdlusvõimalus) kui õpitulemuste mõõtmist. See võimaldab hinnata meeskondade tegevustulemust, ja väidab, et tulemuste muutus ajas näitab ka õpitulemusi. Ajatempleid saab koguda nii traditsiooniliste meetoditega (nt. intervjuud, vaatlused ja küsimustikud), aga ka potentsiaalselt mitte-intrusiivselt võrgulogidest (nt. pcap'id). Metoodika aitab parandada tagasisidet, tuvastada õppuse disaininõrkusi ja näidata kübekaitseõppuste õpiväärtust. Crossed Swords õppuse hindamisel keskenduti eelkõige osalejatele (punaste meeskond) kohese tagasiside andmisele nende tegevuste kohta. Käesolev töö annab olulise panuse küberkaitseõppuste õpitulemuste hindamise teoreetiliste ja praktiliste aluste kohta ning pakub välja praktilised soovitused õpikogemuse parendamiseks.

Võtmesõnad: Küberkaitse – küberkaitseõppused – koolitus ja haridus – tõhusus – õpitulemused – õpitulemuste mõõtmine

CERCS: T120 Süsteemitehnoloogia, arvutitehnoloogia; S270 Pedagoogika ja didaktika

Improving and Measuring Learning at Cyber Defence Exercises

Abstract: Cyber security exercises are believed to be the most effective training for all training audiences from top (military) professional teams to individual students. However, evidence of learning outcomes for those exercises are often anecdotal and not validated. This thesis takes a fresh look at learning in Cyber Defence Exercises (CDXs) and focuses on measuring learning outcomes. As such exercises come in a variety of formats, this thesis focuses on technical CDXs with Red and Blue teaming elements. The review of adult learning theories and current state of learning measurement in CDXs context are presented. The learning measurements are performed at two CDXs: Locked Shields and Crossed Swords. First one is the largest unclassified live-fire CDX in the world with nearly 900 participants (with Blue teams as main training audience). Second one is a small scale exercise designed to train Red teams. Both exercises are organised by the NATO Cooperative Cyber Defence Centre of Excellence (CCD COE). Such top-end CDXs are highly complex, which makes it hard for organisers and participants to handle. Therefore, both learning design and measurement need careful consideration. This work proposes a novel and scalable learning measurement methodology, called the "5-timestamp methodology". This method aims at accommodating for both—effective feedback (including benchmarking opportunity) and learning measurement. The method is capable of assessing team performance, and argues that changes in performance over time equal learning. The timestamps can either be collected using traditional methods, such as interviews, observations and surveys, but also potentially be obtained non-obtrusively from raw network traces (such as pcaps). The method enhances the feedback loop, allows identifying learning design flaws, and provides solid evidence of learning value for CDXs. Crossed Swords measurement focused on providing the training audience (Red team) with instant feedback about their actions to ensure effective learning. This work contributes to theoretical foundations and in practical terms by providing practical recommendations readily applicable for improvement of learning experience in CDXs.

Keywords: Cyber Security – Cyber Defence Exercise – Training and Education – Learning Outcomes – Measuring Learning Outcomes

CERCS: T120 Systems engineering, computer technology; S270 Pedagogy and didactics

Contents

Glossary

AAR After Action Report. 7, 11, 17, 18, 28, 32, 64, 75, 77

BT Blue Team. 16, 17, 22, 28–30, 37, 39–44, 47, 49, 50, 53–56, 59–64, 68, 90

CCD COE NATO Cooperative Cyber Defence Centre of Excellence. 4, 12, 13, 16, 18, 66, 68, 75

CCDC Collegiate Cyber Defence Competition. 15

CDX Cyber Defence Exercise. 4–7, 10–15, 19–26, 29–31, 34, 36–38, 43, 45, 46, 58, 63, 75–77

CTF Capture The Flag. 14, 15, 25

ENISA European Union Agency for Network and Information Security. 11, 14

ET Emerging Threats. 69

EVE Event Visualization Environment. 70, 73

GBL Game-based Learning. 22, 34

GT Green Team. 17, 44, 50, 59, 63, 66

ICS Industrial Control System. 10, 27, 49, 50, 54–56, 59, 62–64, 90, 91

IDS Intrusion Detection Systems. 68, 69

IoC Incident of Compromise. 43, 50

LS Locked Shields. 10, 12, 13, 16–18, 26–29, 39, 47, 49–61, 65, 75–77, 87–89

MNE7 NATO Multinational Experimentation. 14

NATO North Atlantic Treaty Organization. 3, 4, 12, 13, 16, 18, 50, 66, 68, 75

NIST National Institute of Standards and Technology. 37, 38

OODA Observation, Orientation, Decision, and Action. 41, 55, 56

pcap packet capture. 4, 41, 43, 59, 61, 62, 65, 69, 75

RCT Randomised Controlled Trial. 33

RT Red Team. 16–18, 22, 26–28, 30, 39, 41–44, 49, 50, 53–55, 59–61, 63, 64, 66–75, 92, 93

SCADA Supervisory Control And Data Acquisition. 27, 50, 56

SITREP Situational Report. 50

TA Threat Assessment. 50

WT White Team. 16, 17, 28, 63, 70

XS Crossed Swords. 10, 12, 13, 18, 26, 66–70, 72, 74–76

YT Yellow Team. 17, 18, 28, 39, 43, 50, 63, 66–68, 70–72

List of Figures

List of Tables

1 Introduction

Cyber security exercises are quickly gaining popularity as a teaching method for "cyber-readiness". Globally there are over 200 cyber security exercises and more than 50% have a performance objective focusing on learning [6].

The European Union Agency for Network and Information Security (ENISA) survey provides an overview over the state of art for cyber exercises: "...after-action reports and "lessons learned" documents have become increasingly at risk of becoming fantasy documents. There is an increased demand that lessons must have been successfully learned, and that noting such instances of lesson-drawing is all there is to it. Few, if any, controls are actually made to verify that they can even be called lessons by any sensible definition, or that anything has actually been learned" [6].

1.1 Problem Statement

As the importance of cyber security exercises (with learning objectives) increases, research and practical questions related to learning effectiveness arise. However, the evidence provided in After Action Reports (AARs) is limited [6] and evaluation methodologies simply focus on the improvement of one cyber exercise to the next [7]. On one side the literature on cyber exercises and competitions describes the enthusiasm of participants for the knowledge gained and lessons learned [8]. At other end of spectrum, Pusey *et al.* [9] analyses cyber security competitions, and claims that evidence is often anecdotal and little work to validate learning outcomes has been done.

The research area for cyber security exercises learning is wide and covers a ray of questions from design to measurement aspects in wide variety of exercise types. This thesis focuses on cyber defence exercises (CDXs) with Red and Blue team elements and specifically from an organiser's perspective and addresses following areas:

1. Categorisation of the cyber exercises from learning design perspective (Section 2);
2. Analysis of CDXs learning processes in the context of adult learning theories, with special focus on team learning (Section 3);
3. Standardised metrics from learning measurement provide insight and enable comparison of learning effectiveness between teams (Section 5);
4. When measuring learning effectiveness by non-intrusive methods, the existing datasets and digital logs gathered as part of an exercise are often sufficient (Section 4, 6, 7);

5. Learning measurement (including feedback from teams/participants) provide basis for improving learning experience at CDXs (Section 6.6, 7.7, 8).

1.2 Contribution

As part of this thesis work, an extensive literature review is performed about the current state of the learning design with specific focus on team learning in the context of cyber security exercises (Section 3), and learning measurement in cyber security exercises including interdisciplinary analysis of game based learning and team learning measurement aspects (Section 4).

The thesis work uses the NATO CCD COE Locked Shields (LS) and Crossed Swords (XS) as testing platforms to put theory into practice. LS is the largest unclassified and advanced team based live-fire Blue/Red Team technical exercise (nearly 900 participants), which is a hybrid mixture of a competition, assessment and complex scenario-based learning event [5, 10]. XS is an intense hands-on cyber defence exercise for Red Team members developing skills in preventing, detecting, responding to and reporting full-scale cyber operations [11]. Specifically for those exercises, the author has contributed by:

1. a proposal for a novel 5-timestamp methodology (for collection of learning data non-intrusively) and validation of methodology using qualitative methods for LS17

2. a qualitative and quantitative evaluation of learning in LS17 (Section 6); (Section 5, 6);

3. a qualitative and quantitative evaluation of learning in XS17 (Section 7);

4. a contribution to situational monitoring tool Frankenstack [12] by evaluating how an automated feedback system can positively/negatively influence learning in XS17 (Section 7);

5. data collection templates (questionnaires) and analysis/measurement methods to evaluate learning impact for LS17 and XS17, and available for organisers of other CDXs (Section 5, Appendices A, B, C, D).

With work performed in this thesis, the author has attempted to provide practical steps how organisers can evidence the learning value and lessons learned at CDXs and connect the learning theories to practical recommendations for improvement of learning experience (Section 8).

1.3 Research Limitations

Due to the nature of CDXs (often military and confidential) there are some limitations to the research, some of which could be controlled while others needed to be accepted. The first limitation is the information provided by the participants can be confidential or not always disclosed (i.e. relating to incident handling procedures used, etc.). This includes measurement of long-term learning impact (i.e. changes in the behaviour in participants' job) cannot be evidenced by an organiser due to confidential nature of participants' daily tasks, thus not fully in the scope of this thesis. The second limitation is that interviews and surveys with open end questions are qualitative by nature and thus subjective opinion of the participants about their learning experience. The third limitation is the restriction by the organiser allowing limited number of survey questions and thus limiting feedback. To address these limitations, the surveys used scaling system allowing the participants to weight their answers and scope of work is specified.

Also, it should be noted that this thesis will concentrate on the learning effectiveness of training audience (e.g. Blue and Red Teams) of CDXs. However, the value and impact of a CDX is much wider, such as verifying the ability of individuals and teams from organisational perspective, opportunity to test the procedures and policies in safe environment, cooperation and experience provided to Yellow, Green or White team members, collaborations and networking opportunities between participants, organisers and stakeholders, etc. Therefore this work should not be read as a comprehensive value assessment of CDXs, but only as one and critical part of such effectiveness measurement.

1.4 Acknowledgments

This work would not have taken place without the NATO CCD COE open-minded and friendly organising team of LS and XS, including my supervisor Rain Ottis and exercise manager Aare Reintam, who allowed me to experiment on such major and prestigious cyber exercises. Special thanks to Bernhards Blumbergs, Mauno Pihelgas, Markus Kont who gave me first experience in collaborative writing of an article, and Jussi Jaakonaho who shared his immense experience and insight about LS and learning overall. I thank my husband Olaf Maennel, for endless encouragement and keeping on track, when I was loosing hope that I will ever finish this thesis. Enormous thanks to my little boys, Oliver and Martin, for bringing happiness into every day and making emotional boost to my learning experience (refer to brain compatible learning theory in Section 3.1.4).

2 Overview of Cyber Security Exercises

There is a wide variety of cyber security exercises and significant differences exist in regards of training audience and key topics addressed (affecting exercise design and methods). A cyber exercise can be defined as "an interactive engagement (half-day to five days or more) that enables participants to react to a plausible scenario in a risk-free environment [13]". There are also cyber defence experiments (such as MNE7 Limited Objective Exercises) to test process(es), however these are scoped out as thesis focuses on training and learning rather than testing.

Based on the international standard ISO 22398, ENISA singles out following types: 1) capture the flag, 2) discussion-based game, 3) drill, 4) red team/blue team, 5) seminar, 6) simulation, 7) table-top, and 8) workshop. Most of the exercises are simulation, table-top and workshop, representing 81% of the total, whilst Red/Blue team represented 11% of the total exercises collected. Most exercises focus on training and providing participants an opportunity to gain knowledge, understanding and skills (47%), followed by those designed to develop activities, abilities and ideas (31%), evaluate the capabilities of individuals, organisations and systems (14%) and measure knowledge, ability, endurance and/or capacity (8%). [6]

The cyber security exercises can also be categorised as academic, competitive, and collaborational. Many educational institutions have used and implemented cyber exercises as part of their computer science curriculum, and these are important tools to provide hands-on learning and assessment environments for information assurance students in college, universities, and the training industry. Some exercises take form of competitions (sometimes with commercial partners) as capstone exercises, ad hoc hack-a-thons, and scenario-driven competitions. Simulated operations are typically used in competitive (often referred as CDXs) and collaborative exercises to test the preparedness of communities against cyber crises, technology failures, and critical information infrastructure incidents at organisation, state, national and international levels. [14]

From learning aspects, for example ENISA categories of capture the flag, Red/Blue team, simulations and table-tops can all involve Red/Blue teaming feature and/or competitive design—and therefore have similar learning considerations. Furthermore, there are three types of capture the flag (CTF) exercises (a special kind of information security competitions)—jeopardy, attack-defence, and mixed. In attack-defence style competitions every team has their own network (or only one host) with vulnerable services and play "wargame", and in mixed category–the formats vary, but can include "wargame" with special time for task-based elements [15]. Simulations similarly include the attack and defence scenarios, such as automated simulation CyberCiege [16] and others.

This thesis work specifically focuses on cyber security exercises for training pur-

poses and with a general concept requiring a team to defend or offend a computer network, including the hosts and devices that comprise the network. Such exercises usually involve one fictional state or team competing against others who conduct social engineering, network reconnaissance, cybercrime, large scale denial of service attacks, network attacks, system attacks and critical infrastructure attacks. Differing objectives can be achieved through different exercises like defensive cyber exercises, small internal CTFs, red-red team competitions and integrated semester-long exercises, etc. [17]. In cyber security the operational work often takes place in teams (e.g. incidence response teams) and requires effective knowledge sharing and collaboration between individuals, teams and organizations—thus those training events are also team based.

Terminology and practices concerning exercise methodology and cyber security can vary widely, and the above analysis focused to cyber exercises and classification most relevant from learning perspective. The author uses throughout the thesis "Cyber Defence Exercise or CDX", however in various literature other expressions have equally been used, such as "cyber exercise", "cyber security exercise", "drill", "wargame", etc.

2.1 History and State of Play: Competitive Cyber Defence Exercises

In 2001, the US military service academies introduced CDX as an inter-academy competition in which teams design, implement, manage, and defend a network of computers [18]. The first Collegiate Cyber Defence Competition (CCDC) was hosted by the Centre for Infrastructure Assurance and Security at the University of Texas, San Antonio in April 2005. (Other US regions quickly became interested and started their own regional competitions with winners of regional competitions attending the Nationals CCDC). The Northwestern CDCC, the Pacific Rim Regional Collegiate Cyber Defence Competition, was first held in 2008 in Redmond, WA and currently with about eight teams competing [19].

CTFs are run locally in small communities, high schools and universities, but also available nation-wide or as a multinational competence through a year period. There are around one hundred CTFs competitions including UCSB iCTF, Ghost in the Shellcode, RuCTFe, Nuit du Hack CTF, CCCAC CTF, Insomni'hack, DEF CON CTF, Codegate CTF, Hack.lu CTF, PlaidCTF, PHD CTF, HackIM, SECCON CTF and so on tracked by "CTF Time" ranking site [15].

Simulation exercises are also designed to simulate the speed and complexity of cyber breaches and include simulated cyber attacks, and the modern cost-effective trend is to have automated simulation environments where scenarios and networks can easily be modified, such as CyberCiege [16], Arena [20], Tracer FIRE [21],

RangeForce [22] and many other proprietary or academic tools.

From the live cyber defence exercises, the larger ones are Locked Shields (NATO CCD COE) [5] and Cyber Shield for US Army National Guard [23] and several other state/regional exercises or part of large scale war games, such as Millennium Challenge 2002 [24].

2.2 Locked Shields

LS one of the largest real-time defensive international cyber exercise, is organised annually by NATO CCD COE. In LS17, more than 2500 possible attacks were carried out and more than 3000 virtualised systems were deployed. Nearly 900 participants from 25 nations were involved in exercise [10]. The training audience comprises of the national Blue Teams (BT)—computer emergency response specialists playing the role of Rapid Reaction Teams of the fictional country. The primary focus is defence and the BTs are tasked to protect and maintain identical pre-built virtualised networks of fictional, yet realistic organisations against Red Team's (RT) attacks. As part of the exercise BTs also need to handle incidents and share findings with White Team (WT) and other BTs; respond to legal, media and scenario injects; and solve forensic challenges. A game-based approach is used, meaning that the participants do not play in their real-life role and the activities take place in a lab environment. The exercise runs on separate virtualised game-net which is accessed remotely over the VPN. [5]

2.2.1 Training Objectives

The overall goal of LS is to "train teams of cyber professionals [aka BTs] to detect and mitigate large-scale cyber attacks and handle security incidents" [5]. The specific training objectives are defined for IT specialists, including learning the network; system administration and prevention of attacks; monitoring networks, detecting and responding to attacks; handling cyber incidents; teamwork: delegation, dividing and assigning roles, leadership; cooperation and information sharing; reporting/ability to convey the big picture, time management and prioritization [5]. The exercise also includes specialized parts, such as conducting forensic investigation, crisis communication (media play), cyber legal aspects (legal play) [5] and strategic game (from 2017).

2.2.2 Team-based Set-up

The BTs can be described as multidisciplinary and in LS17 average team size was 30 (ranging from 15-56). The individual learning is vital, but important part is how teams overcome individual shortcomings in skills and knowledge and achieve the

best result as a team. In addition to BTs as main training audience other teams involved in the exercise are: 1) RT's objective is to conduct equally balanced attacks on all BTs' networks; 2) WT is responsible for preparing the exercise and controlling it during the execution; 3) Green Team (GT) is responsible for preparing and maintaining the technical infrastructure; 4) Yellow Team's (YT) role is to provide situational awareness about the game (mainly to WT but also to all other teams). [5]

2.2.3 Competition vs. Learning

Due to a competitive set-up, it needs to be taken into account that some participants (teams) consider the exercise not purely a training event and other factors play part in ultimate learning result. As recommended by the author, the BTs were asked during LS16: "Do you see the exercise as a competition or learning event?" 88% (15 out of 17 responses, 3 teams provided no response) of teams responded that they see the exercise as learning event and only 12% (2 out of 17, 3 teams no response) as competition. Some of the comments received about the perception of the exercise were that for technical people it is a learning event, but managers see it as a competition, scoring makes it competitive and scoring rules need to be clear. [5]

2.2.4 Scoring vs. Learning Measurement

For evaluating efforts and directing motivation, point system is used. The scoring provides feedback and options for comparison of BTs. Availability is automatically checked, other scoring comes together from RT reporting, YT observations and WT decisions. In the past, the scoring rules have not communicated to BTs to avoid focusing on higher scores and teams were encouraged to deliver quality in all their actions. One of the points forward from LS16 is full disclosure of scoring rules, so impact of that is still not clear [5]. From the learning measurement aspects, post-exercise survey is conducted, that includes few learning questions.

2.2.5 Feedback

Immediate feedback is provided at the hotwash session after exercise, but "human touch" is challenging to achieve due to number of teams. The AAR workshop is conducted a month later that is open for BTs representatives, and results are made available for all participating teams. In addition to the AAR individualised feedback reports are provided using Cobalt Strike [25] and customised capture tool that provide reasonably good chronological record and statistical analysis (for web-attacks) from RT's perspective [5].

2.3 Crossed Swords

XS is organised annually by NATO CCD COE. The exercise is oriented at penetration testers and aims to train them working as a single united team (commonly referred as RT), accomplishing the laid out mission goals and technical challenges in a virtualised cyber environment [11]. It was created as a generic RT preparation event also for LS [5].

2.3.1 Training Objectives

The main focus is to develop tactical and stealthy execution skills in a responsive cyber defence scenario. Training objectives include practising evidence gathering and information analysis for technical attribution; executing responsive cyber defence scenario for target information system infiltration; applying stealthy execution and attack approaches; exercising working as a united team in achieving the mission objectives; and developing red teaming skills and effective tool usage, information exchange and situational awareness provision [26].

2.3.2 Team-based Set-up

The exercise is built up as one mission and one team that is divided into sub-teams (e.g. network, client-side, web/database, and exploit development) based on the participants specific area of expertise [26].

2.3.3 Learning Measurement

Feedback on some learning observations from the sub-team leaders and YT is provided in casual format during the exercise and at short hotwash session at the end of the exercise. No other specific learning measurement is carried out for the exercise. In XS17, the survey was conducted with a specific aim to evaluate impact of providing instant situational awareness (feedback) to participants (Section 7) and overall learning feedback. Lessons learned are documented in the AAR.

3 Learning–Theoretical Background

Learning is a driving force behind every change or improvement. It can be defined as "the acquisition of knowledge or skills through study, experience, or being taught" [27]. It usually means that changes take place in behavioural potential. Learning happens at individual, team, organizational, and inter-organizational levels [28]. In modern organizations the teams are central in the organizational learning process [29].

Training cyber security professionals at cyber exercises (including CDXs) should be based on sound pedagogical theories, but often these exercises are designed by technical people and learning aspects might be neglected. This section provides an overview of current adult learning theories, importance of hands on experience, considerations from game based learning and individual vs. team (group) learning aspects. Learning objectives classification and learning design analysis is recommended using Bloom's Taxonomy. As summary, these theoretical and pedagogical theories are brought together in analysis of the learning effectiveness from the CDXs perspective.

3.1 Adult Learning Theories

Much research has already been conducted regarding learning models. There are several schools of thoughts such as behaviourism, gestalt theory, cognitivism, etc. Several key models such as Kolb's model of experiential learning, Piaget's theory of cognitive development, group dynamics model by Lewin, brain compatible learning are considered as relevant for the CDXs.

It does not make sense to advocate for one best learning theory as different theories better explain learning within or for different purposes, such as neural processes vs. cultural activity systems which both are relevant when "aim is to understand how individuals or larger social communities are able to cope with rapid change" [30]. Not all modern learning theories are covered (for example competency based learning, Knowles' Adult Learning (andragogy)), and some theories are only included from team learning perspective (Section 3.4).

3.1.1 Kolb's Experiential Learning

It is widely accepted that learning takes place as a result of critical reflection on experiences rather than as a result of formal training in remembering dull theories [28]. Kolb defines learning as "... the process whereby knowledge is created through the transformation of experience." His learning model represents learning as a two dimensional process—one dimension describes methods of grasping, or perceiving, information, while the other defines methods of transforming, or pro-

cessing, information. The grasping dimension represents two different methods for perceiving, or taking in materials—i.e. feeling or thinking (Concrete Experience (CE) and Abstract Conceptualization (AC)), and the processing represents two different methods for processing material—i.e. doing or watching (Active Experimentation (AE) and Reflective Observation (RO)). The two mutually opposite dimensions create four possible learning modes, see Figure 1. While everyone can utilize each learning mode, most people favour a particular mode or combination of modes [1].

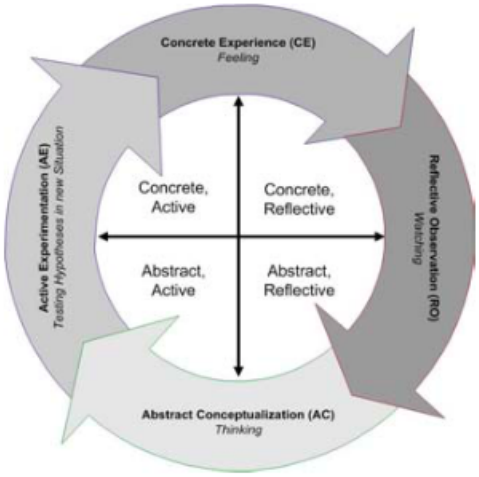

Figure 1: Kolb's learning cycle [1]

In CDXs context, the participants form their own "rules" that is a result of observation and reflection of past learning experiences (that is a basis for future learning). Feedback and reflection are critical, as from a discovery moment new learning is created and applied to different situations. As CDX are complex, it is challenging to provide relevant and detailed feedback and at the same time prepare for unknown future developments. Sections 6.6 and 7.7 give some practical insights how to improve reflection element based on learning measurement efforts.

3.1.2 Piaget's Theory of Cognitive Development

Piaget's theory of learning is based on discovery. He stated, "to understand is to discover, or reconstruct by rediscovery, and such conditions must be complied with if in the future individuals are to be formed who are capable of production and creativity and not simply repetition" [31]. Individuals develop mental structures ("mental maps") and new information is incorporated, or accommodated, into the existing structures. New information can be either rejected or transformed to fit into the learner's mental maps. Human beings have a natural desire to find and

operate in equilibrium—disequilibrium exist if information is "too far from the mental structure to be accommodated but makes enough sense to make it difficult to reject". Some people have less and some more "rigid" mental maps, making learning process different. [31]

From CDXs design perspective ensuring that participants make such discoveries and connections is critical (i.e. that scenario fits into "mental maps" of wide audience). Simply practising learned routines in the simulated exercise is not automatically disruptive enough to result in learning and change of behaviour. For example, to facilitate thinking outside of the box, Israel had a wargame defending against aliens [32]. Another dimension to consider is that for learning to move into more permanent state repetition is required, and one time attendance at a CDX might not give such effect, especially if not appropriately debriefed.

3.1.3 Lewin's Group Dynamics Model

Lewin used the term "group dynamics" to describe complex social behaviours and psychological processes that emerge in groups. Two key group processes ideas emerged out of his theory—interdependence of fate and task interdependence. His research showed that different methods of leading, significantly influenced the groups' dynamics, such as comparing different behaviour of groups with autocratic leaders those with the laissez-faire leaders. [33]

In CDXs context this is an extremely critical aspect, as learning is set up in teams. More analysis and relevance to the exercises are presented in Section 3.4.

3.1.4 Brain Compatible Learning

Brain-compatible learning endeavours to teach subject matter in a manner and format which is naturally complimentary to the brain's physical and psychological processing functions. Few of the most important principles of brain-compatible learning are that "Emotion is the gatekeeper to learning" and that brain stores most effectively what is meaningful from the learner's perspective [34].

From CDXs perspective, for example emotions, such as being in the loosing team, has possible negative connotation on overall learning impact. Also, when teams are just commissioned for the exercise purposes—learning can be affected by team (non)stability. From the meaningfulness perspective, the realistic set-up for scenario and deliverables is vital.

3.2 Hands-on Experience

As in many technical fields, hands-on learning is very important in an information security context. Cyber security students and personnel are expected to have not

only a theoretical understanding of information security concepts, but also practical skills to identify security threats, implement security mechanisms to defend against them, and restore compromised information systems. Such practical skills can only be gained through hands-on experimentation. In the literature, ethical hacking involving RT/BT activities [35, 36, 37] are recommended for teaching advanced skills.

CDXs provide hands-on experimentation that is an effective pedagogy to teach higher order thinking skills as defined within Bloom's Taxonomy (see Section 3.5). A well designed hands-on activity can integrate skills from multiple levels of the taxonomy, thereby enhancing both technical and critical thinking skills [38]. However, CDXs execution provides only hands-on training and omitting that combination of theory and practical is needed—without clear understanding it is not possible to solve technical tasks effectively and learning impact is not realized. For example, Cyber Shield exercise setup overcomes such shortcoming by providing classroom teaching before the exercise execution [39]. Thus pre-exercise phase is critical in CDXs learning context.

3.3 Game-based Learning and Serious Games

CDXs have larger human element in training setup (vs. automated game environments), however there is also connection to computer-based learning and some parallels can be drawn.

Game-based learning (GBL) is a subset of serious games focusing on the use of games for learning, skill acquisition and training [40]. Learning in games provides activities which support learning that is active, experiential, situated, problem based, provides immediate feedback, is consistent with cognitive theory and involves communities of practice which provide collaborative support to players as they learn [40].

Meta-studies conducted by Wouters *et al.* [41] and Connolly *et al.* [42] analysed papers for empirical evidence about the impacts and outcomes of computer games and serious games with respect to learning and engagement. The findings revealed that playing computer games is linked to a range of perceptual, cognitive, behavioural, affective and motivational impacts and outcomes. The most frequently occurring outcomes and impacts were knowledge acquisition/content understanding and affective and motivational outcomes. Serious games are more effective in terms of learning and retention, but they were not more motivating. Most learning impact was achieved when the game was supplemented with other instruction methods, when multiple training sessions were involved, and when players worked in groups—as these aspects enable learners to engage in learning activities from which they would otherwise refrain. To foster learning, design ideas such as think-aloud protocols and prompting players automatically to reflect on

their performance during game play have been proposed. There are only few studies on whether competition is required to make effective and compelling serious games and what specific game features determine learning effectiveness [41].

3.4 Individual vs. Team (Group) Learning

A common belief is that team exercises are valuable learning experiences for participants (individuals), teams (groups) and organizations. In CDXs, learning takes place in small teams consisting of individuals with differing skill sets who need to perform tasks together. Learning occurs at all levels—individual learning in a group context, individual learning to perform successfully in a group, individual learning on how to make groups more effective, and group-learning. However, despite organizers talking about "training the teams (groups)", there is not necessarily clarity what is meant by team (group) learning and how they measure its success, specifically for the teams (groups) in such exercises. Even the exercise design at a fundamental level often fails at having a clear differentiation what learning objectives are with respect to the team learning.

Definitions of team learning vary considerably across studies. It can be defined as a process, in which a team takes action, obtains and reflects upon feedback, and makes changes to adapt or improve. According to Senge, learning involves transforming collective thinking skills so that groups can reliably develop ability greater than the sum of individual member talents. It can be also viewed as dynamic process in which learning steps, environment, individuals in the group, and group behaviours change as the group learns. Some interpretations of group learning confuse levels of analysis by not distinguishing "individual learning in the context of groups" from "group-level learning". If an individual leaves the group and the group cannot access his or her learning, the group has failed to learn—so the other processes like sharing must have happened in learning context. [29, 43]

Three types of group learning can be distinguished: 1) adaptive learning (often called single-loop), which is concerned with developing capabilities to manage new situations by making improvements without necessarily examining the relevant learning behaviours, 2) generative learning (often called double-loop), which is motivated and regulated by the group to acquire, share and use new skills, knowledge, and information with emphasis on experimentation and feedback, 3) transformative learning, which happens when the group needs to make a major change to accommodate outside pressures, respond to opportunities, etc. [44, 29, 43]

From CDXs perspective the teams are critical component and achieving some training objectives, such as effective team communications, is directly dependent on team composition and team dynamics. However, the incident response teams can function together for several hours or a few days, and never meet again. These teams can be described as "extreme action teams—teams whose highly skilled

members cooperate to perform urgent, unpredictable, interdependent, and highly consequential tasks while simultaneously coping with frequent changes in team composition and training their teams' novice members" [45]. Findings of Klein *et al.* [45] study about leadership in the extreme action teams suggest those teams have a hierarchical, de-individualized system of shared leadership. Relevant finding for team learning is that when senior leaders delegate then junior leaders/team members learn by doing. Although some characteristics of extreme action teams have been studied, hardly any literature has been published specifically on learning aspects and cyber defence teams.

Without a focus and understanding the basic mechanisms of team learning, it can be difficult to distinguish learning from other team processes (including performance) in CDXs context. At team level knowledge is aggregated, however it is not a straight-forward accumulation—i.e. information is not used equally, information losses can occur, and knowledge can be unequally distributed [46]. As this is very wide research area, this topic is not covered in this thesis but rather should be addressed as future work.

3.5 Bloom's Taxonomy

Bloom's Taxonomy is a classification of different levels of cognitive learning objectives that educators set for students. These learning objectives describe six progressive levels of learning: knowledge, comprehension, application, analysis, synthesis, and evaluation [2]. Anderson *et al.* revised Bloom's work, as depicted in the right side of Figure 2 [47].

Anderson and Krathwohl's Taxonomy (updated Bloom's Taxonomy) levels are explained as follows:

1. Remembering: Learner's ability to recall information
2. Understanding: Learner's ability to understand information
3. Applying: Learner's ability to use information in a new way
4. Analysing: Learner's ability to break down information into its essential parts
5. Evaluating: Learner's ability to judge or criticize information
6. Creating: Learner's ability to create something new from different elements of information

Not originally incorporated in Bloom's Taxonomy, but if the exercise or serious game is set up in virtual game environment then psychomotor domain is also relevant (e.g. at game interaction level) [2].

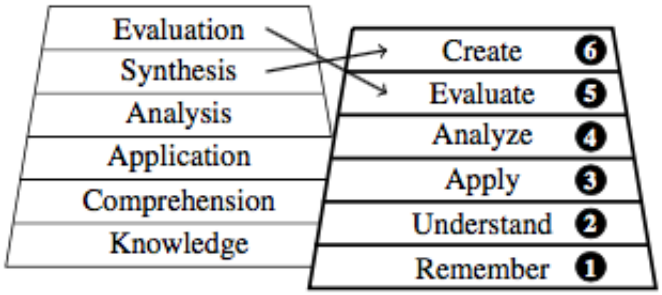

Figure 2: Bloom's learning objectives; the original objectives appear on the left, and Anderson *et al.*'s revision appears on the right [2]

From cyber learning standpoint, security failures are often due to practitioners managing the security problems at the less-sophisticated levels of Bloom's hierarchy. For example, firewalls ignore that transport-layer ports have at best a de facto relationship with services (and that furthermore protocols such as HTTP can represent any number of services); antivirus scans have difficulty catching anything other than known threats; and keeping software up to date does not remove undiscovered flaws, which leaves systems vulnerable to unknown attacks, not breaking system into mutually untrusting, isolated components (e.g. Heartbleed) and decision to build servers from sophisticated software designed for interactive-command-line use (e.g. Shellshock) [47].

Development of serious games and interactive exercises (relevant also for CDXs) must blend subject matter content, instructional design aspects, learning objectives and engaging game design (such as scenario) to encourage learners to practice and develop their skills. Different types of games and tasks are differently suited to certain learning objectives as defined by Bloom's Taxonomy. For example, 3-D games or simulations with avatars are not well suited to basic knowledge acquisition (e.g. CyberCiege learner needs to make strategic choices that demonstrate comprehension of best practice and higher-level concepts), and are more suited for applying and evaluating taxonomy level. [2]

In CDXs context, Moses *et al.* analysed CTF competition of Cyber Security Awareness Week Conference of the New York University Polytechnic School of Engineering—a jeopardy-style competition in which teams race to complete security challenges with differing complexity levels. The vast majority of challenges met objectives corresponding to levels 1–3 (remember–apply), the challenges with the lowest completion rates typically involved multiple learning objectives at levels 3–4 (apply–analyse), and there was the complete absence of challenges mapping to level 5–6 (evaluate–create) [47]. Thus such mapping would give insight and

identify shortcomings in CDXs design.

The Bloom's Taxonomy enables the classification of training objectives and provides a well-known basis for measurement criteria in CDXs.

3.6 Learning Process in CDXs

The cyber exercises are organized according to different methodologies. This section focuses on application of the learning theories and some significant learning design questions from CDXs perspectives.

From the learning perspective, the exercise execution is only a part of the learning journey for participants. The complete participants' experience is a process comprised of preparation, execution, and post-execution phase. Below analysis combines theory and relevant examples from LS and XS analysis.

3.6.1 Pre-execution Phase

Preparation phase is actually where the learning focus should be. As this phase is not wholly under supervision of the organisers, one way of increase by impact is to give sufficient guidance to teams what they need to be able know and do before.

One of the LS team leaders writes that in pre-phase it is important to ensure every team member is familiar with and understand all the information about the exercise. He recommends to use lectures and creative discussions, but also mental maps and brainstorming to define security risks, team members' abilities, and the knowledge areas which were important to be learned or executed. Teams should plan expected situations in the exercise and prepare individuals, technology, and information for all the backup scenarios. The ideal is to have identical infrastructure (i.e. virtual lab) used at exercise for detailed analysis. Risk (threat) identification and preparing detection and mitigation processes should be automated in "rapid development" style and tested. [48]

3.6.2 Execution Phase

There is an ancient Chinese proverb: "I hear and I forget, I see and remember, I do and I understand". Thus the idea of active hands-on learning producing understanding is not a new concept (Section 3.2) and thus the exercise itself is invaluable.

Level of Difficulty/Complexity An exercise needs to cater for different skill levels, and it is a fine balancing art how to make it right difficulty level from learning perspective. Some teams are technical experts and some are novices. For example, in LS16 one team was so "bored" that they noticed RT violated the game

rules. Some other teams were struggling to put up any fight to RT. LS is designed for high-end skilled teams—so potentially additional challenges to keep such highly skilled experts "entertained" might be worthwhile to provide (reflected in scoring by additional points for solving difficult challenge).

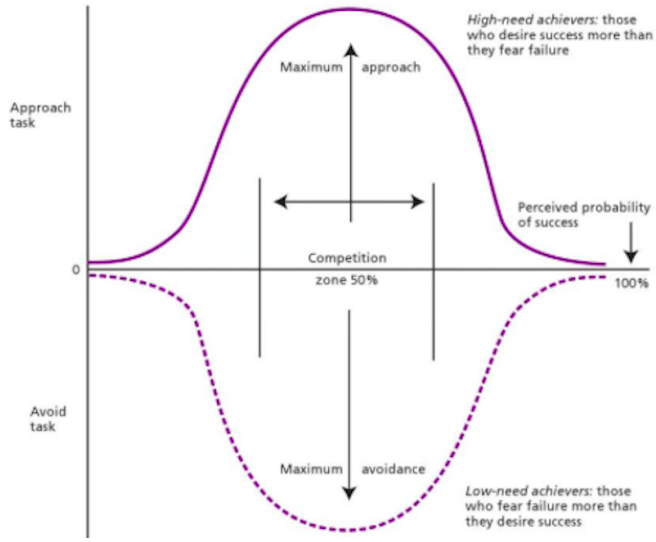

Figure 3: Probability of success, motive predominance and task involvement [3]

An example of feedback for one network component shows that from learning perspective the level of difficulty is about right, but also realization that real-life is more challenging: "In LS15 it was relatively easy to protect the simulated power generator. However, protecting real SCADA/ICS systems is much more complicated and failure will have much bigger consequences. ... . Secondly, in the scenario of LS15, only one variable–the operating frequency–was taken into account. However, a real power plant has a lot more important variables and their behaviour in complex dynamic systems can be unclear. Therefore, it is much harder to defend real SCADA/ICS. The exercise gave us an opportunity to test and improve our skills. We experienced the importance of an Intrusion Detection System based on anomaly detection. The attackers in LS penetrated deep in the network despite of all countermeasures" [49].

However, such feedback is anecdotal and might not reflect teams experience overall. See Figure 3 for considerations in the exercise design of finding the right balance between probability of success, motive predominance and task involvement suggested by Dennis McInerney [3].

Scoring—Motivational or Disruptive There are differing opinions regarding scoring as motivational or disruptive to learning. The comment received from one

of the LS organisers was "if you don't score–they will not do it". From positive side scoring supports teams' enthusiasm from the competitiveness. Measuring the progress and making the results available to all BTs is essential in order to provide feedback—but the scores do not reflect the ultimate truth regarding how good each team is or how much they have learned. Some teams take scoring more seriously than others—i.e. teams viewing the exercise as learning event vs. competition (e.g. Section 2.2.3). The scoring needs to reflect the learning objectives and provide learning insights, not mainly focus on "game" rules aspects. With unclear scoring, a team's effort is unclear and makes identifying their weakness, i.e. further learning needs, challenging. From learning aspects, scoring and performance results cannot simply be equated. In cases, learning does not necessarily lead to improvement in performance, because results of learning processes are not the only determinants behaviour, individual abilities (e.g. skills), personal motivation and team impact. It is possible to learn any kind of behaviour, and learning process can result in a deterioration of performance [50].

Injects The purpose of scenario injects is "to make the game more versatile and load the BTs with additional tasks" [51]. From learning perspective, injects should be used to create reflection and reinforce the key training objectives. MITRE guidance says that "each inject to the exercise presents an opportunity to assess, teach, and learn with the training audience" [52]. Injects need to meet the training objectives, match the skills and capacity of the training audience, and should be analysed that these are not overly disruptive to learning and provide right balance for maximize learning impact.

Instant Feedback to Participants/Teams The on-going feedback is vital in the learning experience. The critical question to ask is whether it is really learning or "adjusting to the game rules". When the exercise has technical objectives, the organisers need to focus on good quality RT reporting—i.e. providing dynamic updates to the campaign progress, what is done and why. Scoring and reporting to WT/YT reveal the progress, but it may be limited due to the complexity and speed of the exercise, thus the capture of exercise timeline is vital to gather baseline for post-exercise feedback.

3.6.3 Post-execution Phase

Reflection and analysis are critical for learning to take place. The teams can do it efficiently and effectively, if meaningful feedback in hotwash/AAR and meaningful reporting are available from the organisers.

One of the considerations is that any big data becomes valuable when refined and can be understood easily—and the organisers can provide such enhancement

and also include baseline metrics. Collecting such metrics and being to able to compare themselves with the best/worst teams, will create learning opportunities for the teams.

An example from BTs perspective in LS: "Every player expects a feedback." Thus the lessons learned should be formulated and written reflecting evaluation of the exercise and suggestions for the next improvements—as a result team gets a package of technologies, processes, activity reports of players, and BTs activity record. Also studying activity reports of other teams and technologies is useful". [48]

3.6.4 Learning in Teams

Cyber security teams operate in a highly uncertain and complex environment with low visibility what is happening in other side. Furthermore, the cyber security environment is collaborative and involving a number of different roles in teams [53].

For CDXs, the teams are often formed few weeks before the exercise and dissolve after the execution. For example, in LS each team can select their necessary capabilities and members, and wide selection of team members is encouraged, as it has been seen that the teams who were able to assign owners to every system were better at detecting and restoring their systems. [5]

A study by Gokhale has shown that group learning supports development of critical thinking through discussion, clarification of ideas, and evaluation of others' ideas [54]. It is one of the main purposes of CDXs to create such environment and provide teams to learn and practice their strategies. Learning is seen as something that can be aided by experience, and in many cases this is true. But, practice does not make perfect, it only makes permanent. If you practice something wrong, you will reinforce doing it wrong [55]. If BTs fail, the reasons needs to be understood why they failed. The team dynamics and learning concepts play a key role in this understanding.

Experiments have demonstrated that when task rules are changed, the teams can still often retrieve the old (and outmoded) learning. Teams often fail to learn because they focus on individual-level changes rather than group-level routines. More collaborative groups are more likely to change their routines, have group discussion about performance discrepancies that increases the probability of group learning. Team members with high centrality (i.e. team leader, key team members) are more involved in indexing stored memories than other members, and this ultimately impacts group learning. [29]

Team learning breaks down when teams fail to reflect on their own actions, or when teams reflect but fail to make changes (due to inability to break routines, lack of resources or motivation). Negative evaluation or criticism is needed to trigger learning, but it may be difficult for teams to have high-quality reflective

discussion about their shortcomings without considerable psychological safety (i.e. without fear of negative consequences to self-image, status or career) [55].

In CDXs often, an intense attack campaign from RT tends to be successful (i.e. BT's defence fails). Does the failure still constitute learning or an experience that reinforces what was or was not learned earlier and elsewhere? 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 the team did not know how to manage it? Was it information overload, technological failure, lack of cues or knowledge, lack of motivation, insufficient procedures or simply an unfortunate misunderstanding? These are questions, for which the current research has not provided any answers. If the causes to an outcome can be identified and analysed, then training needs can be pinpointed. CDXs show that the power of an individual may be multiplied with synergy effects during the exercises, when teams are configured properly and the qualities of individual players are utilised [48]. Applying some simple triggers, e.g. explaining the value of participation, agreeing clear team roles, etc. can increase the motivation to learn and thus increase learning effectiveness both at individual and team level.

3.6.5 Key Points for CDXs Learning Design

To conclude, in CDXs following learning principles should be kept in mind:

1. Learning is not always intentional but it is a motivated behaviour—it might be hard to motivate, as individuals themselves have not realised that. Setting the "right and motivating learning tone" that a CDX is learning event, will play significant role in likelihood of increasing learning;

2. Individuals learn from models and casual inferences—there are differing learning theories, and organisers should ensure sound pedagogical principles are applied in CDXs design. The learning that each participant takes away will be individual and differing factors will contribute into learning impact;

3. Team and collaborative learning—in cases with technical tasks and individualised team roles, team learning aspects might not be considered. Team dynamics and leading play significant role in learning outcome or failure;

4. Feedback and reflection are essential elements that should be integrated into overall design of an exercise and not be limited to follow-up activities only;

5. Learning does not equal performance—a team who wins the competition is not necessarily the one who learns the most. Improvement in performance can be an indicator for learning (a process, a change in behaviour) and achieving the training objectives.

4 Learning Measurement at CDXs

The general guidance, such as [56, 52], on how organisers should look at design and performance (training success) measurements has been published. However, there is currently no clear and widely accepted methodological evaluation methods published and scientifically proven for learning results and measuring learning impact or assessing cyber security skills/competencies obtained through cyber exercises and/or serious games. However, these exercises provide unique opportunity, as non-intrusive digital logs offer grounds for qualitative and quantitative evaluation of learning effectiveness specific to such exercises.

Evaluation is defined in ISO 22398 as "the systematic process that compares the result of measurement in relation to recognized criteria (i.e. training objectives) to determine the discrepancies between intended and actual performance" [57]. Learning (training) objectives define the expected goal of exercise in terms of demonstrable skills or knowledge acquired by a participant as a result of exercise [58]. Without clear objectives it is not possible to design a meaningful exercise [52], or measure the outcomes. A consideration should be given to team learning aspects, and how this learning is transferred to the organisation.

The literature review and analysis in this section focus on learning impact measurement and aims to summarize what methodologies can be used to determine effectiveness of learning outcomes, and what are the results and challenges (limitations) of such learning effectiveness measurements. The literature search did not provide any specific CDXs related learning effectiveness meta-studies, and only limited number of papers on specific cyber training event learning outcome measurements. Thus the review includes papers on learning measurement at cyber exercises, interdisciplinary papers on serious games and game-based learning meta-analysis, team learning measurements and focusing on other measurements in cyber exercises. Due to the commonalities on game-based learning and elements of team learning, the author believes interdisciplinary approach can be transferred to CDXs.

4.1 State of Play: Learning Measurement at Cyber Exercises

Dr Ahmad [14] investigated how a cyber crisis exercise can benefit participants' individual learning and how their experience in the exercises is transferred to their organisation. This research used a post assessment framework that adopts the four-level Kirkpatrick training model to collect, code and categorise the participants interview data in order to investigate the learning outcome from four levels: reaction about the exercise, learning skills experienced during the exercise, behaviour developed during the exercise, and result, i.e. how the benefits are trans-

ferred to their organisation. At the organisational level, the framework provides an assessment of organisation cyber resilience of Critical National Information Infrastructure sectors involved in the exercise. From this thesis perspective an individual part of assessment model is relevant, however the approach lacks team aspects of learning.

U.S. Army Research Institute for the Behavioural and Social Sciences Research [59] measured game-based simulations by different questionnaires: 1) game performance assessment, including measure for any technology-based training simulation effort is the time required to achieve proficiency in using the simulation, 2) game experience measure—to separate the experience and skill from the actual correct knowledge about games, 3) graphical user interface questionnaire—a poor interface can diminish the potential for learning and transfer, 4) exercise questionnaire—questions addressing the training effect and fidelity of the system relevant to the mission(s) performed, 5) AAR questionnaire—to provide participants with a relatively objective ground truth [60], and 6) biographical questionnaire and ancillary information—a collection of baseline information, e.g. age, education, time in career, and experience with computer programs in general. The authors conclude that it is difficult to encompass all aspects of each factor that may training effectiveness with questionnaires—thus recorded interviews with probing questions after the exercises were also completed [59].

These measurements use questionnaires, observation and other traditional measurement approaches and do not utilise digital dataset as non-intrusive data source.

4.2 Game-based Learning Metrics and Methodology

Connolly *et al.* [42] proposes a broad model for the evaluation of games for learning that includes motivational variables such as interest and effort, as well as learners' preferences, perceptions and attitudes to games in addition to looking at learner performance. Outcomes relate to learning and skill acquisition but also affective and motivational outcomes. However, the categorizing and naming of skills and learning outcomes in a useful way is challenging and non unified.

Examples of learning outcomes categorisation relevant include: 1) Connolly *et al.* [42] suggested categorization such as distinct employability skills, learning skills and core skills, 2) Wouters *et al.* [41] provided learning outcomes and factors such as learning and retention, motivation, learning arrangement of the comparison group, serious game combined with other instructional methods, number of training sessions, group size, instructional domain, age, level of realism, narrative and methodological variables, 3) another Wouters *et al.* [61] study proposes four categories of learning outcomes: cognitive, motor skills, affective (change of attitude of the learner and motivation) and communicative.

Data Gathering Considerations Serious games and game-based learning provide excellent environments for mixed-method data gathering (i.e. triangulation), including crowd sourcing, panel discussions, surveys and observations (including video observations). Some of the examples of different forms how to gather data are: 1) use video observations and network analysis to analyse team communication patterns and effectiveness, 2) conduct pre-game, in-game and post-game knowledge tests to measure the increase in knowledge, 3) use validated pre-game and in-game questionnaires on relevant psychological constructs, including team communication and commitment to change, 4) use pre-game and post-game questionnaires on such aspects as learning satisfaction, game play and motivation, in combination with maps and strategic decisions, 5) use in-game logging and tracking on hundreds of events and results, including distances, paths, play time and avoidable mistakes, in combination with questionnaire, 6) use group of international players as an expert panel in a survey [4]. However, not yet explored issues are seamless, or "stealth" data-gathering and assessment as well as performance based evaluation [62]. Stealth assessment (i.e. non-invasive, unobtrusive assessment) could potentially increase the learning efficacy of serious learning given that much of the learning now remains relatively "implicit" and "subjective" (e.g. as noted in personal debriefings) [4].

Evaluation Methodology Qualitative research is more subjective than quantitative since it is more interpretative, but it can provide a broader approach to examining the skills that playing games can support [62]. The methodology comprises a framework, conceptual models, research designs, data-gathering techniques, hypothesis formulation, directions for developing and using evaluation constructs and scales, and procedures for testing structural equation models. The effectiveness of serious games can be evaluated in different ways: constructivist vs. objectivist; theory-based vs. explorative, summative vs. formative, learning vs. accountability type, broad vs. narrow; and rigorous vs. generic evaluation [4]. Methods to evaluate the learning outcomes of serious games include 1) meta-analyses, 2) randomized controlled trials (RCTs), 3) quasi-experimental designs, 4) single case experimental designs—pre and post test, and 5) non experimental designs—surveys, correlational, qualitative [62]. Connolly *et al.* [42] shows that studies of games for learning and serious games did not use many quasi-experimental designs with surveys, also RCTs and qualitative designs were relatively uncommon. The effect of training on learning (acquisition of skills or knowledge) was measured by calculating the difference between the pre-test and post-test scores on the questionnaires or cognitive tests, and comparison to control group [63]. Figure 4 includes an overview what to measure, how and when [4].

How		What?	Pre-Game	In Game	Post-Game
Self-reported	Qual.	Personality, player experiences, context etc.	Interviews, focus group, logbook	Logbook, interviews or small assignments as part of the game	Interviews focus group, after-action review
	Quant	Soc-dem., opinions, motivations, attitudes, engagement, game- quality learning, power, influence, reputation, network centrality, learning satisfaction etc.	Survey, questionnaires, individual or expert panel	In-game questionnaires	Survey, questionnaires, individual or expert panel
Tested	Qual.	Behaviour, skills etc.	Actor role-play, case- analysis, assessment, mental models etc.	Game-based behavioural assessment	Game-based behavioural assessment
	Quant	Values, knowledge, attitudes, skills, personality, power	Psychometric, socio- metric tests (e.g. personality, leadership, team roles, IQ)	Game-based behavioural performance analysis	Game-based behavioural performance analysis
Observed	Qual.	Behavioural performance of student, professionals, player and/or facilitator, others; decisions, strategies, policies, emotions, conflicts etc.	Participatory observation, ethnographic methods	Video, audio personal observation, ethnography, maps, figures, drawings, pictures etc.	Participatory observation, ethnographic methods
	Quant	Biophysical–psychological responses, including stress (heart rate, perspiration)	Participatory observation, network analysis, biophysical–psychological observation	In-game tracking and logging; network analysis, data mining, biometric observation	In-game log file analysis, network analysis

Figure 4: What to measure, how and when? [4]

Limitations and Issues Identified The current evaluation models are dominated by single case-studies, single games, single contexts of application, there is a lack of information on the questionnaires used, there is only little attention to advanced professional learning outside of education and little attention to the learning of teams, groups, organizations, networks or systems [4]. Most of the existing models and frameworks are high-level models and their weaknesses include the following: 1) a lack of comprehensive, multi-purpose frameworks for comparative, longitudinal evaluation, 2) few theories to formulate and test hypotheses, 3) few operationalised models to examine "causal" relations (e.g. in structural equation models), 4) few validated questionnaires, constructs or scales, whether from other fields (e.g. psychology) or constructed especially for serious games and GBL, 5) lack of proper research designs that can be used in dynamic, professional learning contexts (i.e. whether what was learnt truly matters for real-life performance (e.g. emergency management, leadership)), and 6) the absence of generic tools for unobtrusive ("stealth") data gathering and assessment [4]. Rather than just reporting descriptive data, it is possible to carry out more sophisticated analysis with survey data, looking at links between variables [62].

Application for CDXs The evaluation is complex and multidimensional since it involves evaluation not just improvement in performance, but also evaluation of the user acceptance, engagement and satisfaction with the game [62]. Evidence of learning outcome (i.e. behaviour change) is hard to measure—the theory to "assist" the measurement is that behavioural intentions are good predictors of behaviour [40]. An evaluation framework for serious games research (that can be applied also to CDXs) should ideally have the broad scope, be comparative, standardized, specific (measure data precisely by pinpointing variables), flexible, triangulated (i.e. using a mixed-method approach with qualitative and quantitative data), multi-level (consider the individual, game, team, organization and system levels), validated (use validated research methods), expandable, unobtru-

sive, fast and non-time consuming, and multi-purpose (extend their data-gathering efforts beyond the obvious and minimal) [4]. Assessment methods that consider learning context may reveal differences in performance that would be undisclosed with traditional assessment methods [61].

4.3 Team Learning Measurement Aspects

Measuring team learning is a complex task with many factors, such as learning impact has not been identified (i.e. simply there is no similar event in reality), change can be environmental (i.e. not due to learning) and learning could be dysfunctional (i.e. false connections made between actions and outcome) [29]. So far researchers have focused on limited set of learning outcomes, mainly learning of fairly simple concrete knowledge. However, full range of group learning outcomes (e.g. cognitive, behavioural, and emotional outcomes) should be considered based on the group learning definition in Section 3.4. Most common measurement methods are interviews, surveys and observations, and learning maps.

Qualitative research is a useful methodology for investigating phenomena that are not well understood [64]. For example, Edmondson used observation and interviews to study role of teams in organizational learning and surprising observation was that many of the studied teams did not learn in the context of existing literature [64]. However, based on her study half of the teams engaged in reflective discussion about process that led to subsequent changes, and would constitute a team learning process.

One shortcoming to note regarding interviews and surveys is that as the learning is not necessarily consciously accessible, thus asking the group members about what they have learned will not uncover any changes. For example, a study identified learned patterns of behaviour (e.g. using metaphors) that members were not consciously aware of. [29, 64]

To balance data requirements, but also minimize time demands on interviews and surveys "informant sampling approach" can be used (i.e. relying on limited sample of most knowledgeable team members) [64]. A challenge is that an informant needs to evaluate their team rather than their his/her own personal behaviours, and a cyber defence team has a range of differing technical skills, which the "team leader" or observant (who might not be technical) might not actually able to respond.

Measuring long-term learning effect requires detailed and multiple real-time observations of the same group over time. Wilson *et al.* [29] recommend waiting for the average time interval between events/incidents and then giving each member a scenario describing another major attack and asking how the group should respond. Then by observing or surveying it can be identified what is, or is not, ultimately retrieved.

Newman *et al.* measured critical thinking during group learning using a questionnaire and the content analysis method (which relies on identifying examples of indicators of obviously critical and uncritical thinking) [43]. Alternatively, Hay used concept mapping—i.e. participants produce concept maps of the topic before and after [65]. Uzumeri *et al.* and Chiva *et al.* used learning maps or curves at team and organization level [66, 67]. The learning maps describe groups of learners in quantitative statistical terms, provide descriptive model that does not need to assume a specific causal mechanism and can be used for differing measurements [66]. However, some negative aspects include that they concentrate on learning by doing and measure learning in terms of the results obtained (i.e. short-term), focus on explicit outputs (not on learning processes, sources or capabilities) [67, 68]. Team learning behaviour mediates the moderated, non-linear relationship between expertise diversity and team performance [64].

From CDXs viewpoint, all these methods are potentially applicable. However, similar challenges are faced as by researches so far—i.e. separating learning from other factors and that learning might not be necessarily "visible". Also, for incident response teams activities are conducted on computers/network—so observations of behaviour (sitting quietly behind computer screen but at the same time mitigating a significant risk or attack) might not provide sufficient information about the learning. Traditionally psychologists have been observing via looking at the behaviour quietly. Now for CDXs, it should be "seen" with a different kind of eyes–on the network and system-level and to learn to "observe" on those technical levels. This is one of the reasons this thesis looks further into "digital footprint" and applying non-intrusive measurements of the team activities during the CDXs, as for possible additional information source about the team behaviours, learned patterns, improvements in time, etc.

4.4 Other Measurements Conducted at CDXs

Team Effectiveness and Proficiency There are few studies performed of team effectiveness and proficiency in CDXs, such as [53] and [39]. Some of team performance and effectiveness metrics ultimately also affect the learning measurement.

Study about Baltic Cyber Shields 2010 team effectiveness [53] used different interdisciplinary methods including automated availability check, exploratory sequential data analysis, and network intrusion detection system attack analysis. Also, 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 main conclusions found were: 1) a combination of technical performance measurements and behavioural 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.

Incorporating social and behavioural research methods into the cyber security field can give new possibilities of understanding causes to a given effect. [53]

In Cyber Shield 2015 [39], the attempt was made to predict proficiency in the teams and also to identifying the best methods of training, assessing, and educating the cyber security workforce. The assessment consisted of a pre-event expertise survey, proficiency data collection during the event, and a post-event training survey. The following BT proficiency metrics were developed: 1) Time-to-Detect, 2) Time-to-apProval, 3) Time-to-End, and 4) Category Correct. The average for each timing proficiency metric across all teams was found and grouped by National Institute of Standards and Technology (NIST) category. [39]

Situational Awareness Reed *et al.* study [69] evaluated cyber defender situation awareness, and showed that the most pervasive form of competition-based exercise is comprised of jeopardy-style challenges, which compliment a fictional cyber-security event. The competition used challenges containing over twenty attack techniques. The following observations were made: 1) a group of defenders performs better than an individual; 2) situation awareness of the fictional event may be measured; 3) challenge complexity does not imply difficulty. Effective competitions complement training goals and appropriately improve the knowledge and skill of participant. [69]

Use of Tools in CDXs Silva *et al.* study [21] considered factors of successful performance in Tracer FIRE exercise with emphasis on the use of software tools by participants. The speed is often not the main consideration—rather participants who devoted more time to challenges tended to make more correct submissions. Findings relevant for learning design were that: 1) software architecture should include a variety of general purpose tools and allow to download preferred tools, 2) frequently switching between challenges or tools, or within a challenge may be a sign for having difficulties, 3) some browser use is essential, however extended browser use together with frequent switching between the browser and other software tools is indicative that a participant is "lost". [21]

4.5 Summary

The contribution of this section is an extensive literature review with analysis in CDXs context to summarise what measurement methods are used in measuring learning effectiveness. The interdisciplinary approach was used, as there are not many papers on the learning impact measurement specifically on CDXs. However, interdisciplinary research methods, specifically from game-based and team learning, can be successfully applied in cyber security learning events.

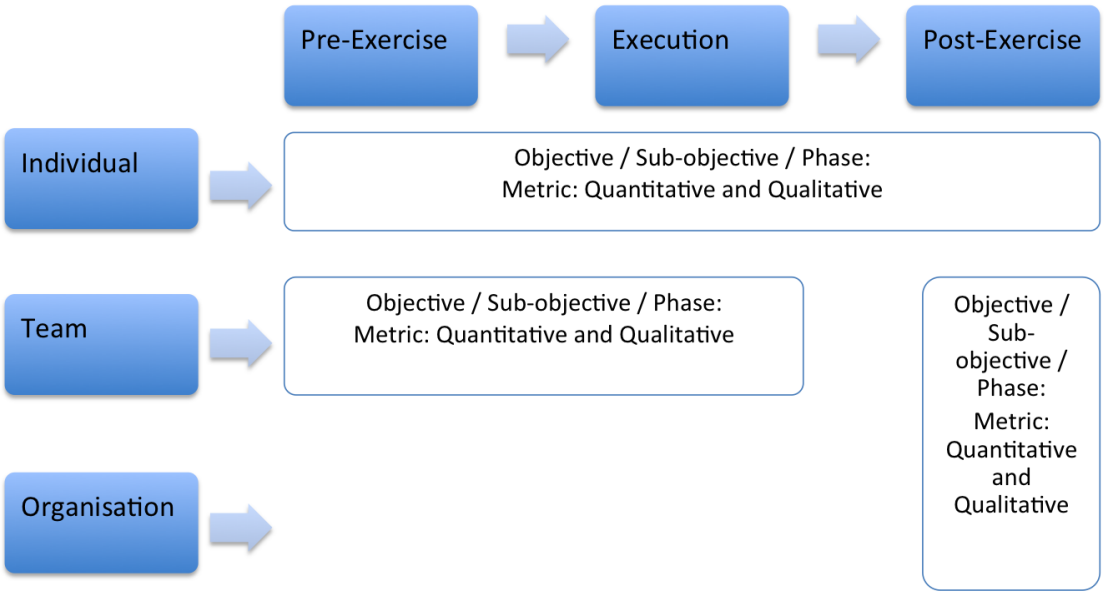

Figure 5: CDX's learning measurement dimensions

5 CDX's Learning Measurement Dimensions

The CDXs in the current form are not sufficiently instrumented for learning measurement and the existing measurements (focusing on scoring) are not using learning related metrics. The author recommends an overall CDX's learning measurement approach that brings together the different phases of exercise and individual/team/organisational aspects. Learning measurement depends on the specific training objectives, however all these dimensions, as depicted on Figure 5 should be considered.

As learning is complex and dynamic process, the measurement in CDXs should include mixture of quantitative and qualitative measurements (i.e. triangulation).

5.1 5-Timestamp Methodology

Learning in CDXs is affected by many variables, however the basic measurements, such as timing and accuracy metrics are still key elements that provide some comparable trends in learning process and benchmarking for the teams. For example, Henshel *et al.* measurements in Cyber Shield 2015 showed that when teams took 20 or more minutes to identify an inject's NIST categorisation, they were more accurate [39].

Such metrics support appropriate exercise learning design. For example, in a very complex high-risk cyber conflict scenario an expectation (game rule) for teams

to respond in 15 minutes to a user's[1] complaint about an unavailable website may prove to be unrealistic and will not contribute to learning, but instead contradict learning objective to prioritize incidents. It rather "forces" teams to learn, share and store wrong behaviours and later retrieve learned, but wrong behavioural models in real life situations.

Furthermore, measuring learning effectiveness and collecting data simultaneously for providing effective feedback can be combined. The learning potential is not fully realised, if the BTs do not know what are their weaknesses (these are key takeaways/action steps to improve in the job) and how much they progressed (was there any value in attending) in the exercise. Scoring might give some indication of how teams compare, but not knowing what is the "baseline" or standard in more detail, the overall score is worthless from learning point of view. For example, a successful RT attack is scored and a BT looses points. But RT score is same for every successful attack and not taking into account how much resistance individual BTs demonstrated and how efficient they were in responding.

As a solution, the author proposes a non-obtrusive methodology to analyse timestamps and time intervals on attack vector/incident/target machine basis. Method focuses on timings and provides metrics for different learning objectives (Figure 6). The idea came from the fact that one of the learning objectives in LS is incident handling. Whether teams are effective and meet that learning objective, needs a basic timing and accuracy metrics—how long does it take to respond to an incident, how long did team take time responding to a significant threat vs. minor defacement issues, what is correlation between teams' detection time and quality of compromise reported to YT, and so on. This can be analysed as change or trend in time.

The 5-timestamps are evidence bringing together RT and BT actions from their digital footprint. The method enables to measure technical skills, but also soft skills (including leadership, team communications, decision making).

The analysis breaks an cyber event into phases to demonstrate in which phase strengths and bottlenecks of an individual and team skills are, and provide the basis for effective feedback. The model follows timeline, and information can be collected non-obtrusively (Table 1) from game-net but also from management network, i.e. from Virtual Local Area Network to that connect participants/teams to the game-net. Even when t_1 and t_2 are intrusive for RT, data collection is not intrusive for BTs. For validation with BTs, a sample using intrusive methods can be selected (Section 5.2).

While data sources, such as RT activity timestamps and scoring are typically collected and analysed in the exercise, then only non-traditional data to further

[1]System user in the gamenet are often "simulated" by humans that are part of the organisation team

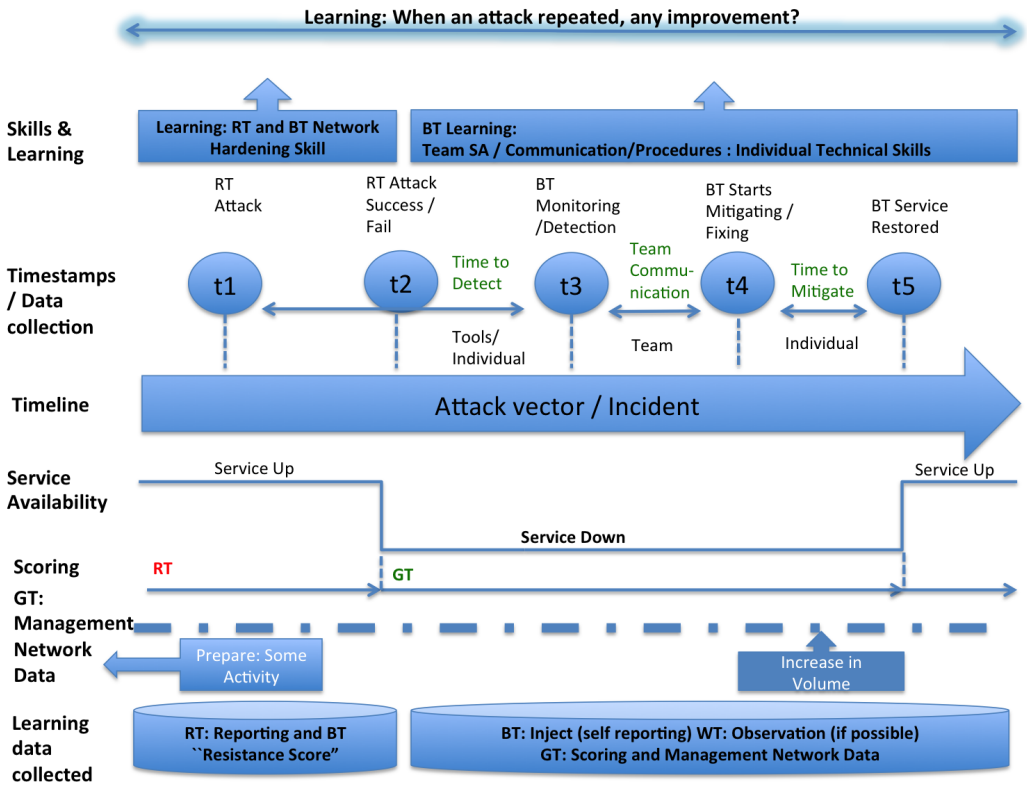

Figure 6: 5-timestamp non-intrusive methodology

analyse is management network—to determine whether a team has accessed the machine and when. Seeing a prolonged higher number of packets per second and shorter packet inter-arrival times exchanged between two machines (source from IP-range where BT is physically located, destination the BT machine under investigation), means someone in the team is working on it.

The idea relies on the fact that organisers are able to collect all raw network traffic (e.g. pcaps) not only from within the game-net, but also from the management network. From those traces it is possible to automatically detect the times of BT activity for each tar- get machine (e.g. when a BT member is working on a machine or not). This can be done by observing a ssh or remote desktop connection from the BT-network through management network. Even if the traffic is encrypted, and the BT member remains logged-in in the background, simply observing the traffic volume and packet inter-arrival times allows automatically detecting times at which someone is working on a specific game-net target. With traditional methods, this can also be achieved by asking the team member to keep a detailed log about timestamps.

The time interval between timestamps provides learning metrics as shown in Table 2.

For example, interval t_5–t_2 is a scored metric. However, note that $t_3 < t_2$, i.e. RT might be detected in the systems before they achieve scored objective. Thus

Time-stamp	Description	Non-intrusive data	Intrusive (for validation only)
t_1	RT starts attacking	RT activity reporting	Not applicable
t_2	RT compromises	RT activity reporting and scoring data	Not applicable
t_3	BT detects	management network, possibly by access patterns	BT observation or self-reporting via inject
t_4	BT mitigates	management network (showing traffic activity)	BT observation or self-reporting via inject
t_5	BT restores	scoring, management network (end of activity)	Not applicable

Table 1: Data sources for 5-timestamps

scoring provides limited insight and further granularity is needed.

The time intervals provide scalable input for different technical and soft skills learning objectives, and also measure team vs. individual effectiveness:

1. Team vs. individual learning—for example, data from management network would also give non-intrusively metrics for individual vs. team aspects, i.e. several people connect to the same machine to work on the problem (intrusively can be validated with inject).

2. Soft skills (leadership and decision making)—teams must make quick decisions (likely to have immediate and significant consequences), thus teams learn also decision-making. OODA (Observation, Orientation, Decision, and Action) loop is a theory of decision-making where time is the dominant parameter [70] and thus supports this framework using time intervals. Teams need to deliver highly reliable performance and adapt their responses to mitigate adverse scenarios, and that can be measured by t_4–t_3, i.e. time between detection to start mitigation (team communication, prioritisation, task allocation, etc.).

In larger and more complex CDXs, there are typically several target machines in a game-network that can be attacked repeatedly using the same attack method. One of the advantages of live-fire Red/Blue team exercise is also defending against a "thinking adversary" that implies that the same target can be attacked using different methods, i.e. repeated.

Using proposed 5-timestamp method, provides several advantages. Firstly, it helps to create a general mental map of the events. For example, in feedback debrief simply providing game-net logfiles and pcaps, and letting the participants search for events that happened is not useful for learning. Similarly as a security

Timestamps Interval	Description	Learning Objectives	Team vs. Individual
t_5–t_2	incident response time	Overall performance (organiser's objectives=scoring)	team
t_5–t_4	time to mitigate	Responding to attacks (technical skills)	individual, sub-team
t_4–t_3	time between mitigation and detection	Time management and prioritization Teamwork: delegation, dividing and assigning roles, leadership Handling cyber incident	team
t_3–t_2 (or also t_3–t_1)	time between compromise and detection	Monitoring networks, detecting of attacks	individual, sub-team
t_2–t_1	time to compromise	Learning the network, System administration and prevention of attacks	individual, sub-team

Table 2: Learning metrics from 5-timestamps and their intervals

camera becomes more effective when combined with a motion sensor—the logfiles become more easily "searchable" when combined with accurate timestamp annotations. Thus debriefing an attack from the high-level objectives together with accurate timestamps, considering also that the participants have already been in the situation during exercise, they understand RT's objectives, and are able to "relive" the events. Useful feedback can only be given, if the exercise can be debriefed in a meaningful way, and accurate timestamps are a first critical step towards achieving this.

Secondly, the timestamps can be used in building a baseline for performance or effectiveness. When grouped by attack method (not target system), those values become comparable. These can be further analysed in several ways: 1) as an average overall performance against defending against this type of attack, 2) viewed over time for the same target machine (for example by looking at repeated attacks using the same attack method) whether anything has been learned during the exercise—or potentially, even between exercises, if similar team composition returns to an event in which the same attack vector is repeated, and 3) for understanding whether the BTs are able to transfer learned knowledge (for example, is the BT able to detect and defend the same type of attack against a different target system provided they have learned it earlier).

Thirdly, analysing the timestamps provides insight into the BT's strategies. Do BTs only focus on certain class or difficulty-level of attacks, and maybe miss some

more important/unknown challenges? Do they invest time during the exercise to understand the system? Such simple metrics allow for a way of getting some basic baseline and benchmarking for the organisers and participants, and identify learning needs that need to addressed in future exercise.

It is important to note that the timestamps themselves only measure effectiveness. However, there is an implicit assumption that measuring changes in effectiveness over time (e.g. repeated comparable events, such as repeated attacks), will allow drawing conclusions about changes of performance over time, which is an indicator for learning (dynamic approach) together with other qualitative data.

Future Work The complete exercise data analysis and projections for full dataset is left for future work, and the scope of this thesis is to demonstrate the suitability of proposed methodology (Section 6). Learning outcome can be measured as changes in performance and/or effectiveness within one exercise or across several exercises. If such data is available over a long range of exercises, then changes in performance and learning effectiveness can be further analysed on the level of detail, e.g. by different types of attack categories (Table 3), for repeated attacks, percentage of not-restored services (showing potential learning need for the failed ones), etc. Based on the team data, learning maps could be drawn for each team using different data categorisation (e.g. network machine, segment, etc.).

5.2 Data Collection and Sources

The data collected as part of the CDXs to enable the learning measurement may vary based on the training objectives and software environment, but it should not be an "additional burden" to the organisers. And as shown by 5-timestamp methodology, the data is collected as part of CDXs execution (to enable scoring, etc.) anyway. As learning is a process, then measurement should above all reflect improvement and change and having performance metrics across timeline will serve to evidence that change.

Quantitative measurement can be performed by analysing non-intrusive learning metrics (and change/improvement in such metrics) in digital data (e.g. pcaps and traffic) depending on learning objectives to be measured. Data is obtained from several data sources from non-intrusive digital logs (quantitative):

1. RT reporting—failed attack, time to resist the attack, number of repeated attacks;

2. YT reporting—reporting about situational awareness (correlation of BT reports with RT campaign reports), stress level, Incidents of Compromise (IoC) statistics (usable vs. non-usable);

3. Scores—scoring for availability, usability scoring, injects (trends over time);

4. Communication channels—chat logs, GT management network traffic (volumes and trends);

5. Pcaps from game-net and management network.

and by qualitative obtrusive methods:

1. Surveys (pre-exercise and post-exercise survey with pre- and post knowledge assessment if possible);

2. Injects—validating quantitative data sample basis and collecting learning feedback during the exercise;

3. Interviews with participants (and management)—assessing the immediate reaction to exercise and long term impact on the job;

4. Information from RT—ratings for resistance level, classification of attack type that can be done semi-automatically, e.g. by using Cobalt Strike [25];

5. Observations of BTs.

Sample Selection for Qualitative Validation It may not be feasible to confirm all incidents qualitatively as it distracts from learning. However, for a sample of attacks qualitative feedback can obtained from the participants in order to validate the metrics. These should be designed into the exercise as enquiries to the BTs via (feedback and self-reporting) injects and observations.

The sample selection depends on the exercise training objectives, however should cover differing aspects, such as complexity, method of attacking, ease of detecting and mitigating the attack.

The existing taxonomies were explored as starting point for what type of attacks to select for learning impact measurement. There is no universal, internationally recognized taxonomy on cyber attacks or incident handling, however several specific taxonomies have been developed by individual CERTs, organizations and academics [71]. Several taxonomies are currently available to classify cyber-security incidents—some focus on the nature of an attack, while others describe the defensive posture of the victim [72]. Common classifications for cyber attacks are done by vulnerability, by lists of terms, by application, and by multiple dimensions [73]. Few examples of such taxonomies to list are eCSIRT.net security incidents taxonomy [71] or case classification (example) by FIRST [74], included in the incidence response guides such as by CREST [75], or developed by cyber labs such as Sandia [76].

Easy to Detect—Easy to Mitigate	Easy to Detect—Difficult to Mitigate
Difficult to Detect—Easy to Mitigate	Difficult to Detect—Difficult to Mitigate

Table 3: Sample selection matrix

There is no widely accepted taxonomy that can be applied from learning perspective in the CDXs context. In order to measure learning impact, a comparison between easy tasks (potentially nothing learned and knowledge is already existing) and complex tasks (more challenging, more potential to learn) is valuable. As teams have differing skillset any such criteria classification is somewhat "forced" and arbitrary, however it would provide comparison and feedback on the appropriateness of level of difficulty and learning opportunities created by the organiser. The proposed classification is based on following criteria: 1) detection and analysis—some attacks and incidents have "visible" signs that can be easily detected, whereas others are almost impossible to detect. This would include priority level determination (and escalation) and impact analysis. 2) mitigation and recovery—responding to an incident involves different skillset and actions to be taken to contain the damage, to eradicate the incident components, and to restore systems to normal operation, and remediate vulnerabilities to prevent similar future attacks. Table 3 can be used for selecting specific events for qualitative learning impact measurement (and also analysed by Bloom's Taxonomy levels).

As an incident is part of whole exercise scenario and teams need to prioritise events, the assigned priority by the team should also be considered (and match the organisers view via scoring).

5.3 Challenges and Potential Limitations

It is important to acknowledge the limitations of the learning measurement results. Data monitoring and collection may fail to capture timing metrics and team actions with perfect reliability, which prevents drawing conclusions with absolute certainty.

Aggregation of data from the multiple reporting tools and data sources can provide reasonably reliable timing and accuracy metrics, a major challenge is to develop clearly defined measures that integrate both qualitative and quantitative inputs.

Some training goals (such as incident handling procedures) may prove difficult to measure due to teams following different operating procedures, standards, and practices. Metrics for future evaluation should include appropriateness and quality of responses and actions.

Limitation specifically with qualitative data is separating learning impact from other behaviour effects (i.e. learning might be not visible straight away or recog-

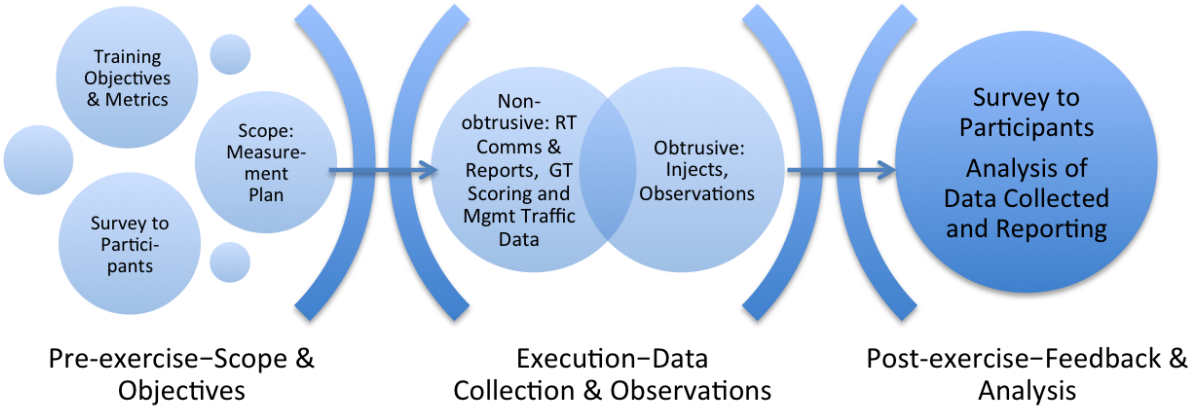

Figure 7: Process for learning measurement at CDXs

nised by participants by themselves, or overestimated and not result in relatively permanent behaviour change).

Finally, exercise design should ensure that intrusive data collection for measurement purposes would not create distraction or out-balance effective participation.

5.4 Summary

The learning measurement process needs to be pre-planned, agreed with the stakeholders, and form an integral part of CDX organisation process. The main decision and action points are depicted on Figure 7.

Selection of what to measure is a challenging task and depends on training objectives. What learning metrics are "must have", "nice to have" and "wasteful" metrics from learning perspective? To have comparative metrics from several CDXs, would enable developing comparable standardized set of learning metrics. The author demonstrated that the data collected as part of the exercise can also reveal learning outcomes. A full learning measurement framework with comparable and standardized metrics is area of future research.

This section introduced a novel 5-timestamp methodology to collect learning metrics non-intrusively (focusing on timing and accuracy) for both individual and team skills measurement, and as benchmark data for feedback in CDXs setting. Also a practical guidance to the organisers with supported research considerations—where to start, what data to gather and validate was presented.

6 Locked Shields—Learning Measurement

LS provides full experience of managing a cyber incident to the BTs. By itself it consists of different attacks and tasks based on a scenario over two days. As the data set is 2500 attacks [10], the measurement plan has to be set in advance what and how to measure, and which quantitative non-obtrusive metrics are validated qualitatively to ensure the balance between obtrusive data collection and participants focus on the learning.

6.1 Measurement Scope

The LS training objectives for IT specialists are included in Table 4. The exercise also includes specialized parts, such as conducting forensic investigation, crisis communication (media play), cyber legal aspects (legal play) [5] and strategic game (from 2017). Due to the complexity and wide scope of the exercise, some of these objectives are scoped out in this thesis due to focus on quantitative measurement methodology using digital footprint.

Technical–IT skills / tools	**Technical–Processes**	**Teamwork and Communication**	**Specialised skills**
Learning the network	Handling cyber incident Time management and prioritization	Teamwork: delegation, dividing and assigning roles, leadership	Crisis communication (scoped out, covered by Communication)
System administration and prevention of attacks	Ability to convey the big picture	Cooperation and information sharing	Conducting forensic investigation (scoped out)
Monitoring networks, detecting and responding to attacks			Legal advisory (scoped out)

Table 4: LS training objectives [5]

The learning measurement is performed for LS17; results of LS16 pre-analysis have been incorporated for comparison where possible.

6.2 Research Questions

This measurement aims to provide answers to following research questions:

1. What and how participants/teams learn in the pre-exercise phase?
2. What common metrics collected as part of the exercise demonstrate learning outcome (have plausible relationships)?
3. What is teams' feedback on exercise learning design and experience, such as complexity/challenge (specific focus on Industrial Control Systems segment, individual vs. team learning)?
4. What are practical challenges in applying non-obtrusive 5-timestamp method proposed in Section 5.1?
5. What design improvements facilitate learning experience of future exercises?

6.3 Methodology

The learning measurement includes mixture of quantitative and qualitative methods with focus on application of the 5-timestamp methodology in practice, combined analysis of participants' feedback and metrics collected in the exercise, and identification of plausible relationships for learning metrics.

6.3.1 Qualitative and Quantitative Learning Measurement

The following approach is applied:

1. Quantitative pre-survey (Appendix A)—the objective is to collect information about the participants, their experiences and learning process in the pre-execution phase, team environment, learning expectations about the execution and evidence of long-term learning from previous exercise participation. The survey was sent out five days before exercise to all participants and requested to be filled individually. The survey consists of multiple choice or ranking style questions with a free-form "additional comments" option;
2. Data analysis—the objective is to understand a) what learning design elements impact learning outcomes, b) whether the level of challenge and complexity are designed at the appropriate level c) evidence of predictors/metrics of learned behaviour for individuals and teams. The data sources vary and overlap; detailed description is provided in Section 6.4;
3. Feedback injects—the objective is to collect feedback and validation information for specific learning objectives relating to complex systems. The inject (Appendix C) close to the end of exercise includes open question about the learning experience that is analysed using qualitative data analysis method with tool HyperResearch 3.7.3 [77] to enhance quality of research conclusions;

4. Quantitative post-survey (Appendix B)—the objective is to obtain feedback from BTs perception of impact to their learning outcome overall. The survey consists of ranking style questions with free-form "additional comments" option.

5. Interviews and enquiries with previous LS participants—the objective is to obtain feedback in regards of long term impact. The interviews are semi-structured and conducted orally or via e-mail correspondence.

6.3.2 The 5-Timestamp Methodology

In order to apply the 5-timestamp method as described in Section 5.1 in practice, the following methodology is applied:

1. Data collection—1) RT timestamps are collected from RT activity report. Timing accuracy is directly confirmed with RT members conducting the attacks, any anomalies discussed; 2) BT timestamps and time intervals are collected via injects. BTs are asked to report timestamp in hours and minutes (UTC) for first notice of compromise, length of team communication and task allocation phase and how long the team resisted attack or gave up; 3) scoring data—the timestamps and intervals of service being available or not available;

2. Data analysis—1) data is cleaned and initial analysis performed for anomalies detection, plausible relations, etc., 2) reconstruction of incident timeline and visualisation, and 3) analysis of the dataset and relations identified.

For validation purposes one complex attack vector for specific network segment all 19 teams was selected—Siemens Industrial Control Systems[2]. As teams have different level of skills, using a complex system as test sample and proof for learning is more equal playing ground as it is "unknown" to everyone. When teams are faced with an unknown system (rather than just dealing with known routine task) already preparation is expected to result in learning; also ICS is one of LS17 learning focus areas.

6.4 Data Sources and Relations Analysed

Data is obtained from several data sources from digital logs (quantitative, mainly non-intrusive) collected during the exercise:

1. RT reporting—attack objectives and reports;

[2] *step7.ics.bluexx.ex*, *plc.ics.bluexx.ex*, *hmi.ics.bluexx.ex* and ICS network segment traffic.

2. YT reporting—Reports (SITREPs and Threat Assessments (TAs)), and IoC statistics;

3. Scores—scoring for injects, attacks, availability, usability and special scoring;

4. Other information channels—chat logs, GT management network traffic volumes, reverts.

and by qualitative and quantitative obtrusive methods:

1. Surveys (pre-exercise and post-exercise survey);

2. Injects;

3. Feedback from RT members;

4. Interviews with organizers and previous LS participants.

6.5 Findings, Discussion and Analysis

6.5.1 Pre-exercise Preparation

The pre-exercise phase analysis is based on 117 individual participants' responses to a pre-survey, Appendix A. The total BTs training audience is approximately 570 participants (19 team with average 30 team members each). The participants' previous experience with cyber exercises is dominant—38% of participants have not attended LS before and 11% have not attended any other exercise. Most common other exercises attended are Cyber Coalition (also organised by NATO) and individual state exercises.

Wilson *et al.* [29] study showed that collaborative teams probability of group learning is increased. The participants describe their teams as hierarchical (with specific roles) 42.7%, collaborative style 37%, and 19.7% military/authoritarian.

The teams have most technical areas covered with skilled persons (at least 80% in each category of media, routing, forensics, legal, system admins, reporting and monitoring), but for ICS and drones only 48%. LS17 focus is on ICSs, but 52% of participants report that their teams have lack of skilled personnel. Few comments from survey to support: "nobody is truly experienced in ICS/SCADA", "ICS is our weakspot", "not so sure about ICS and drones", etc. The appropriate exercise design for upskilling or training people who lack assumed technical knowledge is crucial.

In majority, 53% of respondents, spent 10-50 hours preparing for the exercise followed by who prepared either less (24%: 3-10; 5% 1-2 hours) or more (17.1%; 50-100: 5%: over 100 hours). Individual preparation is substantial—half of the time or more often (73%). Sub-teams preparations are taking place either half

Description	None	Limited	Medium	High	Expert
Knowledge/skill level (in the scale of 5)	1%	12%	43%	37%	8%
Working experience	4%	13%	39%	39%	4%

Table 5: LS17—Participants' self-assessment about their skills and knowledge

of the time (35%) or seldom 31%. This shows relatively balanced combination between individual and sub-team preparations. Whole team preparations were mostly seldom 37% or half of the time 26%, however 22% of participants claim they never attended whole team preparation session–considering that team collaboration is a key element for successful group learning, this is somewhat worrying indicator.

Table 5 summarises the participants' self-assessment about their skills and knowledge, and Table 6 presents data about increase of skills/knowledge obtained in pre-exercise phase.

What new skills/knowledge have you learned in the preparation process?	Significantly increased	Minor improvement	No change	N/A
Learning the network	15%	43%	36%	6%
System administration and prevention of attacks	17%	48%	25%	10%
Monitoring networks, detecting and responding to attacks	13%	47%	30%	10%
Handling cyber incidents	11%	42%	35%	12%
Conduction forensic investigation	9%	17%	51%	24%
Teamwork: delegation, dividing and assigning roles, assignment	17%	48%	26%	9%
Cooperation and information sharing	22%	49%	26%	3%
Ability to convey big picture	19%	34%	42%	5%
Reporting	9%	43%	40%	8%
Crisis communication	8%	35%	41%	16%
Time management and prioritization	15%	46%	34%	5%
Cyber legal aspects	4%	23%	48%	25%

Table 6: LS17—Participants' self-assessment about new knowledge/skills obtained in pre-exercise phase

The training audience assesses their knowledge and skills in majority at medium

(43%) to high level (37%) and working experience has similar levels (both medium and high 39%).

No clear distinction between learning knowledge/skill on either technical or soft skills in pre-exercise phase is visible. The high level of "no change" responses is possibly related to technical specialisation (only certain learning objective are relevant for specific team member). On average minor improvement in each learning areas is 40% and significant improvement is felt by 13%.

As LS is technically complex, the concern is whether participants feel they are ready or confused before the exercise—65.8% report ready and 34.2% feel confused (very strong correlation to having previous LS experience or not). The reasons described for confusion include: "The preparation was hampered by routine work and additional tasks.", "It is hard to break from wanting to harden the systems and just focus on monitoring and reporting. There is so much going on it is hard to concentrate on one lane", "It's pretty hectic...", "Without prior experience or training in the relevant areas, this seems extremely intense.", "Though some of the items seen are very misleading and look like malicious activity.", etc. As the exercise is getting more complex (scenario, network map) every year, then compared LS16 higher percentage of responses reported "confused" (28%).

The survey results support the hypothesis that in order to improve overall learning experience, preparation phase is critical and if the organisers want to increase learning impact, strengthening learning design (and reducing confusion levels) for pre-exercise phase is vital (supported by LS16 pre-survey results where 87% participants assessed that the preparation process for the exercise has been a great learning experience).

The expectation for learning during the exercise is strongly positive (56%) as shown in Table 7. The expectation varies from: "The situation is too stressful for me to really learn new things. I tend to revert back to what I know since there isn't enough time to try new things" to "Already know of new tools and techniques from other team members. Expect to learn more from other blue teams in exercise."

When comparing what the participants have learned in pre-exercise phase and what they expect to learn during the exercise in Figure 8, then there is no clear distinction between training objectives, such as teamwork training objective is more relevant for execution than pre-exercise phase. Top technical skills participants expect to learn are monitoring networks, detecting and responding to attacks (81%) and administration and prevention of attacks (74%) learning objectives, and top soft skills are cooperation and information sharing (76%) and teamwork (74%).

6.5.2 Exercise Execution—Learning Metrics in Existing Measurements

Even though scoring does not equal learning outcome, these reflect performance and looking over time to how the teams perform in the exercise gives indication of

I expect during the exercise:	Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree
Learn nothing new overall	56%	20%	15%	4%	3%
Practice skills I already had or obtained in during preparation	0%	3%	10%	43%	44%
Learn new knowledge and skills	1%	1%	9%	35%	54%

Table 7: LS17—Learning expectations during the exercise

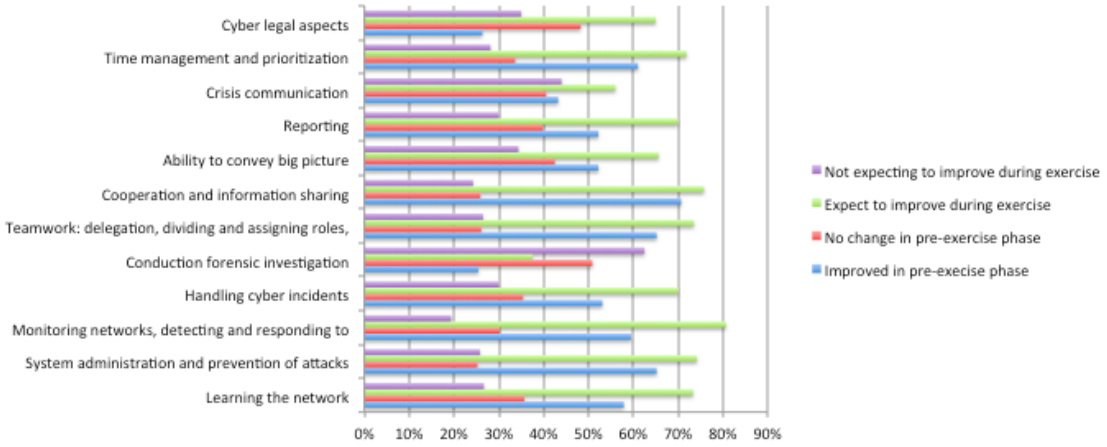

Figure 8: LS17—Comparison of what participants learned in pre-exercise phase and are expecting to learn in exercise phase

change, or learning curve. Certain dimensions of scoring indicate possible weaknesses, i.e. learning needs in specific areas.

Due to confidentiality, the overall graph is not presented for the team scoring timeline or final scoring status, however when analysing the teams' performance over time following overall trends are visible, such as:

1. BT scores against RT attacks timeline—when a team has delay in RT attacks scores at the start, this results in better defence (lower RT scores) throughout the exercise. From learning aspects that shows learning and hardening network efforts done by teams before exercise (and should be correlated).

2. BT scores in RT attacks categories (assuming more complex attacks are scored higher)—highly performing teams resist both (and have knowledge and skills) more complex attack, and easy (low scored attacks, i.e. known

vulnerabilities, etc.). That shows network hardening skills, but also incident handling effectiveness (i.e. correct prioritisation of threats, etc.)

3. Responding to Injects—not only the score is vital, but also delays in acknowledging and responding indicate whether teams are "in control".

4. Reporting scores (Sitreps and Threat Assessments) over time—the teams that achieve better result overall, have trend in increase in scores over time, the teams who achieve lower overall results, seem to have degrading trends. From learning perspective, it can indicate that they have given up or do not improved their skills during the exercise.

5. Stress reporting—it is not a precise measurement but correlation shows slightly higher self-reporting of "in control" (however not conclusive) for the teams who achieved higher competition ranking. From learning perspective provides insight regarding perceived learning environment (calm vs. panic) and links to an assumption that learning can happen when team acknowledges they lack some knowledge or skills and learner/team overly rely on tools and not "sensing more than see".

The existing metrics collected as part of exercise do show some learning trends, but do not enable learning benchmarking or learning curve analysis for the BTs.

6.5.3 Exercise Execution—Feedback from Injects

Learning Outcomes Specific for Complex Systems, i.e. ICS Injects were used to collect information for learning assessment, some specifically for ICS system (focus of LS17 [10]), but also obtain feedback on other parts of the exercise design indirectly (via comparison). The attempt was to collect feedback how teams perceive individual team members/sub teams and whole team learning outcome. Based on pre-survey 52% of respondents felt they do not have ICS capabilities in the team. Information collected via injects shows that teams have dedicated ICS personnel—58% of teams reported they have one person and 33% have 2-3 team members assigned.

Table 8 presents a sample of overall data received. Only half of the teams are presented and data is anonymised. The data analysis below needs to be read with caution as some teams who were compromised did not respond to some questions in inject(s).

Average self-believed resistance level was surprisingly low compared to RT members' assessment—44% believed that their resistance was at medium level, 33% at high level 22% strong. This links positively to an assumption that learning can happen when team acknowledges they lack some knowledge or skills and

Anonymized Team Identification	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
Ranked (in 5 unit ranges out of 19)	6-10	11-15	1-5	11-15	1-5	6-10	1-5	6-10	6-10
Reported IN CONTROL (% of total reportings)	92%	50%	50%	71%	20%	100%	90%	100%	83%
Average level of activity felt (1: low 2: medium, 3: high)	2.08	2.33	2.33	1.97	2.40	1.84	2.60	3.00	1.72
BT self-assessment (no resistance to very strong resistance): (1 to 5, 1: not at all, 5: very strong)	5	3	3	4	3	3	4	4	5
RT assessment (scale of 5, 1=weakest, 5=strongest defence)	5	4	4	4	4	5	5	4	5
RT attack successful?	No	No	No	Partial success	No	No	No	Partial success	No
Rated complexity of ICS attacks compared to other attacks in the exercise (DMZ, LAB, etc.)	About the same level	About the same level	Somewhat more complex	Somewhat more complex	About the same level	Significantly more complex	Somewhat less complex	About the same level	Moderately more complex
Rated complexity of ICS attacks compared to other attacks in special systems (Spectrum5, drones,...)	Somewhat less complex	Somewhat more complex	About the same level	Somewhat more complex	Somewhat less complex	Significantly more complex	Somewhat less complex	About the same level	Moderately more complex
Solved the task:	One person	One person	One person	2-3 persons	One person	One person	One person	2-3 persons	One person
The ICS incidents were:	Easy to detect–easy to mitigate	Easy to detect–easy to mitigate	Difficult to detect–easy to mitigate	Easy to detect–difficult to mitigate	Easy to detect–difficult to mitigate	Easy to detect–easy to mitigate	No answer	Easy to detect–difficult to mitigate	Easy to detect–easy to mitigate
Identified a root cause of the ICS activities (i.e. forged packets, CnC, backdoor,...)? If so, what did you discover?	No	Yes	No answer	Yes	Yes	No	Yes	Yes	Yes
Priority level assigned to the ICS events (Critical, High, Medium, Low, Very low, N/A)	High	Critical	Critical	Critical	High	High	Critical	Critical	High
Learned as individual team member(s) / sub-team working on the ICS systems	Significantly	Slightly	Significantly	Very	Moderately	Significantly	Moderately	Very	Significantly
Learned as whole team	Moderately	Moderately	Slightly	Significant	Slightly	Slightly	Very	Very	Very

Table 8: LS17—Sample feedback on ICS segment

"sensing more than see" (OODA loop). It is also interesting to see how the teams perceive the level of difficulty for this segment: 41% easy to detect–easy to mitigate, 39% easy to detect–difficult to mitigate, 12% difficult to detect–easy to mitigate and 8% difficult to detect–difficult to mitigate. This is going against the "complaints" about the complexity of ICS systems. Priority for ICS attacks was consistently (78% of teams) at critical or higher priority level, as expected by the exercise scenario. Only later changes reported by one team that downgraded from Critical to High, and another team that initially assess Low, changed the priority to High. 52% of the BT reported that they managed to track/hunt the root cause of the activities (i.e. forged packets, CnC, backdoor,...) and 42% not (showing missed learning opportunity without proper feedback).

Table 9 shows self-assessment by the teams how much individual (or sub-team)

and whole team learned. When team learning has quite even distribution from slight to significant improvement, then for individual learning was in majority (59%) assessed as significant. The comments include: "the awareness for the importance of SCADA Systems grew.", "...I learned a considerable amount regarding how reacting to a cyber crisis in an unknown environment is completely different from running a team in daily operations.", "The whole team learned slightly since after the loss of the power systems some members of other subteams tried to support the ICS team.", "For most of us it was first time we dealt with ICS security or even ICS at all. We were preparing for some time before the exercise.", "As individual that was a huge opportunity to get to know how this type of Siemens System works. As a team we have learned a lot. The manual was a huge help.".

How much did you learn?	**Significantly**	**Very much**	**Moderately**	**Slightly**	**Not at all**
- individual team member(s) / sub-team working on the ICS systems?	50%	19%	25%	6%	0%
- as whole team?	24%	24%	24%	29%	0%

Table 9: LS17—Learning outcome self-assessment for ICS segment

How do you rate the difficulty/complexity of ICS attacks, compared to:	**Significantly more complex**	**Somewhat more complex**	**Same level**	**Somewhat less complex**	**Significantly less complex**
- other attacks in the exercise (DMZ, LAB, etc.)?	6%	33%	44%	11%	6%
- other attacks in special systems (Spectrum5, drones)?	17%	28%	33%	22%	0%

Table 10: LS17—Complexity assessment in comparison to other network segments

Table 10 shows that 44% of teams assessed level of difficulty as same level with other attacks in the exercise. As there are more special systems (drones, Spectrum5) then 33% teams assessed difficulty at same level and 28% and 22% somewhat less complex. This indicates that attack vectors on other network segments are considered relatively at same complexity level. In combination with feedback of classifying ICS attacks as easy to detect–easy to mitigate by 41% of the BTs, indicates that the teams feel quite confident and possibly not "sensing more than see" (OODA loop). This could result in a missed learning opportunity or exercise design flaw, that would need further investigation.

Overall Learning Outcomes The final inject included learning question to collect narratives about the teams' learning experience at the exercise to uncover and understand the big picture. Qualitative analysis involves labelling and coding the data to recognise similarities and differences can be recognised. HyperResearch 3.7.3 tool [77] was used for assistance (code, i.e. key words/expressions counting) in content analysis. The groupings and key words emerged from feedback analysis are shown in Table 11.

Key words / expressions used	**Number of counts**
Successes in learning (learning curve)	19
Challenges in learning	12
Complexity / Variety of systems	12
Preparations	11
Team learning	8
Gameplay / Competition	6
Feedback	3
Specialised knowledge	3

Table 11: LS17—Key words/expressions from open question on learning experience

The learning experience was described with high regularity using words such as learning experience, useful, learning curve. Learning was also mentioned in team context. Challenges or failures of learning often related to too many rules (or too few), extreme complexity and competition/game elements (i.e. scoring). For example, as reverting a machine is penalised by points then instead of accepting a mistake, instead of trying to learn participants struggled to restore services. Complexity and wide variety of systems is seen very differently, many team appreciate it and others see it negatively. Pre-exercise phase and preparations are seen as key to success or failure in the exercise.

6.5.4 Post-exercise—Long-Term Learning Outcomes

Long-term learning effects are difficult to measure, as learning outcome is affected by other outcomes. In 2016 and 2017, the author proposed to include specific learning measurement questions (Appendix B). However the rate of responses overall was extremely low (only 21%, i.e. 4 out of 19 teams).

LS16 data shows that out of this 21%, 100% (strongly) agreed they learned something that change how they do their job or help them do their job. The few examples given were quite general, such as "more automation and better integration with other roles" and "the needs of team leader during the operation. How you can prepare for unknown situation." Estimates how much teams job performance

has improved as a result of the exercise, were quite optimistic, 75% reported 1-20% increase(25%: 51-80%). All teams responded that they (strongly) agree that participating in this exercise was useful and the training objectives were met, and overall they are satisfied with the exercise.

Similar questions were asked in LS17 post-survey, however due to timing constraints of this thesis the results are not yet available.

However, as the exercise design in major aspects is similar to previous year, the attempt was made to specify if and what skills do participants recall from LS16. 51% of LS17 participants have attended the exercise before and responded in pre-survey (Appendix A) about skills learnt and maintained, see results in Figure 9. Sadly the survey results were limited in the comments about what exactly that participants feel they have learned. Analysing based on type of training objectives—69% responded that they recall a skill from participating earlier LS, average for technical training objective is 67% and soft skill related training objective 69%.

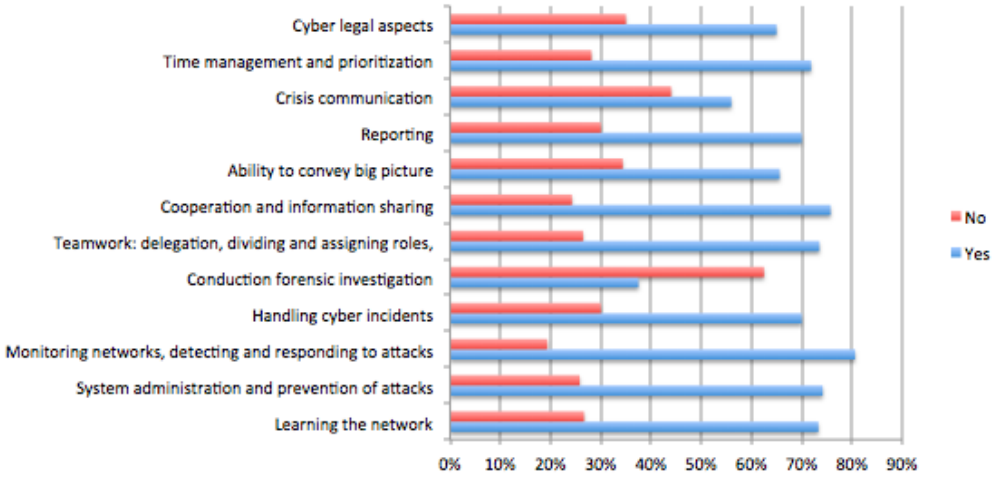

Figure 9: LS16—Long-term learning effect evaluation

The survey also attempted to determine whether teams have continuity from prior year in LS17. Majority 59% of the teams have changed significantly (less than 50% old team members, thus the conclusions regarding teams' long-term impact on teams, as summarised in Table 12, need to be interpreted carefully.

Feedback from some participants who participated over five years back supports long term learning impact of CDXs on mindset (e.g. "to have an emergency procedure in place, as when you're in the middle of the event there is no time to think, just to act.", "...key is thinking and mindset—and learning why something was done, not what."

Did attending LS16 change your team dynamics or team behaviour at your workplace? (select multiple if applicable)	**Strongly Disagree**	**Somewhat Disagree**	**Neutral**	**Somewhat Agree**	**Strongly Agree**
Our team has become more coherent, confident and collaborative	0%	6%	35%	29%	29%
Our team's knowledge has increased (as a result of individuals sharing)	3%	0%	32%	38%	26%
Only individual team member's knowledge increased, however it has not been shared (if team member would leave, knowledge in group is lost)	9%	21%	38%	26%	6%
There is no team dynamics or behaviour changes noted	14%	37%	34%	9%	6%
N/A, as the team was only formed for attending exercise and dissolved after exercise	19%	13%	35%	13%	19%

Table 12: LS16—Long-term impact on team and team learning

Despite of limited evidence, the survey and interview results support the learning value of the LS. Further work needs to be conducted to evaluate long-term impact further for specific training objectives.

6.5.5 5-Timestamp Methodology Experience

This section illustrates how the 5-timestamp methodology works at the example of LS17. This author picked one type of attack to a specific network segment. Due to the high-profile nature, an attack on a Siemens ICS system, was used as featured attack where all timestamps would be recorded. The timestamps were obtained from the BTs self-reporting (through Injects), RT attack reports, and scoring data for all 19 teams. Furthermore, those RT members conducting the attack on those ICS system, were asked to keep a detailed log of all events, as accurately as possible. Regarding pcaps from the management interfaces, there was a technical issue and very unfortunately, GT was unable to record the traffic from the management interfaces; thus, leaving any analysis of the inter-arrival

Incident Timeline	Time	Description	Data source
t_1 RT starts an attack	06:59	Campaign officially opened	RT reporting system
$t_{1.1}$ RT starts 1st attack attempt	07:35	Actual Attack started	RT members
t_2 RT compromises	07:40	Spilling started	Scoring
t_2 RT compromises	07:43	Spilling started	RT members
$t_{2.1}$ BT mitigates	07:44	Spilling stopped	Scoring
$t_{2.1}$ BT mitigates	07:45	Spilling stopped	RT members
$t_{1.1}$ RT starts 2nd attack attempt	07:58	Attack repeated	RT member
t_3 BT detects	09:00	Suspicious activity noted	BT Inject
$t_{2.1.3}$ RT reporting	09:18	RT objective partially scored	RT reporting system
t_4 BT starts mitigating	09:20	Timestamp or interval reported	BT inject
t_2 RT compromises	09:23	Spilling started	Scoring
$t_{2.1}$ BT mitigates	09:30	Spilling stopped	Scoring
t_5 BT fights back	09:30	Timestamp or interval reported	BT inject
t_5 BT resolves	09:40	Suspicious user removed	BT Inject

Table 13: LS17—5-timestamp example reconstructed timeline

times for future work in future exercises.

The attack objective was to control the airport fuelling station via Siemens Step7 attacks, i.e. start spilling fuel. BTs had time to mitigate before "all fuel was spilled". Before the exercise, RT had prepared some potential attack vectors, but which vector would work or not depends on BT defences. Starting the fuel spill is a very "noisy" attack, which means even if the initial compromise remained undetected, the BT had some time to mitigate the full-scale attack.

Four teams were successfully attacked by RT (i.e. all fuel was spilled). RT managed to compromise the systems and start spilling for two more BTs, but they managed to mitigate the attack before all fuel was spilled. The remaining 13 teams defended their systems well (e.g. no spilling started).

While all teams were analysed, for anonymity and clarity reasons only one timeline is presented here. Table 13 shows detailed timeline of events recorded according to the 5-timestamp methodology for BT Z (Z anonymised).

Before RT is allowed to attack in the exercise, the respective objective must be opened. For this specific team and this objective, this was done at 06:59 UTC,

which corresponds roughly to the time the first phase of attacks was allowed to start. The objectives are not opened individually in the RT reporting system, but rather for all teams at the same time. Therefore, a RT member might have to "entertain" several teams at a time, which means opening of objective and actual start of an attack might differ. In this example case, the BT was only attacked at 07:35 UTC (about $1\frac{1}{2}$h later), i.e. timestamp reported from the detailed RT member logs. LS has a comprehensive and automated scoring system, which recorded at 07:40 UTC that the attack has been successful and spilling started. However, RT members reported that spilling started at 07:43 UTC. This small time difference is an artefact of the self-reporting, and understandable, as all teams are very busy during the exercise. It also highlights that self-reporting timestamps should be avoided. This is not only for accuracy reasons, but also to reduce the work-load for various teams during the exercise. Similarly, the scoring system reported that the BT mitigated the attack at 07:44 UTC, while RT member recorded a timestamp of 07:45 UTC. Such minor discrepancies were observed throughout. As this attack was only partially successful, RT does not give up and manages to gain foothold in the systems again at 7:58 UTC (reported by RT member log), but this time RT does not manage to cause any fuel spilling. This is not recorded in the scoring (and should not be scored as the BT successfully defended against the attack), but it is an important factor that hints at resistance and team performance. Having such timestamps can also facilitate a reflective team debrief after the event.

However, when analysing the BT self-reporting, then the BT only reports detecting any suspicious activities for the very first time at 09:00 UTC. Clearly, some BT members must have mitigated the attack already before 07:44 UTC, so this points to an intra-team communication/reporting problem. Therefore asking the BTs to self-report accurate timestamps, while defending systems during a "crisis situation", is not going to work (neither observation). The team's internal reporting systems do not capture such information, or at least not accurately enough. It is therefore of vital importance to obtain such timestamps from the management network (e.g. by observing in pcaps when a BT member logs into the target system, or in case they are already logged in when the activity of system changes by a changed inter-arrival frequency of packets on the management network).

Overall during the exercise, spilling attempts start 7 more times using at least two different attack vectors. The first time spilling was for 3'39" (3 min and 39 sec), the second spilling continued for 6'44". The next day spilling durations were significantly reduced, in the end only taking 0'07" (7 seconds) to mitigate—despite the fact that different attack vectors were used.

To summarise, the main challenges encountered in the validation process and assumptions for data quality are:

1. RT scoring timestamps from the system need to be sufficiently accurate—

when attacking multiple teams the objectives are started for all teams simultaneously and final scoring is often delayed, thus RT scoring timestamps are not necessarily accurate;

2. BTs self-reporting is not reliable and accurate data collection method—this supports the argument that non-intrusive methods for collecting and analysing data from logs (pcaps, network, traffic, etc.) is helpful;

3. Traditional "observations" methods not possible as it's a technical exercise—"there is nothing to see".

Of course, this is a first attempt to understand the feasibility of proposed methodology. Before drawing any conclusion on learning more data and measurements needed to be obtained in future work, however, such initial tests appear to be a promising metric.

6.6 Improving Learning Experience and Effectiveness

Learning Objectives Learning objectives need to be SMART (Specific, Measurable, Agreed, Realistic, Time-bound) [78], in order to provide direction to scenario, learning design, scoring, etc. For example, a learning objective could be written in the format as "Participants are able to use [modern/state of art/latest] tools for identifying [specific issue/risk/problem used in learning design]". Adding Bloom's Taxonomy level helps challenge and reflect the design is appropriate level, as pointed out by Moses et [47] for another CDX, design might lack challenges mapping to level 5–6 (evaluate–create).

The measurements efforts in thesis have found overall evidence that learning is significant, but due to very general objectives, the detailed measurement for all specific training objectives is not in scope for this thesis. The attempt was took look into one specific area, i.e. ICS systems—but that due to lack of basic timing and accuracy baseline metrics it still remains "anecdotal", i.e. feedback and stories from the participants. Therefore collecting benchmark data, i.e. such as suggested by 5-timestamp methodology, would give for example expected time to detect for different categories of the attacks, and provide the "benchmarking".

Use of Interactive Learning Tools in Pre-exercise Phase How to bring participants up to speed quickly and easily in the pre-exercise phase? Firstly, identifying what is absolute critical information that each participants needs to know (such as game rules) and indicate on exercise information sharing site with clear notification absolute minimum critical read and/or make those into quick "performance supports"/"cheat sheets". Secondly, creating short introductory summary videos. These can be used instead of exercise sharing site or lengthy webinars that

provide overviews of information, and are timely/flexible for participants. We live in the age of visual information where visual content plays a role in every part of life, with as much as 65% being visual learners and visuals are processed $60,000$ times faster in the brain than text. Graphic interfaces including photos, illustrations, charts, maps, diagrams, and videos are gradually replacing text-based learning [79].

Learning Design Considerations Due to the unique set-up that BTs are distant from organisers (i.e. YT, WT, RT, GT), it is challenging to understand what is happening in BTs and improve the learning experience. So organisers of the games should look for other methods to bring "closeness" between the teams.

The traditional war games have an instructor who coaches, criticizes and guides BTs. For improving learning effectiveness this is definitely benefit, as only final feedback at the end of the exercise is too late. One of the options in CDXs with training objectives would be on-time coaching, such as "you started too reactive way, what are the reasons, what can you do differently?". In essence that would be role of YT or WT Liaisons, but with speed and complexity of the exercise, and distance between organisers and teams, YT/WT is not equipped to handle that (i.e. stress reporting is very subjective and self-reporting, as also shown by the attempt to validate some BT action timestamps in this thesis is not very reliable).

As the complexity of exercise increases, an option would be a "hotline", i.e. support centre, specifically for complex systems such as Siemens, drones, etc. As clearly learning outcome such as (quote from BT feedback): "The ICS team says they learn never to buy anything from Siemens until they (Siemens) learns how to program.", was not intended. Simply providing lengthy manuals is not effective, as this is unrealistic to train personnel in such short time-frame to become experts on the special systems: "We had to debug and complete the documentation to understand the global picture and details.... But without a deep understanding, the task was hard and even a fail". To avoid such misunderstanding and conclusions as learning outcome, the learning design needs support the training objectives and appropriate messaging.

Feedback As mentioned feedback at the end of exercise might not be sufficient, and the design in the execution should already include some feedback elements.

As overall feedback is biased on RT side—organisers are not able to provide constructive feedback on several other training objectives, such as team communication, incident handling, etc. This partially is due to lack of any benchmarking data and could be tackled by non-intrusive data collection by 5-timestamp methodology and build-up of baseline. This work is out of scope for this thesis.

The suggestion is to provide more examples why certain "game" element was

selected in exercise and how it is relevant (why considered important to train using this specific task) or share some examples how it could be applied in real life. This would apply to all feedback given.

Some simple and no significant additional resource requiring ideas for strengthening the feedback loop from RT are following:

1. provide BTs high level attack plan in advance of the briefing—to be able to better follow "thought process" of briefing;

2. make feedback comments more transferable. For example, reality vs. "game" attack from players/attacks, i.e. way/style of attacking may be nothing similar to participants face later;

3. agree in advance feedback structure and level of detail to ensure consistent quality of feedback by RT members;

4. divide RT sub-team leaders and relevant BTs into smaller groups and give them an hour to discuss in smaller groups (as with 19-20 teams the audience has grown to big), as often questions are not raised in large auditorium (and also level of feedback is more general). In this way, the hotwash session will not be longer, as it can be conducted simultaneously (e.g. set a rotation plan like "speed dating").

Feedback is two way thing—an approach "we tell, you listen" is not effective. RT can learn some tricks of trade from the BTs. Making hotwash sessions more interactive (by suggestions above) will partially tackle this problem. The AAR meeting includes BT leaders who are not necessarily technical and time lag after exercise will diminish the learning impact, thus effective hotwash session is critical.

As lot of learning is taking place in the teams and not under control of organizers, providing short guidance materials (i.e. facilitator guide) to how to run local feedback session within BTs is another step forward with ensuring learning impact is reaching all team members. Such materials would include slides and notes, with key objectives and takeaways from the exercise that each team leader can make specific to his/her team to run a discussion how to apply what they have learned to their everyday tasks/jobs.

6.7 Summary

This section focused on specific research questions (Section 6.2). It demonstrated evidence that pre-exercise phase is significant—both individuals and teams spend considerable time and effort, and report improvement in knowledge/skills. For complexity (including ICS segment) there is mixed feedback, indicating that design

needs to be careful, and allow flexibility to improve learning in teams with different skill levels. This and other design suggestions were raised, and are base on learning theories as described in Section 3. With regards to the 5-timestamp methodology, the challenges in practical application demonstrated that timestamp reporting from teams cannot be relied upon alone but proposed alternative is non-obtrusively collected pcaps.

The contribution of LS learning measurement is: 1) analysis of participant feedback on learning experience, with specific focus on learning experience of complex systems and team/individual aspects, 2) analysis of the existing metrics collected in the exercise for identification of learning metrics and relations, 3) validation of 5-timestamp methodology, and 4) learning design improvements for next year LS.

7 Crossed Swords—Learning Measurement

The successful design and execution of a RT training exercise consists of myriad of interrelated training components–such as scenario, complexity of network map, infrastructure provided by GT, competence and devotion of trainers, YT feedback and interactions, learning environment, appropriate level of technical challenge, communication plan established by RT leader for training event, tools used and available for RT, etc.

One of the learning design challenges in RT training exercises is to provide adequate and timely feedback to learners (RT) during the exercise, so the learning impact could be maximised.

7.1 Measurement Scope

This learning measurement is not meant to be a comprehensive analysis and only aims to measure impact of providing feedback to the learning experience in the RT exercise.

Frankenstack is a monitoring toolset developed at the NATO CCD COE at the coding hackathlon that automates the manual data analysis usually done by YT. It uses open-source tools (such as Grafana, Scirius, Suricata, Kibana, Alerta, etc.), but also an event correlation, a novel query automation tool. Prototype has been made publicly available at *https://github.com/ccdcoe*. Further details and overview of technical aspects is provided in [12].

7.2 Background—Feedback and Frankenstack

The feedback and providing situational awareness are essential for learning and improvement to take place, as discussed in Section 3. However, little research focuses on perceptions and impact that instant feedback has on participants learning in any cyber training exercise. Too much feedback could reveal information about objectives the RT is tasked to discover, while not enough feedback impacts learning about stealthiness of attack methods. This is a delicate trade-off.

In previous XS exercises, YT has provided the information to RT members at the end of the training day and participant feedback showed that feedback was provided too late. In XS17, the organisers attempted to provide instant situational awareness to participants by the use of Frankenstack toolset, in addition to feedback provided by YT members on significant matters.

Some of the assumed benefits of Frankenstack tool from learning perspective include: 1) the information regarding complexity of the attacks used in the training, 2) the level of abstraction (visibility) in which the participant interacts with the attack scenarios, 3) the speed at which the information is presented and participants'

behaviours changed, and 4) providing evidence to support a priori assumptions of participants learning patterns and impact of feedback and having "improved" situational awareness. Whereas from the challenges, having the immediate feedback available would mean that learners might focus on bypassing Frankenstack and/or avoid more risky strategies, and thus not using the learning opportunity provided to maximum extent.

7.3 Research Questions

This measurement aims to provide answers to following research questions:

1. What is the impact of providing instant feedback (i.e. situational awareness) in XS17 to participants' learning experience?

2. What technical metrics provided by Frankenstack (and similar monitoring tools) correlate to learning effectiveness?

7.4 Methodology

The tools and infrastructure are an essential for RT training, but they do not make the RT training event successful by default. Often human factors, such as how YT and RT's perceive and use the tools have significant impact learning. To evaluate the impact of the situational awareness tool (Frankenstack) the combination of quantitative and qualitative approach is applied:

1. Data analysis—the objective is to understand if the level of difficulty in the exercise is designed at the correct level for the participants. In order to achieve this: 1) selection of four machines based on various level of difficulties: easy/hard to exploit, and easy/hard to reach by cyber security experts, who had constructed exercise, 2) validation and analysis of data available for the selected targets to determine any learning impact;

2. Qualitative interviews with RT members—the objective is to gain deeper insight from participants about the learning experience. Interviews are conducted in casual setting and format during the breaks. Among several questions, focus is their reaction to Frankenstack and overall learning experience;

3. Quantitative survey (Appendix D)—the objective is to get independent and anonymous feedback from RT with specific focus on situational awareness tools and perception of impact to their learning process. The survey consists of multiple choice or ranking style questions. There is a freeform "additional comments" question for each question and the survey concludes by asking

some general questions about skills improvement and satisfaction with exercise;

4. Observations—the objective is to observe the learning behaviour of RT members, and also their interaction with YT and/or monitoring the situational awareness tools to gain further insights to learning process (e.g. impact on team communication).

7.5 Data Sources and Relations Analysed

Note that the analysis presented in this section is based on joint work with Markus Kont, Mauno Pihelgas and Bernhards Blumbergs from NATO CCD COE. The author has contributed with initial idea of selection by matrix, suggestions of possible learning metrics and evaluation from learning perspective.

The level of information that is collected by the various Intrusion Detection Systems (IDSs) can be overwhelming and the objective of this section is to understand what amount of feedback is appropriate for the RT to facilitate their learning. In order to approach that question a sample of attack vectors and machines is select and analysed for whether and how RT successfully reached and exploited the machine. Furthermore, what situational awareness is made available during the attack to learners. Prior to exercise start, an exercise developer was asked to select four systems based on two difficulty criteria (exploitability and reachability) and two difficulty levels (easy and hard). Based on the network map and learning scenario the targets summarised in Table 14.

Metrics	Easy to Reach	Difficult to Reach
Easy to Exploit	mail.clf.ex	fw.clf.ex
Difficult to Exploit	srv1.dev.clf.ex	git.dev.clf.ex

Table 14: XS17—Learning measurement matrix

Two of these targets were successfully compromised—*mail.clf.ex* (BT e-mail server) and *fw.clf.ex* (BT network central firewall). Both systems were considered easy to exploit, but firewall was also deemed difficult to reach. Information in central alert dashboard was cross-referenced and correlated with RT documentation. Reachability assessment could be verified on this data, as little confirmation could be found regarding direct attacks against firewall. Furthermore, firewall served as Network Address Translation (NAT) interface for internal network segments, obscuring alert source and destination addresses. In other words, large amount of attacks originating from or going toward firewall IP address were instead connections between BT workstations and RT simulated internet.

Signature	Count
ET SCAN Nmap Scripting Engine User-Agent Detected (Nmap Scripting Engine)	21
ET SCAN NMAP OS Detection Probe	8
ET POLICY Suspicious inbound to PostgreSQL port 5432	7
ET POLICY Suspicious inbound to MSSQL port 1433	5
ET POLICY Suspicious inbound to mySQL port 3306	5
ET POLICY Suspicious inbound to Oracle SQL port 1521	4
ET SCAN Potential SSH Scan	3
ET SCAN Potential VNC Scan 5800-5820	2
ET SCAN Potential VNC Scan 5900-5920	2
ET SCAN Potential SSH Scan OUTBOUND	1

Table 15: XS17—Suricata alerts for mail.clf.ex

As Frankenstack was not designed with post-mortem analysis in mind, it is difficult to reproduce and study the exact information available to the RT during the exercises. We focus here mainly on data-output from the Suricata IDS which would be realistic to assume RT has looked at this information during the game. Suricata with Emerging Threats (ET) open-source rule-set was specifically chosen because it shares a common rule format with its main competitor Snort, but ET open provides higher number of rules and Suricata has better parallelization. Furthermore, log data requires temporal correlation with other data sources to provide comparable results, and were thus not considered in this analysis. Various IDS alerts reflecting malicious RT probing are presented in Table 15.

Data source	Distinct	Total
IDS	16	176
Snoopy	2	8
Apache access logs	3	1642

Table 16: XS17—Correlated Frankenstack alerts for mail.clf.ex

Table 16 presents Frankenstack results for the e-mail server. Overall, 1826 events were identified from 3 distinct data sources. To some extent the higher number alerts is caused by the measurement infrastructure— for example, a real event may start in one pcap file and end in another, and are likely not reassembled properly. However, this is also a realistic picture network operators would face when dealing with alerts. RT might also "hide" in a flood of legitimate events. The question arises, if it is useful for RT training to show them the same flood of information a system administrator would have to deal with? A more skills network administrator might have detected and proved RT lateral movement, which could be identified in the log files—just harder to spot. While it is important for the

RT to learn remaining stealthy (one of XS's learning outcomes), it is questionable what the right level of information is that they should be presented with. Clearly, not every RT member is or has to be an experienced system administrator as well.

This initial data analysis shows that there are still several technical shortcomings and challenges to be addressed that YT faces. From learning perspective also a few questions arise. Firstly, is the technical complexity or exercise set-up designed appropriately? RT reached only two easy to exploit targets, and two difficult to exploit targets were not reached (50% of the selected sample)—this indicates that the exercise network design is potentially overly complex for the learners' capabilities. Instant feedback provided to participants did not accelerate the exercise process in maximizing learning potential—i.e. exploiting more difficult targets. Secondly, looking at the distinct number of alerts for just one easy to reach/easy to exploit target and considering that this information was presented on five separate screens—it indicates possible "information overload", i.e. that learner was not able to obtain relevant situational awareness and sensible instant feedback from the monitoring screens.

7.6 Feedback Analysis

For the survey, 14 responses (out of 27 participants, 52%) were obtained. As an overall respondent profile, 46% of participants had attended other RT exercise(s), but none of those exercises situational awareness tools such as Frankenstack (remainder 54% had not previously attended any RT exercise).

The training room had 4 large screens with dedicated screens to Alerta, Grafana, Scirius and Suricata (fifth screen displaying Event Visualization Environment (EVE) was only visible to YT and WT members). RT members preferred to view the main screens displayed in training room, and 38% responded they checked the screen about every 60 minutes or less or another 38% about every 30-50 minutes. However, when they tried out a new attack vector, learners preferred to monitor on their own screen(s).

When surveying about learning impact, 79% agreed (of those 57% strongly agreed) that the instant feedback / situational awareness received during exercise is useful for their learning process (21% were neutral and none disagreed). This is supported by feedback that: 1) 77% of the learners considered the speed of feedback received is at correct level, 15% said it is still too slow and only 7% that it is too fast (too early), 2) 57% agreed that the alerts and information provided are accurate and sufficient for their learning process. A large number of participants (43%) are rather neutral about this question. This raises the question, if the automated feedback is appropriately given in this format.

In relation to learning behaviour, 45% of the participants agreed that they learned a lot about how their actions can be visible (i.e. it is useful to see si-

multaneously what attack method could be detected how), 30% said that they were more careful with their attacks and thus tried to be more stealthy than they normally would have been (i.e. more careful because of all the monitoring). A few comments from RT members were as follows: "yes, looked at the monitor—every time I got spotted, I changed my IP" and "from screens, I got some insights about network map, otherwise I would not have known". In training set-up, there is a trade-off between the short period of time, how stealthy the learners can be, and how much help they have from the trainers on how to be stealthy. Some comments were that over time the RT was becoming less stealthy, which suggests that the task of being too stealthy didn't work out under the time pressured exercise learning environment.

Were some learning situations simply avoided? 64% of respondents confirmed that the situational awareness tools and monitors are not distracting them or having negative impact; 30% did agree they are somewhat distracted and believed they performed worse (remainder 6% being neutral). This feedback confirms the challenges of monitoring and providing instant feedback to learning.

One of the key training objectives is working as a united team in achieving the laid out mission objective, and thus team communication and cooperation is vital. Overall 83% of participants indicated some improvement of the skills for this specific training objective. However, feedback specifically for impact of situational awareness tools and team communication and cooperation is mixed—50% perceived positive impact, whereas 21% negative and remainder being neutral. Some comments from RT members are the following: "yes, it impacts definitely affects team communication", "can talk less as can see from screen, otherwise would need to ask", "the biggest problem was due to lack of communication". The main challenge is to get individuals and sub-teams "talking" and screens might be used as inhibitor or enabler for team communication—but the right messages to learners must be ensured.

To provide comparison to other RT exercises (without such situational awareness tools), 50% responded that they needed to ask less information from YT members, as they got relevant situational awareness information via Frankenstack. At the same time only 37% believed they learned more effectively (compared to 63% with believing no difference) with such tools used. The risk here is also that learners might not have interpreted the results of tools correctly or in essence actually miss out on expert YT advice—that means that the tool, wrongly used in exercise learning design, could actually have negative impact on learning outcome. There were several observed cases of YT discussions with RT members—"ok, we know that you are visible, how can you "hide"/ be more stealthy?" These are the moments that bring learning insight, and might be reduced reading situational awareness information by the learner alone.

Human interaction and guidance is a critical success factor for learning, especially in team setting. 64% of respondents said they had sufficient help for their learning process, i.e. when they got stuck and did not know how to proceed, their team members or sub-teams leaders provided guidance. It's somewhat disappointing result, and could be increased with improved learning design. For example, one comment received was: "no clue what was going on in other teams and stared at server that weren't vulnerable/didn't know how to crack open with advanced sql injections, very frustrating in conclusion". The exercise time is very compact and limited, and therefore it should not be wasted sitting or digging into the "wrong direction" (aka "try harder") and thus not progressing.

7.7 Improving Learning Experience and Effectiveness

Given the amount of work that goes into preparing such exercises, the level of learning potential needs to be maximized. Some of the learning design recommendations are included in this section, however due to constraints of timing (XS is an annual exercise) impact is not been validated.

From the learning perspective, it cannot be assumed participants know how to use or interpret the results of situational awareness tools at the training setting. Based on lack of knowledge learners might not interpreted the results of tools correctly or miss out on expert YT advice during the exercise, which has negative impact on learning. The training for tools needs to take place to prior training session for those RT participants who do not have the knowledge. A solution proposed is a pre-survey to evaluate the learners' familiarity with the tools and level of training needed before exercise. The training does not necessarily need to be excessive and can for example be conducted beforehand.

The selection of tools included in Frankenstack (usefulness vs. distraction) and the information display needs further analysis, as switching between multiple screens (either on own computer or large training room display) was not convenient and somewhat distracting. One of the options is to combine into one monitor/display.

In RT exercises, such as XS, there are specific main objectives to be achieved by the team. It is challenging to evaluate achieving objectives, since there are many steps involved in reaching a specific objective. Often the tasks or sub-objectives are divided between sub-teams (network, web and client-side) and between individuals in those sub-teams. The difficulty of a specific exploitation or attack of a machine directly depends on the individual skillsets, which varies widely. That means there is a trade-off between assigning tasks to participants, who will just do what they already know, and actually learning something new, which implies that the learners need discover unknown territory.

The division of tasks and sub-objectives between sub-teams and individuals

also diminishes the learning potential, as each individual focuses on specific objectives and will not necessarily gain knowledge from the vast pool of different attack vectors and strategies used by other RT members. Thus a reflective team sharing is crucial for the learning success of each individual, and would overcome project management approach—each team member looks mainly into machines assigned to him/her, and assumed self-initiative and/or time availability of each team member to analyse and see what other are doing. From feedback quite many RT members felt they were stuck or "alone". A recommendation is planning into training schedule the regular "time-out"/reflection sessions, when team can share what they have learned or when they are stuck, obtain tips and further insights how to proceed, and thus collectively improve the knowledge.

Another common learning design approach to enhance transfer of learning and reflection is "buddy-system" where RT members are not assigned a sub-task individually, but in groups of two or three. This design also supports team communication training objective. Participants would then have to share their knowledge and can benefit from different background the various participants of the group have.

The EVE with overall view on all those tools visible on the network map that was not shared with RT during the exercise), as it revealed the network map. Too much gives hints to trails that would otherwise not be there and thus make the exercise rather unrealistic, too little and there is no feedback. Potentially having some parts of the EVE greyed-out and be discovered as the team discovers the topology, is likely not to work well. Showing the attack visualization as the end is also not adding much value from learning perspective, as this is un-reflected. Rather it should be incorporated into regular feedback / time-out sessions as suggested above (for example show status of the penetration by only "compromised" machines).

Finally, it is important to have a better time planning during the execution. While it is certainly appropriate to allow for flexibility in the paths RT can take to solve the objectives, it should also be avoided that participants spend too much time digging in "a wrong direction". Time is very limited during the exercise and therefore it needs to be planned how long someone can work on a target and when help is needed to get someone "unstuck". To some extent this was already present depending on the individual coaches/facilitators, but no pre-exercise plan exists.

7.8 Summary

Overall learning impact of the exercise in this format (with awareness tools) is very positive. There are 13% of total participants responses, which reported no improvement of their skills has taken place. Overwhelming 87% perceived different levels of the improvement skills (44% responses in between of 10-50% improvement

level) and 93% agreed they are satisfied with exercise.

There is mainly positive feedback on use of the awareness tools to the learning impact, however the critical questions to be answered in design phase of RT exercise include: what is the right balance of information for RT, is the behaviour changed due to monitoring or information visible (i.e. learners unconsciously limit themselves by not trying out more risky strategies, etc.). Some learning design changes, and not necessarily only limited to situational awareness, can increase the return on significant investment into preparing such RT team exercises.

The contribution of XS learning measurement effort described in this section is: 1) analysis of participant feedback on learning experience, with specific focus on situational awareness impact as part of the overall learning experience, 2) analysis of the data obtained from Franckenstack tool for the purpose of learning design assessment, 3) learning design improvements for next year XS, and 4) assistance to evaluate and improve Frankenstack tool.

8 Recommendations for Learning Improvement at CDXs

Learning design of the CDXs can be improved as a result from effective learning measurement and feedback from the participants.

8.1 Learning Design Enhancements

The author has described throughout the thesis various learning design enhancements, such as practical considerations based on adult learning theories (Section 3.6.5), and specifically for LS (Section 6.6) and XS (Section 7.7).

Learning design can only be improved when the exercises are measured and evaluated. This thesis outlined several shortcomings of existing measurements based on studying two CDXs, and recommended the 5-timestamp methodology (Section 5.1) as a starting point how to collect quantitative performance and learning information non-obtrusively from pcaps, scoring logs, etc. Successful learning requires effective feedback, including trade-off between automated and human feedback. Considering the size and complexity of CDXs, feedback at individual and team level needs to become scalable. Both the 5-timestamp methodology and NATO CCD COE's Frankenstack tool focus on overcoming this challenge.

8.2 Strengthening AARs

Useful information should be made public because it will help the participants realize what was really going on during the exercise and will help the organizers improve other editions of the exercise [56]. There are several AARs available, however they do not have control processes to ensure that lessons learned were actually validated or implemented [6] and such information is very varied and not allowing comparison. The learning measurement related information in the AARs made available should include average information (i.e. can be disclosed for all teams in total to preserve anonymity) that will create comparative measurement basis between the exercises to also prove overall learning curve from such exercises. For CDXs, such information (i.e. time to detect, time to mitigate, percentage of RT attacks mitigated, etc.) also shows possible areas for improvement. Such studies and related metrics start to emerge, for example [39] and [69] showed that accuracy might suffer from the speed. Having such information available will help improve learning effectiveness and also measure that positive learning outcomes of such exercises over time. The evaluation of learning impact forms parts of wider impact analysis of the exercise that would cover all stakeholders.

9 Conclusion

Learning is such a complex and intractable process that has made its study difficult and contentious. Senge makes a good point that we can give participants learning experience but without analysis, feedback or guidance it is wasted effort [29]. It applies at individual, team or organizational level and understanding the basic learning processes and implications can shape how learning is designed and delivered at the cyber exercises.

Small-scale exercises, such as developed by RangeForce [22], are excellent learning environments, but often lack complexities for more advanced participants and/or to develop teams. On the other hand, high-end exercises, such as LS, aim at addressing this shortcoming. Unfortunately, their complexity makes it very hard to incorporate learning and skill assessment aspects appropriately. In the current designs those aspects are sadly too often neglected and thus invalidating one of the main purposes that brought such exercise in existence. In this study the author has outlined a set of ideas on how to evolve learning and skill assessment in the future.

In this thesis author explored the modern learning theories and related theoretical aspects applicable for CDXs, including team learning aspects (Section 3). Theoretical answers were sought for a question: what factors contribute to learning success or failure in such exercises? In Section 4 interdisciplinary analysis and literature review (specifically from team learning and game-based learning studies) was performed what are the current methodologies to measure the learning in CDXs. In Section 5 the author proposed novel non-obtrusive 5-timestamp methodology for timing and accuracy metrics for technical skills, but the model also suggests metrics for team aspects and soft skills. Some practicalities of data collection and validation approaches with qualitative measurements were also explored and proposed. To put theory into practice, validation of the 5-timestamp methodology and selected learning analysis and measurements were performed on LS17 (Section 6) and XS17 (Section 7) exercises. To benefit from the extensive literature review and measurements made, practical insights to organisers were summarised in Section 8 based on the relevant learning theories.

Main conclusions reached are:

1. Team learning dimension needs to built into measurement frameworks for the CDXs (Section 3.4);

2. Methodological measurement of achieving the training objectives is required to conclude whether an exercise design was appropriate and effective. The training objectives should be evaluated using Bloom's taxonomy to ensure also higher levels of taxonomy, i.e. evaluate—create, are addressed. This can be used to improve and optimize future cyber exercises;

3. For learning measurement, it is challenging and resource consuming to "observe" the teams and self reporting might prove to be inaccurate or fail, thus non-intrusive methods such as the 5-timestamp methodology. Data collection for learning measurement and effective feedback can be combined (Section 5.1, 6.5.5);

4. Learning design for highly complex exercise (scenario and systems), should be include supportive learning tools (e.g. support "helpline", visuals includes in the guidance materials, etc.) (Section 6.6);

5. The exercises need to provide adequate and timely feedback to learners during the exercise, so the learning impact could be maximised. The hotwash and AAR session at the end of exercise are too late (Section 6.6, 7.7);

6. Individual and team feedback needs to be scalable. Learning metrics from training events will help teams and organisers to benchmark themselves (and identify further learning needs) (Section 8).

This thesis presented an idea for non-obtrusive data collection and measurement, i.e. the 5-timestamp methodology. Future work should continue with performing the data analysis of an exercise to compile learning metrics and trends benchmark. Identification and analysis of data trends will provide solid baseline and demonstrate learning improvement achieved in CDXs. This will complement often anecdotal and positive feedback obtained via traditional methods (surveys, interviews) that participants have actually learned.

This thesis attempted to evaluate long-term effect by surveying LS17 participants who attended LS16—69% of them reported recalling skills learned in previous exercise in specific LS training areas. Future work should be conducted regarding the generality and maintenance of behaviour change produced by the CDX and identifying the conditions under which exercise gamification is effective.

Future work should also be how team learning can be effectively transferred to the organization. As even when teams learn it may not translate to organizational learning (e.g. because teams fail to communicate with others in organizations) [65]. This process should be further investigated to gain practical insights and improve efficiency.

In overall, incorporating non-intrusive, social and behavioural research methods into the cyber security field can give new insights and possibilities in effective training for cyber defence teams in the future.

"Learning without thought is labour lost; thought without learning is perilous." (Confucius)

References

[1] E. Crowley, "Experiential learning and security lab design," in *Proceedings of the 5th conference on Information technology education*, pp. 169–176, ACM, 2004.

[2] L. Buchanan, F. Wolanczyk, and F. Zinghini, "Blending Bloom's Taxonomy and serious game design," in *International Conference on Security and Management Las Vegas, Nevada USA, July 18-21, 2011*, pp. 518–521, CSREA Press, 2011.

[3] D. M. McInerney, *Educational psychology: Constructing learning*. Pearson Higher Education AU, 2013.

[4] I. Mayer, G. Bekebrede, H. Warmelink, and Q. Zhou, "A brief methodology for researching and evaluating serious games and game-based learning," in *Psychology, pedagogy, and assessment in serious games*, pp. 357–393, IGI Global, 2014.

[5] NATO CCD COE, "Locked Shields 2016 After Action Report." NATO Cooperative Cyber Defence Centre of Excellence Publication, 2016.

[6] A. Ogee, R. Gavrila, P. Trimintzios, V. Stavropoulos, and A. Zacharis, "The 2015 Report on National and International Cyber Security Exercises." `https://www.enisa.europa.eu/publications/latest-report-on-national-and-international-cyber-security-exercises`.

[7] A. Ahmad, C. Johnson, and T. Storer, "A cyber exercise post assessment: Adoption of the Kirkpatrick Model," *Advances in Information Sciences and Service Sciences*, vol. 7, no. 2, p. 1, 2015.

[8] J. A. Mattson, "Cyber defense exercise: A service provider model," in *Fifth World Conference on Information Security Education*, pp. 81–86, Springer, 2007.

[9] P. Pusey, M. Gondree, and Z. Peterson, "The outcomes of cybersecurity competitions and implications for underrepresented populations," *IEEE Security & Privacy*, vol. 14, no. 6, pp. 90–95, 2016.

[10] NATO CCD COE, "Locked Shields 2017." `https://ccdcoe.org/locked-shields-2017.html`. Accessed: 2017.04.25.

[11] NATO CCD COE, "Crossed Swords 2017." `https://ccdcoe.org/cyber-defence-exercise-focuses-vulnerability-testing.html`. Accessed: 2017.03.14.

[12] M. Kont, M. Pihelgas, K. Maennel, T. Lepik, and B. Blumbergs, "Frankenstack: Monitoring framework for real-time red team feedback," NATO CCD COE, 2017. Under Submission.

[13] "Cyber definitions." `https://ccdcoe.org/cyber-definitions.html`. Accessed: 2017.22.03.

[14] A. Ahmad, *A cyber exercise post assessment framework: In Malaysia perspectives*. PhD thesis, University of Glasgow, 2016.

[15] "CTF Time." `https://ctftime.org/ctfs`. Accessed: 2017.03.21.

[16] M. F. Thompson and C. E. Irvine, "Active learning with the CyberCIEGE video game," *CSET*, vol. 11, pp. 10–10, 2011.

[17] F. Szalai, "Does cyber security exercise information sharing work? Review and analysis of technical cyber security exercises and their information sharing," 2016. MSc thesis, Tallinn University of Technology.

[18] L. J. Hoffman, T. Rosenberg, R. Dodge, and D. Ragsdale, "Exploring a national cybersecurity exercise for universities," *IEEE Security & Privacy*, vol. 3, no. 5, pp. 27–33, 2005.

[19] Y. Bei, R. Kesterson, K. Gwinnup, and C. Taylor, "Cyber defense competition: a tale of two teams," *Journal of Computing Sciences in Colleges*, vol. 27, no. 1, pp. 171–177, 2011.

[20] M. E. Kuhl, J. Kistner, K. Costantini, and M. Sudit, "Cyber attack modeling and simulation for network security analysis," in *Proceedings of the 39th Conference on Winter Simulation: 40 years! The best is yet to come*, pp. 1180–1188, IEEE Press, 2007.

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

[22] M. Ernits, J. Tammekänd, and O. Maennel, "I-tee: A fully automated cyber defense competition for students," *SIGCOMM Comput. Commun. Rev.*, vol. 45, pp. 113–114, Aug. 2015.

[23] Stephanie A. Hargett, "Cyber shield 2016 expands training opportunities." `https://www.army.mil/article/166711/cyber_shield_2016_expands_training_opportunities`. Accessed: 2017.03.26.

[24] United States Joint Forces Command, "Millennium challenge 02." https://web.archive.org/web/20070928005405/http://www.jfcom.mil/about/experiments/mc02.htm. Accessed: 2017.05.02.

[25] R. Mudge, "Cobalt Strike." https://www.cobaltstrike.com. Accessed: 2017.05.07.

[26] NATO CCD COE, "Cross Swords 2017 RT Objectives," 2017. Accessed: 2017.03.21.

[27] https://en.oxforddictionaries.com/definition/learning. Accessed: 2017.04.04.

[28] A. Drejer, "Organisational learning and competence development," in *The Learning Organization*, vol. 7, pp. 206–220 Issue 4, ACM, 2000.

[29] J. M. Wilson, P. S. Goodman, and M. A. Cronin, "Group learning," *Academy of Management Review*, vol. 32, no. 4, pp. 1041–1059, 2007.

[30] A. Nicolaides and V. J. Marsick, "Understanding adult learning in the midst of complex social "liquid modernity"," *New Directions for Adult and Continuing Education*, vol. 2016, no. 149, pp. 9–20, 2016.

[31] J. Piaget, "Piaget's theory of cognitive development," *Childhood cognitive development: The essential readings*, pp. 33–47, 2000.

[32] E. Maiberg, "Israeli military prepares for cyberwar by staging an alien invasion." https://motherboard.vice.com/en_us/article/israel-prepares-for-cyberwar-by-staging-an-alien-attack. Accessed: 2017.05.02.

[33] K. Lewin, "Frontiers in group dynamics: Concept, method and reality in social science; social equilibria and social change," *Human relations*, vol. 1, no. 1, pp. 5–41, 1947.

[34] R. Reid, J. V. Niekerk, and R. V. Solms, "Guidelines for the creation of brain-compatible cyber security educational material in Moodle 2.0," in *2011 Information Security for South Africa*, pp. 1–8, Aug 2011.

[35] B. A. Bratosin, "Cyber defense exercises and their role in cyber warfare," *Journal of Mobile, Embedded and Distributed Systems*, vol. 6, no. 2, pp. 70–76, 2014.

[36] J. Mirkovic, P. Reiher, C. Papadopoulos, A. Hussain, M. Shepard, M. Berg, and R. Jung, "Testing a collaborative ddos defense in a red team/blue team exercise," *IEEE Transactions on Computers*, vol. 57, no. 8, pp. 1098–1112, 2008.

[37] W. J. Schepens, D. J. Ragsdale, J. R. Surdu, J. Schafer, and R. New Port, "The cyber defense exercise: An evaluation of the effectiveness of information assurance education," *The Journal of Information Security*, vol. 1, no. 2, 2002.

[38] R. Richards, A. Konak, M. R. Bartolacci, and M. Nasereddin, "Collaborative learning in virtual computer laboratory exercises," *Network, Security*, vol. 155, p. 9, 2015.

[39] D. S. Henshel, G. M. Deckard, B. Lufkin, N. Buchler, B. Hoffman, P. Rajivan, and S. Collman, "Predicting proficiency in cyber defense team exercises," in *Military Communications Conference, MILCOM 2016-2016 IEEE*, pp. 776–781, IEEE, 2016.

[40] E. Boyle, T. M. Connolly, and T. Hainey, "The role of psychology in understanding the impact of computer games," *Entertainment Computing*, vol. 2, no. 2, pp. 69–74, 2011.

[41] P. Wouters, C. Van Nimwegen, H. Van Oostendorp, and E. D. Van Der Spek, "A meta-analysis of the cognitive and motivational effects of serious games.," *Journal of Educational Psychology*, vol. 105, no. 2, p. 249, 2013.

[42] T. M. Connolly, E. A. Boyle, E. MacArthur, T. Hainey, and J. M. Boyle, "A systematic literature review of empirical evidence on computer games and serious games," *Computers & Education*, vol. 59, no. 2, pp. 661–686, 2012.

[43] D. R. Newman, B. Webb, and C. Cochrane, "A content analysis method to measure critical thinking in face-to-face and computer supported group learning," *Interpersonal Computing and Technology*, vol. 3, no. 2, pp. 56–77, 1995.

[44] Y. Altman and P. Iles, "Learning, leadership, teams: corporate learning and organisational change," *Journal of Management Development*, vol. 17, no. 1, pp. 44–55, 1998.

[45] K. J. Klein, J. C. Ziegert, A. P. Knight, and Y. Xiao, "Dynamic delegation: Shared, hierarchical, and deindividualized leadership in extreme action teams," *Administrative Science Quarterly*, vol. 51, no. 4, pp. 590–621, 2006.

[46] https://en.wikipedia.org/wiki/Organizational_learning. Accessed: 2016.11.04.

[47] K. V. Moses and W. M. Petullo, "Teaching computer security," 2014.

[48] B. Kulich, "Lessons learned from military cyber defence exercises," *Science & Military Journal*, vol. 9, no. 1, p. 47, 2014.

[49] A. Dijk, J. Meulendijks, and F. Absil, "Lessons learned from NATO's cyber defence exercise Locked Shields 2015," in *Militaire Spectator, No 2*, pp. 68–74, 2016.

[50] L. Randmann, "Organizational learning and learning organization." HPP8410 Organizational theory and psychology, TTU, 2016 Fall Semester.

[51] NATO CCD COE, "Locked Shields 2015 After Action Report." NATO Cooperative Cyber Defence Centre of Excellence Publication, 2015.

[52] J. Kick, "Cyber exercise playbook," tech. rep., DTIC Document, 2014.

[53] M. Granasen and D. Andersson, "Measuring team effectiveness in cyber-defense exercises–a cross-disciplinary case study," in *Cognition, Technology & Work, No 2*, vol. 18, pp. 121–143 Issue 1, Springer-Verlag London, 2016.

[54] A. A. Gokhale, "Collaborative learning enhances critical thinking," *Digital Library and Archives of the Virginia Tech University Libraries*, 1995.

[55] J. Silberman, "5 Organizational learning disabilities which can limit success, Blog entry 12 March 2013." http://trainingstation.walkme.com/5-organizational-learning-disabilities-which-can-limit-success. Accessed: 2016.12.07.

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

[57] "ISO 22398:2013 Societal Security—Guidelines for Exercises." http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=50294. Accessed: 2016.11.18.

[58] "Learning objective." http://edutechwiki.unige.ch/en/Learning_objective. Accessed: 2016.05.22.

[59] M. J. Singer and B. W. Knerr, "Evaluation of a game-based simulation during distributed exercises," 2010.

[60] J. Vince and C. Best, "An evaluation of new After-Action Review tools in exercise Black Skies 10 & exercise Black Skies 12," tech. rep., DTIC Document, 2013.

[61] P. Wouters, E. D. Van der Spek, and H. Van Oostendorp, "Current practices in serious game research: A review from a learning outcomes perspective," in *Games-based learning advancements for multi-sensory human computer interfaces: techniques and effective practices*, pp. 232–250, IGI Global, 2009.

[62] J. B. Hauge, E. Boyle, I. Mayer, R. Nadolski, J. C. Riedel, P. Moreno-Ger, F. Bellotti, T. Lim, and J. Ritchie, "Study design and data gathering guide for serious games' evaluation," 2014.

[63] C. Girard, J. Ecalle, and A. Magnan, "Serious games as new educational tools: how effective are they? a meta-analysis of recent studies," *Journal of Computer Assisted Learning*, vol. 29, no. 3, pp. 207–219, 2013.

[64] A. C. Edmondson, "The local and variegated nature of learning in organizations: A group-level perspective," *Organization science*, vol. 13, no. 2, pp. 128–146, 2002.

[65] D. B. Hay, "Using concept maps to measure deep, surface and non-learning outcomes," *Studies in Higher Education*, vol. 32, no. 1, pp. 39–57, 2007.

[66] M. Uzumeri and D. Nembhard, "A population of learners: A new way to measure organizational learning," *Journal of Operations Management*, vol. 16, no. 5, pp. 515–528, 1998.

[67] I. Svetlik, E. Stavrou-Costea, R. Chiva, J. Alegre, and R. Lapiedra, "Measuring organisational learning capability among the workforce," *International Journal of Manpower*, vol. 28, no. 3/4, pp. 224–242, 2007.

[68] P. Jerez-Gomez, J. Céspedes-Lorente, and R. Valle-Cabrera, "Organizational learning capability: a proposal of measurement," *Journal of business research*, vol. 58, no. 6, pp. 715–725, 2005.

[69] T. Reed, K. Nauer, and A. Silva, "Instrumenting competition-based exercises to evaluate cyber defender situation awareness," in *International Conference on Augmented Cognition*, pp. 80–89, Springer, 2013.

[70] M. R. Stytz and S. B. Banks, "Addressing simulation issues posed by cyber warfare technologies," *SCS M&S Magazine. n (3)*, 2010.

[71] "Existing taxonomies." https://www.enisa.europa.eu/topics/csirt-cert-services/community-projects/existing-taxonomies. Accessed: 2017.01.07.

[72] "TERENA Incident Taxonomy and Description Working Group, Work in Progress." https://www.terena.org/activities/tf-csirt/iodef/docs/i-taxonomy_terms.html. Accessed: 2017.01.07.

[73] W. Jiang, Z.-h. Tian, and C. Xiang, "DMAT: A new network and computer attack classification," *Journal of Engineering Science & Technology Review*, vol. 6, no. 5, 2013.

[74] "CSIRT Case Classification (Example for Enterprise CSIRT)." https://www.first.org/_assets/resources/guides/csirt_case_classification.html. Accessed: 2017.04.24.

[75] J. Creasey and I. Clover, "Cyber Security Incident Response Guide." https://www.crest-approved.org/wp-content/uploads/2014/11/CSIR-Procurement-Guide.pdf, 2013. Accessed: 2016.11.18.

[76] J. D. Howard and T. A. Longstaff, "A common language for computer security incidents," *Sandia National Laboratories*, 1998.

[77] "Qualitative analysis with HyperResearch." http://www.researchware.com/products/hyperresearch.html. Accessed: 2017.05.07.

[78] "SMART definition." https://en.wikipedia.org/wiki/SMART_criteria. Accessed: 2017.14.04.

[79] "Studies confirm the power of visuals in elearning." http://info.shiftelearning.com/blog/bid/350326/Studies-Confirm-the-Power-of-Visuals-in-eLearning. Accessed: 2017.14.04.

A Appendix: Locked Shield 2017 Pre-Exercise Survey

Thank you for taking the time to complete this confidential questionnaire. Your response is anonymous and will be treated in confidence. Please answer each item honestly and thoughtfully. The information will be used to academic research purposes and to improve the quality of the exercise and learning.

You will need approximately 5 minutes to complete the survey.

1. Please select your team number

2. Please enter your team size
less than 10
10-15
16-20
21-30
more than 31

3. How many technical exercises have you participated?
Locked Shields Other Exercises
1
2
3
4
5
Other (please specify)

4. How many hours have you spent for team preparation?
1-2
3-10
10-50
50-100
more than 100
Other (please specify)

5. How did you prepare for the exercise?

	Never	Seldom	About half the time	Usually	Always
Individual					
Sub-team					

Whole team
Other (please specify)

6. Do you have the following areas covered by skilled personnel?
Media
Routing
Forensics
Legal
Industrial Control System
System admins
Reporting
Monitoring
Drone system
Please comment

7. Would you describe your team as:
Military/authoritarian (clear command line)
Hierarchical (specific roles)
Friendly and collaborative (buddies)
Other (please specify)

8. How do you assess your own knowledge/skill level and experience before the exercise in your technical area?

	None	Limited	Medium	High	Expert
	0	0-1	1-2	2-5	more than 5
Knowledge/skill level (in the scale of 5)					
Working experience					
Please comment					

9. What new skills/knowledge have you learned or improved in the preparation process?

	Significantly increased	Minor improvement	No change	N/A
Learning the network				
System administration and prevention of attacks				
Monitoring networks, detecting and responding to attacks				
Handling cyber incidents				
Conduction of forensic investigation				
Teamwork: delegation, dividing and assigning roles, assignment				
Cooperation and information sharing				
Ability to convey big picture				

Reporting
Crisis communication
Time management and prioritization
Cyber legal aspects
Please bring examples

10. Do you feel ready for the exercise, or are you confused?
Ready
Confused
Both, ready and confused
Please specify

11. I expect during the exercise:

	Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree	N/A
Learn nothing new overall (please specify below)						
Practice the skills I already had or obtained during the preparation for this exercise						
Learn new knowledge and skills						
Other (please specify)						

12. If you attended Locked Shields in 2016, can you recall any skill you learned from participating at LS?

	Yes	No	N/A
Learning the network			
System administration and prevention of attacks			
Monitoring networks, detecting and responding to attacks			
Handling cyber incidents			
Conduction of forensic investigation			
Teamwork: delegation, dividing and assigning roles, assignment			
Cooperation and information sharing			
Ability to convey big picture			
Reporting			
Crisis communication			
Time management and prioritization			
Cyber legal aspects			
Please bring examples			

13. If you attended LS in 2016, is your team composition:
75%-100% the same team as in LS16

50%-74% the same team as in LS16

Changed significantly, less than 50% old team members

Completely changed / new team

Other (please specify)

14. Did attending LS in 2016 change your team dynamics or team behaviour subsequently in workplace? (can select multiple if applicable).

	Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree	N/A
Our team has become more coherent, confident and collaborative						
Our team's knowledge has increased (as a result of individuals sharing)						
Only individual team member's knowledge increased however it has not been shared (if that team member would leave, the knowledge in the group is lost)						
There is no team dynamics or behaviour changes noted						
N/A, as the team was only formed for attending exercise and dissolved after exercise						
Other, please specify						

B Appendix: Locked Shields 2017 Post-Exercise Survey

NOTE: Questions only regarding participants' learning experience included in the post-exercise survey for LS16 and LS17.

5.10. Did you learn anything that will change how you do your job or help you do your job?

Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree	N/A

Please comment:

5.11. Give an example of how you will apply what you learned in this course back in your role.
Please enter your text here:

5.12. Please estimate how much your job performance has improved as a result of the exercise.

0%	1 - 20%	21 - 50%	51 - 80%	81 - 100%

Please comment:

5.13. Participating in this exercise was useful and the training objectives were met.

Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree	N/A

Please comment:

C Appendix: Locked Shield 2017 Injects

The below questions were included in the ICS scenario and final feedback injects.

The teams were provided pre-exercise notice to collect timestamps and questions related to ICS segment, including *hmi.ics.bluexx.ex*, *plc.ics.bluexx.ex*, *step7.ics.bluexx.ex* and ICS network segment.

1) What malicious/ suspicious activities did you see? What tool/rule/other allowed you to spot the activities?
Please provide exact timestamp time did you discover the attack , i.e. MMDDHHMMZ (for example 04261315Z)

2) Did your BT track/hunt the root cause of the activities (i.e. forged packets, CnC, backdoor,...)? If so, what did you discover?

3) What was the priority level you assigned to this event (Critical, High, Medium, Low, Very low)?
Please comment if your initial priority level assigned to this event, changed after initial impact assessment:

4) What countermeasures did you put in place? Were they successful? Why/Why not?
How do you assess your resistance level to the attack was (no resistance to very strong resistance): (1 to 5, 1: not at all, 5: very strong)?

5) How long (minutes) did it take from detection until taking actions against the attack on the systems? (Note: We are interested in the duration that team communication or thinking about the problem took.)

6) How long (minutes) did you "put up a fight" for each incident from detection until kicking out Red Team or giving up?

7) Did your strategies evolve during the game? How?

8) This incident for our team was (please select most appropriate):
a) Easy to detect—easy to mitigate
b) Easy to detect—difficult to mitigate
c) Difficult to detect—easy to mitigate
d) Difficult to detect—difficult to mitigate

9) Who solved the task?
a) One person
b) 2-3 members
c) 4 or more
d) Collaborated with other teams

10) How much did you learn (Not at all, Slightly, Moderately, Very, Significantly)?
a) individual team member(s) / sub-team working on the ICS systems?
b) as whole team?

11) How do you rate the difficulty/complexity of ICS attacks, compared to: (1 to 5, 1: significantly less complex, 5: significantly more complex).
a) other attacks in the exercise (DMZ, LAB, etc.)?
b) other attacks in special systems (Spectrum5, drones,...)?

12) Do you have additional (positive/negative) remarks for any of the ICS systems (gameplay, management, infrastructure,...)?

OVERALL exercise:

13) Do you have any other feedback on learning experience during the exercise? (i.e. complexity, gameplay vs. learning, opportunity to learn, etc.)

14) Any other comments for the exercise as a whole?

D Appendix: Crossed Swords 2017 Survey

Thank you for taking the time to complete this feedback questionnaire. Your response is anonymous and will be treated in confidence. Please answer each item honestly and thoughtfully. The information will be used to academic research purposes and to improve the quality and learning experience of the exercise. You will need approximately 10-15 minutes to complete the survey.

1. What was your role and sub-team?

	Network team	Client-side team	Web team
Team leader			
Sub-team leader			
Team member with task to monitor situational awareness			
Team member			
Other			

2. Have you participated in other RT exercises?

Yes, but they didin't had Frankenstack tools (Scirius, Grafana, Alerta, etc screens shared)

Yes, and that exercise had better feedback tools than Frankenstack (Scirius, Grafana, Alerta, etc screens shared)

No

3. Which of the situational awareness tools provided was the MOST/LEAST useful?

Alerta (incl SEC, TICK)
Grafana
Scirius
Suricata
Moloch
Other

4. I monitored/checked the situational awareness information screens (choose appropriate):

	On the large screen	On my own laptop
Not at all		
No so often (every 60 minutes or less)		
Often (30-50 minutes)		
Very often (10-30 minutes)		
All the time (every 1-10 minutes)		
Every time I tried an attack vector		

Sometimes when I tried a new attack vector

5. Speed of feedback received during the exercise using Frankenstack situational awareness information screens for my learning purposes was:

Too slow
Slow
Correct level
Fast (too early)
Too fast (too early)

6. Did feedback/instant situational awareness tools impact your decisions and further actions in selecting attack vectors?

Yes, I was more careful with my attacks and thus tried to be more stealthy than I normally would have been. However, I didn't check the monitors often, I was just careful because of all the monitoring.

Yes, I learned a lot about how my actions can be visible. It was useful to see simultaneously what attack method could be detected how

Maybe. but I didn't consciously change my decisions.

No impact, it did not have impact on my decisions and further actions in selecting attack vectors.

It had negative impact on my learning, as I was too careful and did not try everything I could have as was too afraid to be detected.

7. Please select the best answer for below statements:

	Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree	N/A
Instant feedback / situational awareness received during exercise useful for my learning process.						
The alerts and information provided were accurate and sufficient for my learning process.						
The situational awareness tools provided in the training were easy to use (even without earlier familiarity with the tools).						
Having instant feedback/situational awareness had positive impact on team communication and cooperation.						
The situational awareness tools and monitors distracted me from doing my actual work and I therefore performed worse than without those tools.						
In comparison to the other RT exercises without instant situational awareness provided, I learned more effectively.						
In comparison to the other RT exercises, I had to ask less information from						

Yellow Team members, as I got relevant situational awareness information via Frankenstack screens myself.

I had sufficient help for my learning process, i.e. when I got stuck and did not know how to proceed my team members or sub-teams leaders provided guidance.

8. My skills have improved as a result of this training:

0%	1-10%	10-25%	25-50%	51-75%	76-100%

TO1: in evidence gathering and information analysis for technical attribution

TO2: in executing responsive cyber defence scenario for target information system infiltration (identifying the origins of malicious activities and stopping those)

TO3: in stealthy execution and attack approaches; evaluating special execution tactics applicability for fast paced operations

TO4: in working as a united team in achieving the laid out mission objectives (attribution evidence gathering and malicious service takedown)

TO5: in red teaming skills needed for target information system takeover (client side targeting, web based attacks, malware and system exploitation, network and service based attacks)

9. Overall I am satisfied with the exercise:

Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree

Non-exclusive licence to reproduce thesis and make thesis public

I, Kaie Maennel (date of birth: 25th of April 1976),

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to:

1.1 reproduce, for the purpose of preservation and making available to the public, including for addition to the DSpace digital archives until expiry of the term of validity of the copyright, and

1.2 make available to the public via the web environment of the University of Tartu, including via the DSpace digital archives until expiry of the term of validity of the copyright,

Improving and Measuring Learning Effectiveness at Cyber Defence Exercises

supervised by Rain Ottis, Liina Randmann and Raimundas Matulevičius

2. I am aware of the fact that the author retains these rights.

3. I certify that granting the non-exclusive licence does not infringe the intellectual property rights or rights arising from the Personal Data Protection Act.

Tallinn, 24.05.2017