frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 20 September 2022
DOI 10.3389/feduc.2022.958405

Check for updates

OPEN ACCESS

EDITED BY
Basel Katt,
Norwegian University of Science and Technology, Norway

REVIEWED BY
Pinaki Chakraborty,
Netaji Subhas University of Technology, India
Enrico Russo,
University of Genoa, Italy

*CORRESPONDENCE
Sten Mäses
sten.mases@taltech.ee

SPECIALTY SECTION
This article was submitted to Higher Education, a section of the journal Frontiers in Education

RECEIVED 31 May 2022
ACCEPTED 23 August 2022
PUBLISHED 20 September 2022

CITATION
Mäses S, Maennel K and Brilingaitė A (2022) Trends and challenges for balanced scoring in cybersecurity exercises: A case study on the example of Locked Shields.
Front. Educ. 7:958405.
doi: 10.3389/feduc.2022.958405

COPYRIGHT
© 2022 Mäses, Maennel and Brilingaitė. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Trends and challenges for balanced scoring in cybersecurity exercises: A case study on the example of Locked Shields

Sten Mäses[1]*, Kaie Maennel[1] and Agnė Brilingaitė[2]

[1]Department of Software Science, Centre for Digital Forensics and Cyber Security, Tallinn University of Technology (TalTech), Tallinn, Estonia, [2]Institute of Computer Science, Vilnius University, Vilnius, Lithuania

Cybersecurity exercises (CSXs) enable raising organizational awareness, testing capabilities, identifying strengths and weaknesses, and gaining hands-on practice in building resilience against attacks. Typical CSX execution is designed as a competition or a challenge with gamification features to increase participant engagement. Also, it requires a significant amount of human resources to ensure up-to-date attack simulation and proper feedback. The usual concerns related to CSXs are how many points the team or participant received and the reason behind the particular evaluation. Properly balanced scoring can provide valuable feedback and keep CSX participants engaged. An inadequate scoring system might have the opposite effect—spread disorder, cause discontent, decrease motivation, and distract the participants from the event's primary goal. Combining both technical and soft sides in CSX makes it increasingly complex and challenging to ensure a balanced scoring. This paper defines scoring challenges and trends based on the case study of one of the largest international live-fire cyber defense exercises, Locked Shields (LS). It reviews the CSX scoring categories of the recent LS executions and provides the most common participant concerns related to scoring. The feedback shows that clarity and transparency of the scoring system together with providing feedback and justification to the scores are one of the top concerns. The design choices of the scoring system are explored to demonstrate the subtle variations of balanced category scoring and make a basis for future discussions. The chosen contrast and comparison approach enabled distinguishing four parameters for design decision categories: complexity, transparency, level of competition, and automatization. The research results demonstrate that learning facilitation requires system simplification and decisions regarding trends of the scoring curve. Even though transparency is a critical issue, concealing some scoring logic details can ensure more flexibility during the event to stimulate participants, support learning experiences, and cope with unexpected situations. Time as a central dimension enables the implementation of complex scoring curves for automated assessment. Our

study contributes to the community of higher education institutions and all organizers of cybersecurity challenges for skill development and assessment.

KEYWORDS

cyber defense exercises, cybersecurity training, scoring system, capacity building, team performance, incident response, performance feedback, cybersecurity exercise design

1. Introduction

Cybersecurity should be a priority in all society and economy domains to build resilience against emerging cybersecurity risks in digital economies influenced by expanding ubiquitous technologies (World Economic Forum, 2022). Globally, organizations experience a shortage of cybersecurity-related workforce and skills. Cybersecurity exercises (CSXs) are becoming increasingly popular in developing organizations and community cyber-resilience as a complex tool that reflects incident response situations.

The European Union Agency for Cybersecurity (ENISA) (Nurse et al., 2021) promotes cybersecurity skill-building challenges and competitions to address the skills shortage because they additionally increase interest in choosing a cybersecurity career. During defense-oriented CSX, intricate cyber ranges are constructed to provide realistic and holistic experiences to the participants. Often, some scoring system is implemented to maximize the learning achievements and keep the exercise participants engaged. Scoring enables rapid feedback which is a great motivator and essential for optimal learning (Chou, 2019).

A simple scoring system could count only the completed tasks or measure relevant times (Mäses et al., 2017). Defenders reacting quickly and attackers compromising a system fast can get extra points. There is a wide range of task types in large-scale exercises, and it makes sense to consider some tasks more critical. However, a linear scoring model might feel unrealistic. Additionally, some participants might not be motivated by points but aim to achieve specific learning outcomes (Cheung et al., 2012).

There is no universal solution to achieving a harmonious scoring balance, assessing participants objectively, and keeping every exercise participant happy. However, over the years of experience, CSX organizers have noticed some repeating patterns regarding scoring challenges. Discussing those challenges can hopefully facilitate the development of more balanced, practical, and enjoyable CSXs.

This paper aims to formalize CSX design aspects related to scoring. It presents the case study based on the authors' experience in Locked Shields (LS)—a large-scale CSX organized by the NATO Cooperative Cyber Defence Centre of Excellence (NATO CCDCOE) and gathering more than 2,000 experts and 5,000 virtual systems.[1]

The case study methodology includes narrative inquiry gathering, feedback analysis, and generalizing expert observations. During the case study, we address three research questions with respect to the paper's goal:

- What are the current trends and challenges relating to scoring in CSX?
- What are the factors to ensure a balanced scoring in CSXs?
- How can these factors be applied in practice for a CSX?

The discussion includes an overview of scoring categories, technical implementation specifics, and mathematics-based insight into the aggregated scoring. This paper contributes to the community of higher education institutions and all organizers of cyber challenges to consider workforce cyber-capacity building—skill development and assessment.

The paper is structured as follows. Section 2 covers related work and presents the CSX case study as a research approach. Section 3 explores scoring balancing factors identified during the case study. Section 4 discusses challenges when supporting trainees with valuable feedback and keeping them engaged. Section 5 concludes the paper.

2. Materials and methods

This section starts with an overview of the research approach and methodology applied. Afterward, the previous work related to CSXs and their scoring strategies is discussed. The LS exercise is then described as a single-case study.

2.1. Research approach and methodology

This paper investigates and formalizes the CSX design aspects impacting scoring, typically chosen and implemented based on intuition or experience and commonly not explained

1 https://ccdcoe.org/exercises/locked-shields/

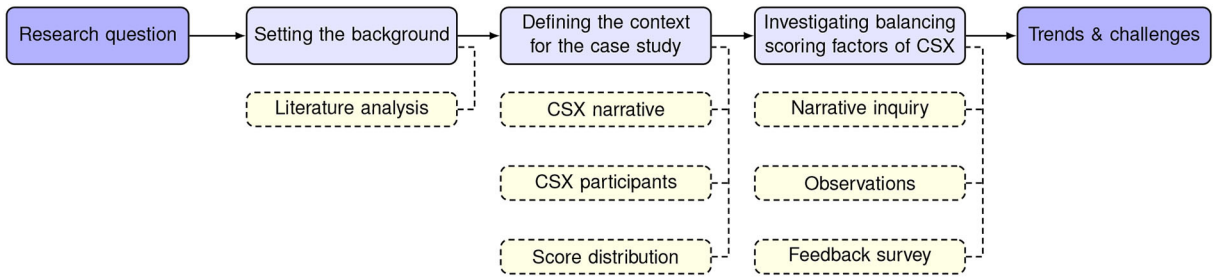

FIGURE 1
Research approach and methodology.

in more detail. We follow an overall research approach and methodology as shown in Figure 1.

The literature analysis sets the background to answer research questions. Then, the case study approach was chosen for exploring processes or behaviors that are new or little understood (Hartley, 1994). The case study format also helps to focus on descriptive and explanatory questions (Yin, 2012) about the scoring approaches in CSXs. The triangulation is done by other methods, such as narrative inquiry, observations, and feedback survey analysis.

We used narrative inquiries for gathering various events and happenings from the organizers (including the authors) and participants of the CSX and using narrative analytic procedures to provide explanatory stories (Polkinghorne, 1995). In addition, we applied observational methodology as "one of the most suitable research designs for evaluating fidelity of implementation, especially in complex interventions" (Portell Vidal et al., 2015). The observations were conducted by actively participating in the organization of LS CSX. The authors have several years of experience in evaluating numerous reports written by the participating trainee teams, designing and updating the scoring system, and gathering feedback about CSX.

A commonly used pretest-posttest approach for measuring change was not used as the access to the participating teams before and after the exercise was limited and repeating a complex set of hands-on tasks was not feasible. A survey was conducted to collect feedback from the target training audience. The designed question set combined closed and open questions to gather more in-depth and comprehensive feedback. Specifically, free-form comments raised on *scoring* or *points* were used as part of this research to validate the data from narrative inquiries and observations obtained from the organizers.

The data collected was analyzed and synthesized (clustered) to define a list of dominant themes. The analysis consisted of comparing and contrasting to establish the similarities and the differences (Macfarlane, 2006) within the scoring balancing factors to identify possible trade-offs for specific cases and approaches. The case study of LS CSX demonstrated the application of the identified factors, including implementation and design challenges. Based on the analysis results, the conclusions were drawn to answer the raised research questions and justify the scoring balancing factors and their applicability in practice for CSX.

2.2. Related work

Recent research presents various scoring systems at a high level, typically described as motivational, linked to learning objectives, and assisting in monitoring performance, progress, and feedback (Patriciu and Furtuna, 2009; Čeleda et al., 2015; Çalişkan et al., 2017; Maennel et al., 2017; Vykopal et al., 2017, 2018; Seker and Ozbenli, 2018; Ernits et al., 2020; Mäses et al., 2021). Due to CSX complexity, a single score or measurement cannot capture every ability learned, and different learning objectives call for different scoring approaches (Andreolini et al., 2020). However, as the CSX scoring consists of various elements, often the research focuses only on a specific element or its technical solution(s), such as availability scoring (Hempenius et al., 2019; Pihelgas, 2019).

The past research mainly focuses on the Capture-the-Flag (CTF) scoring rather than more complex scenario-based blue-red team exercises. CTFs often use a weighted system distributed among confidentiality, integrity, and availability (Werther et al., 2011) that can also be used as a guiding principle behind functional CSX scoring. However, the CSX scoring system includes more parameters due to team aspects, the magnitude of the cyber range, and elaborate holistic scenarios that need more sophisticated scoring principles. Based on the current research, Koutsouris et al. (2021) distinguish a list of specific measurements and performance indicators, including the quantity of successfully mitigated attacks and information sharing quality. They review scoring rubrics, e.g., incident reporting, and suggest a dimension-wide performance score that depends on parameters, e.g., session and dimension to compute a total score for all dimensions—average, weighted, and root mean square scores.

Inspired by attack trees and attack graphs, Andreolini et al. (2020) have developed a scoring approach based on graph operations to evaluate the trainee's performance during an exercise. This study suggests looking at the shortest path (speed) and symmetric difference (precision) from a reference graph. Those will sum up to aggregated scores to aid in the construction of more elaborate trainee evaluation models. The authors emphasize scalability, fault tolerance, and ease of use in implementing such a scoring system. However, there is no "correct" reference graph for novel attack vectors. Furthermore, the values are not justified but can be modified according to the specific learning objective, and measuring speed may hinder stealthiness. Also, the team aspects are not covered at all.

Diakoumakos et al. (2021) provide a scoring system agnostic to the data sources in the federated cyber range. It uses pre-specified metrics to measure trainee's performance under a scenario, and various defined indicators enable measuring a participant's performance through methods, tools, and metrics. The authors define the scenario in a network topology diagram. To achieve the action's goal, the user typically needs to perform a series of tasks, with each task including several steps. Depending on the complexity, each task is assigned a difficulty level: easy, medium, hard, extreme, or very hard. However, the authors do not go deeper into the reasoning. Normalization of the scores is left as an arbitrary choice of the exercise designer, and no insight or holistic approach is given for deriving or reasoning the overall CSX scoring logic.

Overall, the designer of the scoring systems for CSXs should follow a core principle that an understanding of "the differences in approach from competing in an event vs. designing, building, administering, and scoring an event offers a deeper insight into the actual goals of the event" (Mauer et al., 2012), and it provides accurate and validated feedback or evaluation. Therefore, it is crucial to focus on reflective scoring instead of purely numerical scoring (Weiss et al., 2016), i.e., focusing on factors and understanding why something happened while assessing multiple types of knowledge. An overall scoring system should support a more comprehensive competence evaluation in CSXs (Brilingaitė et al., 2020).

Relatively few studies have objectively validated that more interactive simulations and competitions are associated with the higher scored performance of a cyber defender, i.e., are more efficient and evidenced by scored results (La Fleur et al., 2021). The recent work provides some novel approaches but focuses only on selected measures (speed, precision, and difficulty) that do not give the reasoning for all aspects of CSXs scoring. It is essential to understand the bigger picture of the various performance aspects (including the recent trends and challenges) and to have a holistic scoring concept that links to competencies and performance.

2.3. Context of LS

NATO Cooperative Cyber Defence Centre of Excellence organizes annually one of the largest live blue-red team CSX called LS (Maennel et al., 2017). The naming of this defense-oriented exercise is inspired by the military formations where connecting shields enables the unit to be stronger than individual soldiers. LS aims to provide a realistic high-stress environment where each participating team is forced to collaborate among themselves and coordinate actions with other teams.

The exercise scenario concerns two fictional island countries, Berylia and Crimsonia, which are located in the northern Atlantic Ocean. Berylia is facing increasing tensions with neighboring Crimsonia. Those tensions lead to several effects, including cyber-attack campaigns targeting the critical infrastructure of Berylia and information warfare in simulated social media platforms. The blueish mineral *beryl* and *crimson* color inspired names are Berylia and Crimsonia. Therefore, blue teams (BTs) represent rapid-response teams (RRT) to assist Berylia in handling cyber incidents, while the red team (RT) performs attacks on behalf of Crimsonia.

Blue teams are the primary training audience. They are presented with networks of the same topology representing systems in different areas of the fictional Berylia. The machines vary from simple office workstations to complex cyber-physical systems. BTs have 1 day to get acquainted with their network, followed by 2 days of cyber-attacks, each consisting of two attack phases. It has to keep different services running, mitigate the impact of cyber-attacks, form regular reports about the situation, solve forensic challenges, and deal with additional tasks received in time-dependent assignments—game injects. Additionally, some BT subnets form the interconnected network of the critical infrastructure. Therefore, BTs have to collaborate with neighboring teams to provide services, e.g., open connections for power distribution.

Other teams are considered exercise organizers and do not compete with each other. However, they can also learn a lot from the exercise. The RT is conducting coordinated attacks testing the defenses of the BTs. The Yellow team (YT) deals with general situational awareness and reporting—which evaluates reports and their correspondence toward the situation and existing threat intelligence. The Green team (GT) develops the core infrastructure and manages it during the exercise. The White team (WT) deals with exercise control and consists of several sub-teams dealing with specific areas such as media, legal, and user simulation. Part of the WT deals with special investigations and assigns special bonuses and penalties to ensure that all the teams are treated equally. A bonus score could be given when one BT has gathered a negative score due to issues with the core infrastructure that is not under their control. A particular negative score could be assigned when one BT is found to be using technical measures to trick the scoring system, e.g., using

a dummy service satisfying automatic checks but not providing a usable service (Pihelgas, 2019).

All BTs start the exercise with an initial score budget. The LS scoring aspects cover technical and non-technical categories, as presented in Figure 2 (approximate weight distribution in 2022).

Technical categories focus on the attack, defense, machine administration, technical end-user support, monitoring computer networks, and digital forensics. Non-technical categories focus on the so-called soft side requiring oral and written communication skills. Note that, despite the naming, the non-technical categories can get quite technical. For example, BTs submit technical reports and perform complex legal analyses.

Locked shields has a strong emphasis on technical capabilities. Therefore, the score distribution is usually 65–75% and 25–35% for technical and non-technical tasks. Each year the organizers define the specific category weight schema. In addition to the schema, the organizers also subtract points for machine reverts, assign special bonuses and penalties, and execute scenario injects that have a minimum impact on the result. The latter aspects are not the paper's focus as they are more gamification details than association with teams' capabilities.

As it can be seen from Figure 2, the total score for uptime of services and the total score for attacks are equal. The reasoning behind this is that the score for the uptime of services should be at least as big as the total score for defending against the attacks to discourage the strategy of taking the systems offline to protect them. During the LS, the attacks usually target the confidentiality and integrity of the BT systems first and leave the attacks against the availability to the very final stages of the exercise. Therefore, the loss of uptime of services is usually due to overly protective measures by the BTs.

2.3.1. Uptime of services

Category *uptime of services* includes *availability* and *usability* scoring sub-categories. Their score is updated continuously throughout time. Availability deals with automated checks. For each discrete game time tick, the running processes check the functionality and setup of the BT network, considering machines and their importance. The scoring-bot subtracts budget points for unavailable services, for example, closed ports are required for essential functionality (Pihelgas, 2019).

The user simulation team (UST) performs usability checks with actual humans simulating the end-users of the systems. UST is part of WT. It maintains a list of capabilities to check regularly in a synchronized manner. For functionalities inaccessible to users, tickets are opened with a request to solve the problem. The longer the ticket stays open, the more points are lost.

2.3.2. Attacks

Category of *attacks* covers three sub-categories—*web*, *network*, and *client-side*. Web attacks target application-layer vulnerabilities of the application layer, for example, poorly sanitized input fields on a web form. Network attacks focus on the network layer attacks, for example, taking advantage of poor IPv6 rules in firewalls. Client-side attacks simulate the threats connected to the human factor.

Red team collaborates with UST to enable factors required for its attacks, primarily client-side. It asks UST to carry out specific activities, for example, uploading a picture to a particular website or executing a file on a workstation. For simplicity, the scenario assumes that humans are successfully tricked or convinced to carry out their actions. Social-engineering-based motivational aspects of simulated users are not practiced during the LS exercise.

2.3.3. Forensics

Locked shields organizers execute the *forensics* part using the CTF system. BT forensics sub-teams use digital forensics tools to perform incident investigations. They analyze the data files and find evidence of threat actor behavior, for example, find traces of malicious activity in provided evidence files. The participants submit answers and findings into the system that automatically evaluates most of the tasks.

2.3.4. Reporting

The purpose of reporting is to inform managers, higher-rank officers, and decision-makers at the Cyber Operation Centre about the cyber situation, including threats, the status of capabilities, key events, threat actors, and technological challenges. Scoring category *Reporting* involves *SITREP* and *CTIREP*—situation reports and cyber threat intelligence reports. The pieces are submitted once per exercise phase and evaluated manually at discrete time points.

SITREPs should correlate with registered key events and the team's posture regarding the defense of the infrastructure. BTs build CTIREPs, referencing incidents recorded in the information-sharing platform (MISP Project, 2022) to provide evidence and support assessment of the risk as attack implications.

2.3.5. Legal and media

Blue teams demonstrate their capacity to manage public relations and solve legal problems related to cyber incidents in critical infrastructure. They try to avoid hasty steps that would potentially amplify the crisis.

Injects of the *Legal & Media* category require timely answers to questions that consider the scenario and provided material. The questions are neither trivial nor solved straightforwardly.

10.3389/feduc.2022.958405

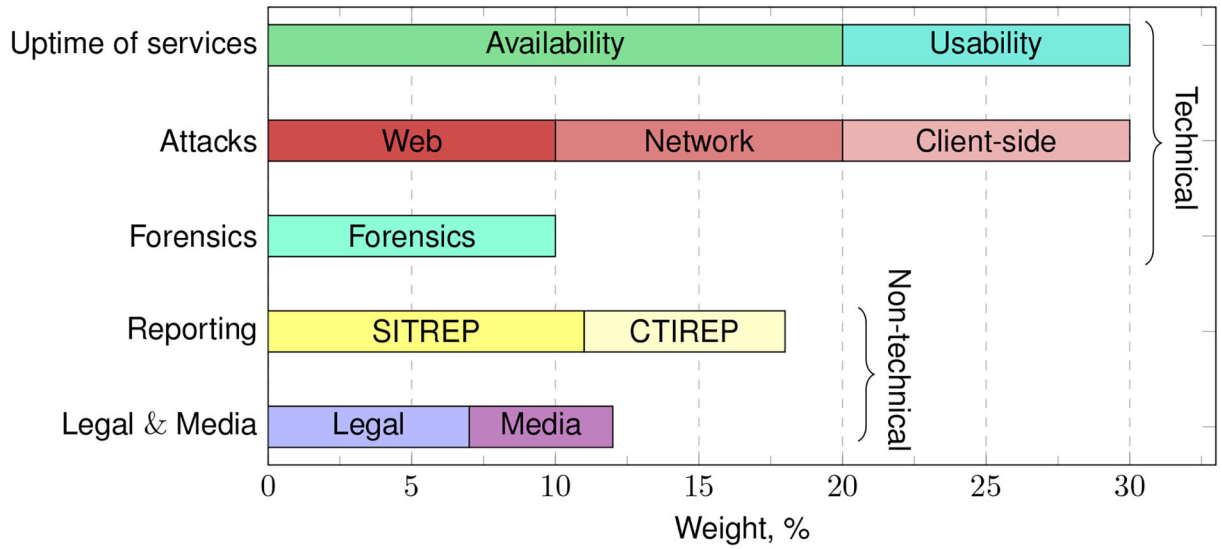

FIGURE 2
Weights of scoring categories.

Therefore, BTs should address problems and argument decisions based on the guidelines considering legislative documents and communication aspects. Additionally, media teams should work appropriately on social media and participate in live interviews if needed. The injects are evaluated manually at specific time points.

3. Results

First, this section summarizes the target training audience (BTs) feedback. Afterward, the synthesized themes are described in the example of LS.

3.1. Feedback from the training audience

In 2022, LS participants were asked to submit feedback, as described in Section 2. Overall, 93% of respondents would recommend LS to their colleagues, the most critical being BTs and especially sub-team leaders. Therefore, the feedback showed dominant positivism about the exercise, including the scoring aspects.

We examined the responses of BTs in more detail to compile the main themes related to scoring (238 responses; app. 10% response rate). Specific feedback regarding scoring emerged from the open question: *"What did not meet your expectations? Room for improvement, etc."* 13% of respondents explicitly commented on the scoring system or points in the free-style open-ended comments.

TABLE 1 Locked shields (LS) scoring feedback themes.

Theme	Percentage
Clarity/Transparency	42
Feedback/Justification for scoring	26
System Set-up/Technical issues	16
Visibility/Comparability to others	6
Proportionality	3
"Game-ism"	3
Visualization	3

Table 1 summarizes the emerged generalized themes. The main concerns related to scoring were clarity and transparency (mentioned by more than 40% of respondents) and providing feedback and justification (mentioned by more than 25% of respondents) to support the learning experience and satisfaction.

The transparency and well-justified feedback are linked to learning and improving performance, with some examples as follows:

- "Not all the reviews of the injects were well argumented. And to train and do better, we need those points."
- "More clear and faster feedback regarding scoring so one get feedback on what one do right and what one do wrong."
- "There was insufficient feedback in general—this is supposed to be a training exercise, not a game to be scored."
- "I believe points and marks should be better explained."

3.2. Balancing factors

3.2.1. Complexity for realism vs. simplicity for learning

Locked shields participants often wish for a more transparent and straightforward system to easily interpret the score. At the same time, there is a continuous demand for realistic environments.

Different scoring categories could use different types of scoring scales and scoring functions. Some tasks or activities can be assessed using discrete functions. For example, in LS, situation reports are delivered once per phase. Therefore, evaluators provide scores at discrete time values.

Typically, in cyber defense exercises, some categories use continuous functions. One of the common examples is availability scoring. Usually, the predefined list of services is checked once per period using a fine time granularity, e.g., 2 min. Therefore, there is a change in the total availability score per time tick. Each tracked service's (un-)availability can represent a different value due to its importance.

The change in score according to the service uptime could be linear, exponential, or more complex. A non-linear approach has been chosen for the uptime score multipliers in LS. Before 2021, an exponential curve was used in LS (Pihelgas, 2019), and then Richard's curve (generalized logistic function) was selected. The Richard's curve was chosen due to its S-shape, which is hypothesized to be characteristic of the perceived value of a service uptime: first, the users might not mind much; then, after some time, their patience runs out, and the frustration grows fast; finally, they do not care much whether the service is up 5% of the time or 10% of the time—it is still essentially unusable. Richard's curve is also flexible to be modified to fit any specific needs. Still, more research is needed to determine whether the choice of Richard's curve is realistically characteristic of real-life user perception. Nevertheless, Figure 3 illustrates different curves and Richard's curve variations in Figures 3A,B, respectively.

In LS, reports and legal and media tasks are evaluated using a scale of 0–12. Each task has several grading aspects, each graded using a scale. During the LS design phase, different evaluating teams wanted different scales—a 2-point, 3-point, or 4-point Likert scale. To unify the approach, a 12-point scale was chosen to accommodate all the needs mentioned above. A team wishing to use a 3-point Likert scale can give out grades of 0, 6, and 12. A team wishing to use a 4-point Likert scale can give out grades 0, 4, 8, and 12.

3.2.2. Transparent vs. obscure

If a team finds a loophole in rules that goes against the general exercise training goals, then in LS, the organizing team has the right to assign *special points* to keep the game fair. For example, teams could restrict access to the infrastructure for the UST ensuring good service availability (gaining points) but making the systems nonfunctional to users (maybe limited access will stay unnoticed).

Although for learning purposes, it is beneficial to have the ability to analyze the score in-depth, having already some categories can tell us interesting information about the teams.

Figure 4 shows how three BTs with a similar final score can have somewhat different performance profiles (a higher score indicates better performance). Teams T_1 and T_3 have demonstrated significantly higher performance in ensuring service uptime (availability and usability) compared to T_2. At the same time, they demonstrate less success in keeping the RT attacks contained. Team T_2 has been much more successful in protecting against the RT, but their service uptime has also suffered (most likely as a result of their defensive activities).

3.2.3. Learning and training vs. competition

Locked shields is a mix of both training and competition. It has strong competition influences, and the best team is brought out publicly each year. On the other hand, participating BTs have reported considerable learning from experience.

Nevertheless, BTs want to know how they compare to other teams. This case was demonstrated during LS 2021 when organizers tried to focus more on individual feedback and not share the general rankings of the BTs. As a reaction, most BTs published their total scores to the general chat making it possible to form a ranking table still.

3.2.4. Manual vs. automated

Locked shields scoring process is a hybrid of automated scripts and manual work. Availability checks are automated, but a separate UST does service uptime checks. The combination of automated checks and human checks has worked out quite well. Automated checks are scalable and provide the general picture. Manual checks enable finding the teams that might want to try tricking the game by creating mock systems only to satisfy the automated checks but not the actual users. Manual checks also enable checking more complex systems where it might not be feasible to develop reliable automated checks.

In the last 2 years, LS has taken steps to segregate evaluation grades from actual scores. In earlier years, manual evaluations had different grading schemes and could result in scores such as 1,600 or 350, but only a more profound analysis could reveal whether a higher score was due to good performance or the high importance of the task. For example, a score of 1,600 out of 3,200 could indicate a much lower performance than 350 out of 360 points. At the same time, there is still the need for evaluating the performance considering the priority systems. Achieving only half of the points from a high-value task while succeeding in lower-value tasks could indicate problems with prioritization in the team or a lack of some specific competence.

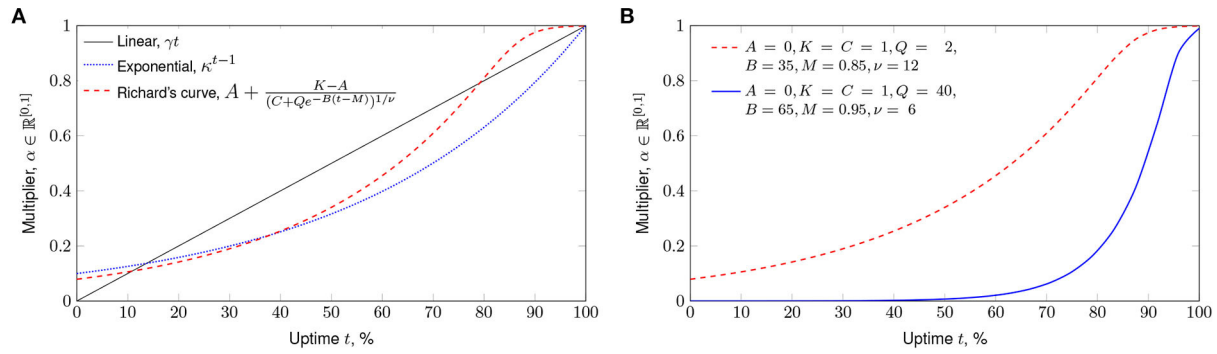

FIGURE 3
Different algorithms for scoring uptime. **(A)** Different curve types. **(B)** Two Richard's curves.

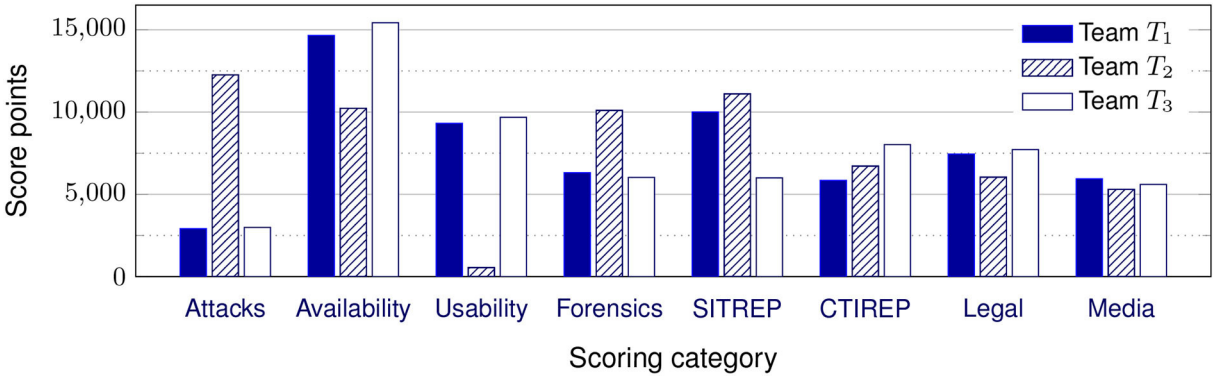

FIGURE 4
Comparing three teams with very similar total scores.

Now, LS implements a unified 12-point grading scale for manual grading. Each grade is then automatically calculated into points according to the importance of the particular grading aspect, task, and category. This approach enables later analysis of the grades separately from the final score (grades multiplied by their importance).

In LS, a more comprehensive overview is aggregated and shared after the exercise in the after action report. Additional comments about different exercise aspects are also communicated during the hot-wash sessions, where the representatives of different teams present their feedback.

4. Discussion

This section analyses and further discusses previously described aspects of the LS exercise design. Each of these aspects requires finding the right balance between extremes in the scoring process and implementation considerations specific to the exercise.

4.1. Complexity for realism vs. simplicity for learning

Some aspects of the exercise are more intense or simplified to facilitate learning. A simple system is more reliable and easier to interpret. At the same time, the lessons learned should be transferable to real situations. Therefore, a more complex and realistic approach seem beneficial. The feedback from the participants confirms this controversy.

Usually, exercise participants or organizers prefer to know the actual score at any time during the game. Even though some categories are based on discrete evaluations, the total accumulative score is a continuous curve within the time intervals of the game phases.

Figure 5 illustrates the vital aspects of the design of the scoring curve. The x-axis represents the time of the game phase, and the y-axis represents points (score). The functions could represent a single category or the total score. First, the curve can be a decreasing or increasing one (see Figures 5A,B). Even though the choice does not impact the participant, it might

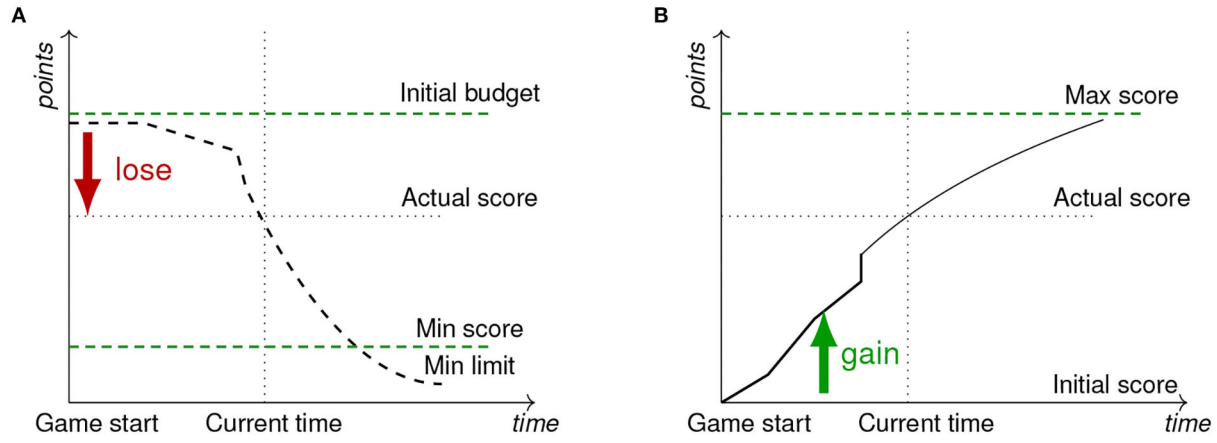

FIGURE 5
The increasing and decreasing score functions curves. **(A)** A decreasing curve. **(B)** An increasing curve.

present the game focus and attitude. The increasing curve means participants gain points for a good solution and action. The better the result is, the more significant points are gained. It is similar to a regular assessment in the education environment when the assessment starts with 0 and ends with some results showing skill level—a positive aspect. The team starts with an initial score and could reach the maximum possible score. There are several different ways to score the uptime of services. Usually, it makes sense to have regular automated checks that either fail or pass. Every passing check can give the uptime (availability) scoring category bonus points. Alternatively, every failing check can give negative points. The team starts with an initial budget, and the scoring curve approaches the minimum limit during the event. While mathematically, those approaches can be considered equal, several studies argue that loss aversion tends to motivate people more than potential gains (Gal and Rucker, 2018).

The minimum score could be a political decision. The figure emphasizes the minimum limit based on the mathematical function limits, and the Min score indicates the minimal possible number of points (no possibility of reaching lower values).

Choosing the grading scale to evaluate non-technical injects is also a challenge. A simplistic approach would be to use a binary scale depending on whether a task was completed successfully. Nevertheless, the strategy should support more nuances when more complex tasks and several grading aspects are considered.

4.2. Transparent vs. obscure

Cybersecurity exercise scoring includes gamification features to stimulate participants' engagement. Every player's dream is to know the game rules in detail. At the same time, knowing all the parameters is not realistic. Strict, specific, and transparent rules enable a more objective performance evaluation. At the same time, more general rules can provide the required flexibility to cope with unexpected situations.

4.3. Learning and training vs. competition

Some CSXs are undoubtedly defined as competitions with awards for the best performers. Other exercises are used in a purely educational context where the main emphasis is on learning. A crisis response team should be able to operate under stress. When optimizing learning, the different tasks could be modified during the exercise according to the performance of the BT. From the competition perspective, it is essential to have strict and fixed rules and settings so that the comparison of different teams is fair and as objective as possible.

The time dimension plays a crucial role in exercise scoring. First, time is used to set up deadlines to support the gameplay, with external factors increasing the stress and realism of the scenario. Second, time can be the dimension to ensure the trend change of the scoring curve of the category.

There are tasks to be submitted before the assigned deadline during the exercises. For example, the team has to submit the solution to the scenario inject by providing a screenshot of some SCADA system to prove it is in a proper state before the planned *maintenance*. Therefore, the organizers must ensure the deadlines are considered in the scoring strategy.

The team's ability to submit assigned tasks in time can be solved in several ways. The easiest way is to give 0 points to those who are late. But this approach does not show the team's capacity to report or understand the SCADA system's internal workings. The team might have problems with time management, task sharing, and coping with stress. Still, nothing

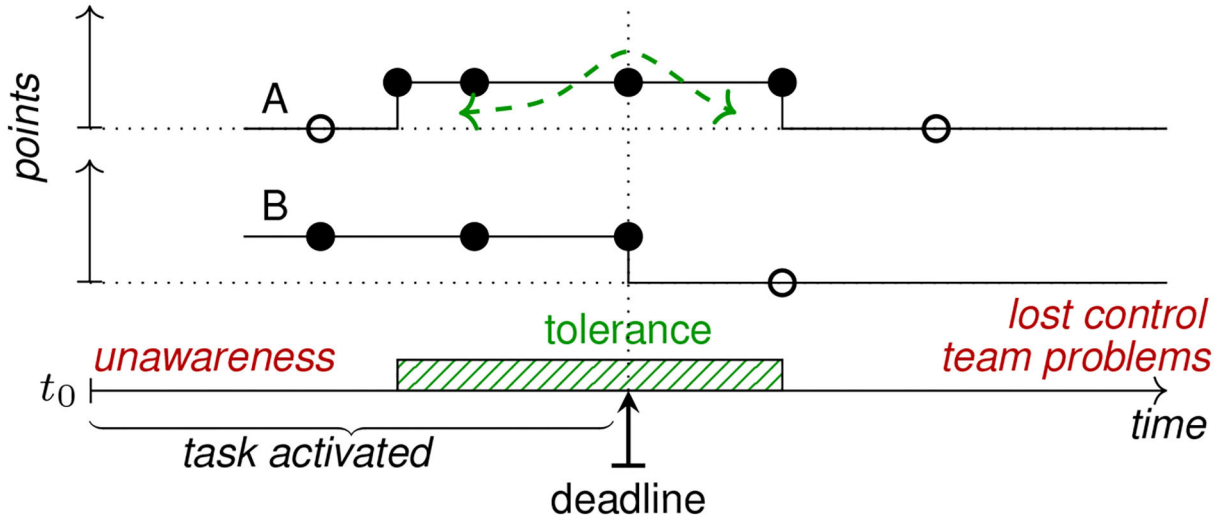

FIGURE 6
Task scoring along the time line.

can be said about sharing qualitative information with decision-makers or colleagues in the responsibility chain.

During the CSX event, the BT is considered a rapid (or incident) response team. In actual circumstances, the delay in submitting the task could be appropriate. For example, the team restores the services and checks service availability based on procedures. At the same time, the team lead is requested to report on the team status and attack impact on the infrastructure. The lead might be late by 5–10 min to provide the attack assessment. Therefore, during the exercises, the tolerance interval also could be supported. At the same time, the opposite situation occurs when the team submits the task far in advance. Figure 6 illustrates two scoring strategies for meeting the deadline. The curves show the maximum possible points along the time dimension. Filled circles show points assigned, and empty circles represent no points scored. Case B represents the assessment during the task activation time with the maximum points possible, disregarding the submission time when the deadline is met. The strategy is easy to implement, as any submission (or team, more precisely) can be assigned 0 points after the deadline. Case A represents the tolerance shift of the deadline and considers the submission time—too early submission gets fewer (or 0) points. The case can be modified to support smooth assessment curves, as shown by the dashed two-sided arrow.

Early reports or system views would mean outdated information that cannot be trusted to make a decision. Finally, the case when the team does not submit the report might signal an ineffective team and loss of control (as listed in the figure at extreme points on the time-axis).

In actual circumstances, the rotation of the lead and adding more staff might be needed. However, in the game, it could be the issue of task prioritization. If the exercise goals focus on the task with a deadline, the scoring should include specific extrinsic motivation to remind about the importance of the task.

4.4. Manual vs. automated

Timely feedback is excellent to keep participants engaged and helps them learn fast as they can understand the impact of their actions on the score. At the same time, it is challenging to provide speedy and objective feedback when the process is not automated. Therefore, the people who manually evaluate various reports must find the balance between the feedback speed and the level of detail.

Attackers have specific targets and apply automated tools but are still conducted by a live RT. Automated activities enable greater comparability and objective evaluation. The manual approach enables a more realistic (expert-oriented) and flexible system.

Red team members have a list of objectives, but treating all teams equally is impossible. Some RT members have more experience and can carry out more sophisticated attacks than others. The RT workshops can be conducted before the exercise to unify the RT members' skill levels, synchronize the attacks, and ensure comparability among the different BTs (e.g., such an approach is also practiced at LS). Still, some objectives might require significantly more effort to complete than others. The RT makes notes to be analyzed after the exercise to provide more

insights to BTs in the post-exercise stage as part of the after action report.

Additionally, there are several categories where having automated checking is not (yet) realistic. For example, evaluating a legal analysis and press releases contain multiple nuances requiring human experts to give justified grades presenting reasoned solutions. YT provides the assessment by checking correlations of artifacts in information sharing platforms and reports. The feedback speed and objectivity concerns are also somewhat connected to manual and automated checks, although they can be partly seen as broader topics.

4.5. Single case limitations

We use a single case (LS, see Section 2.3) approach as a primary data source for several steps in the research process. However, we recognize the limitation of the approach in generalizability and several information-processing biases (Eisenhardt, 1989). The triangulation was done by other methods, such as narrative inquiry, observations, and feedback survey analysis. However, when applying the conclusions to other CSXs, the specific exercise's relevant factors should be considered and balanced.

5. Conclusions and future work

It is challenging to construct a holistic exercise that involves the technical and non-technical areas connected in a meaningful way. It is much easier to form a set of tasks that follow the same theme (an island state is under a cyber attack) but are not strongly connected. Ideally, different parts of the exercise would be connected, requiring the BT specialists from different teams to communicate, which would require the injects to be created in close collaboration between the different sub-teams of the organizers. Also, this approach would require the evaluators (people grading the responses to the injects) to have good situational awareness of the exercise.

In this paper, we aimed to formalize the exercise design aspects that often are driven by intuition and not always well justified. Based on the analysis of the current trends and challenges, we derived and synthesized the factors that help to ensure a balanced scoring. We used a contrast and comparison approach to determine the range of decision challenges related to the factors impacting the scoring system. The categories covered include complex realism vs. simplistic learning, transparency, learning vs. competition, and manual vs. automated. The factors were practically used and discussed in the context of a large blue-red team CSX, LS.

There is no magic solution to achieving a harmonious scoring balance and keeping every participant happy. However, some repeating patterns regarding scoring factors have appeared from past exercises that can help the community of higher education institutions and all organizers of cybersecurity challenges for skill development and assessment. Balancing the scoring factors and making pedagogically justified design choices can ensure rapid objective feedback and improve the overall learning experience and motivation for effective learning in CSXs.

Although the scoring aspects of the general exercise design are likely relevant for any CSX, we acknowledge that different exercises can have vastly different requirements. Therefore, although the final exercise design choices can vary significantly for different exercises, we hope that discussing and sharing the experience of the specific design choices and challenges helps inspire more informed and conscious trade-offs in other CSXs.

Future work could look at the different ways to achieve a more robust connection of different parts of the exercise to encourage collaboration between different specialists. The challenge here is to achieve this without completely sacrificing the comparability of the results.

For example, there could be two teams, T_X and T_Y, with similar crisis communication capabilities but different technical capabilities: T_X has good skills in protecting the network while T_Y has significantly lower capability to ensure network security. In a strongly interconnected exercise, the team T_X might have fewer opportunities to demonstrate their crisis communication capabilities because they manage to avoid the crisis. The more complex an exercise is, the more difficult it is to balance different factors regarding the scoring. Thus, another possible direction is future education research on mapping the score and status to team capability or joint resilience level with clear indications of missing skills.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the participants or participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

Author contributions

SM, KM, and AB contributed to the conception and design of the study and wrote sections of the manuscript. SM provided the main insight into the scoring system of the LS exercise. KM was the main contributor to analyzing related work. AB created

the illustrating figures. All authors contributed to manuscript revision, read, and approved the submitted version.

Funding

This work was supported by the ECHO project which has received funding from the European Union's Horizon 2020 Research and Innovation Programme under the grant agreement no. 830943.

Acknowledgments

The authors thank the rest of the organizers of the LS exercise.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Andreolini, M., Colacino, V. G., Colajanni, M., and Marchetti, M. (2020). A framework for the evaluation of trainee performance in cyber range exercises. *Mobile Netw. Appl.* 25, 236–247. doi: 10.1007/s11036-019-01442-0

Brilingaitė, A., Bukauskas, L., and Juozapavičius, A. (2020). A framework for competence development and assessment in hybrid cybersecurity exercises. *Comput. Security* 88, 101607. doi: 10.1016/j.cose.2019.101607

Çalişkan, E., Topgül, M. O., and Ottis, R. (2017). Cyber security exercises: a comparison of participant evaluation metrics and scoring systems. *Strategic Cyber Defense* 48, 180. doi: 10.3233/978-1-61499-771-9-180

Čeleda, P., Čegan, J., Vykopal, J., and Tovarňák, D. (2015). "Kypo-a platform for cyber defence exercises," in *M&S Support to Operational Tasks Including War Gaming, Logistics, Cyber Defence* (Munich: NATO Science and Technology Organization).

Cheung, R. S., Cohen, J. P., Lo, H. Z., Elia, F., and Carrillo-Marquez, V. (2012). "Effectiveness of cybersecurity competitions," in *Proceedings of the International Conference on Security and Management (SAM)*, (Las Vegas, NV: Steering Committee of The World Congress in Computer Science; Computer Engineering and Applied Computing), 5.

Chou, Y.-K. (2019). *Actionable Gamification: Beyond Points, Badges, and Leaderboards*. Milpitas, CA: Packt Publishing Ltd.

Diakoumakos, J., Chaskos, E., Kolokotronis, N., and Lepouras, G. (2021). "Cyber-range federation and cyber-security games: a gamification scoring model," in *2021 IEEE International Conference on Cyber Security and Resilience (CSR)* (Rhodes: IEEE), 186–191.

Eisenhardt, K. M. (1989). Building theories from case study research. *Academy Manag. Rev.* 14, 532–550. doi: 10.2307/258557

Ernits, M., Maennel, K., Mäses, S., Lepik, T., and Maennel, O. (2020). "From simple scoring towards a meaningful interpretation of learning in cybersecurity exercises," in *ICCWS 2020 15th International Conference on Cyber Warfare and Security* (Norfolk, VA: Academic Conferences and publishing limited), 135.

Gal, D., and Rucker, D. D. (2018). The loss of loss aversion: Will it loom larger than its gain? *J. Consum. Psychol.* 28, 497–516. doi: 10.1002/jcpy.1047

Hartley, J. F. (1994). "Case studies in organizational research," in *Qualitative Methods in Organizational Research: A Practical Guide* (London), 208–229.

Hempenius, N., Chou, T.-S., and Toderick, L. (2019). "Automatic collection of scoring metrics in competitive cybersecurity lab environments," in *2019 Conference for Industry and Education Collaboration, CIEC*(New Orleans, LA; Advances in Engineering Education), 10. Available online at: https://peer.asee.org/31523 (accessed May 7, 2022).

Koutsouris, N., Vassilakis, C., and Kolokotronis, N. (2021). "Cyber-security training evaluation metrics," in *2021 IEEE International Conference on Cyber Security and Resilience (CSR)* (Rhodes: IEEE), 192–197.

La Fleur, C., Hoffman, B., Gibson, C. B., and Buchler, N. (2021). Team performance in a series of regional and national us cybersecurity defense competitions: generalizable effects of training and functional role specialization. *Comput. Security* 104, 102229. doi: 10.1016/j.cose.2021.102229

Macfarlane, A. (2006). *To Contrast and Compare*. Irvine, CA: UC Irvine; Working Papers Series.

Maennel, K., Ottis, R., and Maennel, O. (2017). "Improving and measuring learning effectiveness at cyber defense exercises," in *Nordic Conference on Secure IT Systems* (Tartu: Springer), 123–138.

Mäses, S., Hallaq, B., and Maennel, O. (2017). "Obtaining better metrics for complex serious games within virtualised simulation environments," in *European Conference on Games Based Learning* (Graz: Academic Conferences International Limited), 428–434.

Mäses, S., Maennel, K., Toussaint, M., and Rosa, V. (2021). "Success factors for designing a cybersecurity exercise on the example of incident response," in *2021 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)* (Vienna: IEEE), 259–268.

Mauer, B., Stackpole, B., and Johnson, D. (2012). "Developing small team-based cyber security exercises," in *International Conference on Security and Management, SAM* (Las Vegas, NV), 5. Available online at: http://world-comp.org/p2012/SAM9727.pdf (accessed May 7, 2022).

MISP Project (2022). *MISP Open Source Threat Intelligence Platform & Open Standards For Threat Information Sharing*. Available online at: https://www.misp-project.org/ (accessed May 5, 2022).

Nurse, J. R., Adamos, K., Grammatopoulos, A., and Franco, F. D. (2021). *Addressing skills shortage and gap through higher education*. Technical report, European Union Agency for Cybersecurity (ENISA).

Patriciu, V.-V., and Furtuna, A. C. (2009). "Guide for designing cyber security exercises," in *Proceedings of the 8th WSEAS International Conference on E-Activities and Information Security and Privacy* (Puerto De La Cruz: World Scientific and Engineering Academy and Society, WSEAS), 172–177.

Pihelgas, M. (2019). "Design and implementation of an availability scoring system for cyber defence exercises," in *Proceedings of the 14th International Conference on Cyber Warfare and Security* (Stellenbosch), 329–337.

Polkinghorne, D. E. (1995). Narrative configuration in qualitative analysis. *Int. J. Qualit. Stud. Educ.* 8, 5–23. doi: 10.1080/0951839950080103

Portell Vidal, M., Anguera Argilaga, M. T., Chacón Moscoso, S., and Sanduvete Chaves, S. (2015). Guidelines for reporting evaluations based on observational methodology. *Psicothema* 27, 283–289. doi: 10.7334/psicothema2014.276

Seker, E., and Ozbenli, H. H. (2018). "The concept of cyber defence exercises (cdx): planning, execution, evaluation," in *2018 International Conference on Cyber Security and Protection of Digital Services (Cyber Security)* (Glasgow, UK; IEEE), 1–9.

Vykopal, J., Ošlejšek, R., Burská, K., and Zákopčanová, K. (2018). "Timely feedback in unstructured cybersecurity exercises," in *Proceedings of the 49th*

ACM Technical Symposium on Computer Science Education (Baltimore, MD), 173–178.

Vykopal, J., Vizváry, M., Oslejsek, R., Celeda, P., and Tovarnak, D. (2017). "Lessons learned from complex hands-on defence exercises in a cyber range," in *2017 IEEE Frontiers in Education Conference (FIE)* (Indianapolis, IN: IEEE), 1–8.

Weiss, R., Locasto, M. E., and Mache, J. (2016). "A reflective approach to assessing student performance in cybersecurity exercises," in *Proceedings of the 47th ACM Technical Symposium on Computing Science Education* (Memphis, TN), 597–602.

Werther, J., Zhivich, M., Leek, T., and Zeldovich, N. (2011). "Experiences in cyber security education: the MIT lincoln laboratory capture-the-flag exercise," in *4th Workshop on Cyber Security Experimentation and Test, CSET* (San Francisco, CA: USENIX Association), 12.

World Economic Forum (2022). *Global Cybersecurity Outlook 2022*. Insight Report. Available online at: https://www3.weforum.org/docs/WEF_Global_Cybersecurity_Outlook_2022.pdf (accessed May 25, 2022).

Yin, R. K. (2012). A (very) brief refresher on the case study method. *Appl. Case Study Res.* 3, 3–20. Available online at: https://www.amazon.com/Applications-Case-Study-Research-Robert/dp/1412989167