Control Effectiveness: a Capture-the-Flag Study

Arnau Erola
arnau.erola@cs.ox.ac.uk
University of Oxford
UK

Louise Axon
louise.axon@cs.ox.ac.uk
University of Oxford
UK

Alastair Janse van Rensburg
alastair.jansevanrensburg@cs.ox.ac.uk
University of Oxford
UK

Ioannis Agrafiotis
ioannis.agrafiotis@cs.ox.ac.uk
University of Oxford
UK

Michael Goldsmith
michael.goldsmith@cs.ox.ac.uk
University of Oxford
UK

Sadie Creese
sadie.creese@cs.ox.ac.uk
University of Oxford
UK

ABSTRACT

As cybersecurity breaches continue to increase in number and cost, and the demand for cyber-insurance rises, the ability to reason accurately about an organisation's residual risk is of paramount importance. Security controls are integral to risk practice and decision-making: organisations deploy controls in order to reduce their risk exposure, and cyber-insurance companies provide coverage to these organisations based on their cybersecurity posture. Therefore, in order to reason about an organisation's residual risk, it is critical to possess an accurate understanding of the controls organisations have in place and of the influence that these controls have on the likelihood that organisations will be harmed by a cyber-incident. Supporting evidence, however, for the effectiveness of controls is often lacking. With the aim of enriching internal threat data, in this article we explore a practical exercise in the form of a capture-the-flag (CTF) study. We experimented with a set of security controls and invited four professional penetration testers to solve the challenges. The results indicate that CTFs are a viable path for enriching threat intelligence and examining security controls, enabling us to begin to theorise about the relative effectiveness of certain risk controls on the face of threats, and to provide some recommendations for strengthening the cybersecurity posture.

CCS CONCEPTS

• Security and privacy → Systems security; Network security; Intrusion/anomaly detection and malware mitigation.

KEYWORDS

cyber threat detection, cyber threat prioritisation, control effectiveness, CTF

ACM Reference Format:

Arnau Erola, Louise Axon, Alastair Janse van Rensburg, Ioannis Agrafiotis, Michael Goldsmith, and Sadie Creese. 2021. Control Effectiveness: a Capture-the-Flag Study. In *The 16th International Conference on Availability, Reliability and Security (ARES 2021)*, August 17–20, 2021, Vienna, Austria. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3465481.3470095>

This work is licensed under a Creative Commons Attribution International 4.0 License.

ARES 2021, August 17–20, 2021, Vienna, Austria
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9051-4/21/08.
<https://doi.org/10.1145/3465481.3470095>

1 INTRODUCTION

Organisations today face numerous cyber-attacks, accelerating in volume and sophistication, resulting in ever-increasing chances that they will suffer a security incident. It was reported that there was an 11% increase from 2017-2018 (and 67% increase over the last five years) in the number of security breaches experienced by organisations, with the average cost of cybercrime in 2018 at \$13.0m [20]. This rise in breaches, resulted in risk transfer by organisations through the purchase of cyber-insurance to become more common: in 2019, 47% of surveyed organisations reported having cyber-insurance, increased from 34% in 2017 [16]. Given that this is the current landscape, it is of critical importance that the cyber-risk exposure of organisations is well-understood and can be reasoned about. Organisations are required to determine how different controls might help reduce their exposure to cyber attacks and how best to target them towards the most significant losses, as well as, guide decisions on cyber-insurance provisions for where risk cannot be avoided or entirely mitigated.

The security controls used by an organisation are crucial to their risk exposure. Controls are deployed by organisations with the aim of reducing this exposure, with a residual risk likely to remain, even if well-evidenced decisions are followed during this process. Currently, there is a lack of evidence of the way in which risk controls are being deployed by organisations and a lack of supporting evidence for the effectiveness of these controls. This includes not only which risk controls are selected, but also the way they are set up; even when standards and guidelines are rigorously followed, mistakes may be made during configuration that result in sub-optimal effectiveness.

The aim of our paper, given the lack of empirical data on security controls, is to explore if experimental set-ups are viable means to improve threat intelligence management and reduce risk exposure. There is usually the belief that when controls are implemented they will protect against any possible threats. Any reasoning about residual risk must be informed by an understanding of how threats can circumvent these controls.

To that end, we created a set of scenarios to be tested by penetration testers in a Capture-the-Flag (CTF)-style study with the aim of empirically assessing the security of a simulated organisation with different sets of controls and with different configurations. In particular, we are interested in exploring if we can measure the relative effectiveness of the deployment of some risk controls in reducing the overall risk exposure of the organisation, especially when other risk controls are in place. We present methodologies

based on CTFs for studying control effectiveness and deployment, report on the findings of our exercise, reflect on its value and make recommendations for strengthening the evidence base.

In Section 2, we present the foundations of our research, before elaborating on the study methodologies, results and implications in Sections 3, 4 and 5, respectively. We conclude with recommendations for future work in Section 6.

2 FOUNDATIONS

2.1 Controls and Organisational Risk

The relationship between organisational risk and risk controls, supports the experimentation and analysis presented in this article. The residual risk that an organisation faces is considered to result from a set of factors, namely the assets it possesses (and their value), the risk controls it deploys, the threats it faces, and the harms that may result [2].

Focusing on harms, various types of harm can result from a cyber incident (e.g., financial, physical, reputational) [3] and a harm that occurs following a cyber incident may lead to further harms (the loss or exposure of customer data may lead to financial harms in the form of regulator fines, for example) [5]. Harm propagates, and risk controls may address various branches of a harm-propagation “tree”. Our examination of risk controls in this article is guided by this organisational risk context, and in Section 5 we reflect on how our findings can be used to refine estimations of organisational risk and Cyber Value at Risk (CVaR).

2.2 Control Effectiveness and Deployment

Throughout this article, we use the SANS 20 Critical Security Controls (CSCs), a risk-control guidance source widely used by organisations that covers the key classes of controls, as the basis for our examination of security-control effectiveness and deployment [11]. The full list of the SANS CSCs can be found in Table 1. Elsewhere in this work, we refer to the CSCs by their number and, for convenience, may include their description – these descriptions may be shortened and therefore serve only as a reference.

2.2.1 Prior work on controls. Scarce evidence exists on the effectiveness of risk controls and their deployment, highlighting various factors affecting effectiveness. It has been shown, for example, the study of controls must take into account the extent to which controls are impacted by the deployment of other controls, and consider which sets of controls must be deployed together to protect an organisation’s architecture [2]. There is industry guidance on how to assess the effectiveness of a particular risk control’s deployment. The SANS CSC supplementary document: “*A Measurement Companion to the CIS Critical Security Controls*” [10] includes effectiveness tests for each of the 20 CSCs. These are descriptions of how the effectiveness of an instance of each type of control might be tested, for example, how to test that an organisation’s hardware or software has been securely configured, which can inform risk-control research by clarifying the optimal deployment of each type of control.

Such et al. evaluated the effect of using the Cyber Essentials controls on four theorised SME networks to mitigate 200 randomly

selected Common Vulnerabilities and Exposures (CVE) vulnerabilities [23]. The networks were modelled through interviews and survey work with SMEs on their deployment of networks. The assessment was made qualitatively: the researchers considered each component of the controls in order to ascertain whether each vulnerability would be mitigated, partially mitigated or not mitigated. Dietrich et al. used interviews and an online survey to study the perspectives of system operators of security misconfigurations: the frequency with which and ways in which they occur and their impact on the security of organisations [9]. The results showed that a range of misconfigurations, including the use of weak or default passwords, delayed or missing updates and faulty storage configuration, are perceived to occur frequently leading sometimes to incidents, and that human errors may be driven by institutional, organisational, and personal factors.

Other research has taken a practical approach to exploring the effect of deploying controls on network security. Mirkovic et al. ran a Red-Team/ Blue-Team exercise specifically aimed at evaluating and comparing DDoS defences, which supported the identification of reliable DDoS-defence characteristics such as avoiding reliance on timing mechanisms [17]. Kewley and Lowry reported on a set of three Red-Team exercises ran on the Defense Advanced Research Projects Agency (DARPA) testbeds to assess the effects of defence in depth (in particular adding middleware and boundary defence layers) on adversary capabilities, and concluded that the addition of security layers may not necessarily reduce risk without proper design, since new layers have the potential to introduce new vulnerabilities or attack surfaces [14]. Another DARPA experiment was reported on, with results indicating that “intrusion-detection and response systems provide improved assurance (defence immunity) over present static mechanisms” [13].

These studies have provided insights into the value of a limited number of specific types of control and some of the factors that are important in their deployment, yet none of this prior work fully achieves our aim of understanding the relative effectiveness of security controls in reducing the risk exposure of an organisation through experimental events such as CTFs.

2.2.2 Examining coverage in practitioner sources. The importance of various risk controls can be implied by their treatment in practitioner-community materials that influence the adoption of risk controls by organisations, such as, industry standards, guidelines and insurance-application forms. We posit that identifying the controls that might be considered important according to these sources may provide an indication of both control effectiveness and control deployment.

We examined how each SANS CSC is covered by practitioners through a number of lenses described below, and assigned a raw numerical score to enable ranking. The results are presented in Table 1.

- **Popularity in relevant standards:** This refers to the frequency with which controls are advised by key standards and guidelines. We aimed to understand the control which is considered most important by key industry bodies, and to gain a realistic interpretation of the controls that are likely to be deployed by the organisations adhering to standards.

	<i>Description</i>	Standards		Insurance		SANS		Characteristics			
		Rank	Raw	Rank	Raw	Rank	Raw	Human	Technical	Longitudinal	Individual
CSC1	“Device inventories”	15	4	17	0	1 ★	3	✓	✓	✓	✗
CSC2	“Software inventories”	15	4	17	0	1 ★	3	✓	✓	✓	✗
CSC3	“Vulnerability assessment”	1 ★	6	16	1	1 ★	3	✗	✓	✗	✓
CSC4	“Admin privileges”	1 ★	6	6	8	1 ★	3	✗	✓	✗	✓
CSC5	“Secure hosts”	1 ★	6	17	0	1 ★	3	✗	✓	✗	✓
CSC6	“Log monitoring”	10	5	13	3	1 ★	3	✓	✗	✓	✓
CSC7	“Web and email defence”	1 ★	6	17	0	7	2	✗	✓	✗	✓
CSC8	“Malware defences”	1 ★	6	2	26	7	2	✗	✓	✗	✓
CSC9	“Protocol controls”	10	5	8	6	7	2	✗	✓	✗	✓
CSC10	“Data recovery”	15	4	1 ★	29	7	2	✗	✓	✗	✓
CSC11	“Secure network devices”	1 ★	6	14	2	7	2	✗	✓	✗	✓
CSC12	“Boundary defences”	1 ★	6	5	9	7	2	✗	✓	✗	✓
CSC13	“Data protection”	10	5	11	4	7	2	✗	✓	✗	✓
CSC14	“Access control”	1 ★	6	3	17	7	2	✗	✓	✗	✓
CSC15	“Wireless access control”	15	4	14	2	7	2	✗	✓	✗	✓
CSC16	“Account monitoring”	1 ★	6	9	5	7	2	✓	✓	✓	✓
CSC17	“Skills and training”	10	5	4	10	17	1	✓	✗	✓	✓
CSC18	“Application security”	15	4	9	5	17	1	✗	✓	✗	✓
CSC19	“Incident response”	10	5	7	7	17	1	✓	✗	✓	✓
CSC20	“Penetration testing”	20	3	11	4	17	1	✓	✗	✓	✗

Table 1: SANS 20 CSC controls together with their ranking and raw scores according to each of the three lenses. For details of how each raw score was calculated, see Section 2.2.2.

In order to evaluate which controls are most frequently present across key standards, we used a set of charts created by AuditScripts [4] (a company responsible for security assessments of information systems), which map the SANS CSCs (v7) to various cybersecurity frameworks and standards. Alongside the CSCs, we examined the controls which are present in five other key UK, US and international standards and guidelines: the NIST Cybersecurity Framework [18], ISO 27002 [1], GCHQ 10 Steps to Cybersecurity [8], Cyber Essentials [7], and COBIT 5 [12].

We assigned each CSC a raw value corresponding to the number of these six selected standards in which it is referenced. There are nine controls that were present in all six; one control, CSC20 “Penetration testing”, was present the fewest times, in only three of the standards.

- **Insurance community:** This refers to the frequency with which controls are asked about by the insurance community, when gathering data for the calculation of cyber-insurance premiums (the pre-screening process). By incorporating this insight we aimed to understand the key concerns of the insurance community with regard to the protection of organisational assets.

To understand the frequency with which risk controls are referenced by the insurance community, we drew on prior work. Studies of insurance application forms have identified which areas of information security the insurers collect information about [24, 25]. We drew on the results presented by Woods et al. on the frequency with which the subcontrols within each of the SANS CSC control is mentioned in

insurance forms [24]. The analysis examined 24 insurance forms offered by insurers based in the UK and the US.

Each CSC was assigned a raw score directly from this work (Table 4 of Woods et al. [24]), thus the score corresponds to the average (rounded) percentage of subcontrols that were addressed by 24 insurance assessment forms. A higher score therefore indicates a greater degree of coverage by insurance assessors.

- **SANS order of importance:** This is a classification created by SANS of their 20 CSCs according to importance. SANS CSC v7 assigns its 20 controls three levels of importance, as follows [10]:

- CSC 1–6 are labelled “Basic” and described as: “should be implemented in every organisation for essential cyber defense readiness”
- CSC 7–16 are considered “Foundational”: “the next step up from basic; these technical best practices provide clear security benefits and are a smart move for any organisation to implement”
- CSC 17–20 are categorised as “Organisational”: “while they have many technical elements, CIS Controls 17–20 are more focused on people and processes involved in cybersecurity”

To assign CSCs numerical raw scores, we simply mapped these categories to nominal values; “basic” being assigned the value 3, “foundational” to 2 and “organisational” to 1. This assignment is intended as an indication of the importance of the deployment of each control.

In Section 5 we reflect on whether our study results support the inferences made from these sources.

2.3 Studying Risk Controls

To support the development of methodologies for studying our research, we consider the characteristics of controls that might impact their effectiveness and deployment. We examined the following characteristics, which we consider relevant to our methodology, for each of the SANS 20 CSCs. Our findings are presented in Table 1.

- **Human and Technical:** It may be more difficult to practically assess the effectiveness of controls that require a “human” in the loop (i.e. based on a procedure that humans must carry out, such as CSC17 “*Skills and training*”) rather than “technical” (i.e., based on a technological measure, such as CSC12 “*Boundary defences*”). Designing an experiment to assess the effectiveness of “human” controls would require qualitative metrics and participation by humans, whose competence varies.
- **Longitudinal:** It may be more difficult to practically assess controls whose effectiveness relies on being successful over an extended period of time, such as monitoring of systems. This would require conducting longitudinal studies; the alternative is controls whose effectiveness can be assessed with an one-time implementation (e.g. email/web browser protection), in a fixed state.
- **Individual:** Some controls do not protect or mitigate threats but support others. Inventories, for example, do not actively prevent threats but provide information that can improve the deployment of other threat-prevention measures. Considering this property of the risk controls may indicate which controls can meaningfully be tested individually or in groups.

2.3.1 Practical Experimentation. Scarce prior experiments have indicated the potential Red-Team/ Blue-Team exercises for studying cybersecurity research questions. Reflective research has also advocated the use of cybersecurity exercises and competitions as a basis for experimentation in cybersecurity, noting that such competitions make it possible to control a number of variables relevant to security, therefore rendering them viable for a meaningful assessment of the impact of controls on the security of a system [21].

A range of network sizes and types have been devised for studies, through variations in the number of machines and sub-networks, range of operating systems and applications, and generated background traffic [17, 22]. In some cases, systems and applications were deliberately rendered more vulnerable (e.g., had not been updated) than those found in a typical enterprise network, in order to create a meaningful exercise for participants [22]. The balance between creating an informative setup for the researchers and an achievable exercise for the participants is a challenge and we describe our treatment of it in Section 3.

Other studies have represented particular types of organisations and adversaries, through scenario designs. A military organisation was represented in a scenario based on a critical image server containing photos, radar and infrared images [14]. The Red Team in this study was modelled as a well-resourced national agency comprising experts who were able to perform at this level, the definition of the attackers’ mission (including confidentiality, integrity and

availability flags from both outside and inside the base firewall, and exfiltration or modification of a sensitive image without being detected), and the provision of reconnaissance information (network maps and traffic information) representative of the type of information that a well-resourced adversary would obtain in advance of an attack. Studies have also included active Blue Teams to represent network defenders, i.e., “human” controls [14, 17, 21].

Various approaches to data collection and a variety of analysis metrics have been used. Data has been collected through flag reporting, logs of successful and failed reconnaissance activities and attacks, the use of screen recorders, and the labelling of attack packets (using the IP headers’ Type of Service field) to facilitate analysis of collected network traffic [17, 22]. Performance metrics include the success and failure of tasks as well as the time taken by participants to perform these, the work factor as perceived by participants, the fractions of legitimate traffic maintained (for DDOS activities), rates of false-positive and negative detection, and rates of and time to response [14, 17, 21].

In designing our CTF-study methodology, we were influenced by a number of the aforementioned approaches: the collection of captured flags, attacker logs and adversary work-factor reports as study data, the definition of attacker aims through the placement and description of flags, and the selection of expert participants (penetration testers) to act as adversaries.

3 METHODOLOGY

To address the gap presented in the literature, we opted for a CTF exercise to empirically study the impact of security controls on attacker capabilities. In our (CTF)-style exercise, participants were given access to virtual network and were tasked with infiltrating it in order to reach checkpoints (or “flags”) throughout the system. This enabled us to record the flags reached by each participant and hence track their successful exploitation of the system. By providing different virtual networks with varied control deployments, we aimed to explore empirically the effect that each of our selected controls had on the ability of attackers (in this experiment, represented by participants) to compromise and navigate the network.

3.1 Control selection

We selected the controls for testing based on the characteristics and rankings identified in Section 2. First, we excluded CSCs that required longitudinal assessment, or that acted as an intermediate step for a risk control, as these would require a longitudinal study or would be impractical to assess individually (we assumed effective mechanisms in place supporting our selected controls, i.e. accurate hardware and software inventories in our experiment). We scoped our examination to only include purely “technical” controls, due to the challenges in setting up and running an exercise involving “humans”, as we will have been validating the competence of human actors instead.

For the remaining controls, we used our rankings (as presented in Table 1) to select those which appeared most ubiquitously through the three lenses; this resulted in a shortlist of five CSCs:

- CSC3 “*Vulnerability assessment*”
- CSC4 “*Admin privileges*”
- CSC8 “*Malware defences*”

- CSC12 “Boundary defences”
- CSC14 “Access control”

All five shortlisted controls can be considered “technical”; can be assessed as one-time implementations; and act as risk controls rather than as intermediate steps to risk control. All five are therefore viable for experimentation.

3.2 Scenarios and controls

We designed a virtual network setup with a number of security controls in place. Clearly, there is a huge space of possible network-security setups due to the range of variables: network topologies and assets; security controls deployed; configuration of these controls; vulnerabilities present on the network (i.e., recency of asset and control vulnerability patches). For this initial study, we selected two contrasting setups, and we discuss how this might be expanded to consider a larger part of the problem space in Section 5.

Two scenarios were created to represent two organisations with different maturity levels. The first one, shown in Figure 1, represents a mature company, where machines are segregated and updated. The second scenario, Figure 2, shows a less experienced company where all machines connect to the same network and patching is not carried out routinely.

Further to the scoping and control selection detailed above, we aimed to specifically explore subcontrols of CSC14 (*Controlled Access Based on the Need to Know*); in particular segmentation of the network based on sensitivity of the data, and encryption of sensitive data at rest, by segregating sensitive database servers onto a critical network and encrypting their contents. We also explore the impact of variations in CSC3 (*Continuous Vulnerability Assessment and Remediation*) by varying the patch recency of machines on the network. Also present were varying levels of password security, and a number of artefacts representing human error: insecure SSH keys, a list of network users, and scripts allowing database connection. These were intended to represent violations of CSC14.

Scenario 1. Figure 1 shows a network topology of a small company where machines are segregated onto three different networks: 10, 110, and 200. Network 10 contains client machines and the Active Directory. Network 110 contains critical services/data for the company. Network 200 is a DMZ. Two firewalls/gateways control access to the networks. The critical gateway restricts connections to port 3306 on the Database Server, on which the database contents are AES-encrypted, and only Admin2 has complete access to the 110 network. *inet* gateway provides Internet access to the 10 network, and restricts access to/from the DMZ. Only Admin2 is allowed to connect to the 200 network. Lastly, the Graylog server collects logs from networks 10 and 110.

Any machines not mentioned in this section are updated and do not contain known vulnerabilities. Any vulnerability found during the CTF in these machines was not intended to be part of the scenario, but provides important insights onto the effectiveness of the security controls, i.e., if that vulnerability can be exploited to bypass the security controls.

Balancing the investigation of research questions with the need to create a meaningful exercise with achievable goals is challenging; our approach was to introduce realistic weaknesses exploiting

which would enable some flags to be captured. The following intentional weaknesses were present:

- Repository (192.168.10.205): Insecure (unauthenticated) NFS; stored credentials.
- Admin2 (192.168.10.215): Weak password; privilege escalation via script.
- User2 (DHCP, network 10): Unpatched, vulnerable to known remote exploit; weak password; stored insecure access keys.
- Old Database (192.168.110.9): Unpatched and containing unencrypted database contents.
- Graylog (192.168.10.253 and 192.168.110.2): Acts as bridge into secure network.

Scenario 2. Figure 2 depicts a network topology where machines are segregated onto two networks: 10 and 200. Network 10 connects all the internal machines of the corporation. Network 200 is the DMZ, and contains a public-facing Web Server. Internet access is provided to the internal machines by *inet* gateway, and traffic is freely allowed between networks 10 and 200.

The same weaknesses were present as in Scenario 1 on Repository, Admin2, User2 and Old Database, although the structure of the network rendered access to Old Database, on which the database contents were not encrypted, more straightforward and removed the need to use Graylog as a bridge.

Flags. The aim of flags on the network was to facilitate the collection of a reliable record of compromised network infrastructure by the participants. The placement of flags was intended to indicate which parts of the network (subnets, individual servers, etc.) attackers could access and with which privilege levels. We therefore littered the network with flags at each individual location (each machine; piece of network infrastructure), that could be captured (i.e. the flag file could be read) only by users with specific privilege levels (users and root). We also placed flags associated with specific network aims: given the presence of sensitive databases on the network, we planted flags in the contents of each database, in order to indicate whether a participant was able to read the contents in plaintext. All flags were weighted equally (in terms of the points assigned) to avoid biasing participants towards a certain course of action.

3.3 Running the CTF

Study materials. We prepared the following materials for the participants of our study.

- **Participant information sheet and written consent form.** The information sheet outlined the study procedure, our data anonymisation and storage methods, and the processes to ask questions or withdraw data after the study.
- **Flag-reporting platform.** This was created using CTFd, an open-source tool for creating reporting platforms for CTF events¹, and was hosted by a server in our research lab.
- **Post-task questionnaire.** After each CTF round, participants responded to the following questions online in free-text boxes.
 - Please describe the security of the network and your experience penetration testing it.

¹<https://ctfd.io/>

Figure 1: Scenario 1 topology and results

Figure 2: Scenario 2 topology and results

- How difficult was it to compromise/capture flags on this network, and why?
- **Participant notebooks.** We asked participants to use a notebook to record the “types of attack you attempt, tools

you use to help you (if you use any), timestamp, network location and any rationale or strategy you apply”.

CTF rounds. Based on personal contacts, we managed to find four participants for our study, all of whom were security professionals who had extensive experience of working as senior penetration testers in three different organisations. Each network had two participants being randomly assigned to it. Participants were given approximately four hours to try and acquire as many flags as possible. Participants worked as individuals, capturing flags and reporting them using the flag-reporting interface, and recording their decision processes in the provided notebooks, guided by the instruction sheet.

Post-CTF focus group. Upon completion of the CTF exercise, we conducted a focus group with all four participants. The aim of the focus group was to promote group interaction, allow individuals to delve in details and to balance free-flowing debate against the need to cover the specified questions [19].

Participants were guided to discuss their experiences in the study, their perceptions of the security and how realistic the networks were, as well as their views on how the experimental setup could be improved. Other topics arose naturally, either in response to new questions asked by the researchers, or through steering of the discussion in new directions by participants. Such topics included, for example, real-world security practice and participants' perceptions of risk-control deployment in organisations, and approaches to adapting the study design to account for the impact of humans as defenders and network users.

3.4 Analysis

We began by analysing the data collected using the flag-reporting interface, by gathering the flags captured by each participant and noting the order of capture. We supplemented this data with comments made by participants in their notebooks, the post-task questionnaire, and the post-CTF focus group.

We manually transcribed the focus group data and analysed it using template analysis, a technique for qualitative data analysis in which the researcher has some understanding of the concepts to be identified [15]. The following a-priori themes were identified: how realistic the network setup is; data-collection methods; views on the potential of our methodology. We then coded the focus-group dataset initially, attaching relevant parts of the transcriptions to the a-priori themes. We assigned new codes to relevant sections of data that did not fit into these themes, thus producing an initial template of codes. This template was developed through iterative application to the dataset, and modified as appropriate to the data. Through this refinement we produced a final template and dataset coded according to it.

3.5 Ethics

We recruited participants by emailing our existing contacts in organisations that employ penetration testers. For each, our recruitment email described the study, its aims, and information on how to participate. Ethical approval to conduct the study was granted by the Computer Science Department Research Ethics Committee at the University of Oxford under reference: CS_C1A_19_033.

4 RESULTS

Our results of the CTF comprise an overview of the flags captured in each scenario, supported by the data collected in the study notebooks and post-task questionnaire, and the comments made in the focus group that were relevant to performance in the study. We present our observations of how the controls in place may have impacted the performance of participants. The four participants are referenced as **P1-P4** in the study, to ensure their anonymity.

4.1 Scenario 1

As Figure 1 shows, two flags were captured by participants on Scenario 1: on the Repository server and on Admin2. The capture of the Repository flag “*via the NFS share hosting a user directory*” was described by (**P1**) and **P4** as captured “*easily*”. The Admin2 flag was captured after participants accessed the Admin2 machine by brute-forcing the weak password of a user account found in the NFS.

Participants expressed the view that the patch level of Scenario 1 was reasonable: “*there weren’t that many really big out-of-date vulnerabilities that we found*” (**P4**). While the User2 machine (Windows 7) was present on this network, and its patch state meant that it was vulnerable to EternalBlue, participants stated that their “*enumeration didn’t pick up a Windows 7*” (**P1**). This was likely due to the host-based firewall on the Windows 7 machine blocking the necessary ports, and **P1** stated that they “*didn’t do any non-ping scans, just to speed up the process*”. In general, participants described the host-based firewall defences as being “*rather relaxed... with some hosts having SMB Signing disabled which facilitates [relay attacks]*” (**P1**). It was noted that there were “*numerous ports available where they are not required*” (**P1**), but “*minimal known vulnerable services*” (**P4**).

Focusing on the performance of participants in compromising the critical network and its databases, it is evident from the participants’ notebooks that both accessed the Database Server MySQL database via the Admin2 machine’s scripts. Since the database’s contents were encrypted, they were unable to capture the flags contained within the database. It was noted that the “*DB is encrypted... no easily accessible information*” (**P4**), and the researchers observed the participants attempting to decrypt the database. While the Database Server was accessible from the 10 network (“*the MySQL database was accessible from a 10 range, although it was apparently firewalled, depending on the ACLs*” (**P1**), the firewall blocking of the Old Database Server prevented participants from accessing it (“*not live*”) (**P1**).

The Graylog server, with connections to the .10 and .110 networks, allowed participants to observe the .110 network (“*so we were seeing another subnet through the Graylog server*” (**P4**)). This observation, however, did not allow them to use the Graylog as a bridge, so they were not able to communicate with machines in .110 through this path.

4.2 Scenario 2

As shown in Figure 1, three flags were captured by participants in Scenario 2: the first on the Repository server, then on the User2 machine (Windows 7), then on the Old Database. As in Scenario 1, participants captured the Repository server flag by using “*insecure*

file shares available on the network giving anonymous access to user directories" (P2). The patch level and host-based firewall rules of the User2 machine meant that it "had open SMB and it was vulnerable to EternalBlue" (P3), and participants leveraged this to carry out "direct remote code execution on the machine" (P2): "I executed a payload creating a new user called 'newadmin' and then I mounted a share folder using that new user" (P3).

Once the User2 machine had been accessed as root, and the "root" flag captured, participants located insecure SSH private keys that "could be used to gain access to .212" (P2). It was noted that this does not relate to "a technical control, it's more of a people thing. You just need to tell people to secure their keys" (P2). Participants thus connected to the Old Database (.212) and captured its flag. Participants did not access the MySQL database, but it was noted that the server was "running outdated mysql" (P2). It was also noted that "Fail2Ban was not present on any of the SSH machines", and that therefore "brute-forcing of passwords could be attempted" (P2). It is evident from P2's notebook that the Database Server (.213) was also scanned and found to have the SSH port open, although they did not try to SSH into this server using the keys found in the User2 machine. The credentials to query both databases were located on the Admin2 machine and could have been accessed had the participants attempted to brute force the weak password of the test user. They, however, focused first on the EternalBlue vulnerability which then determined their path through the network.

With regard to the patch level, participants noted variations between machines, stating: "the use of outdated software was quite prominent in the network that I was testing, so that allowed for some pretty easy exploitation in some regards, but other machines were quite up-to-date and that made it a little bit harder" (P2). The variation in machine patch levels prevented lateral movement even after a foothold had been gained: "EternalBlue not patched but in all the others it was patched, so you cannot leverage that and move laterally", as P3 explained. However, combining the circumvention of technical controls with the exploitation of human errors, such as the insecure SSH keys, allowed participants "to leverage those and move laterally – and with that, you actually bypass some of the patching and updated versions." (P3).

4.3 Impact of controls on performance

We present our observations of the way in which control deployment may have impacted on the performance of participants. It is important to note that these observations are tentative, and cannot constitute hard evidence given the small scale of the study (in terms of the timeframe of the experiment, the number of networks tested and the number of participants). It is more fitting that our results contribute towards working hypotheses to be tested as future work.

- (1) **Patch levels.** Participants could easily compromise unpatched machines (particularly due to the EternalBlue vulnerability).
 - (a) While participants were able to compromise some unpatched machines, they were not able to leverage this to move laterally, due to the improved patching level of other machines. This cannot be seen as a security feature, as the time constraint likely prevented the lateral movement.

- (b) Participants needed to leverage the misconfiguration of other control types in combination with the unpatched-machine vulnerabilities in order to compromise the machines. For example, User2 in Scenario 2 was only compromised because of the combination of EternalBlue and open SMB, while User2 in Scenario 1, it was not spotted, and the EternalBlue vulnerability therefore was never discovered, due to the host-based firewall blocking scans (ping).
- (2) **Network segmentation.** Participants did not access machines in isolated network segments unless using some in-built access-control rules.
 - (a) Web Server in the DMZ was not accessed by any participant. This was probably due to time constraints, as firewall rules allowed direct access to this network in Scenario 2.
 - (b) In Scenario 1, database servers were segregated from the critical network. Participants accessed the database on Database Server via Admin2 which contained a script that provided access to the database, but they did not access Old Database. In Scenario 2, on the other hand, in which the database servers were not isolated, participants did access Old Database, and scanned Database Server, finding that it had open SSH.
- (3) **Encryption of data at rest.** Participants were unable to access the flag included within the Database Server database's content in Scenario 1, even after accessing the database, since it had been AES-encrypted. This represents a level of database security in which certain parties (the administrators) are able to connect to the database, but require a separately stored password to read its contents.
- (4) **Password security.** Participants gained access to one machine by compromising its authentication credentials, and this was the Admin2 machine which had a weak (brute-forceable) password. This represents a valid method of account takeover, which is still one of the most predominant threats today, but there is a wide range of means by which credentials can be compromised, including social engineering, use of common passwords, shoulder surfing, and phishing emails.
- (5) **Human and technical controls.** Participants leveraged artefacts that represent human errors, alongside the omission of, and mistakes in the configuration of, technical controls, to achieve compromises, move laterally and capture flags on the network. It was noted, for example, that the presence of insecure SSH keys on Scenario 2's User2 represents a human error, and this allowed bypassing of some technical controls such as the strong patch state of other machines. The importance of leveraging user mistakes to attacker actions was emphasised in the focus group: "most of the time, the attackers leverage the user level, all the mistakes that are done there, and all the artefacts that are left" (P3).

5 DISCUSSION

We believe the CTF results enabled us to reason about the effectiveness of the tested controls and were in line with our expectations:

the setup of controls affected the performance of participants as anticipated. Given that senior penetration testers are high in demand, we had time constraints to conclude our experiment. Therefore, we could not compare the performance of the same participants across the two networks. However, the level of expertise of the penetration testers is similarly exceptional and it is therefore unlikely that our results have been biased by the differences in the experience of participants and in their areas of expertise. Participants in the focus group stated that the short timeframe posed challenges to fully understanding the network and may have affected the attack techniques used (e.g., a participant reported not having tested all open ports, in order to save time).

This must be taken into account in the interpretation of our results. Designers of similar studies in the future should note the considerable length of time required for a full network-security assessment and balance this with the need to design an experiment that can be completed within the time participants are able to offer. In an experiment that will run over a longer time period with a greater number of participants, our intended randomisation of the networks presented to each participant, appears a viable approach to mitigating the impact of learning effects.

The conflict between a CTF event and a security experiment was highlighted by participants in the focus group. In particular, while CTFs are generally gamified, built on vulnerabilities inserted by the developers which can be exploited to capture flags, the research questions we aimed to address – on the effect of deploying a set of risk controls on network security – were not based on creating particular vulnerabilities but on protecting networks through the deployment of controls. This mismatch between concepts, resulted in inserting realistic vulnerabilities into the CTF scenarios, alongside our deployment of risk controls in line with the research questions, in order to ensure that flags were somehow “capturable”. This matter has been commented in prior work [22].

The CTF setup may have affected the actions of the participants in attacking the networks, altering their actions from those they would usually take during a penetration test, as was commented by participants in the focus group. In particular, the desire to capture flags could cause participants to move more quickly between different machines, searching for exploitable vulnerabilities, rather than obtaining a slower and more holistic view of the network security. The use of flags may also lead participants towards an attack pattern not representative of what they would usually follow.

It was noted that while penetration testers are good representations of attackers, they differ in that they aim to avoid any damage to the network. Participants noted the advantages that assigning a threat context to exercises could create: by modelling the capabilities of particular types of attacker through definition of the attacks permitted and provision of representative resources (approaches to which have been presented in prior work [13]), it would be possible to explore the impact of controls in the face of particular threat types. Organisational context is another aspect that could be incorporated into modelled scenarios and the methods of prior practical studies in which organisational contexts have been created could be drawn on [14].

In order to enhance how realistic the networks represented in this type of exercise are, approaches to the representation of various facets of the user level might be taken. Specifically, humans might

participate in live exercises, either as defenders (Blue Teams) or as network users. In-memory artefacts (such as Kerberos tickets) could also be created through human interaction with the network prior to the event, as noted in the focus group. Various pieces of prior research have included and made methodological recommendations for the inclusion of user interaction in practical security studies [17, 21, 22], as described in Section 2.3. In a similar way, different network configurations should be implemented in order to mimic different types of organisations. As noted previously, the effectiveness of one control in a specific network can be different to the same control in another network. This implies that this methodology can generalise insofar that different CTF events should be run for different configurations.

The interactions and dependencies between controls were discussed by participants across studies: various examples were given of cases in which controls depend on other controls, compensate for others, and work in combination to create security. The need to consider a broad control context from a holistic standpoint, rather than individual controls in isolation, was emphasised by our participants. This is also considered in [6], where automated compliance assessment of services is proposed using a logic compliance analysis.

The collection of data in future practical security studies can be refined based on our findings. CTF participants highlighted the potential utility of collecting traffic captures at network gateways and endpoints, records of participants’ bash history and complete or partial penetration testing reports. Other promising approaches to data collection in prior work, which fell outside the scope of this study, include attack-tree descriptions prepared by attackers in advance of the study, labelling of attack packets to facilitate the analysis of traffic captures, and qualitative descriptions of the adversary work factor.

Finally, we reflected on the nature of controls, including whether they are performed by humans, or are pieces of technology. While we chose to focus on technical controls in this research, it is clearly important to also study the effectiveness of cybersecurity controls performed by humans. Participants in the post-CTF focus group highlighted the criticality of human actions on a network, and described the use of both human- and technical-level exploits in combination by attackers in exploiting a network.

6 CONCLUSION AND FUTURE WORK

To understand the risk that an organisation faces, it is imperative to take into account the impact of the risk controls on their security. However, evidence is lacking on the way in which the use of risk controls affects risk exposure across organisational contexts. In this paper we explore the use of empirical security studies as a viable pathway to provide evidence for the effectiveness of controls. We conducted a Capture-the-Flag study with four penetration testers, followed by a focus group, where we examined a small subset of SANS 20 CSC risk controls. Our initial experimentation indicates that this research method is worthy of further exploration, especially in modelling the risk control effectiveness in specific contexts. As it was highlighted during the interviews, CTFs present an interesting approach, not only for security analysts, but also for organisations, as they could set up a test environment to enrich

their threat intelligence and to measure how the development of specific controls would affect their security risk.

Although scoping our CTF scenarios to a limited number of controls was necessary to run a successful experiment, it allowed us to understand the tested controls in finer detail. Future studies could explore a broader scope of controls, by involving a greater number of participants over a longer period of time (in the case of practical experimentation especially). The use of online CTFs could be a viable approach to achieving this wider experimentation.

Based on the performance of participants in the CTF, our preliminary observations and their comments in the focus group, we have identified a set of working hypotheses on the effectiveness of the controls studied in the CTF. We anticipate that these working hypotheses can be used tentatively to inform the estimation of cyber risk and should be further explored through experiments on a larger scale.

- (1) Patching is a critical control for all organisations, and can compensate for the omission of, or weaknesses in, the configuration of a range of other controls (based on observation 1b).
- (2) The segregation of sensitive systems at the network level is a highly effective method of protecting sensitive data, but only if combined with appropriate segregation at the user level. For example, segregating critical machines onto a critical network is not effective if attackers can access these networks by posing as users (based on observation 2).
- (3) Encryption of sensitive data is an effective control only if the encryption credentials are inaccessible by attackers.
- (4) Minimising the security mistakes made by users is more effective than technical security controls, since many well-implemented technical controls can be bypassed using artefacts left by users.

Future work might expand our analysis of the perceptions of security practitioners. There may be tacit knowledge within the community that is shared and helping to deliver cybersecurity, but is not formalised or documented fully on relevant factors, such as the interdependencies between controls. Capturing and communicating this knowledge will not only help to promote best practice, but also could lead to more accurate cyber risk estimations. There is a need to assess the reliability of this source of evidence by examining the preconceptions that may underpin practitioners' views: for example, there may be a convergence towards commonly held beliefs that echo the recommendations of standards. It would be valuable, therefore, to study converges in perceptions and the preconceptions that may underlie them across a large sample size of practitioners.

ACKNOWLEDGMENTS

This research was sponsored by AXIS Insurance Company, whose support is gratefully acknowledged.

REFERENCES

- [1] 2013. ISO/IEC 27002 Code of practice for information security controls. <https://www.iso27001security.com/html/27002.html> [accessed on 25/05/2021].
- [2] Ioannis Agraftiotis, Sadie Creese, Michael Goldsmith, Jason RC Nurse, and David Upton. 2016. The Relative Effectiveness of widely used Risk Controls and the Real Value of Compliance. (2016).
- [3] Ioannis Agraftiotis, Jason RC Nurse, Michael Goldsmith, Sadie Creese, and David Upton. 2018. A taxonomy of cyber-harms: Defining the impacts of cyber-attacks and understanding how they propagate. *Journal of Cybersecurity* 4, 1 (2018), ty006.
- [4] AuditScripts. 2018. AuditScripts Critical Security Controls. URL: <https://www.auditscripts.com/> [accessed on 25/05/2021].
- [5] Louise Axon, Arnau Erola, Ioannis Agraftiotis, Michael Goldsmith, and Sadie Creese. 2019. Analysing cyber-insurance claims to design harm-propagation trees. In *2019 International Conference on Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA)*. IEEE.
- [6] Francesco Buccafurri, Lidia Fotia, Angelo Furfaro, Alfredo Garro, Matteo Giacalone, and Andrea Tundis. 2015. An Analytical Processing Approach to Supporting Cyber Security Compliance Assessment. In *Proceedings of the 8th International Conference on Security of Information and Networks*. ACM, 46–53.
- [7] National Cyber Security Centre. 2014. Cyber Essentials. <https://www.cyberessentials.ncsc.gov.uk/> [accessed on 25/05/2021].
- [8] National Cyber Security Centre. 2021. *10 Steps to Cyber Security*. <https://www.ncsc.gov.uk/collection/10-steps-to-cyber-security> [accessed on 25/05/2021].
- [9] Constanze Dietrich, Katharina Krombholz, Kevin Borgolte, and Tobias Fiebig. 2018. Investigating System Operators' Perspective on Security Misconfigurations. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 1272–1289.
- [10] Center for Internet Security. 2015. A Measurement Companion to the CIS Critical Security Controls. URL: <https://www.cisecurity.org/white-papers/a-measurement-companion-to-the-cis-critical-controls/> [accessed 25/05/2021].
- [11] SANS/Center for Internet Security. 2021. 20 Critical security controls. <https://www.cisecurity.org/controls/> [accessed on 25/05/2021].
- [12] ISACA. 2021. COBIT 5. <https://www.isaca.org/cobit/> [accessed on 25/05/2021].
- [13] Dorene L Kewley and Julie F Bouchard. 2001. DARPA information assurance program dynamic defense experiment summary. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 31, 4 (2001), 331–336.
- [14] Dorene L Kewley and John Lowry. 2001. Observations on the effects of defense in depth on adversary behavior in cyber warfare. In *Proceedings of the IEEE SMC Information Assurance Workshop*. Citeseer, 1–8.
- [15] Nigel King. 1998. Template analysis. *Qualitative Methods and Analysis in Organizational Research: A Practical Guide* (1998).
- [16] Marsh. 2019. Global Cyber Risk Perception Survey Report. <https://www.marsh.com/uk/insights/research/marsh-microsoft-cyber-survey-report-2019.html> [accessed on 25/05/2021].
- [17] Jelena Mirkovic, Peter Reiher, Christos Papadopoulos, Alefiya Hussain, Marla Shepard, Michael Berg, and Robert Jung. 2008. Testing a collaborative DDoS defense in a red team/blue team exercise. *IEEE Trans. Comput.* 57, 8 (2008), 1098–1112.
- [18] National Institute of Standards and Technology. 2018. Cybersecurity Framework. <https://www.nist.gov/cyberframework> [accessed on 25/05/2021].
- [19] Jane Ritchie, Jane Lewis, Carol McNaughton Nicholls, Rachel Ormston, et al. 2013. *Qualitative research practice: A guide for social science students and researchers*. sage.
- [20] Accenture Security. 2019. The Cost of Cybercrime. https://www.accenture.com/_acnmedia/pdf-96/accenture-2019-cost-of-cybercrime-study-final.pdf [accessed on 25/05/2021].
- [21] Teodor Sommestad and Jonas Hallberg. 2012. Cyber security exercises and competitions as a platform for cyber security experiments. In *Nordic Conference on Secure IT Systems*. Springer, 47–60.
- [22] Teodor Sommestad and Fredrik Sandström. 2015. An empirical test of the accuracy of an attack graph analysis tool. *Information & Computer Security* 23, 5 (2015), 516–531.
- [23] Jose M Such, John Vidler, Timothy Seabrook, and Awais Rashid. 2015. Cyber security controls effectiveness: a qualitative assessment of cyber essentials. Lancaster University.
- [24] Daniel Woods, Ioannis Agraftiotis, Jason RC Nurse, and Sadie Creese. 2017. Mapping the coverage of security controls in cyber insurance proposal forms. *Journal of Internet Services and Applications* 8, 1 (2017), 8.
- [25] Daniel Woods and Andrew Simpson. 2017. Policy measures and cyber insurance: a framework. *Journal of Cyber Policy* 2, 2 (2017), 209–226.