How Ready is Your Ready? Assessing the Usability of Incident Response Playbook Frameworks

Rock Stevens, Daniel Votipka*, Josiah Dykstra†, Fernando Tomlinson‡,
Erin Quartararo, Colin Ahern◇, and Michelle L. Mazurek
University of Maryland, *Tufts University, †National Security Agency, ‡Mandiant, ◇New York City Cyber Command
USA

ABSTRACT

Incident response playbooks provide step-by-step guidelines to help security operations personnel quickly respond to specific threat scenarios. Although playbooks are common in the security industry, they have not been empirically evaluated for effectiveness. This paper takes a first step toward measuring playbooks and the frameworks used to design them, using two studies conducted in an enterprise environment. In the first study, twelve security professionals created two playbooks each, using two standard playbook design frameworks; the resulting playbooks were evaluated by experts for accuracy. In the second, we observed five personnel using the created playbooks in no-notice threat exercises within a live security-operations center. We find that playbooks can help simplify and support incident response efforts. However, playbooks designed using the frameworks we examined often lack sufficient detail for real-world use, particularly for more junior technicians. We provide recommendations for improving playbooks, playbook frameworks, and organizational processes surrounding playbook use.

CCS CONCEPTS

• **Security and privacy → Usability in security and privacy**.

KEYWORDS

incident response; security operations; usability of frameworks

ACM Reference Format:
Rock Stevens, Daniel Votipka*, Josiah Dykstra†, Fernando Tomlinson‡,, Erin Quartararo, Colin Ahern◇, and Michelle L. Mazurek. 2022. How Ready is Your Ready? Assessing the Usability of Incident Response Playbook Frameworks. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3491102.3517559

1 INTRODUCTION

Digital security *playbooks* — structured action plans for incident response — are designed to help organizations prepare for security breaches and enable quick and appropriate responses. High-stress situations such as an ongoing data breach may disrupt technicians' cognitive abilities [25, 37, 46]; playbooks are designed to present documented best practices to prompt action and momentum during incident response.

Playbooks are touted for defending high-value systems, conducting investigations, and keeping defensive systems up to date [8, 16, 63]. Others highlight the usefulness of checklists and communications templates for business continuity [33] or promote playbooks as part of cybersecurity training [42, 48]. The U.S. Cybersecurity and Infrastructure Security Agency (CISA) regularly recommends playbook use for organizations operating critical infrastructure, such as healthcare [20] and the chemical sector [19]. Further, in the U.S. President's recent Executive Order on Improving the Nation's Cybersecurity, playbook development for incident response is listed as an important tool for national security [40]. Playbooks are also commonly used by security professionals; 140 of 200 surveyed security-operations personnel reported using playbooks [72]. In many cases, playbooks created for one organization are made available to the larger security community, where they can be reused by others (potentially after customization) [5, 39].

But where do playbooks come from, and how do organizations know they include all the necessary information? *Playbook frameworks* are guidelines designed to help organizations identify core problems and systematically design playbooks tailored for their environments. Playbook frameworks focus on breaking down complex security processes into steps easily understandable and implementable by practitioners. Several different playbook frameworks have been proposed by organizations such as the U.S. National Institute of Standards and Technology (NIST) and MITRE [1, 14].

Unfortunately, there has been little to no empirical measurement of playbooks' effectiveness for incident response, or of the frameworks used to design them. As a result, it can be difficult to assess the benefit of playbooks and playbook frameworks in practice. Do playbook frameworks effectively support the design of usable and useful playbooks that improve incident-response outcomes?

We take a first step toward answering this question with two studies conducted in real-world networks: one evaluating the usability of two frameworks during the playbook design phase, and one examining the process of implementing security controls based on playbooks generated from these frameworks and then using the playbooks during incident response. These small but in-depth studies allow us to observe the entire playbook process, from design to use, in an ecologically valid context. This type of close examination of specific cases can be valuable during early investigation for building understanding, generating theories and hypotheses, and providing insights in settings — like incident response — where larger samples are difficult to access [45].

For these studies, we partnered with two organizations that did not previously use playbooks but are required to secure trade secrets

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3517559

worth millions of dollars. We examined two exemplar playbook frameworks that have federal-level recognition within the United States: the Integrated Adaptive Cyber Defense (IACD) framework and NIST Computer Security Incident Handling Guide [14, 38, 55, 69] and have been used to create playbooks across the public [19–21, 29, 74] and private [4] sectors.

First, we introduced 12 personnel to the IACD and NIST frameworks and asked them to design local playbook instances using each framework. We measured participants' perceptions of each frameworks' usability and usefulness. Three experts also evaluated each playbook's completeness and correctness.

We selected the highest scoring playbooks for evaluation in a second study: a field deployment. One partner organization implemented the security controls called for by the selected playbooks in their production networks and adopted the playbooks for use by their security-operations technicians. Then, with cooperation from organizational leadership, we conducted three no-notice insider-threat exercises that would be expected to trigger use of the playbooks. We observed how well the playbooks supported incident response in practice. While this field deployment was necessarily small (n=5), to our knowledge it is the first real-world assessment of playbook utility. Together, our two studies provide significant insight into the benefits and challenges of using playbook frameworks, and the resulting playbooks, in practice.

We found that both frameworks had pluses and minuses for designing playbooks. The NIST framework elicited more fine-grained details for incident response, but IACD allowed technicians to more quickly identify, in general terms, tasks that needed to be performed. Both frameworks were considered easy to learn by participants, but experts determined that the resulting playbooks had high error rates. Experienced technicians were able to use high-scoring playbooks successfully during incident-response exercises, but our novice participant struggled. After redesigning the playbooks — and the organizational infrastructure surrounding them — based on lessons learned in the first two exercises, novices in the third exercise were more successful.

Our observations provide hope that playbooks can, when well designed and implemented, achieve their stated goal of simplifying and supporting incident response; however, significant improvement and investment, as well as further systematic evaluation, are needed. We distill from our results recommendations for improvement in playbooks and frameworks, as well as associated organizational procedures.

2 BACKGROUND AND PRELIMINARIES

First, we detail the playbook frameworks we chose to evaluate, our two partner organizations, and the two incident-response scenarios we selected as targets for playbook design.

2.1 Selected frameworks

We study two frameworks that have U.S. government support, but guide users toward different focuses: The NIST Computer Security Incident Handling Guide (hereafter: the NIST framework) [14], and Integrated Adaptive Cyber Defense (hereafter: the IACD framework) [38]. As the following paragraphs detail, the IACD framework asks users to consider how tasks could be automated, while NIST instructions emphasize speed of recovery and restoration of availability. While other proposed frameworks, such as SOTER [57], exist for digital-security incident-response playbooks, we selected IACD and NIST because they are currently the most commonly used frameworks in practice, offer free guides and examples, and focus specifically on creating playbooks. In fact, IACD and NIST currently dominate the playbook framework landscape. IACD is used by several prominent financial institutions [74] and U.S. state governments [29] and the NIST framework is used by CISA to develop example playbooks for critical infrastructure and healthcare [19, 21], as well as private enterprises such as Amazon [4] and Microsoft [31]. Additionally, in our review of several information security vendor incident response guides, the process for constructing specific playbooks was either left to the reader or given by reference to NIST or IACD. Lastly, we chose NIST and IACD over the framework developed by MITRE because the MITRE framework shares many similarities with IACD (such as a focus on automation) and is still under development [1]. As noted in Section 1, little to no publicly-available data supports playbooks' effectiveness for incident response, for any framework.

2.1.1 IACD. The IACD framework was created by the U.S. Department of Homeland Security, National Security Agency, and Johns Hopkins Applied Physics Laboratory to leverage automation in incident response [38]. IACD playbooks' defining feature is a visual flowchart with essential response actions for humans and automated systems (Figure 2 in Appendix E).

The IACD framework breaks playbook design into 10 steps. (Section 2.3.1 provides a running example in greater detail.) The first step is to identify the initiating condition: the event or situation triggering playbook use (e.g., a database breach) and how that event is detected (e.g., an automated email alert sent to an administrator). The second step involves listing all possible actions that could occur in response to the initiating condition, typically via mind mapping. Practitioners should reference existing best practices to identify possible actions. Next, playbook designers designate each identified action as required or optional. For example, generating a written report that details the incident from beginning to end — which may provide invaluable insight after the event but does not contribute directly to response efforts — is expected to be labeled optional. Steps 4-8 involve grouping actions by function, ordering required actions sequentially, and interleaving optional actions where appropriate. The designer produces a diagram showing these ordered relationships and noting steps that should be automated. In step 9, the designer verifies that the playbook terminates either in a desired end state or in a new initiating condition that flows into another playbook. The final step ensures the playbook satisfies applicable regulatory controls and requirements.

2.1.2 NIST. The NIST framework focuses on quick recovery after a security incident [14]. Using this framework, designers break a security incident down into three phases and create playbook content for each. The preparation phase occurs before an incident and requires analysts to identify critical assets that must be protected from a particular threat. Playbook content for the detection and analysis phase should help defenders identify the incident's entry point, breadth of impact, potential consequences, and containment methods. Phase three content — containment, eradication,

and recovery — should guide defenders in patching or isolating the attacker's entry point and other similar potential entry points, increasing monitoring, and safely bringing services back online. Each phase of a NIST playbook should emphasize communication and metrics tracking: ensuring essential personnel are informed, victims are notified, and the scope of impact is thoroughly documented. While playbook designers may or may not deem communications as required actions in IACD playbooks, communication is required throughout NIST playbooks.

Unlike IACD, NIST does not typically result in a visualization of response actions (although it could). Instead, NIST playbooks typically provide detailed textual descriptions, intended to be drawn from institutional procedures or best practices. Section 2.3.2 provides a detailed running example.

2.2 Our partners

To evaluate playbooks and frameworks in organizations that had not previously used them, we partnered with two organizations specializing in digital security. Each organization is responsible for securing confidential information and trade secrets worth millions of dollars.[1] We spent a year developing relationships with these organizations, agreeing to memorandums of understanding and data-handling protocols as well as legal reviews to enable our access. For anonymity, we refer to them as the network defense center (NDC) and the security development team (SDT).

NDC manages networks spanning multiple countries and 600 user accounts, with a service-level agreement to maintain availability levels at or above 98.9% while securing highly-sensitive customer intellectual property. NDC had 12 employees during the first study and 13 during the second study.

SDT develops secure applications for nearly 1500 worldwide customers, often building custom solutions for niche requirements. SDT employs 28 developers.

Both NDC and SDT must secure their development and production environments from malicious attacks, insider threats, and natural disasters. Both organizations have personnel with a range of security experience: a few entry-level and the majority with more than 10 years' experience. Employees from both are expected to secure their respective networks, perform incident response duties, and perform basic triage.

Prior to the study, one co-author reviewed one year's worth of incident response reports to better understand the threats both organizations regularly face as well as their missions, cultures, customers, and risks. Outside of unexpected service outages, NDC and SDT collectively experienced three security incidents during that year. Technicians' incident response efforts to these events were ad-hoc, rather than drawing on predetermined plans, policies, or procedures.

2.3 Selected scenarios

As playbooks are designed to address specific incident scenario, we needed to select two scenarios to use in our studies. We collaborated (over nine weeks of discussion) with leaders from NDC and SDT to select two scenarios from the MITRE ATT&CK database [49], using the following three criteria: (1) both organizations could realistically encounter them; (2) each organization should be able to quickly and consistently respond to them at any time; and (3) neither organization had a standard policy or procedure in place to handle them. We selected brute-force login attempts [50] and valid credential compromises [52]. Leaders from both organizations considered insider attacks and password attacks as high-probability, severe-impact concerns. While we might have preferred to choose a wider range of scenarios, allowing organizational leadership to select issues of most concern to them both helped to secure cooperation and ensured strong motivation for incorporating the resulting playbooks into existing workflows. Further, these scenarios represent in some sense a best case for playbook design: because these scenarios are familiar, we can evaluate which aspects of playbooks and frameworks can fail even for well-understood scenarios.

2.3.1 Brute-force login attempts. In a brute-force login attack, an adversary attempts to gain unauthorized access by guessing commonly used or randomly generated passwords. Here, we focus on protecting user-level domain accounts from locally-originating attacks. Using the IACD framework, we briefly detail some essential tasks associated with detecting and responding to brute force attempts from within the network.

The initiating condition is the detection of multiple password-guessing attempts against one or multiple systems. A centralized log repository must continuously audit and correlate login failures from across the network. If a brute-force pattern is detected, the system should generate an alert (e.g., a dashboard push notification or email to a technician).

Required actions, in sequential order, might include: identify the system(s) being attacked; identify the potential attack source; isolate source and/or victim nodes; install new sensors for traffic monitoring; identify compromised accounts; conduct root-cause analysis; perform root-cause mitigation; and restore accounts/services. Two optional action groups might be prioritizing assets (determining which resources are most important to isolate first) and producing reports (helping responders understand the situation and make better decisions).

After these steps, we recognize that the playbook terminates in a desired end state: the root cause has been patched and affected services and accounts have been restored. The final IACD step is to validate that the playbook satisfies regulatory controls and requirements, such as log retention policies.

2.3.2 Valid credential misuse. Valid credential compromise can occur, e.g., when a data breach reveals credentials from one account that can be reused at another site. Here, we focus on protecting user-level domain accounts from local abuse.

We next walk through a sample NIST playbook that uses *honeywords* — usernames and passwords for valid but fictional accounts — to detect credential misuse [41].

The preparation phase includes the creation of honeyword accounts, ensuring security team members understand the significance of the honeywords, and deploying an automated log-event parser to scan for login attempts associated with the honeyword account and generate an alert if found.

[1] We note that although these are large organizations which are willing and able to devote significant resources to information security, neither had an incident-response plan for the target scenarios in place prior to our study.

The detection and analysis phase starts when a human analyst receives an alert (e.g., a dashboard push notification or an email). The analyst should then investigate breadth of impact; for example, if the honeyword was created on a domain controller, then the analyst may assume there has been a compromise of all accounts on the domain controller.

The final containment, eradication, and recovery phase involves root-cause analysis to determine how the domain controller was initially compromised and generate a "fingerprint" to check for similar compromises on other systems. Next, all affected accounts must be denied access until they have changed their password. Affected users and compliance entities (as applicable) must be notified of the breach. Finally, the incident must be fully documented, and there may also be regulatory requirements for follow-on security assessments.

3 PLAYBOOK DESIGN AND EVALUATION

In this section, we detail our first study, exploring the usability of playbook frameworks. For this study, we sought to understand whether participants could use the two frameworks to effectively and efficiently generate playbooks that experts would consider to be complete and correct.

For this study, we recruited twelve participants from NDC and SDT, familiarized them with the IACD and NIST frameworks, and asked them to each design two playbooks *specific to their organizations* using the two frameworks and our two selected scenarios. We measured participants' perceptions of the process, and three external experts evaluated the designed playbooks for thoroughness and accuracy. We found that although most participants considered the frameworks reasonably easy to use, about half of the designed playbooks were rated as insufficiently detailed for real-world use.

3.1 Method

We sought to understand participants' perceptions of the frameworks' usability, as well as whether the resulting playbooks would be usable in a real-world setting. In this study, we measure usability following Nielsen [56]: in terms of learnability (ease of first-time use), efficiency (timely task completion), errors, and satisfaction.[2]

This study occurred from September through December 2019 and was approved by our organization's ethics review board. To protect our participants and partner organizations, we generalize or redact details about sensitive information, including job descriptions and identified vulnerabilities.

3.1.1 Recruitment. We partnered with NDC and SDT to recruit employees performing daily security functions. Because of their direct role in operations and familiarity with the environment, these individuals are representative of the type of employees commonly tasked with developing playbooks. Leadership from both organizations announced the study during group meetings, describing our motivation and goals while emphasizing that participation was voluntary. Employees were told that participants would be introduced to new techniques that could be useful in their work, and that playbooks from the study would be adopted into daily practice. Employees and contractors were permitted to participate during regular work hours but were not otherwise compensated. We asked NDC/SDT leaders to emphasize that participation in the study would have no impact on performance evaluations.

[2]We exclude memorability — usability over time, after periods of disuse — for this initial exploratory study.

ID	Org.	Role	Years Exp.	Study Phase[1,2]	Order[3]
P1	NDC	Manager	11+	D	NBF : ICM
P2	NDC	Technician	0-4	D, IR1/2/3	ICM : NBF
P3	NDC	Manager	11+	D, I, IR1/2	NBF : ICM
P4	NDC	Manager	11+	D, IR1	IBF : NCM
P5	NDC	Manager	11+	D	NCM : IBF
P6	SDT	Manager	11+	D	NCM : IBF
P7	SDT	Technician	11+	D	IBF : NCM
P8	NDC	Technician	5-10	D, IR2	NCM : IBF
P9	SDT	Technician	11+	D	ICM : NBF
P10	SDT	Technician	11+	D	NBF : ICM
P11	SDT	Technician	5-10	D	IBF : NCM
P12	SDT	Manager	11+	D	ICM : NBF
P13	NDC	Technician	0-4	IR3	–
E1	–	Senior Mgr	11+	E	–
E2	–	Senior Mgr	11+	E	–
E3	–	Senior Mgr	11+	E	–

[1]D: Design, E: Evaluate, I: Implement, IR: Incident Response Exercise

[2]IR1: Dec 2, 2019; IR2: Jan 13, 2020; IR3: March 2, 2020

[3]Assigned order for playbook design (N: NIST, I: IACD, BF: Brute Force, CM: Credential Misuse)

Table 1: Participant and expert demographics

3.1.2 Participants. In total, 15 people participated in this study, including 12 NDC/SDT employees who designed playbooks and three expert evaluators (discussed in Section 3.1.4). Participant details are given in Table 1. Our population size was consistent with previously published studies with similar participant types and goals [9, 11]. Our sample represented 58% of NDC's workforce and 21% of SDT's workforce at the time of the study. Prior to the study, all design participants (P1-12) said they knew that playbooks were an industry best practice, but none had used a playbook to respond to an incident. All design participants had completed at least one year of on-the-job training for their role; overall, they averaged 10.6 years of digital security experience. All participants had at least five years experience in secure computing and three in network engineering. All SDT participants currently work as developers, but are also expected to perform basic incident response duties, including triage if they identify an issue placing their systems at risk. In fact, in the year prior to our study they did respond to multiple incidents. While they may not perform incident response tasks as frequently or at the same scale as full-time incident response personnel, this as-necessary incident-response requirement means that all SDT participants are able to perform common incident response tasks. Further, the frameworks used in this study did not specify minimum incident-response skill qualification for use, and in some cases implied that anyone with security knowledge can design a playbook [8, 38, 55, 66]. As such, we considered SDT employees eligible to design and use playbooks and representative of junior incident-response professionals. We note that several developers outperformed full-time incident response personnel according to our expert review, and their perspective added richness to our analysis.

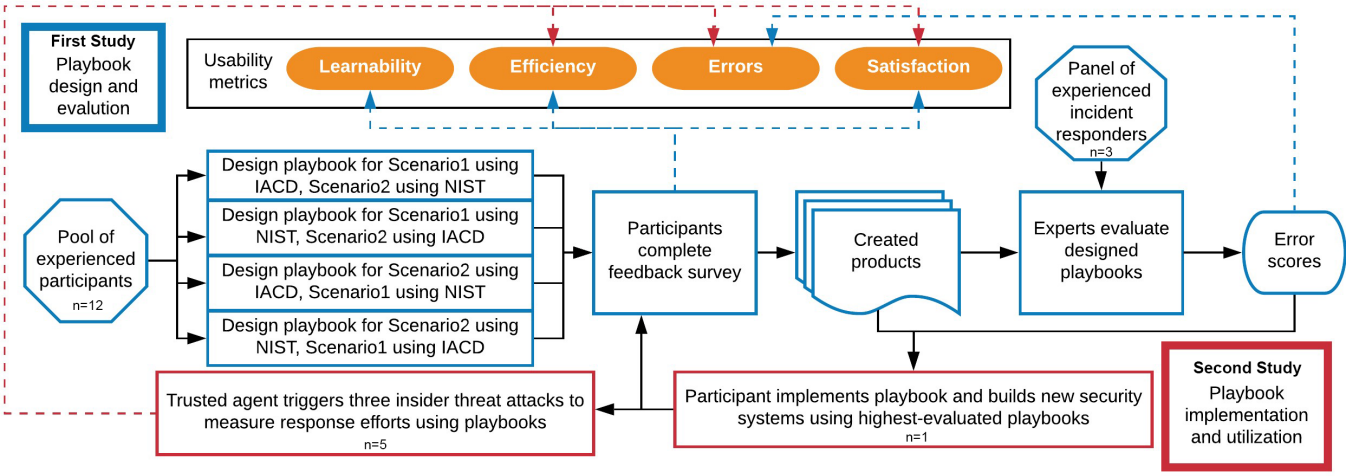

Figure 1: Our study protocol using two studies to measure the usability of playbooks and frameworks.

We recruited three expert evaluators (E1-3) via email, based on participant contact lists aggregated during previous research. We sought experts who had (1) used playbooks within real-world environments, (2) more than five years of experience working with incident response management, and (3) management experience within a security operations center. Our three participants more than met these criteria: E1 is the director of a security operations center with more than 300 employees; E2 is the Deputy Chief Information Security Officer (CISO) of one of the largest cities in the U.S.; and E3 is the CISO of a major U.S. financial institution. Collectively, they averaged 16.7 years of digital security experience. The use of three experts to evaluate playbooks is consistent with other studies relying on expert reviews for ground-truth evaluation [32, 65].

3.1.3 Playbook design. We first gave participants group-based, in-person instruction to familiarize them with using IACD and NIST frameworks, using an exemplar scenario (not in our main study): responding to spearphishing links [51]. The session included a slide presentation with step-by-step instructions for each step or phase of the relevant framework (see Section 2.1), collaborative development of response plans for spearphishing using each framework, a discussion on how to improve the co-developed plan, and a final questions-and-answer period.

We based these 30-minute introductory sessions on fundamentals from adult learning research, including learning through examples and hands-on implementation [6, 44]. The introductory session was jointly developed by two co-authors who were previously familiar with both frameworks; both co-authors had used the NIST framework as part of their full-time jobs and both performed an in-depth analysis of IACD's workflow as part of a month-long internal security audit. One of these co-authors, who has five years of experience designing incident-response scenarios for organizations as a consultant and four years of university-level teaching experience, also served as the instructor. The instructor communicated this experience to each class to establish credibility. Additionally, the instructor and co-developer of the introductory session conducted a pilot with two security professionals to ensure the framework concepts were thoroughly taught. Results from the pilot sessions indicated that the session was sufficient and no revisions were required.

Next, we asked each participant to design two incident response playbooks, one for each threat scenario (Section 2.3), using publicly-available references and relevant entries in the MITRE ATT&CK database [49]. We randomized the assignment of frameworks to scenarios, as well as the order in which framework/scenario pairs were assigned, to mitigate ordering effects. Each participant used each framework and each scenario once. As playbooks are generally customized to a particular environment, participants were instructed to design the playbook specifically to support their organization and its network.

After designing each playbook, participants completed an online survey to understand their impressions of playbook design and collect participant demographics. We then conducted open-ended follow-up interviews, averaging half an hour, with each participant. Questionnaires and interview guides for all segments of both studies are given in Appendices A and B, respectively. The survey and interview were designed to address perceived learnability as well as user satisfaction.

For all surveys and interviews in both studies, two co-authors jointly analyzed all open-ended questions using iterative initial coding [13, 54], building the codebook incrementally. Both coders had prior experience with incident response management as well as qualitative coding methods. For each response, coders independently applied category labels, then jointly discussed. When a new label was created, the coders re-coded previously coded responses accordingly. If coders disagreed, the disagreement was logged and then resolved by agreeing on a final label. We continued this process until we coded all responses, resolved all disagreements, and the codebook was stable (Appendix D).

3.1.4 Playbook evaluation. Participant perceptions are valuable, but to fully understand usability, error rate measurement is also needed. We recruited three expert evaluators with extensive playbook experience in enterprise environments to grade the playbooks. This expert review addresses learnability (actual), efficiency (whether participants were able to fully complete the task in the

	Brute Force		Credential Misuse	
	NIST	IACD	NIST	IACD
P1	E1, E2	–	–	E1, E2
P2	E3	–	–	E3
P3	E1, E3	–	–	E1, E3
P4	–	E1, E2	E1, E2, E3	–
P5	–	E3	E3	–
P6	–	E1, E3	E1, E3	–
P7	–	E2	E2	–
P8	–	E1	E2	–
P9	E2, E3	–	–	E2, E3
P10	E3	–	–	E3
P11	–	E3	E3	–
P12	–	E2	E3	–

Table 2: Participants' playbooks were evaluated by one of three experts (E1–E3).

alotted tiime), and errors (correctness). We measure playbook correctness in two dimensions: completeness (addresses the incident-response task beginning to end) and accuracy (no unnecessary or incorrect tasks).

For each playbook (anonymized before review), evaluators completed an online survey with closed- and open-ended questions about whether the playbook accomplishes its goals, contains enough detail to be implemented in a real environment, and contains any likely sources of error.

Because of limited time availability, each evaluator examined a subset of playbooks. To ensure consistency, 10 of the 24 total playbooks were assigned to two different evaluators. In the event of disagreement on any key attributes, the third evaluator was asked to review the playbook, and their response was used to break the tie; one playbook required a third evaluation. The remaining 14 playbooks received a single evaluation each. Table 2 shows the distribution of evaluations.

3.1.5 Limitations. All qualitative research should be interpreted in the context of its limitations.

Our recruitment materials explained the study's purpose. This may have caused self-selection bias: those most interested in the study topic opting to participate. Therefore, our participants may have been more engaged in the process than would be expected in general, allowing us to identify mistakes made in playbook development even by motivated practitioners.

Our results may also exhibit demand characteristics, in which participants are more likely to respond positively due to close interaction with researchers [36, 58, 70]. We mitigated this using online surveys to promote candid feedback. We also used both positively- and negatively-framed questions to ensure our participants could provide both perspectives.

In-depth, qualitative studies like ours are not intended to be generalizable. In particular, although NDC and SDT use organizational structures and technological resources common to many security-conscious organizations of similar size, specific details vary across organizations. Further, because our partners are U.S. organizations, we focus on playbook frameworks recommended by the U.S. government. Our two selected scenarios both focus on authentication; these were chosen to reflect needs and priorities of our partner organizations, but does narrow our scope somewhat. Nonetheless, we believe our approach can illuminate systemic issues that organizations must account for when adopting playbook frameworks.

This study is not a direct comparison of two frameworks, but rather an observational study attempting to identify benefits and shortcomings for each and for playbook frameworks in general. Our sample (n=12) is small, but it represents 58% of NDC's total workforce during the study period and 21% of SDT's employees, and aligns with common practices in HCI [11]. Because of the small sample and observational nature of the study, we do not attempt statistical comparisons.

For each qualitative finding, we provide a participant count for context. However, participants who did not mention a specific concept when responding to survey or interview questions may simply have failed to state it, so we do not use statistical hypothesis tests for these questions.

To limit biases, we selected incident response scenarios for which our partners did not have existing policies or procedures. This forced participants to build plans around technologies not yet in place, which may have contributed to a lack of detail in many playbooks (see results). However, this limitation is also realistic: Evaluator E2 said his organization often faces similar situations, and IACD cites the identification technology gaps as a key function of playbook design [38].

3.2 Results

Below we present the results of our first study, including participant feedback on the playbook design process, and expert evaluations of the accuracy and completeness of the designed playbooks.

All participants completed both assigned tasks, averaging 32.8 minutes (σ = 6.1) for IACD and 42.1 minutes (σ = 7.4) for NIST, which we consider acceptable for learnability and efficiency for a complex task of this type. We did not observe a noticeable difference in task completion time based on order.

Overall, participants reported a somewhat favorable perception of playbook design frameworks and their ability to assist with incident response efforts. In general, they appreciated thinking proactively and identifying solutions to realistic threats they might face in the future. However, expert evaluation suggested about half of designed playbooks were insufficient.

3.2.1 IACD feedback. Participants identified a variety of positive and negative features of the IACD framework.

Visualization is a key benefit. Most participants (n=10) identified the graphical depiction of required tasks as IACD's most beneficial attribute. P1 and P4 both indicated that visually distinguishing between human and automated tasks helped them focus on their roles during incident response, better understand how to leverage automated systems, and ensure they are compliant with mandatory controls.

Playbooks can help make up for lack of experience. P10 noted that IACD "helped me organize my thoughts and guided me through problem-solving"; even though he had never responded to the given scenario in a real event, he believed the framework was helpful in eliciting the necessary steps to handle it. P12 said, "it allows

…junior defenders to execute something without the guidance of a senior defender," especially when a speedy response is critical.

Grouping activities is difficult. Several participants (n=7) had difficulty grouping similar activities and functions, a core step many following steps build on. IACD does not provide a list of common groups to choose from, requiring users to determine their own groupings. P8 said it took him approximately one hour to develop a playbook, and a majority of that time was spent attempting to identify appropriate groupings to use.

Not have enough detail, not enough contingencies. Two participants felt the resulting products were too abstract for technicians to follow during incident-response events. P7 said "the diagram is nice and easy to follow, but probably also needs a document to go with it explaining in more detail what each action entails." Three others said they were never confident they had provided enough information. These comments foreshadow many of the difficulties junior technicians faced using the playbooks for incident response (Section 4.2.2).

P4 wanted more emphasis on loops (and their exit conditions), parallel activities conducted simultaneously by both humans and systems, and multiple possible end states based on conditional transitions. P12 felt similarly: "The point of a playbook is to recapture the initiative from the attacker by having several iterations of the OODA loop[3] unrolled," but felt the IACD did not support multiple paths. P12 suggested more modeling akin to the cyber kill-chain framework [77].

Identifying the initiating condition is most important. Ten participants agreed the "identify the initiating condition" step was the most important. All 10 described this first step as setting conditions for all follow-on steps, and noted that failure to recognize the initiating condition would significantly delay or even prevent incident response; this again foreshadows complications observed in the second study.

P11 mentioned the initiating condition will be in the playbook "table of contents," which technicians will reference when selecting a playbook during an event. The technician will therefore "need to be able to correlate what they believe to be occurring with how you laid out the [initiating condition] entry into the playbook." All participants used five or fewer words to describe their initiating conditions; playbook designers must use concise yet descriptive terms to cue defender actions.

Identifying regulatory requirements is least important. Eight participants indicated that "identify regulatory controls and requirements" was the least important section of the playbook, noting that regulatory compliance was not relevant to their job role or was someone else's responsibility. P9 said: "Compliance is less of an issue than actually solving problems." This sentiment aligns with prior work suggesting technicians view compliance as inhibiting security [2, 15].

3.2.2 NIST feedback. As with IACD, participants identified benefits and drawbacks to the NIST framework.

The framework was easy to understand. The most prevalent positive feedback was that NIST playbooks were easy to understand (n=5). P6 stated NIST was "[v]ery clear on what steps I needed to follow and what outputs are expected after each step," and P4 noted the "[r]esulting text could be passed on to anyone to help them perform initial triage."

The framework prompted for detail. Participants liked that the NIST framework prompted them to include as much detail as desired for response actions. They felt that fine-grained details would reduce uncertainty during response actions taken by junior defenders, rather than requiring novices to figure out on the fly how to implement abstract instructions. Two self-identified managers said the detail-oriented design of the framework would help security engineers to understand and implement controls or systems required by the playbook. For example, the NIST framework prompted these two participants to describe in detail the expected content for an alert email, providing guidance to security engineers who would be tasked with building or configuring the alert system.

The framework supports proactive planning. Two participants appreciated that the NIST framework allowed them to think about problems before a full-blown crisis occurred. P6 noted that NIST offered him an option to "identify possible solutions, identify gaps in technology, and have at least an initial plan in place for handling the situation."

More organization may be needed for novices. Participants (n=5) were concerned that it might be difficult for a novice to quickly orient themselves to a NIST playbook given the lack of visual aids, reflecting the importance of accommodating various learning styles [44]. "It's all just a bunch of words. During a crisis, you need something concise and clean to follow," P8 stated, after using NIST but prior to using IACD. "I liken it to if IKEA's instructions were text only. They wouldn't be as valuable." Participants suggested adding a headline-style title, executive summary, and visual cues to NIST playbooks.

Even more detail may be needed. Five participants wanted the NIST framework to require even more fine-grained detail. They felt the NIST framework was too open-ended, and would have liked the framework to prompt for exact commands that an analyst should execute, rather than requiring them to reference another guide or have the commands memorized. Two other participants felt the framework did not adequately prompt for decisions and branching plans to account for incident variability. Two participants believed NIST did not account for partially-complete tasks. P11 felt that if there is not a check on task completion, it could result in unnecessary actions just because it is in a playbook or missed opportunities to do something in parallel. These comments were similar to comments about the IACD framework, suggesting that our participants were looking for detailed, pre-planned responses that account for branching investigation paths.

Examples and instructions were again a challenge. Two participants wished for multiple NIST playbook examples as a reference during the design process. P4 noted that the NIST framework was too abstract in places, making it hard to understand what is required for each step.

Detection and analysis is most important, but no strong consensus. A plurality of participants rated "detection and analysis" the most critical response step (n=5). All five indicated that knowing a security event is underway and they need to take action, even

[3] A common decision process in operational environments: observe the situation (Observe), determine possible responses (Orient), decide which actions to take (Decide), and take these actions (Act).

if it is a false positive, is invaluable for a defender. P9 stated: "if you miss it, your plan for responding is useless." This finding parallels the importance participants placed on the initiating condition in IACD playbooks and suggests the importance of user interfaces that deliver critical information without overwhelming the analyst [10].

Similarly, participants narrowly scored containment, eradication, and recovery as the NIST framework's least important phase. Three participants said this step is only important if a problem is detected, investigated, and confirmed to be a true positive.

3.2.3 Expert evaluation. Overall, the playbooks participants created lacked sufficient detail and suggest that amount of experience did not meaningfully impact accuracy. Our expert evaluators assessed six of 12 IACD playbooks and five of 12 NIST playbooks as insufficiently detailed for incident response use; when asked if the playbook would be likely to adequately respond to the associated scenario (see Appendix A.2), IACD playbooks averaged 2.71 (σ = 1.40) out of 5 while NIST averaged 3.0 (σ = 1.57).

We next report on some common themes observed by the evaluators across playbooks from both frameworks.

Missing "implied" tasks. Our experts noted that many playbooks were missing what one evaluator referred to as "implied" tasks: necessary for the defensive strategy to succeed, but not codified within the playbook (IACD=4, NIST=3).

As one example, E1 noted that P3's IACD playbook was missing investigative steps necessary to confirm whether an alert is a true or false positive. To do this, the analyst must determine (in the case of credential misuse) where the login attempt originated from and why it occurred. In particular, when receiving an alert related to a honeyword, the analyst should check whether an administrator is performing a standard periodic login to ensure the account does not expire, before assuming a valid attack; this step was not included in the playbook. Further, E1 suggested that if the login attempt does appear to be a true positive, the analyst should then take steps such as searching Internet forums for the honeyword to investigate how and when the credentials were leaked.

Imprecise language may cause delay. Evaluators identified three IACD and four NIST playbooks with imprecise language or instructions that could delay response efforts. For example, some playbooks rely on client applications running on all workstations. If a technician pushes commands without first ensuring all clients are running, the technician may have to re-run the commands once they identify abnormalities in the results (wasting minutes or even hours).

Missing essential communications. Our experts agreed most playbooks missed at least some essential communications; the six lowest-scoring playbooks per framework all lacked this information. A sufficiently detailed playbook should include specific information (name, position, email address, phone number) about who to contact in different circumstances. Expert E2 compared this to a bomb threat checklist, which allows users to collect essential information and communicate it to the right people (e.g., calling 911) [71]. Business continuity and disaster preparedness experts recommend that incident responders have essential contact information, as well as fill-in-the-blank forms for communicating must-know information, readily available in case of crisis [73].

Missing humans in the loop. A few playbooks (IACD=2, NIST=2) relied too heavily on automation rather than including humans in the decision process. For example, several playbooks included automatic account disabling during a brute-force attack; while superficially sensible, this could result in locking out administrators, hampering the response. E2 emphasized that when decisions affect critical services or accounts, a human decision maker must be involved.

Too linear. Evaluators noted that, especially in the case of IACD, playbooks did not account for parallel actions that humans and automated systems could accomplish concurrently, increasing efficiency and reducing overall investigation time. For example, while a technician responding to a brute-force attack searches network traffic for any attempts occurring in real time, automated systems could query log data and locate attempts from the prior hours or days. We note that Steps 3 and 4 of the IACD framework ask the playbook designer to order tasks sequentially and group by functions, which may inhibit designers from planning parallel tasks.

Some include details and best practices. Our experts did identify multiple playbooks (IACD=2, NIST=4) with high-quality, fine-grained detail. P10 included references to best practices such as data loss prevention (preventing users with insufficient privileges from accessing sensitive data via credential misuse). Other playbooks detail steps for determining what information, if any, was stolen from the network in the event of a successful brute force attack or credential misuse.

One playbook was not just incomplete but incorrect. The evaluators identified one playbook (P7, NIST, credential misuse) as potentially impeding a technician's ability to respond to the incident. E2 believed this playbook's response events were ordered incorrectly, which could lead an incident responder to miss valuable information in one step that would be required later. Both E1 and E2 noted that this playbook lacked root-cause analysis and therefore could not lead to a successful resolution. Without root-cause analysis, it is possible for an attacker to regain access or spread throughout a network undetected while responders focus on inconsequential details.

After reviewing these expert findings, we conducted a follow-up interview with P7, a software developer with more than 10 years of experience in reverse engineering and secure code development but less exposure than any other participant to network defense or incident response. P7 said they understood the scenario, but did not find the NIST framework particularly usable. The playbook framework guidelines and reference material about each tested scenario from the MITRE database were not enough to help P7 avoid the ordering mistake that doomed the playbook.

3.3 Summary

This study suggests that the two frameworks we examined have only moderate usability. Although all participants completed each playbook design task in under 45 minutes, only about half were considered by experts sufficiently complete and correct for real-world use. Participants found the idea of playbooks, and some individual features of the two frameworks, useful, but also identified key weaknesses. In particular, participants placed high importance on identifying the triggering action (in both frameworks) and appreciated the visual process overview associated with IACD, but also

wanted detailed checklists. Participants appreciated that the NIST framework came closer to prompting for this amount of detail, but wanted the framework to go even further, such as requiring exact syntax for system queries and commands.

4 PLAYBOOK IMPLEMENTATION AND USE

We next detail our second study, which investigates playbook performance in incident response. The goal of this study was to understand whether playbooks developed using the two frameworks we examined could be implemented and used in practice to improve incident-response outcomes. In the process, we also examined *how* organizations implement and use playbooks, allowing us to make observations and recommendations that apply to playbooks more broadly.

To address these questions, we designed a field-deployment study focusing on implementation processes required to support the playbook's incident detection and response plan, then execution of the response plan during an incident. In a field deployment, researchers introduce a prototype in situ, "within the intended context of use" [61]. A field deployment enables otherwise unavailable "rich data about how closely a concept meets the target population's needs and how users accept, adopt, and appropriate a system in actual use over time" [61].

In our study, one participant implemented technical controls called for by the two playbooks, and five personnel from NDC participated in no-notice exercises that called for use of these implemented playbooks.

4.1 Method

For the field deployment study (December 2019–March 2020), we selected two playbooks that received high scores from the expert evaluators and spanned both frameworks and scenarios: P4's IACD playbook for brute-force login and P11's NIST playbook for credential misuse. We chose the highest scoring playbooks to provide an upper bound for playbook efficacy; we wanted to identify challenges that persist even with well-crafted playbooks. We asked one participant to engineer security solutions based on the playbooks, and then evaluated their usability during three controlled insider-threat events. We examine efficiency, errors, and satisfaction to measure usability.

Despite spending one year obtaining legal approvals and establishing study protocols with our partners, midway through our study the SDT legal advisors revoked our permission to modify network monitoring solutions at SDT. Therefore, no SDT employees were permitted to participate in this second study. We consider it acceptable to use P11's playbook, which was designed for SDT, within NDC, because (1) it scored well in the expert evaluation and was noted for containing fine-grained detail, and (2) it could easily be implemented as-is within NDC, because neither organization had pre-existing solutions in place, so implementation could start from a blank slate. As in the first study (Section 3), NDC leaders informed potential participants about the study, while emphasizing its voluntary nature. Again, participants were allowed to participate during work hours but not otherwise compensated and were assured that participation (or not) would have no effect on performance evaluations.

4.1.1 Playbook implementation. In this phase, one participant implemented new technical controls, based on the selected playbooks, to detect and respond to brute-force attacks and the misuse of valid credentials within their live network.

It was infeasible for more than one participant to perform implementation while interacting with the live network; however, this phase was necessary to enable evaluation during controlled events (Section 4.1.2). NDC leaders nominated one participant (who subsequently volunteered) for this phase. As such, we do not attempt to generalize any findings from this process, but instead briefly comment on our observations from this (previously unexplored in the literature) process.

After the participant implemented the two playbooks' controls, they completed a survey about the experience (Appendix A) and we conducted an in-depth interview (Appendix B). In this phase, we focus on the usability of the playbooks themselves, in the context of the frameworks used to design them.[4] The questions focused primarily on satisfaction from the point of view of the implementer. We used both positive and negative framing for our survey and interview questions to mitigate social desirability bias [26].

4.1.2 Playbook use during incident response. The second study's main goal was to evaluate the selected playbooks' usability during actual incident response. Given actual attacks' unpredictability, we worked with NDC leadership to conduct no-notice incident response exercises to trigger playbook use. Similar approaches are used in compliance programs, but to our knowledge have never been used to investigate playbooks.

After our first study (Section 3), NDC leadership instructed all employees to use the selected playbooks in their daily duties. Copies of each playbook were provided to participants and placed in a binder in NDC for easy access. Each technician received a 30-minute orientation by NDC leaders on how and when to use the playbook. Technicians were also asked to review the differences between the playbook they designed in the first study and the ones selected for use.

To maximize ecological validity, technicians were not informed the study would include incident-response exercises testing the playbooks. We discuss this deception further below.

NDC leaders identified one employee as a trusted agent to simulate the insider threat. We then coordinated directly with the trusted insider to schedule the simulated attacks; neither NDC leadership nor technicians received any advance notice of when they would occur. We triggered three no-notice incident response exercises on December 2, 2019 (IR1, brute-force); January 13, 2020 (IR2, credential misuse); and March 2, 2020 (IR3, credential misuse) to evaluate NDC's ability to use the playbooks over time. During these events, responders were required to log their usage of approved playbooks as well as any issues they encountered. (We note that NDC technicians are required to log daily actions in detail generally, not just during incident response.) Further, our trusted insider conducted and logged debriefs with each participant. Due to security concerns, we were not permitted to retain copies of the action or debrief logs, but one co-author reviewed all logs to verify the expected playbooks were used throughout.

[4]The implementer, P3, was familiar with both frameworks after participating in our first study.

To initiate the brute-force attack, the trusted agent used a script to rapidly attempt logins against actual user-level domain accounts throughout the enterprise, using randomly generated passwords. This attack included 50 total login attempts against two domain accounts. To initiate the credential misuse attack, the trusted agent successfully logged into a designated user-level domain account configured as a honeyword account. Appendix C provides additional details on how the trusted insider implemented these two attacks during the incident response exercises.

After each exercise, we asked each participant to complete a survey about the experience (Appendix A). We conducted in-depth follow-up interviews with each participant and with the trusted insider (Appendix B). We also reviewed NDC network and system logs related to the exercises. Together, the surveys, interviews, and logs allow us to evaluate the usability of the playbooks themselves; as part of our analysis process, we place the results in the context of the frameworks (and what we learned about the frameworks in the first study). We are able to evaluate satisfaction, errors, and efficiency; we obtain limited insight into learnability when individual participants first attempt to use the playbooks. We again analyzed the interviews as described in Section 3.1.3.

For this study, we consider incident response efforts taking less than 140 minutes to be a success. According to a CrowdStrike analysis, this is the time required for access expansion by many nation-state threat actors and criminals [18].

4.1.3 Recruitment and participants. Five people participated in our second study: one implemented security controls based on the selected playbooks and participated in two exercises, one participated in three exercises, and three participated in one exercise each (Table 1). Four participants had also participated in designing playbooks for the first study; P13 joined NDC in late February 2020 and only participated in the final incident response event (IR3). None of the participants had designed the playbooks we selected for implementation. As before, all participants knew about playbooks as an industry standard, but none had used a playbook to respond to an incident. Participants averaged 8.2 years of digital security experience.

Ethical considerations. No-notice exercises simulating real-world attacks carry several potential risks: they may create unnecessary participant stress or cause senior personnel to make unnecessary decisions based on a fictional threat. To mitigate these risks, we directed (in consultation with NDC leadership) the trusted insider to immediately inform participants who detected the event that it was an exercise. All subsequent written and verbal communications began with "EXERCISE" to indicate it was not a real event; this practice is common at NDC. Although participants were told the incident was an exercise, they were not told it was part of the playbook study.

Further, NDC leaders agreed not to consider participants' performance in the exercises (good or bad) in annual performance reviews, and instead treat the exercise as a learning opportunity to improve institutional practices. Finally, responding to a controlled event may detract from NDC's ability to respond to an actual threat. To mitigate this, we ensured the events occurred only on days when NDC was fully staffed. Only 2-3 participants engaged in each response effort. As is standard in deception studies, after the final exercise, we debriefed participants, explained the true nature of the study, and provided them with an opportunity to withdraw their data from the study; no participants withdrew. This study was approved by our institution's ethics review board.

Limitations. Due to security and legal concerns, we were only able to conduct this study with one partner organization and five participants. Small samples like ours are not atypical in HCI for in-depth observational studies in a field setting [11]. Further, four of the incident response participants also participated in developing playbooks during our first study, and one participant in the first no-notice exercise (P4) developed the playbook that was being tested. This may have introduced biases about what a playbook should look like in general. Specifically, it might be expected that P4 would be motivated to defend the usefulness of the playbook; however, this did not prove true in practice (see results below). Because of these factors, we provide observations from this unique opportunity to observe playbooks in use but do not attempt to generalize our findings.

Due to our sample size, we were only able to test a few of the playbooks crafted in the prior study. We chose the two best playbooks, so that they would have the best chance to demonstrate strengths and weaknesses even of well-written playbooks. Poorly written playbooks should be expected to have additional weaknesses; our study does not draw any conclusions as to the average quality of already-existing playbooks used by organizations currently.

Ethically, it was necessary to inform participants (after initial threat detection) that they were participating in an exercise. While this may somewhat degrade ecological validity, it does not impede our primary objective: evaluating playbook use without prior notice. On the other hand, our unique vantage from inside a real organization planning to respond to attacks of real concern — and updating these plans after each exercise — provides unusually strong ecological validity overall.

Together, our two studies provide a holistic, end-to-end view of designing, implementing, and using playbooks for incident response. This approach increases validity, while allowing us to take maximum advantage of our difficult-to-negotiate organizational partnerships. However, it does introduce limitations because the second study depends on the results of the first. We nonetheless believe that our overall observations provide a significant first contribution in this area.

4.2 Results

Below we present our observations about the use of playbooks to (1) implement new security controls based on the playbook and (2) use a playbook during an incident-response event. These results are based on observations, survey answers, and logged digital security artifacts throughout the network. We report participant demographics, participant feedback on implementing security controls, and participant feedback on using playbooks during three incident response events. These observations provide the first structured evaluation of playbook usability — and the effectiveness of playbooks designed using the NIST and IACD frameworks — from within a live security operations center.

4.2.1 Playbook implementation. Participant P3 — the most experienced defender at NDC, with 18 years of hands-on and management

experience — implemented the security controls called for by the two selected playbooks. P3 did not design either playbook selected for implementation; they spent approximately 30 minutes becoming familiar with the playbook requirements before implementing the requisite controls.

After assessing the playbooks, P3 determined that all of the requisite logging mechanisms (e.g., account login failures) and recorded network traffic already existed; all that was needed was a way to aggregate this data and correlate events to obtain meaningful information. After a three-week acquisition and change-oversight period (detailed below), P3 created within one hour a new alert dashboard and a data-analysis plan to populate the dashboard with events. For the first time, NDC had a real-time system to continually monitor the network for brute-force attacks and credential misuse. The dashboard is visible on a large monitor displayed in the front of the NDC workspace and is accessible from each analyst workstation. If either scenario is detected by the automated system, the dashboard shows an alert that investigation is needed.

Oversight requirements, change control, and purchasing — mainly associated with buying new equipment capable of storing and processing required amounts of network data — added about three weeks to the implementation process. Potential delays of this kind should be taken into account when planning to adopt playbooks and new security controls. P3 suggested that playbooks explicitly include implementation requirements such as equipment specifications and change control procedures to make this more transparent.

Implementation feedback. Overall, our participant responded neutrally regarding playbook usefulness. P3 felt playbooks might be useful for more complex problems, but were not especially useful or time-saving for smaller-scale issues, like those in our scenarios.

P3 also reported needing to rely heavily on their security engineering background, as they found both playbooks too abstract to directly guide the development of new security controls. P3 slightly preferred the NIST playbook, citing previous familiarity with the framework (which they had seen but never used prior to the study). They reported spending more time with the IACD playbook to ensure an effective outcome, but attributed this primarily to lack of familiarity with the framework. P3 hypothesized that IACD's visual presentation would be easier for less experienced technicians to work with, but found the resulting playbook too generic for direct implementation. P3 reported making many notes to expand on each step and recommended adding complementary reference sheets providing detailed instructions for each step.

4.2.2 Playbooks in use. We next present observations about the use of incident response playbooks by five participants in an enterprise environment during three no-notice incident-response events. During the first two events, the playbooks had mostly negative results: experienced security professionals did not feel the playbooks added much value to their response efforts, and a junior analyst struggled with detecting the events. After modifying the playbooks based on feedback from participants and our experts, as well as lessons learned from the first two incident response events, the perceived utility of the playbooks increased noticeably during the third event.

IR1 outcomes. During the first event, our trusted agent initiated a no-notice brute-force login attack against two user accounts. P4 (who had developed the playbook being tested) was the first to detect the event, notifying his supervisor of a potential incident 10 minutes after attack execution. P3 independently detected the event two minutes later. The supervisor informed both participants that this was an exercise, and that they were to finish investigating the breach independently and without informing other technicians. Within one hour of detecting the threat, both participants successfully identified the point of origin for the attack, recommended removing the infected system from the network, identified the person using the now-quarantined system, and notified the physical security team about the (notional) insider threat.

P2, however, did not detect the attack until 14 days later. P3 and P4 left all attack logs in place after their investigation concluded, to provide P2 with more chances to detect the attack in the future (and allow us to assess P2's response decisions). According to the NDC supervisor, it is common for multiple technicians to check the same security logs and dashboard for alerts for redundancy. The brute-force alert appeared on P2's dashboard at least 19 different times during morning and evening checks, but P2 did not recognize it.

Once P2 realized an event had occurred, they made an initial report to a supervisor within 10 minutes. After that, it took P2 four hours to successfully identify the root cause of the attack (and the associated user) and submit an incident report to the physical security team. Altogether, 335 hours elapsed between the initiation of the attack and P2's report.

IR1 feedback. We asked P3 and P4 how the playbook contributed to their success. Both said the brute-force playbook (IACD, by P4) clearly guided them to correct actions. However, both largely credited their past security experience rather than the playbook for the successful outcome. P3 noted they "got to a point where I knew what to do and looking at the playbook just slowed me down." In particular, both said they relied on knowledge from past experience to make up for missing details, such as access log query syntax to determine who was logged in at the system that initiated the attack. P4 said the playbook "would have been more useful during IR if it had the commands instead of having to Google them."

P2 confirmed that they used the playbook, but nonetheless missed the alert associated with the brute-force attack all 19 times. They said, "the playbook did not have enough information for us to conduct a step-by-step walk-through. Because I was unfamiliar with the new [brand] dashboard, I didn't know what the alerts would look like compared to normal data." This suggests the 30-minute orientation session was insufficient for this novice defender to learn how and when to use the playbook. Because P2 missed the alerts, they never identified the initiating condition that requires a technician to use the incident response playbook. This aligns with participants' comments in Section 3.2 that the triggering event is the most important step in a playbook design framework.

Further, P2 commented that since it was their "first time using [the playbook], we needed to work out who to inform and when. Identifying critical information for each step and who needs to know it would have saved time." P2's supervisor rejected three reports during the 10-minute initial response window because they lacked sufficient detail.

P2 also noted that having two playbooks available delayed their response: faced with a stressful situation, P2 read through both playbooks to ensure they were using the correct one. There was no table of contents and no easily identifiable markings in the playbook headers (like bolded or colored text) to help an analyst quickly choose the correct playbook. "It would help to more clearly identify which playbook is for which event." P2 commented that this problem could become worse with more playbooks for other kinds of incidents.

IR2 outcomes and feedback. P2 and P3 participated in IR2, a credential-misuse event conducted in January 2020. (P4 was unavailable due to off-site training.)

P3 again successfully responded to the incident, performing nearly identically to their response in IR1 and resolving the situation in 65 minutes. P2 again failed to recognize the significance of alerts generated during the attack; we chose to end the incident after 11 days with no recognition.

As with IR1, both participants noted the NIST-framework credential-abuse playbook (by P11) lacked sufficient detail, and P3 again relied heavily on past experience. P2 provided two possible explanations for failing to detect the incident. Primarily, P2 said they took "several weeks off from work for the holiday season," causing playbook familiarity to atrophy. Second, P2 did not believe they would be evaluated with the playbook a second time. These comments align with findings from previous adult learning theories about the importance of continual, hands-on practice with new concepts [6, 44].

Resetting after failure. After IR1 and IR2, we worked with NDC to revise the playbooks, applying feedback received in the first study and in this study so far. Refining a prototype during a field deployment is common and valuable [61]. This also fits NIST guidelines for refining incident response procedures based on lessons learned after an event [14].

In particular, we sought to address three interrelated concerns that surfaced repeatedly: 1) playbooks contained insufficient detail for use during incident response, 2) too much experience was required to use the playbooks properly (making things difficult for novices), and 3) identifying a playbook trigger was the most critical challenge.

First, we asked all NDC participants to collectively improve both playbooks by adding details appropriate for use by an entry-level technician, including click-by-click instructions for GUIs and specific text for command-line interfaces. As the playbooks expanded in detail, technicians documented lengthy processes by creating complementary guides alongside the playbooks. Technicians also made changes focused on recognizability: creating a table of contents for all playbooks, using bold-font titles on each playbook, and including summaries for what the playbooks are intended to help with.

Collectively, NDC participants walked through both scenarios in a tabletop exercise, annotated playbook gaps, and later made updates accordingly. For example, this process revealed that communication instructions were not yet sufficiently detailed. After the exercise, technicians made cheat sheets documenting which information must be reported to whom for each playbook step. P2 said it "became more clear that communication was critical during these stressful events."

Next, we asked NDC to implement a more collaborative model in which technicians could work together while using playbooks. In particular, junior technicians were encouraged to ask questions and seek advice from senior leaders and technicians. After these changes, we allowed one month to pass before initiating our final incident-response event.

IR3 outcomes and feedback. The trusted insider initiated IR3 (credential abuse) in March 2020. Volunteers P2 and P13 were selected by NDC leadership as participants. P13 joined NDC two weeks prior to IR3 and completed all the on-board training related to playbooks that NDC had implemented. P3 and P4 were unavailable for IR3 due to other job obligations.

Both participants successfully detected (P2=3 min, P13=5 min) and responded to (P2=90 min, P13=104 min) the threat within our 140-minute threshold. P2 said that the more detailed steps added to the playbooks and the new mentorship program helped drastically with their understanding of how to respond to events and communicate more effectively with their supervisors. P13 said, "As a new employee, it helped me better understand our mission and how to do my job if a supervisor is not available." By completing the on-boarding training using playbooks, P13 felt they more completely understood their role and responsibilities within NDC: "This is what I do, this is what is required of me." This supports previous claims regarding the usefulness of playbooks for helping professionals learn new responsibilities and technologies [33].

While we cannot generalize from this one experience, IR3 suggests that, when they include sufficient detail as well as additional practice and orientation, playbooks may be useful to help junior technicians with incident response. Future work is needed, however, to investigate the extent to which different elements of the implemented improvements are useful.

4.3 Summary

Even the top-rated playbooks designed using both frameworks required significant modifications to be useful, especially for junior technicians. During IR1 and IR2, experienced technicians used the playbooks — created within 45 minutes in the first study — successfully, but credited their success to prior experience rather than the playbooks. A junior technician was unable to respond within the expected time window in either case.

After updating the playbooks (and associated organizational processes) using lessons learned from our studies, two junior technicians used them to mitigate a credential misuse attack within 110 minutes.

These findings suggest that the playbook frameworks we assessed are not sufficient on their own, but may be useful as part of a larger process for developing and institutionalizing playbooks; further research is required for validation.

5 RELATED WORK

We are not aware of other research exploring the application of playbooks or the use of playbook frameworks within real-world information security settings. We report here on other research examining security operations in general, as well as how playbooks — and guidelines for developing them — are used in other domains.

5.1 Security operations

In an ethnographic study, Sundaramurthy et al. find that frequent "burnout" and turnover among security analysts, caused in part by poor management and communication as well as overly repetitive tasks and poorly aligned metrics, decreases the overall effectiveness of security operations [67]. Kokulu et al. found that analysts identify similar issues, particularly poor communications and disjoint priorities, as key challenges in their work [43]. Dietrich et al. asked system administrators — a separate but often overlapping group from security analysts — about security misconfigurations, finding again that management issues, lack of communication, and overly repetitive tasks impede the application of known security fixes [23].

In follow-up work, Sundaramurthy et al. developed and tested tools to improve security operations, concluding that tools must be customize to specific operational environments [68]. A case study within the New York City Cyber Command found that teaching staff to use threat modeling improved communication and enabled proactive planning to strengthen security [64].

In this work, we studied playbooks rather than operations in general, but we similarly found that tailoring to the environment and emphasizing communications are critical aspects for success.

5.2 Playbooks in other domains

Business continuity plans (BCPs) help minimize financial losses, ensure the continuation of core functions, ensure resource availability, and train employees. Many organizations are required by insurance or regulations to have BCPs. Numerous references provide reporting templates for communicating essential information, how-to guides for audits, and training scenarios for a vast array of situations that may cause damage to a business [7, 30, 35]. BCPs typically contain fine-grained detail to assist with implementation and auditing (similar to the playbooks used during IR3). BCP training varies, but typically involves intricate exercises [35].

U.S. government agencies maintain playbooks for natural disaster continuity and health emergency preparedness, among other crises [73]; libraries of pre-made disaster response playbooks are available for reference [27, 53, 76].

In the medical field, crisis resource management combines standard medicinal practices with non-technical skills to ensure exposure to best practices for likely emergency situations [12]. Studies found that simulated rehearsals with response action "playbooks" gave participants confidence that the lessons learned would transfer to real-life situations [59].

Pilot training uses simulation to allow aviators to experience dangerous situations (even cyber attacks) prior to entering a real cockpit [28, 62]. This readiness and preparedness goes beyond the cockpit, with much emphasis on future readiness when presented with large-scale disruptions like natural disasters or acts of terrorism [24, 34]. Allowing international organizations and pilots alike to rehearse and refine their playbooks improves their ability to handle threats.

6 DISCUSSION AND CONCLUSIONS

Using two studies derived from partnerships with two multi-million-dollar security organizations, we provide the first structured evaluation of playbook framework usability and playbook effectiveness within an enterprise environment. Overall, our findings suggest that playbook frameworks are moderately usable for technicians designing playbooks, but do have important areas for improvement. We find that even the top rated playbooks generated using these frameworks may require significant modification to meet their goals of helping technicians implement the associated security controls and then respond to security incidents. Perhaps the most significant drawback, observed in all phases of our evaluation, is that the frameworks do not elicit playbooks written in sufficient detail for real-world use. More experienced technicians were able to rely on their prior knowledge to fill in these gaps, but junior participants struggled to make use of the playbooks. Additionally, our findings align well with usability concerns from other domains: notifications within life-or-death naval interfaces must prompt user response actions [60] and communication (both internal and external) during disaster management is critical [30]. Based on these results, we make several suggestions for information that should be included in playbooks — and therefore elicited by frameworks — and changes to associated organizational processes.

6.1 Improvements to playbooks and frameworks

Our expert reviews and field study demonstrated the need for specific information or playbook structures which are not currently suggested in the NIST or IACD frameworks. Here we suggest several elements which should be included in playbooks to make them more useful; accordingly, frameworks should therefore prompt for their inclusion.

Playbooks should be easy to quickly select and comprehend. Playbooks must be usable during stressful situations. Minor changes such as using boldface titles, using a table of contents to organize multiple playbooks, and affixing summaries atop playbooks seemed to help technicians in our field deployment study quickly select the appropriate playbook for a given situation. Playbook frameworks could provide well structured templates to support standardization and easy reference across playbooks.

Playbooks should provide detail. Technicians of all experience levels indicated that highly-detailed instructions (from best practices) are critical. Playbook designers should not assume user expertise with various technology platforms or command-line interfaces. Instead, they should provide detailed instructions both for implementing required security controls and for incident response. These recommendations are congruent with incident response playbooks released by Microsoft after this study [47]. Frameworks should emphasize the importance of detailed instructions and provide examples of sufficient and insufficient levels of detail.

Links to outside resources may be helpful to provide detail without information overload. Our participants consistently asked more detail; however, there can be a fine line between enough and too much information. Too much information could slow response time as technicians sift through details to determine appropriate next actions. One possible mitigation could be to include links and references to external resources such as best-practice repositories, allowing designers to convey important information without overly cluttering the playbook itself. Further research is needed to explore the appropriate balance.

Playbooks should prompt regular communication. The NIST playbook framework emphasizes communication throughout incident response, but IACD allows the designer to determine which communication is essential or optional. Our studies suggest that playbooks should prompt technicians about what information to record as well as who to inform and when. Incident response communication issues (e.g., inaction or incomplete information sharing) were also seen in disaster management training scenarios [30]. Adapting best practices such as fill-in forms (like those found in DHS bomb threat checklist [71]) could be useful for improving communication, and frameworks could support this approach with templates and examples.

Frameworks should provide structured guidance for task grouping. Playbook designers using IACD had difficulty grouping tasks together, in part because the instructions left the choice of groupings open-ended and provided little guidance for how to identify and label groups. Playbook frameworks might consider providing multiple-choice options for category selection, guides with more detailed prompts, or a large corpus of training examples annotated with explanations.

Playbooks should support non-linear actions. Playbook frameworks should prompt designers to plan for non-linear actions: accomplishing tasks in parallel to expedite investigation and accounting for multiple scenarios that may occur during response. Offering best practices for a variety of likely encounters and adversarial actions could allow responders to maintain momentum during an investigation.

Playbooks should state intent, not just actions. We also suggest including the intent associated with every task within the playbook. Helping users understand why a task is relevant may allow them to exercise initiative and improve overall response efforts [22]. Additionally, our experts suggested that intent specification may help security engineers who are implementing automation solutions based on playbook design to better understand and meet requirements. Again, frameworks have a role to play in prompting playbook designers to include intent specification.

6.2 Improvements to organizational processes to support playbook adoption

Our results also suggest potential changes beyond the playbooks themselves, which would likely improve incident response using playbooks.

Initiating alerts should be meaningful and noticed. Technicians must understand the initiating condition for incident response and be able to detect it – everything that follows the initiating condition is irrelevant if defenders do not recognize the need for action. Engineers who implement detection mechanisms must generate meaningful alerts and should consider requiring a technician's acknowledgement [17]. This is consistent with usability concerns for life-or-death naval interfaces [60].

Tabletop exercises should be conducted to identify playbook gaps. During the evaluation phase, all three experts recommended using tabletop exercises [75] to iteratively update each playbook until it is sufficiently detailed and tailored specifically for the local environment. These exercises can be used to validate that the playbook is complete and that all necessary policies and procedures are in place to support incident response. These exercises were perceived as helpful to the playbook revision process we observed.

Improving perceived playbook usefulness can fuel adoption. Expectancy theory [3] suggests that if playbooks do not feel useful, it is unlikely they will be used. We hope that improving playbooks themselves, as described above, will improve perceived usefulness, but organizational culture around playbooks may also play an important role. In our second study, organizational improvements such as mentorship programs, peer partnering, and continual reinforcement of the playbook process were cited by our participants as helpful in improving their perception of playbook usefulness.

Organizational constraints' impact on playbook design and use should be considered. Finally, we argue that playbook designers must consider organizational concerns and processes. Particular constraints, such as requiring approval to make changes to a network or limiting hardware purchases to approved vendors, may shape an organization's incident response strategy and therefore its playbook design. Designing a playbook that meets best practices and conforms to local constraints may require significantly more effort and time than the averages in Section 3.1.2.

6.3 Lasting playbook adoption at NDC

One year after our study concluded, NDC employees briefly described to us their continued adoption of incident response playbooks, which are in sustained use within daily network defense operations. NDC found considerable benefit in using the playbooks to on-board new employees. They internally developed seven additional playbooks for scenarios defenders are likely to encounter, to complement the two from our study. These playbooks are used for routine familiarization training every month. However, due to high operational demands, NDC had not (as of our communication) conducted any additional incident response exercises since the study concluded.

6.4 Future work

As a first, exploratory step, the goal of this work is not to provide a final answer. Instead, our results provide insights grounded in a real operational environment, which can be tested in follow-on work. Our preliminary evaluation of the usability of playbook frameworks and resulting playbooks suggests many potential directions for future work. First, further research is needed to understand whether and how our results generalize to organizations with different sizes, cultures, existing security technologies, and levels of technician experience. Additionally, future work should consider whether our results generalize to other attack scenarios less familiar to the participants. Direct comparison of frameworks — which may require sacrificing some ecological validity to obtain larger sample sizes and more experimental control — is also a critical area for continuing research.

Another possible research direction is to evaluate whether adapting existing playbooks, drawn from numerous community contributions [8, 38, 42, 48, 55], is more or less effective than using a framework to generate a brand-new playbook tailored to a specific environment.

Additionally, we suggest exploring how security automation can support many of the manual tasks executed throughout these studies. IACD specifically was designed to help automate response actions; understanding effective ways to share vendor-specific automation methodologies may assist organizations in using playbook frameworks.

ACKNOWLEDGMENTS

We thank the participants for contributing to our research, and the anonymous reviewers whose suggestions improved the paper. The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

REFERENCES

[1] Andy Applebaum, Shawn Johnson, Michael Limiero, and Michael Smith. 2018. Playbook oriented cyber response. In *2018 National Cyber Summit (NCS)*.
[2] Hala Assal and Sonia Chiasson. 2018. Security in the software development lifecycle. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS)*.
[3] John W Atkinson. 1957. Motivational determinants of risk-taking behavior. *Psychological review* 64, 6, Pt.1 (1957), 359.
[4] Amazon AWS. 2020. AWS incident response runbook samples. https://github.com/aws-samples/aws-incident-response-playbooks/tree/0d9a1c0f7ad68fb2c1b2d86be8914f2069492e21 (Accessed 01-08-2022).
[5] Tucker Bailey, Josh Brandley, and James Kaplan. 2013. How good is your cyberincident-response plan? *McKinsey on Business Technology* 31 (2013), 16–23.
[6] Albert Bandura and Richard H Walters. 1977. *Social learning theory*. Prentice-Hall.
[7] Michael Blyth. 2009. *Business continuity management: Building an effective incident management plan*. John Wiley & Sons.
[8] Jeff Bollinger, Brandon Enright, and Matthew Valites. 2015. *Crafting the InfoSec playbook: Security monitoring and incident response master plan*. O'Reilly Media, Inc.
[9] David Botta, Rodrigo Werlinger, André Gagné, Konstantin Beznosov, Lee Iverson, Sidney Fels, and Brian Fisher. 2007. Towards understanding IT security professionals and their tools. In *Third Symposium on Usable Privacy and Security (SOUPS)*.
[10] Cristian Bravo-Lillo, Lorrie Cranor, Saranga Komanduri, Stuart Schechter, and Manya Sleeper. 2014. Harder to ignore? Revisiting pop-up fatigue and approaches to prevent it. In *Tenth Symposium On Usable Privacy and Security (SOUPS)*.
[11] Kelly Caine. 2016. Local standards for sample size at CHI. In *ACM Conference on human factors in computing systems (CHI)*. 981–992.
[12] Belinda Carne, Marcus Kennedy, and Tim Gray. 2012. Crisis resource management in emergency medicine. *Emergency Medicine Australasia* 24, 1 (2012), 7–13.
[13] Kathy Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis*. Sage.
[14] Paul Cichonski, Tom Millar, Tim Grance, and Karen Scarfone. 2012. Computer security incident handling guide. *NIST Special Publication* 800, 61 (2012), 1–147.
[15] Robert Clark. 2018. Compliance != security (Except when it might be). In *Enigma 2018*. https://www.usenix.org/node/208142
[16] Allan Cook, Helge Janicke, Richard Smith, and Leandros Maglaras. 2017. The industrial control system cyber defence triage process. *Computers & Security* 70 (2017), 467–481.
[17] Lorrie F Cranor. 2008. A framework for reasoning about the human in the loop. In *Usability, Psychology, and Security Workshop (UPSEC)*.
[18] CrowdStrike. 2019. 2019 CrowdStrike global threat report: Adversary tradecraft and the importance of speed. https://go.crowdstrike.com/rs/281-OBQ-266/images/Report2019GlobalThreatReport.pdf
[19] Cybersecurity and Infrastructure Security Agency. 2019. *Chemical sector-specific agency incident management and coordination playbook*. Technical Report. U.S. Department of Homeland Security.
[20] Cybersecurity and Infrastructure Security Agency. 2020. *Technical Approaches to Uncovering and Remediating Malicious Activity*. Technical Report. US Department of Homeland Security.
[21] Cybersecurity and Infrastructure Security Agency. 2021. *Cybersecurity Incident & Vulnerability Response Playbooks*. Technical Report. US Department of Homeland Security.
[22] Richard Dempsey and Jonathan M Chavous. 2013. Commander's intent and concept of operations. *Military Review* 93, 6 (2013), 58–66.
[23] Constanze Dietrich, Katharina Krombholz, Kevin Borgolte, and Tobias Fiebig. 2018. Investigating system operators' perspective on security misconfigurations. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. https://doi.org/10.1145/3243734.3243794
[24] Jacques Dopagne. 2011. The European air traffic management response to volcanic ash crises: Towards institutionalised aviation crisis management. *Journal of Business Continuity & Emergency Planning* 5, 2 (2011), 103–117.
[25] Josiah Dykstra and Celeste Lyn Paul. 2018. Cyber Operations Stress Survey (COSS): Studying fatigue, frustration, and cognitive workload in cybersecurity operations. In *11th USENIX Workshop on Cyber Security Experimentation and Test (CSET)*.
[26] Allen L Edwards. 1957. *The social desirability variable in personality assessment and research*. Dryden Press.
[27] Federal Emergency Management Agency. 2020. FEMA Playbooks.
[28] Christina Rosa Filipowski. 2017. *A qualitative case study of airline pilot leadership behaviors and practices during crisis situations*. Ph.D. Dissertation. Grand Canyon University.
[29] Charles Frick. 2020. New cyber defense feed protects government systems in live trial across four states. https://www.jhuapl.edu/PressRelease/201203-APL-leads-cyber-defense-feed-protecting-live-active-government-systems
[30] Julia Graham and David Kaye. 2015. *A risk management approach to business continuity: Aligning business continuity and corporate governance*. Rothstein Publishing.
[31] Microsoft Enteprise Cybersecurity Group. 2017. *Incident response reference guide: First aid for major cybersecurity incidents*. Technical Report. Microsoft.
[32] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G Shin, and Karl Aberer. 2018. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th USENIX Security Symposium*.
[33] Bob Hayes and Kathleen Kotwica. 2013. *Business continuity: Playbook*. Elsevier.
[34] Joan C Henderson. 2008. Managing crises: UK civil aviation, BAA airports and the August 2006 terrorist threat. *Tourism and Hospitality Research* 8, 2 (2008), 125–136.
[35] Andrew Hiles. 2010. *The definitive handbook of business continuity management*. John Wiley & Sons.
[36] Allyson L Holbrook, Melanie C Green, and Jon A Krosnick. 2003. Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias. *Public opinion quarterly* 67, 1 (2003), 79–125.
[37] Jesper F Hopstaken, Dimitri Van Der Linden, Arnold B Bakker, and Michiel AJ Kompier. 2015. A multifaceted investigation of the link between mental fatigue and task disengagement. *Psychophysiology* 52, 3 (2015), 305–315.
[38] IACD. 2019. About IACD. https://www.iacdautomate.org/aboutiacd
[39] Incident Response Consortium. 2019. Playbooks. https://www.incidentresponse.com/playbooks/ (Accessed 2021.).
[40] Joseph R. Biden Jr. 2021. Executive order on improving the nation's cybersecurity. https://www.whitehouse.gov/briefing-room/presidential-actions/2021/05/12/executive-order-on-improving-the-nations-cybersecurity/
[41] Ari Juels and Ronald L Rivest. 2013. Honeywords: Making password-cracking detectable. In *ACM SIGSAC Conference on Computer & Communications Security (CCS)*.
[42] Jason Kick. 2014. *Cyber exercise playbook*. Technical Report. MITRE Corporation.
[43] Faris Bugra Kokulu, Ananta Soneji, Tiffany Bao, Yan Shoshitaishvili, Ziming Zhao, Adam Doupé, and Gail-Joon Ahn. 2019. Matched and mismatched SOCs: A qualitative study on security operations center issues. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*.
[44] Alice Y Kolb and David A Kolb. 2005. Learning styles and learning spaces: Enhancing experiential learning in higher education. *Academy of Management Learning & Education* 4, 2 (2005), 193–212.
[45] J. Lazar, J.H. Feng, and H. Hochheiser. 2017. *Research Methods in Human-Computer Interaction*. Elsevier Science.
[46] Dimitri Van Der Linden, Ger PJ Keijsers, Paul Eling, and Rachel Van Schaijk. 2005. Work stress and attentional difficulties: An initial study on burnout and cognitive failures. *Work & Stress* 19, 1 (2005), 23–36.
[47] Microsoft. 2021. Incident response playbooks. https://docs.microsoft.com/en-us/security/compass/incident-response-playbooks
[48] Devesh Mishra. 2018. Cybersecurity playbook – An executive response. *Available at SSRN 3240285* (2018). https://ssrn.com/abstract=3240285
[49] MITRE. 2019. ATT&ACK. https://attack.mitre.org
[50] MITRE. 2019. Brute Force. https://attack.mitre.org/techniques/T1110/
[51] MITRE. 2019. Phishing: Spearphishing Link. https://attack.mitre.org/techniques/T1192/
[52] MITRE. 2019. Valid Accounts. https://attack.mitre.org/techniques/T1078/
[53] Paula Murrain-Hill, C Norman Coleman, John L Hick, Irwin Redlener, David M Weinstock, John F Koerner, Delaine Black, Melissa Sanders, Judith L Bader, Joseph Forsha, and Ann R Knebel. 2011. Medical response to a nuclear detonation: Creating a playbook for state and local planners and responders. *Disaster Medicine and Public Health Preparedness* 5, S1 (2011), S89–S97.
[54] Johnny Salda na. 2014. *The coding manual for qualitative researchers* (2 ed.). Sage.
[55] National Institute of Standards and Technology. 2014. NIST Cybersecurity Framework. https://www.us-cert.gov/ccubedvp/cybersecurity-framework

[56] Jakob Nielsen. 2012. Usability 101: Introduction to Usability. https://www.nngroup.com/articles/usability-101-introduction-to-usability/ Accessed: 2022-01-08.
[57] Cyril Onwubiko and Karim Ouazzane. 2020. SOTER: A playbook for cybersecurity incident management. *IEEE Transactions on Engineering Management* (2020).
[58] Martin T Orne. 1962. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American psychologist* 17, 11 (1962), 776.
[59] Martin Reznek, Rebecca Smith-Coggins, Steven Howard, Kanthi Kiran, Phillip Harter, Yasser Sowb, David Gaba, and Thomas Krummel. 2003. Emergency Medicine Crisis Resource Management (EMCRM): Pilot study of a simulation-based crisis management course for emergency medicine. *Academic Emergency Medicine* 10, 4 (2003), 386–389.
[60] Dylan D Schmorrow. 1998. *A benchmark usability study of the tactical decision making under stress decision support system.* Technical Report. Naval Postgraduate School, Monterey, CA.
[61] Katie A Siek, Gillian R Hayes, Mark W Newman, and John C Tang. 2014. Field deployments: Knowing from using in context. In *Ways of Knowing in HCI*, Judith S Olson and Wendy A Kellogg (Eds.). Springer, 119–142.
[62] Matthew Smith, Martin Strohmeier, Jonathan Harman, Vincent Lenders, and Ivan Martinovic. 2020. A view from the cockpit: Exploring pilot reactions to attacks on avionic systems. In *The Network and Distributed System Security Symposium (NDSS).*
[63] Rock Stevens. 2017. Calcifying crisis readiness. In *31st USENIX Large Installation System Administration Conference (LISA).*
[64] Rock Stevens, Colin Ahern, Daniel Votipka, Elissa Redmiles, Patrick Sweeney, and Michelle L Mazurek. 2018. The battle for New York: A case study of applied digital threat modeling at the enterprise level. In *27th USENIX Security Symposium.*
[65] Rock Stevens, Josiah Dykstra, Wendy Knox Everette, James Chapman, Garrett Bladow, Alexander Farmer, Kevin Halliday, and Michelle L Mazurek. 2020. Compliance cautions: Investigating security issues associated with U.S. digital-security standards. In *The Network and Distributed Systems Security Symposium (NDSS).*
[66] Blake E Strom, Andy Applebaum, Douglas P Miller, Kathryn C Nickels, Adam G Pennington, and Cody B Thomas. 2018. *MITRE ATT&CK: Design and philosophy.* Technical Report. MITRE Corporation.
[67] Sathya Chandran Sundaramurthy, Alexandru G Bardas, Jacob Case, Xinming Ou, Michael Wesch, John McHugh, and S Raj Rajagopalan. 2015. A human capital model for mitigating security analyst burnout. In *Eleventh Symposium On Usable Privacy and Security (SOUPS).*
[68] Sathya Chandran Sundaramurthy, Michael Wesch, Xinming Ou, John McHugh, S Raj Rajagopalan, and Alexandru G Bardas. 2017. Humans are dynamic – our tools should be too. *IEEE Internet Computing* 21, 3 (2017), 40–46.
[69] Eric C Thompson. 2018. *Cybersecurity Incident Response: How to Contain, Eradicate, and Recover from Incidents.* Apress.
[70] Roger Tourangeau and Ting Yan. 2007. Sensitive questions in surveys. *Psychological Bulletin* 133, 5 (2007), 859.
[71] U.S. Department of Homeland Security. 2019. DHS bomb threat checklist. https://www.cisa.gov/sites/default/files/publications/dhs-bomb-threat-checklist-2014-508.pdf
[72] Virtual Intelligence Briefing (VIB). 2017. The state of incident response 2017. https://a51.nl/sites/default/files/pdf/The%20State%20of%20Incident%20Response%202017.pdf
[73] Andy Wapling and Chloe Sellwood. 2016. *Health Emergency Preparedness and Response.* CABI.
[74] Shaun Waterman. 2017. Looking to fit it all together, banks adopt standards for cyber automation and integration. https://www.cyberscoop.com/banks-fs-isac-jhu-apl-ais-iacd/ (Accessed 01-08-2022).
[75] Gregory B White, Glenn Dietrich, and Tim Goles. 2004. Cyber security exercises: Testing an organization's ability to prevent, detect, and respond to cyber security events. In *37th Hawaii International Conference on System Sciences (HICSS).*
[76] Rick Wilson, Patrick Lynett, Kevin Miller, Amanda Admire, Aykut Ayca, Edward Curtis, Lori Dengler, Michael Hornick, Troy Nicolini, and Drew Peterson. 2016. Maritime tsunami response playbooks: Background information and guidance for response and hazard mitigation use. *California Geological Survey Special Report* 241 (2016), 48.
[77] Tarun Yadav and Arvind Mallari Rao. 2015. Technical aspects of cyber kill chain. In *International Symposium on Security in Computing and Communication (SSCC).*

A SURVEY INSTRUMENTS

A.1 Design phase

After using the *FRAMEWORK* playbook design framework, please answer the following:

Rate each step in order of importance for completing the playbook, with #1 being the most important step. <Drag and drop list of steps based on framework>

Please explain why you ranked <TOP CHOICE> step most important. <open response>

Please explain why you ranked <LOWEST CHOICE> step least important. <open response>

Please provide any positive feedback you may have on using the *FRAMEWORK* playbook design framework.

Please provide any negative feedback you may have on using the *FRAMEWORK* playbook design framework, especially any parts that you felt were confusing or needed additional information.

Please provide any neutral feedback you may have on using the *FRAMEWORK* playbook design framework. Do you feel anything was missing? Anything that could be better designed?

Demographics:

What is the highest level of school you have completed or the highest degree you have received?

Please estimate the number of years experience you have in the digital security and information technology fields:

Please indicate which role most accurately reflects your current position:

Please estimate the organization size that you work in:

A.2 Evaluation phase

Is this playbook sufficiently detailed to implement and actually detect the event? <Yes, no, unsure>

How likely is the playbook to adequately respond to the scenario event <1 to 5, with 1 being least likely>?

Please explain why this playbook would or would not adequately respond to the event. <open response>

Are there errors in the provided playbook that would hinder response efforts? <Yes, no, unsure>

Are there critical elements of a response plan missing from the playbook? <open response>

Do you have any other feedback for this playbook? Explain. <open response>

Demographics:

What is the highest level of school you have completed or the highest degree you have received?

Please estimate the number of years experience you have in the digital security and information technology fields:

Please indicate which role most accurately reflects your current position:

Please estimate the organization size that you work in:

A.3 Implementation phase

Based on your experiences, please indicate which framework was more useful for each task:

<Matrix ranging from NIST much better, NIST better, no difference, IACD better, IACD much better>

Identifying assets at risk:

Identifying required response tasks:

Building a comprehensive plan:

Being easily understandable:
Being easily implementable:
Please provide any positive feedback you may have for NIST with respect to implementing a playbook. <open response>
Please provide any negative feedback you may have for NIST with respect to implementing a playbook. <open response>
Please provide any positive feedback you may have for IACD with respect to implementing a playbook. <open response>
Please provide any negative feedback you may have for IACD with respect to implementing a playbook. <open response>
Were there any unexpected modifications you had to make to implement your plan using NIST? <open response>
Were there any unexpected modifications you had to make to implement your plan using IACD? <open response>

A.4 Utilization phase

From intrusion event to detection, how much time do you assess passed? How did you determine an event occurred? <open response>
Were there any unexpected issues associated with detecting the event? What decisions did you have to make during detecting the event that were not covered in the playbook? <open response>
From detection to initial response using a playbook, how much time do you assess passed? <open response>
Were there any unexpected issues associated with initial response? <open response>
What decisions did you have to make during responding to the event that were not covered in the playbook?<open response>
From initial response to threat neutralization, how much time do you assess passed? <open response>
How did you determine the event was stabilized/quarantined to a sufficient level?<open response>
Were there any unexpected issues associated with threat neutralization?<open response>

B INTERVIEW GUIDE

For each survey response across all phases that required follow-up questions:
In your survey, you indicated *TOPIC*. Could you please explain more information about *TOPIC*?

For each expert evaluation survey response required follow-up questions:
In your response, you indicated *TOPIC*. Could you please explain more information about *TOPIC*? How would you handle this in your organization? Have you encountered this situation in your organization before? Do you have any insight that would not necessarily be intuitive for people following playbook frameworks?

For our trusted insider during the utilization phase:
What time did you initiate your attack?
Were there any special considerations when you conducted the attack?
What times did participants report detecting the attack?
How long until they initiated initial response actions?
How long did they take to neutralize the threat?
Were there any observations that stuck out to you?

For each participant during the utilization:
In your survey response, you indicated *TOPIC*. Could you please explain why you felt *TOPIC* presented a unique challenge? Was there anything that could have prepared you more for *TOPIC*?

C ATTACK SCENARIOS

Study participants developed and used playbooks to detect and respond to two attack scenarios: brute-force login attempts and valid credential misuse. Our trusted insider used Powershell scripts to automatically execute these attacks from a trusted virtual machine logically located within the local NDC network. Our trusted insider changed the IP address of the attack system for each incident response event to ensure defenders could not develop signatures based on the attack source, but instead focused on the behavior of the attack. Additionally to enhance realism, the network address range used for attacks by our trusted insider provided legitimate services within the local NDC network and could not have been blocked without the loss of critical services.

When executing brute-force login attempts, the attacker's Powershell script would initiate the attack using valid domain usernames and randomly-generated passwords against two domain accounts; the script randomly selected legitimate systems within the network as targets and repeated login attempts 50 times per incident response event.

Our trusted insider leveraged the automated credential misuse script to initiate one successful login to a randomly-selected domain-connected system using the honeyword account credentials. Next, the script would then initiate one more successful login from the "exploited" system into a neighboring domain-connected system to simulate malicious lateral movement and access expansion using compromised credentials within the target network.

D CODEBOOK

Codebook for both case studies is available for viewing at: https://controlc.com/c5e40525.

E PLAYBOOK EXAMPLES

Exemplar playbooks made from study participants using the IACD (Figure 2) and NIST (Figure 3) playbook frameworks.

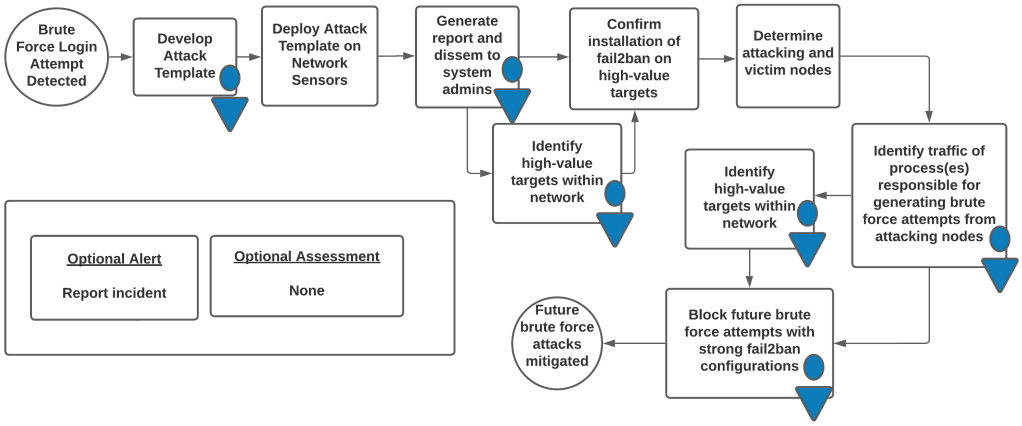

Figure 2: Participant P4's IACD playbook for brute force login attempts.

Scenario: Detecting and responding to the login of a fake account (honeyword).
https://attack.mitre.org/techniques/T1078/

1. Preparation:
Preparation helps speed up response time. In this step, a list of critical assets and critical endpoints associated with a threat is compiled. This list is ranked by level of importance and monitored.

a. Passwords for "honeyword" fake user account
b. Logging Services
c. Alerting mechanisms
    i. Dashboard
    ii. Email

2. Detection and Analysis:
Now that the threat incident has been identified, information must be gathered on the threat and an analysis is done.

Where is the entry point of this breach?

a. Identify the location that the "honeyword" account login originates from by reviewing logs.
b. Dashboard alerts cannot be removed until after the breach is resolved.

What is the breadth of this breach?

a. Correlate the time and origin of the "honeyword" account login with other attempted logins.
b. Determine what access each compromised account had and if failed login attempts also occurred.
    i. If there were no failed login attempts, this could be indicative that the password used to login was not guessed, but acquired through some other means.
c. Identify if any data was exfiltrated from the network.

Analyze the threat to the best of your ability. Think about these questions and add anything else you can think of.
-What are the consequences of not resolving this incident?

Figure 3: A sample of Participant P11's NIST playbook for credential misuse attempt.