How to Hack Compliance: Using Lessons Learned to Repeatably Audit Compliance Programs for Digital Security Concerns

Rock Stevens*, Josiah Dykstra†, Wendy Knox Everette‡, Michelle L. Mazurek*
*University of Maryland, College Park; †Independent Security Researcher; ‡Leviathan Security Group, Inc.
{rstevens,mmazurek}@cs.umd.edu, josiahdykstra@acm.org, wknox@wellesley.edu

***Abstract*—Digital security compliance programs and policies help protect organizations' intellectual property, sensitive resources, customers, and employees through mandated security controls. Organizations place a significant emphasis on compliance and often conflate high compliance audit scores with strong security; however, we find that security concerns may exist even within fully-compliant organizations. In this paper, we detail the methodology and processes we used to obtain the results found in "Compliance Cautions: Investigating Security Issues Associated with U.S. Digital-Security Standards." The Network and Distributed Systems Security Symposium 2020 version of this research focused on the results, but here we emphasize how we leveraged social-science methods to perform our analysis and developed partnerships with organizations to validate our findings. We highlight lessons learned that may assist with replicability.**

I. Introduction

Many digital-security guidelines present best practices for system owners and digital-security technicians to improve their overall security posture [25]. These guidelines are designed to protect intellectual property, sensitive resources, customers, and employees from security risks. Exemplar protection mechanisms include installing anti-virus applications on all systems and conducting background checks on employees before providing privileged access.

Over the years, governments and organizations have elected to adopt these guidelines as compliance controls: mandatory policy and technical controls that must be enforced across applicable organizations. Non-compliance with these controls is typically followed by significant fines, revocation of access, or employment termination [3]. To illustrate this fact, one energy company was recently fined $10 million for non-compliance [24].

Compliance standards are often presented as a proven metric for improving security. The International Organization for Standardization routinely provides metrics on how compliance standards keep users and businesses safe online [16]. Some federal-level programs and businesses develop and deploy systems that are fully compliant with established standards as an implicit seal of security [11], [1], and some organizations actively use digital compliance standards to shape their defensive strategies [17], [29]. Because compliance itself is treated as a first-class security property, standards are often used as word-literal checklists.

Despite the significant emphasis placed on compliance with these standards, their actual efficacy is not well understood. While they may provide important security benefits, it is also possible that they lull security practitioners into a false sense of security, conflating high compliance audit scores with strong security. It is also possible that standards which are useful as general guidelines can become problematic when interpreted legalistically as checklist requirements. In our previous research published at the Network and Distributed Systems Security Symposium 2020, we reported on a two-part study investigating these questions [30].

In this paper, we provide an in-depth discussion of the methods we used in that paper, in which we audited three publicly-available compliance standards: Internal Revenue Service (IRS) Publication 1075 (P1075) that protects taxpayer information, the Payment Card Industry Data Security Standard (PCI DSS) that protects credit card data, and the North American Electric Reliability Corporation (NERC) Critical Infrastructure Protection for system security management (CIP 007-6) that protects the electric grid. We explain our data collection methods to enable asynchronous researcher collaboration across multiple time zones, and we discuss the steps we took to ensure repeatability. Additionally, we detail the content-analysis techniques (such as codebook development) that we adapted from social science to identify security concerns and categorize them based on root cause and estimated risk levels. (We define security concerns as *any security control or policy within a compliance standard that can lead to sub-optimal security conditions when implemented as written*.)

We also discuss our strategy for partnering with security industry experts. These partnerships were essential for validating our findings and responsibly disclosing them to the appropriate entities. Our experience reveals that no viable, standardized process for reporting exists. Despite this, we discuss successes in directly coordinating with panel members to help improve the PCI, DSS, and CIP standards. We also discuss our failed attempts to disclose our findings to the IRS, several vulnerability disclosure officers, and the Department of Homeland Security.

Our methodology offers recommendations to future researchers and may assist them with auditing other compliance standards or assessing the unique impacts of compliance standards on their organizations.

II. Background

Digital security compliance programs within the United States date back to the Computer Security Act of 1987, which required agencies to protect sensitive systems and conduct security training [23]. Many programs implement a "carrot-and-stick" approach to compliance, in that organizations are

Learning from Authoritative Security Experiment Results (LASER) 2020
23 February 2020, San Diego, CA, USA
ISBN 1-891562-65-7
https://dx.doi.org/10.14722/laser.2020.23003
www.ndss-symposium.org

rewarded for successful programs and levied with sanctions for compliance deviations. In this section, we briefly review past studies involving digital security compliance and its impact on organizations.

Compliance audits force organizations to balance being "inspection ready" and sustaining daily operations, such as providing essential services or selling goods. Because of this careful balance, many organizations choose to perform compliance actions only before a pending audit, and then neglect further security maintenance until another audit requires them to repeat the process [26]. This behavior meets the security minimums for compliance standards, but fails to adhere to the spirit of secure practices. Moreover, evidence shows that fully-compliant organizations can still suffer data breaches. Auditors certified Target as PCI-compliant in September 2013, just before it suffered a massive data breach in November 2013 [26]. We highlight sections of compliance standards that may permit similar incidents to occur again and provide recommendations for mitigation.

Previous studies highlight cultural disconnects between developers, engineers, and compliance officials that create issues when digital security measures are "bolted on" after software development is complete [7], [4]. To combat these issues, entities must find ways to overcome organizational behaviors and factors that affect secure software development [33]. Some organizations have embedded compliance experts within development teams to encourage grass-roots-style compliance integration [7]. Other organizations found that threat modeling could proactively identify security gaps that may exist in compliant solutions [7], [2]. Some organizations have even overhauled their physical network topology to meet federally-mandated requirements, restructuring their teams and network architecture to limit the scope of auditable systems within their environment [15]. This, too, meets the letter of compliance requirements while seeming to contradict the intended goals. In this study, we identify several unintended security implications within technical controls and implementation processes that could affect organizations as they alter their normal business practices for compliance adherence.

Numerous studies focus on how humans perceive compliance standards and modify their behaviors based on those perceptions. Julisch highlighted numerous factors that shape organizational decision-making when investing in compliance measures, often seeking new security technologies that are out-of-the-box compliance ready [17]. Beautement et al. describe the "compliance budget," the human factors behind the implementation of compliance controls; their research illuminated ways to improve security and compliance readiness through resource allocation optimization [5]. Building upon previous works, Puhakainen and Siponen found that training employees to better understand compliance standards can improve organizational behaviors and shift employees toward implementing more secure practices [28]. Additionally, Hu et al. found that managers who "lead by example" and implement top-down management initiatives encourage employees' compliant security behaviors [13]. Our study is a significant departure from previous studies, as we do not focus on improving adoption rates within organizations. Instead, in this study we assume organizations are 100% compliant with the letter of the standard and focus on the insecure practices and security concerns that may exist anyway.

ID	Employment[1]	Role[2]	Org Size	IT Exp (yrs)	Edu[3]	Docs[4]
R1	A, G	M, R	500	18	MS	I,P,N
R2	G	M, R	10k+	16	PhD	I,P
R3	A, G*, I	M, R	100	20	BS	I,N
R4	I	M, R	35	15	JD	I,P
R5	A, G*, I	M, D	100	8	BS	I,N
R6	G	M, D	100	5	BS	I,N
E1	G, I	M	150	10	BS	I
E2	G	M	150	15	MS	I
E3	G*, I	M, D	1k	18	MS	P
E4	A, G*, I	R	5k	20	MS	N

[1] A: Academia, G: Government, I: Industry, *: Previous experience
[2] M: Management, R: Research, D: Development
[3] BS: Bachelor's, MS: Master's, PhD: Doctorate, JD: Juris Doctorate
[4] I: IRS P1075, P: PCI DSS, N: NERC CIP

TABLE I: Researcher and expert demographics

III. Method

In the first step of our study, researchers comprehensively audited three compliance standards to identify potential security concerns. To validate these concerns, we then recruited four compliance experts to provide their assessment of our findings. We performed quantitative and qualitative analysis on expert responses to identify discrepancies and also derive additional context for applicability within enterprise environments.

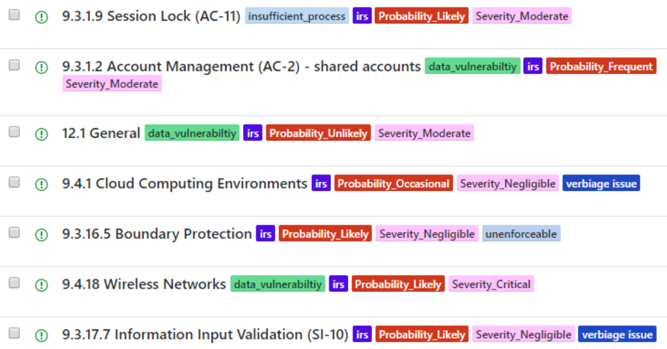
This study occurred originally occurred from October 2017 through September 2018 and was ruled not human subjects research by our ethics-compliance office, as we communicated with experts in their professional capacity and did not collect personally identifiable information. Due to the sensitive nature of unmitigated data vulnerabilities within real environments, we generalize many of our findings to protect networks and systems. Additionally, due to several memorandums of understanding and non-disclosure agreements, we generalize many of our partnered relationships to protect the organizations.

A. *Beginning the audit*

Our team of six researchers designed the audit to systematically evaluate three unrelated compliance standards in a repeatable manner. This cross-section of standards was intended to evaluate the breadth and scope of security concerns across three domains. Each researcher audited a subset of the standards, with at least three researchers per standard (shown in Table I). Our objective was to identify issues that might negatively affect digital security, including policies that expose sensitive information and processes that create issues due to ambiguous implementation guidance.

B. *Enabling asynchronous collaboration*

Best practices suggest that empirical research should be conducted by personnel with extensive domain knowledge [27]. At the time of data collection, our group of researchers possessed an average of 14.3 years of digital security experience within academia, the federal government,



□ 9.3.1.9 Session Lock (AC-11) insufficient_process irs Probability_Likely Severity_Moderate

□ 9.3.1.2 Account Management (AC-2) - shared accounts data_vulnerabiltiy irs Probability_Frequent Severity_Moderate

□ 12.1 General data_vulnerabiltiy irs Probability_Unlikely Severity_Moderate

□ 9.4.1 Cloud Computing Environments irs Probability_Occasional Severity_Negligible verbiage issue

□ 9.3.16.5 Boundary Protection irs Probability_Likely Severity_Negligible unenforceable

□ 9.4.18 Wireless Networks data_vulnerabiltiy irs Probability_Likely Severity_Critical

□ 9.3.17.7 Information Input Validation (SI-10) irs Probability_Likely Severity_Negligible verbiage issue

Fig. 1: Example of recorded security concerns using GitHub.

and industry. We relied heavily on this past experience when identifying security concerns throughout the compliance documents; we required researchers to cite their findings based on past security encounters or publicly-available current events.

One facet of building a cadre of experienced experts is managing available time. All researchers had full-time employment unrelated to this study – and these jobs often required frequent global travel. These time constraints necessitated the ability to collaborate asynchronously: describe researchers' independent findings, indicate disagreement with others' findings, and describe their assessments for root cause of the issue.

Based on our requirements, we chose to collaborate using private repositories in GitHub. GitHub provided us with the ability to host documents, use its markup language to generate detailed write-ups, apply labels, and record comments/feedback (Figure 1). We maintained a running list of codified definitions and instructions for researchers to reference throughout the study.

For the actual audits, all six researchers conducted a complete audit of IRS Publication 1075, following a content-analysis process drawn from social-science research. Each researcher independently examined each line of the standard. At each of several predetermined milestones within the document (e.g., the end of a section), the researcher would log their findings, including the section title where the issue was found, the exact phrase deemed problematic, a short description of the perceived issue, and references to related, publicly-known issues. If a researcher found multiple issues within one phrase or section, they logged each separately and each issue was given a unique identification number (this assisted greatly in performing post-collection analysis). For every logged issue, all other researchers would indicate (1) if they found the same issue independently and (2) whether they concurred with the finding. If there was not unanimous consensus on an issue, we discarded it but maintained a record of the disagreement (used to calculated inter-rater reliability, as discussed in Section III-C). We chose to outright discard issues without unanimous agreement instead of resolving disagreement due to time constraints; future studies using this methodology may choose to mediate disagreements within the group or use an external expert for conducting tie-breakers. Mediation may provide addition real-world discoveries that would have been discarded otherwise.

	R1	R2	R3	R4	R5	R6
Control1	1	1	1	1	1	1
Control2	0	1	0	0	0	1
..						
ControlN	0	0	0	0	0	0

TABLE II: Example matrix used for calculating inter-rater reliability

The ecological validity of our study relied on the ability to consistently apply our audit methodology to any compliance program by any experienced security professional. We believe our lessons learned from independent auditing may assist other researchers with repeatability in performing similar efforts in the future.

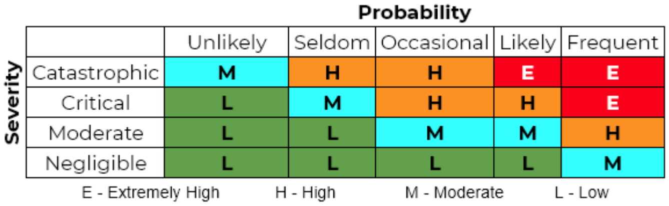
When auditing IRS P1075 together as a group, all of our findings were first independently logged to prevent each researchers' results from influencing each other's. Given that P1075 is a densely packed standard with 180 pages of detail, we needed to break the work into chunks. To this, we set milestones and established agreed-upon deadlines for reaching them. At the end of each milestone, we all submitted our findings and then reviewed the findings of others. This allowed for us to check for independent discovery and agreement. At times, it was necessary to conduct conference calls to gather all researchers, discuss progress, address concerns or lingering issues, and establish future near-term milestones.

C. *Calculating agreement*

After each researcher logged all of their independently-discovered security concerns, we then calculated the intercoder reliability — a measure of consistency among independent auditors — for IRS P1075. We calculated our Krippendorff's Alpha (α), which accounts for chance agreements [12]. To do this, we exported all GitHub comments and labels to a CSV file and normalized the data for ingest into ReCal3, an online inter-rater reliability calculator. We considered each individual compliance control as an independent item that researchers could agree (or disagree) upon. For example, each individual compliance control fell into one of three different categories: (1) all researchers identified and agreed that the control contains security concerns, (2) all researchers agreed that the control did not contain security concerns, or (3) there is a disagreement whether the control contains security concerns.

The first step of normalizing our data for inter-rated reliability was converting IRS P1075 into a binary matrix, listing each technical control in the document as a row. Columns in this matrix indicated agreement levels from each of our six researchers for corresponding control listed in the row ('1' indicates research believes a security concern is present, '0' indicates a researcher believes a security concern is not present). Using the example shown in Table II, there would be unanimous agreement that 'Control1' has a security concern but a disagreement for 'Control2.' We obtained reliability $\alpha = 0.815$ for P1075; an α value above 0.8 indicates high reliability [18], [19].

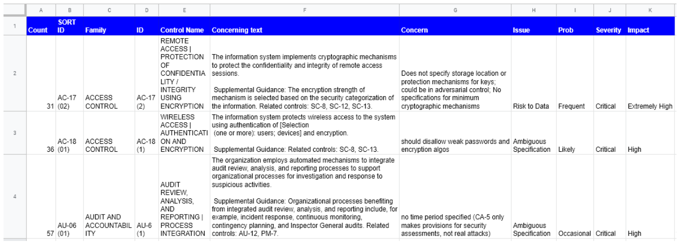
We further analyzed the identified issues using iterative open coding, a process for creating and applying category labels (known as a *codebook*) to data [32]. In particular, the



Severity	Probability: Unlikely	Seldom	Occasional	Likely	Frequent
Catastrophic	M	H	H	E	E
Critical	L	M	H	H	E
Moderate	L	L	M	M	H
Negligible	L	L	L	L	M

E - Extremely High H - High M - Moderate L - Low

Fig. 2: Security concern risk levels. Levels were assigned based on a Composite Risk Management risk-assessment matrix that includes both probability of occurrence and impact severity.



	A	B	C	D	E	F	G	H	I	J	K
1	Count	SORT ID	Family	ID	Control Name	Concerning text	Concern	Issue	Prob	Severity	Impact
2	31	AC-17 (02)	ACCESS CONTROL	AC-17 (2)	REMOTE ACCESS | PROTECTION OF CONFIDENTIALITY / INTEGRITY USING ENCRYPTION	The information system implements cryptographic mechanisms to protect the confidentiality and integrity of remote access sessions. Supplemental Guidance: The encryption strength of mechanism is selected based on the security categorization of the information. Related controls: SC-8, SC-12, SC-13.	Does not specify storage location or protection mechanisms for keys; could be in adversarial control; No specifications for minimum cryptographic mechanisms	Risk to Data	Frequent	Critical	Extremely High
3	36	AC-18 (01)	ACCESS CONTROL	AC-18 (1)	WIRELESS ACCESS | AUTHENTICATION AND ENCRYPTION	The information system protects wireless access to the system using authentication of [Selection (one or more): users; devices] and encryption. Supplemental Guidance: Related controls: SC-8, SC-13.	should disallow weak passwords and encryption algos	Ambiguous Specification	Likely	Critical	High
4	57	AU-06 (01)	AUDIT AND ACCOUNTABILITY	AU-6 (1)	AUDIT REVIEW, ANALYSIS, AND REPORTING | PROCESS INTEGRATION	The organization employs automated mechanisms to integrate audit review, analysis, and reporting processes to support organizational processes for investigation and response to suspicious activities. Supplemental Guidance: Organizational processes benefiting from integrated audit review, analysis, and reporting include, for example, incident response, continuous monitoring, contingency planning, and Inspector General audits. Related controls: AU-12, PM-7.	no time period specified (CA-5 only makes provisions for security assessments, not real attacks)	Ambiguous Specification	Occasional	Critical	High

Fig. 3: Example compliance audit format using Google Sheets to log security concerns instead of GitHub. Switching platforms saved approximately three hours during data analysis.

researchers who audited each standard coded each identified issue in that standard for perceived root cause, probability of occurrence, and severity. We resolved all disagreements among coders and developed a stable codebook by establishing a unanimously agreed-upon definition for coded terms, adapting many terms from the Composite Risk Management (CRM) framework [35] and the Information System Risk-based Assessment framework [9]. After any revisions to these definitions, we re-coded previously coded items, repeating this process until we coded all responses, resolved all disagreements, and the codebook was stable.

Our final codebook described four root causes for security concerns. A *data vulnerability* is an issue that will result in a data breach or compromise of sensitive information. An *unenforceable security control* cannot be enforced as written; these controls should be reworded or removed from the compliance standard. An *under-defined process* is an issue explicitly missing instructions or details that are required for a secure implementation, resulting in security gaps. An *ambiguous specification*, in contrast, is vague or ambiguous about some implementation details, such that different readers could interpret it differently. Some interpretations could potentially result in either an inappropriate action or inaction. We describe numerous examples of these issues in our previous research [30].

We used the following terms and definitions for probability: *frequent* occurs often and is continuously experienced; *likely* occurs several times; *occasional* occurs sporadically; *seldom* is unlikely, but could occur at some time; and *unlikely* we assume it will not occur. We used the following terms for severity: *catastrophic* results in complete system loss, full data breach, or the corruption of all data; *critical* results in major system damage, significant data breach, or corruption of sensitive data; *moderate* results in minor system damage or partial data breach; and *negligible* results in minor system impairment. Using a risk assessment matrix adopted from the CRM framework (Figure 2), we then calculated each issue's risk level — a function of probability and severity — as extremely high, high, moderate, or low [35].

After months of independent auditing and aggregation of our findings, we completed our audit of IRS P1075 and had established a strong inter-rater reliability. Strong agreement meant that each researcher was likely to independently assess compliance controls in a similar manner. At this point, we felt confident that we could split our group into smaller teams and audit the PCI and NERC standards in parallel. Four researchers audited NERC CIP 007-6 and three researchers audited PCI DSS. Splitting our group of researchers to assess these two compliance programs concurrently saved us approximately 2.5 months (based on time-completion estimates for serialized efforts). We ensured one researcher was a member of each sub-group (in this case, it was the lead researcher) to help provide consistency across the research efforts and also share lessons learned across sub-groups as the study progressed. An important fact to note is that we achieved this 2.5 months of time-saving benefits without a significant detriment to our inter-rater reliability score. The subgroups attained $\alpha = 0.801$ for PCI DSS and 0.797 for NERC CIP 007-6. Had one of the resulting inter-rater reliability scores dropped below acceptable levels, we would have considered re-auditing the standard with the entire group of researchers.

When exporting data from GitHub to calculate inter-rater reliability, it immediately became apparent that the data extracted from GitHub would require extensive work for normalization (e.g., standardizing the order that labels appeared). In a follow-on study on different compliance standards, we opted to use collaborative tools such as Google Sheets for logging and tracking security concerns (shown in Figure 3) and we saved approximately three hours in data wrangling when compared to previous efforts.

D. Expert validation process

To obtain external validation of our findings, we established partnerships with real-world organizations and compliance subject-matter experts to confirm or reject our findings. We felt this was critical to the validity of our study because, despite having vast experience with security assessments, our group of researchers did not possess that compliance expertise that others could bring to the study.

1) Recruiting experts: Just as discussed in Section III-B, researchers should strive to use subject-matter experts for empirical studies. We wanted to recruit experts in compliance programs to validate or reject our findings based on how they use compliance programs in real-world situations.

We established the following criteria for partnering with organizations: (1) the organization must regularly be subjected to audits, must regularly audit other organizations, or must contribute to the content of the relevant compliance standard, (2) the provided validators must have at least two years of experience with the relevant standard, and (3) the organization must be able to mediate responsible disclosure of our findings.

Using years of accumulated business cards and personal contacts from academic and information security conferences, our group of co-authors sent emails and text messages to approximately 20 experts that we believed met our criteria. These messages described the nature of our study and how their unique insight could contribute to the community's understanding of compliance programs. For those experts that indicated interest, we scheduled follow-on phone calls to provide more context and describe in detail what we would ask of them.

Many of these experts offered to help but would require non-disclosure agreements (NDAs) and anonymity. Two of our recruited experts worked for the same organization and we completed all requisite NDAs within three months. We spent six months negotiating with one expert and his organization before all NDAs were complete. Our fourth offered to provide assistance without an NDA. Researchers partnering with experts from large organizations should plan for similar delays during NDA negotiation and should forecast these delay within your research timeline.

After six months of negotiating an NDA with "Company B," we encountered three considerable setbacks. Our first point of contact unexpectedly departed the organization and we spent a month trying to get in contact with someone else that was qualified to assist with our study. The second point of contact retired approximately one month after agreeing to help us and we spent another month trying to find Point of Contact #3. When relying on external support for studies, we highly recommend being proactive with milestone management and establishing alternate points of contact prior to needing them.

Table I shows the qualifications of our four volunteer experts that assisted us in completing the study. Each expert completed their surveys during regularly scheduled work hours and did not receive any additional monetary incentives for participating.

2) Expert methods: We asked the experts to classify our identified issues in one of three categories: confirmed, plausible, or rejected. A confirmed issue indicates that the expert has previously observed security concerns associated with the issue or that observable consequences from the issue actively exist within an enterprise environment. A plausible issue occurs when the expert has not personally observed security concerns related to the issue but agrees such security concerns could manifest within other organizations. A rejected finding indicates that there is no observable evidence of security concerns related to the issue within a live environment, or that there are related security factors that we had not considered.

We used a series of closed- and open-ended survey questions to elicit information from each expert. In addition to directly validating or rejecting each issue, the experts were asked to provide additional context from their personal experience. We presented the issues to the experts in a randomized order, providing the referenced section title, exact text from the section, and a short narrative describing the perceived issue.

After collecting data from each expert and removing rejected findings, we used the Wilcoxon signed-rank test to compare researchers' assessment of probability and severity with our experts' responses for PCI DSS and NERC CIP 007-6; we used the Friedman test (omnibus) with planned pairwise Wilcoxon signed-rank tests for comparing IRS P1075 responses, for which we had two expert validators [36], [8]. We also conducted open-ended discussions with the experts to discuss similarities and differences in assessments.

We note that an essential tenet for partnering with experts is minimizing disruption to their daily responsibilities. Research suggests that the quality of survey responses decreases over time, and excessive time away from work may result in an expert terminating their participation in the study [14]. To this end, we designed our surveys for experts to complete within 60-90 minutes of focused effort; actual completion time averaged 84.8 minutes. Given our limited pool of experts, this required us to select only a subset of our findings to validate; we selected the issues to validate semi-randomly, while prioritizing the extremely-high-risk and high-risk issues.

Prior to deploying our protocol with partnered organizations, we piloted surveys to pre-test relevance and clarity with security practitioners familiar with auditing and compliance standards. Given that we wanted to minimize disruption to our experts' time, we felt it was critical to elicit information as clearly and concisely as possible; our two piloted iterations played an important role improving our questionnaires. Even within these two pilots, we recruited experts in digital-security penetration testing with experience with compliance programs.

3) White-box validation: One special aspect of the IRS P1075 expert audit of our findings is that our recruited experts had administrator-level access to the networks at their organization. This allowed them to search the network to confirm or deny the findings we believed would be present because of their compliance with IRS P1075. This methodology resembles a white-box penetration test, where trusted insiders have privileged access to systems and are permitted to search for vulnerabilities with impunity [10]. This proved to be considerably better than our original proposal.

Our original study proposal involved a subset of our researchers conducting our own penetration test of live systems to confirm the presence of IRS P1075 security concerns. This posed a considerable risk to live systems that were essential for critical services and presented numerous legal obstacles that could have taken upwards of 12 months to overcome.

We found that the white-box penetration test that our experts conducted was much faster (as it skipped non-essential steps such as gaining initial access into systems) and was much safer considering these administrators knew how to interact with the systems without harming service availability.

For future research, we highly recommend a similar approach for obtaining results from live production systems.

E. Limitations

One of the key strengths of our audit method is the extensive experience and backgrounds of our researchers. Their expertise brought decades of first-hand knowledge into a previously unexplored areas. We recognize that it is entirely possible that a different group of researchers could have conducted the same audit using the same method and discovered different security concerns. We believe our cadre of experts with diverse backgrounds helps mitigate this to some degree, and the requirement for unanimous agreement mitigates it even further. We observed that forcing unanimous agreement among a diverse group of researchers pushed findings towards a lower bound and helped remove potential false positive results. Lastly, requiring strong inter-rater reliability ensured our group was trained to look for the same types of security concerns, understood key definitions/terms, and increased the likelihood of researchers discovering the same security concerns independently from one another.

IV. Results Overview

While the intent of this paper is to focus on the methodology and processes involved with auditing compliance standards, we felt it necessary to include a high-level description of our results to describe the effectiveness of our methodology in this particular instance.

The first crucial finding from our study was that we confirmed all four of our recruited compliance experts reported first-hand experience with auditors using compliance documents as line-by-line checklists, which support our decision to treat them as such for analysis purposes. This finding is counter to claims from organizations such as NIST, who insisted that compliance programs were never intended to be used as audit checklists [22].

In total, we identified 148 unique security concerns that would exist, based on past experiences from our team of researchers, when organizations follow compliance programs "by-the-letter." We broke these security concerns into four distinct root causes: data vulnerabilities, unenforceable security controls, under-defined processes, and ambiguous specifications.

These security concerns range in risk (assessed based on probability of occurrence and associated severity) from low to extremely high and include issues relating to vague requirements, outdated technology, and improperly protecting sensitive information. Some security concerns could potentially be addressed with straightforward rewrites of the standards and minor changes at compliant organizations, while others likely cannot be remediated without significant, potentially impractical, investment by affected organizations.

The compliance experts validated our findings, confirming 36 of 49 as definite security concerns and 10 as plausible, while rejecting only three. Further, compliance experts confirmed that problems like poorly defined time windows and unclear division of responsibility — trends we observed across the three standards we examined — can manifest in real-world ways that increase risks. We include four exemplar findings below.

Under-defined process. We identified another issue in IRS P1075 Section 9.3.5.8, which outlines a procedure for establishing an Information System component inventory (i.e., a listing of authorized devices that operate within an organization). As written, this procedure does not require the inventory process to be tied to a "ground truth," meaning there is no comparison of which devices should be operating within an organization with which devices actually are. This is dangerous, as it could permit a rogue system to persist on a network or even be inventoried as a legitimate system. Providing a rogue system with legitimate access within a sensitive environment obviates the need for an attacker to deploy malware within the environment and reduces the likelihood that any defensive sensors would ever detect anomalous activity from the attacker. We assessed this moderate-risk issue to have an occasional probability and moderate severity. Industry recommendations integrate asset inventory with supply acquisition, ensuring that only company-purchased, legitimate systems are on the network [6].

Data vulnerability. The "Network Segmentation" section of PCI DSS scopes the standard's safeguards to only the network segment that contains cardholder data. Effectively, this provision would allow an organization with no security controls outside of the CDE to pass an audit as long as the CDE itself is protected in accordance with PCI DSS specifications. Allowing vulnerable servers and systems within the same network as the CDE could provide attackers with a landing point into internal portions of the network and establish conditions for lateral movement into the CDE from adjacent network segments (through well-known attacks such as VLAN hopping). Due to the series of security holes that must be present for such an attack to occur, we assessed that exploitation of this vulnerability would be seldom but critically severe for affected systems.

Unenforceable security controls. IRS P1075 Section 4.7 provides several measures for secure telework access to FTI. P1075 provides many requirements for physical data protections, such as badge-based control and on-premises guards; these are infeasible in the case of telework, as most personnel with FTI access at their private homes cannot abide by these types of controls. Additionally, IRS inspections of private residences for physical security compliance seems fraught with complications. We recommend that either the IRS ban residential-based telework programs until it can verify that all locations with FTI access are compliant with physical security requirements, or that the standard acknowledge that these physical controls are not actually required. We assessed this high-risk issue to have a frequent probability and moderate severity.

Ambiguous specification. PCI DSS Section 11.3.3 discusses corrective action plans for vulnerabilities discovered during penetration tests. The section does not specify how soon after a penetration test vulnerabilities must be addressed, nor the party responsible for fixing the vulnerabilities. Based on the researchers' past experiences with organizations delaying remediation, we assess this security concern to have a high risk level with a frequent probability of occurring and a moderate severity. Moreover, the non-validator assessor we spoke to confirmed that in his experience, organizations often delay remediation, and typically dedicate one to two full-time employees for 30-40 days prior to an inspection to ensure remediation is complete just in time [21]. We recommend this section specify a time limit (based on vulnerability severity) for addressing issues discovered during a penetration test and clarify the party responsible for fixing the vulnerable condition.

V. Disclosures

We made an effort to disclose our findings responsibly. Compliance standards typically have a request-for-comment

(RFC) period that allows for the submission of comments, concerns, and recommendations during a fixed window. During this study, none of the standards we assessed had an open RFC, and we found that no clearly defined channel existed for reporting security concerns, either directly to affected organizations or at the federal level. Using our partners as mediators, we turned over all of our findings to the IRS; the PCI Security Standards Council; a contributing author of the NERC CIP standards; the United States Computer Emergency Readiness Team (US-CERT); the MITRE Corporation's Common Vulnerabilities and Exposures (CVE) team; and the Department of Homeland Security. Even though we conducted this study between October 2017 to September 2018, as of June 2020, we are still actively working with the U.S. Government to help organizations understand the impact of our findings. Overall, we have had varying levels of success with our disclosure attempts, as described below.

A. *Our vision*

Before disclosing any of our findings, we envisioned that our research could help the U.S. federal government establish a centralized repository of best-practices and lessons learned associated with compliance controls. We felt this information could (1) help authors of compliance programs adopt language that has been proven to be effective, (2) help organizations understand potential risks they could inherit, and (3) allow compliance programs to incrementally evolve at speed with technology to remain secure and relevant.

To achieve our vision, we sought contacts at the highest levels of federal government, at organizations responsible for creating compliance programs, and directly at the organizations that use the compliance standards we analyzed. This involved extracting contact information from the audited standards, extracting information from publicly-available official sites, contact-chaining through our personal contacts, and even sometimes searching through social media for appropriate contact information. Below we describe our attempts and shortcomings in trying to achieve our vision.

B. *IRS P1075*

We contacted the IRS, NIST National Vulnerability Database (NVD), US-CERT, and the MITRE Corporation to disclose our P1075 findings. US-CERT was the first organization to respond to our disclosure attempt. After exchanging several emails, their technicians concluded that "CVEs are assigned for specific vulnerability in implementations. Each issue that requires a 'separate patch' can get a CVE [34]." In a series of email and phone exchanges, we argued that each of the recommendations we provided are "patches" for the vulnerable portions of the compliance standards, but US-CERT stated that the "patches" we identified must be tied to a specific piece of software. Future research that correlates security concerns to compliant software may be eligible for CVE identification numbers using US-CERT's definition.

Both NIST NVD and the MITRE Corporation indicated that compliance documents are outside their scope of responsibility, with MITRE stating "that a reasonable person can conclude that IRS Publication 1075 was never intended to have a level of abstraction that was sufficient to direct secure coding [20]." Contradicting this argument, our partners confirmed that auditors are indeed using compliance standards such as P1075 as a line-by-line checklist to confirm controls at levels as granular as access control lists on firewalls.

We attempted to disclose our findings directly to the IRS nine times via personal contacts, emails, and phone calls over the span of three months. To date, we have not received any form of acknowledgment other than the automated responses from SafeguardReports@irs.gov, the only point of contact listed in IRS P1075.

C. *PCI DSS*

Unlike P1075, we had success in responsibly disclosing our findings to members of the PCI Security Standards Council. We established a memorandum of understanding with a PCI SSC member organization; in turn, this organization provided our findings to the PCI DSS Version 4 Working Group.

We received notification that our recommendation for improving the "Network Segmentation" section of PCI DSS has already been implemented within Version 4, prior to the opening of their RFC submission window. This change will apply PCI DSS guidelines to the entire networked environment and not only an isolated subnet with cardholder data – this change could help reduce the likelihood that an attacker could gain access via unprotected portions of the network. Additionally, the v4 Working Group is considering incorporating all feedback associated with our ambiguous specification findings.

D. *NERC CIP 007-6*

Expert E4, after providing feedback, noted that our recommendations would be included at future working groups for CIP revisions. However, it could be years before the next CIP update (potentially taking our recommendations into account) is released. Additionally, our partnered organization for CIP disclosure is incorporating our feedback into a comprehensive evaluation of electric grid security. More than any other expert, E4 provided years' worth of lessons learned from CIP audits and helped explain why the standard was written the way it is. Given that our group of researchers had little experience with industrial control systems or the electric grid prior to this study, Expert E4's insight was truly invaluable for assessing the validity of our findings.

E. *Federal-level recognition*

To approach problems with federal-level compliance standards in a top-down manner, we met with representatives from the NIST National Cybersecurity Center of Excellence (NCCoE) to discuss our findings [22]. We highlighted that IRS P1075 Section 9 (which contains 49% of the P1075 security concerns we discovered) is copied from older versions of NIST SP 800-53 (NIST has since updated SP 800-53 twice). NCCoE offered to incorporate our findings into future document revisions. In ongoing revisions that began before our meeting, NIST acknowledged in draft SP 800-53v5 that organizations may inherit risk when implementing mandated security controls; that is, standards may create security problems [25]. Specifically, NIST describes deliberate efforts to remove ambiguity, improve understanding of responsibility, and keep controls up to date, corroborating many findings from our study.

Next, we contacted the Department of Homeland Security (DHS) National Protection and Programs Directorate. Several personnel within the Federal Network Resilience Division expressed interest in assisting with our findings; however, the DHS Office of External Affairs for Cybersecurity and Communications directed our contacts to cease communication and did not provide any alternative mechanisms for disclosure. This decision continues to provide friction between our agent contacts at DHS and the organization – the agents are truly motivated to help remedy the issues we have discovered. Through open publication, these agents are now able to use our findings and shape future compliance development on their own.

VI. Conclusion

We find that using compliance standards as checklists, with "by-the-letter" implementation of security controls, can create security concerns. As detailed in this paper, our systematic approach was applied across multiple compliance frameworks to identify security issues spanning multiple root causes and varying levels of risk.

We believe our framework for auditing compliance programs presents a repeatable methodology for assessing other compliance programs and for assessing the specific impacts of compliance programs on specific organizations. Techniques such as independent auditing, establishing strong inter-rater reliability, leveraging diverse groups of security experts, and requiring unanimous agreement on findings allowed us to establish a methodology that could be applied across multiple, diverse compliance frameworks. We found that our techniques reduced false positives and improved the likelihood that one researcher could independently apply our methodology and discover the same findings as other researchers.

Three researchers from this study used the described audit methodologies and applied lessons learned to audit another federal-level compliance standard (without expert validation this time) [31]. From start to submitting a paper for publication (currently undergoing editorial review), the process took three weeks. We recognize that familiarity with the process may have played a role in this expedited execution, but it may indicate that our methodology is repeatable across compliance programs and should be considered for future researchers and organizations that desire to self-assess the risk impact of compliance programs.

The lessons learned throughout the execution of our compliance study should allow future research to bypass obstacles, assist with time management, and expedite study completion.

References

[1] Amazon, "IRS Publication 1075," 2018. [Online]. Available: https://aws.amazon.com/compliance/irs-1075/

[2] Amazon Web Services, "Compliance and top security threats in the cloud – are you protected?" 2018. [Online]. Available: https://www.youtube.com/watch?v=Rc55aYODnMI&feature=youtu.be&t=18m10s

[3] H. Assal and S. Chiasson, "Motivations and amotivations for software security," 2018.

[4] ——, "Security in the software development lifecycle," pp. 281–296, 2018.

[5] A. Beautement, M. A. Sasse, and M. Wonham, "The compliance budget: managing security behaviour in organisations," in *Proceedings of the 2008 New Security Paradigms Workshop*. ACM, 2009, pp. 47–58.

[6] B. Beyer, C. Jones, J. Petoff, and N. R. Murphy, *Site Reliability Engineering: How Google Runs Production Systems.* " O'Reilly Media, Inc.", 2016.

[7] R. Clark, "Compliance != security (except when it might be)," in *Enigma 2018 (Enigma 2018)*. Santa Clara, CA: USENIX Association, 2018. [Online]. Available: https://www.usenix.org/node/208142

[8] G. W. Corder and D. I. Foreman, *Nonparametric statistics for non-statisticians: a step-by-step approach*. John Wiley & Sons, 2009.

[9] S. Elky, "An introduction to information system risk management," *SANS Institute InfoSec Reading Room*, 2006. [Online]. Available: https://www.sans.org/reading-room/whitepapers/auditing/introduction-information-system-risk-management-1204

[10] D. Geer and J. Harthorne, "Penetration testing: A duet," in *Computer Security Applications Conference, 2002. Proceedings. 18th Annual*. IEEE, 2002, pp. 185–195.

[11] Government Services Administration, "Federal Risk and Authorization Management Program," 2018. [Online]. Available: https://www.fedramp.gov/about/

[12] A. F. Hayes and K. Krippendorff, "Answering the call for a standard reliability measure for coding data," *Communication methods and measures*, vol. 1, no. 1, pp. 77–89, 2007.

[13] Q. Hu, T. Dinev, P. Hart, and D. Cooke, "Managing employee compliance with information security policies: The critical role of top management and organizational culture," *Decision Sciences*, vol. 43, no. 4, pp. 615–660, 2012.

[14] L. Hugick and J. Best, "Questionnaire length," *Encyclopedia of Survey Research Methods*, 2008.

[15] J. Humble, "Continuous delivery sounds great, but will it work here?" *Queue*, vol. 15, no. 6, p. 70, 2017.

[16] International Organization for Standardization, "Are you safe online? New ISO standard for cybersecurity," 2012. [Online]. Available: https://www.iso.org/news/2012/10/Ref1667.html

[17] K. Julisch, "Security compliance: the next frontier in security research," in *Proceedings of the 2008 New Security Paradigms Workshop*. ACM, 2009, pp. 71–74.

[18] K. Krippendorff, "Reliability in content analysis," *Human Communication Research*, vol. 30, no. 3, pp. 411–433, 2004.

[19] M. Lombard, J. Snyder-Duch, and C. C. Bracken, "Content analysis in mass communication: Assessment and reporting of intercoder reliability," *Human communication research*, vol. 28, no. 4, pp. 587–604, 2002.

[20] MITRE Corporation, Personal communication, 2018.

[21] S. Nangle, Private Communication, Feb 2019.

[22] National Institute of Standards and Technology, Personal communication, 2018.

[23] M. Nieles, K. Dempsey, and V. Yan Pillitteri, "NIST Special Publication 800-12: An Introduction to Information Security," 2017.

[24] North American Electric Reliability Corporation, "NERC Sanction Guidelines," 2012. [Online]. Available: https://www.nerc.com/FilingsOrders/us/RuleOfProcedureDL/Appendix_4B_SanctionGuidelines_20121220.pdf

[25] N. I. of Standards and Technology, "Sp 800-53 rev. 5 (draft) security and privacy controls for information systems and organizations," *Special Publications*, 2019. [Online]. Available: https://csrc.nist.gov/publications/detail/sp/800-53/rev-5/draft

[26] A. Peterson, *Cracking Security Misconceptions*. O'Reilly Media, Inc., 2013.

[27] C. Potts, "Software-engineering research revisited," *IEEE software*, vol. 10, no. 5, pp. 19–28, 1993.

[28] P. Puhakainen and M. Siponen, "Improving employees' compliance through information systems security training: an action research study," *Mis Quarterly*, pp. 757–778, 2010.

[29] R. Stevens, C. Ahern, D. Votipka, E. Redmiles, P. Sweeney, and M. Mazurek, “The battle for new york: A case study of applied digital threat modeling at the enterprise level,” USENIX Association, 2018.

[30] R. Stevens, J. Dykstra, W. K. Everette, J. Chapman, G. Bladow, A. Farmer, K. Halliday, and M. L. Mazurek, “Compliance cautions: Investigating security issues associated with us digital-security standards,” *Network and Distributed System Security Symposium*, 2020.

[31] R. Stevens, J. Dykstra, W. Knox-Everette, and M. L. Mazurek, “It Lurks Within: A Look at the Unexpected Security Implications of Compliance Programs,” *IEEE Security & Privacy (In Draft)*, 2020.

[32] A. Strauss, J. Corbin *et al.*, *Basics of qualitative research*. Newbury Park, CA: Sage, 1990, vol. 15.

[33] T. W. Thomas, M. Tabassum, B. Chu, and H. Lipford, “Security during application development: an application security expert perspective,” in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, p. 262.

[34] United States Computer Emergency Readiness Team, Personal communication, 2018.

[35] U.S. Department of the Army, “Field Manual 100-14 Risk Management,” 1998.

[36] F. Wilcoxon, “Individual comparisons by ranking methods,” *Biometrics bulletin*, vol. 1, no. 6, pp. 80–83, 1945.