ABSTRACT

Title of dissertation: QUALITY AND INEQUITY IN DIGITAL SECURITY EDUCATION
Elissa M. Redmiles
Doctor of Philosophy, 2019

Dissertation directed by: Professor Michelle L. Mazurek
Department of Computer Science

Few users have a formal, authoritative introduction to digital security. Rather, digital security skills are often learned haphazardly, as users filter through an overwhelming quantity of security education from a multitude of sources, hoping they're implementing the right set of behaviors that will keep them safe. In this thesis, I use computational, interview, and survey methods to investigate how users learn digital security behaviors, how security education impacts security outcomes, and how inequity in security education can create a digital divide. As a first step toward remedying this divide, I conduct a large-scale measurement of the *quality* of the digital security education content (i.e., security advice) that is available to users through one of their most cited sources of education: the Internet. The results of this evaluation suggest a security education ecosystem in crisis: security experts are unable or unwilling to narrow down which behaviors are most important for users' security, leaving end-users—especially those with the least resources—to attempt to implement the hundreds of security behaviors advised by educational materials.

QUALITY AND INEQUITY IN DIGITAL SECURITY EDUCATION

by

Elissa M. Redmiles

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2019

Advisory Committee:
Professor Michelle L. Mazurek, Chair/Advisor
Professor Ross Anderson
Professor Ashok Agrawala
Professor Lorrie Cranor
Professor John P. Dickerson
Professor Anne Simon

Acknowledgments

With much gratitude to the many mentors who collected and fostered me along the way, the family and friends who supported me and learned deeply about the work, and the dog who faithfully slept through the past four years.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

It is easy to identify how most people learn the majority of basic skills: learning to ride a bicycle from parents, algebra from school, and how to make paper airplanes from childhood friends. How users learn security behaviors, however, is less clear. As the Internet quickly evolves, users must cobble together a digital-security education by collecting and evaluating a plethora of ever-changing digital security and privacy advice.[1] Previous research related to educating users about security has primarily focused on changing user behavior directly within a system [56,176] or teaching users individual behaviors such as phishing awareness through targeted interventions [23, 44, 87, 89, 139, 194, 202, 208, 225].

In this thesis, we focus instead on *organic* security education: how users learn security practices in the first place, outside of a specific system or educational intervention. Specifically, we investigate:

1. *How do users learn digital security practices?* (Chapters 3 and 4)

2. *How does users' security education relate to the security incidents they experience?* (Chapter 5)

[1] Hereafter, we refer to security and privacy advice as security advice, given that the majority of currently available educational material focuses on security (see Figure 6.2).

3. *What is the quality of one of the largest channels of digital security education: online security advice documents?* (Chapter 6).

While answering these questions, we additionally identify and interrogate inequities in access to security education (Chapters 4 and 5) and illuminate the ways in which issues of advice quality (Chapter 6) exacerbate existing structural inequities. We also present two methodology studies (Chapters 7 and 8) that validate the approaches used in the previous chapters.

In this thesis, we answer these questions and address issues of inequity and quality in security education through a series of research studies. First, in Chapter 3 we describe 25 qualitative interviews with a diverse group of U.S. Internet users to develop an initial understanding of users' security education channels and decision-making strategies for accepting and advice provided by these channels. In this chapter, we compare digital security with the better studied domain of physical safety, to understand how digital security education may be unique. Further, we explore how exposure to security-sensitive content (e.g., HIPAA or FERPA regulated data) and workplace trainings impact users' use of different education channels and users' reasons for accepting or rejecting advice.

In order to generalize the initial results of this qualitative exploration, in Chapter 4, we quantify the prevalence of security education channels and users' evaluation strategies for the advice they receive from these channels. Specifically, we describe a census-representative survey with 526 U.S. Internet users to validate our taxonomy of educational channels and advice evaluation strategies. Further, we explore potential differences in both channels and strategies among different groups of users, including security-sensitive users, as well as those with different sociodemographic backgrounds and Internet skills. This work constitutes the first large-scale empirical analysis of how users' security beliefs, knowledge, and demographics correlate with their channels of security education. Through the work presented in these two chapters, we make several contributions.

- We develop a comprehensive taxonomy of users' eight security education channels: the media (websites, TV, radio), family and friends (some of whom are IT professionals), negative experiences, the workplace (including both trainings and IT staff), teachers service providers (e.g., Verizon, Bank of America), and being prompted or forced to do a behavior by a system [2]. Of these, system prompts, service providers, and the workplace are unique to digital security.

- We find that digital-security advice is evaluated primarily based on the trustworthiness of the advice source. The degree to which this is true varies based on users' familiarity with and understanding of a particular security behavior. For passwords, respondents rely slightly more heavily on their own evaluation of the content, rather than their trust of the advice source. This significantly differs from two-factor authentication (2FA), for which about 50% of respondents rely solely on their trust of the advice source to reason about advice regarding 2FA. Updating and antivirus advice is even more likely to be evaluated solely based on the trustworthiness of the source, presumably because

[2] Users generalized what they learned from a specific prompt on a specific system to other systems they thought were similar: e.g., one application requiring the use of an 8-character password inspired users to use an 8-character password on other systems, where this was not required.

users have the least understanding of these behaviors. This contrasts sharply with physical security, where the trustworthiness of the source is less important because the majority of users feel comfortable independently evaluating the content and value of the advice.

- Unsurprisingly, respondents most often rejected security behaviors because they were inconvenient. However, they rejected advice nearly as often for containing too much marketing material or because they had not yet had a negative experience. These findings support the need not just to make advice simpler and less intrusive, but also to minimize marketing messaging. They also confirm and extend prior work about the importance of negative experience stories and teachable moments [136, 176], which we find to be not only an effective teaching tool but in some cases are a prerequisite for behaving securely.

- Finally, we document the first empirical evidence of a **digital divide in security**. We find that higher-skilled users, who tend to be socioeconomically advantaged, are significantly more likely to take advice from their workplace and to have the skills necessary to learn from a negative experience. In contrast, those with lower skills tended to take advice from family, friends, and service providers (e.g. TimeWarner).

In addition to understanding *how* users learn digital security behaviors, we seek to understand how this education relates to actual security outcomes. In Chapter 5 we explore the relationship between reported security incidents and security advice sources and more broadly examine the impact of the digital divide identified in Chapter 4. To do so, we analyze a probabilistic random- digit-dial (RDD) telephone survey of 3,000 U.S. residents, the results of which are statistically weighted [28] to be accurate within 2.7% of the true values in the entire U.S. population. We find that advice matters: users' security outcomes are significantly correlated with the authoritativeness of their advice sources. Further, we confirm a second-level digital divide in security: with SES affecting respondents' advice sources, which in turn relate to outcomes. Finally, we find that 49% of users in our survey report experiencing a serious security or privacy incident, despite the majority reporting that they have sought out advice regarding digital security.

This suggests a security education ecosystem unable to help most users protect themselves, with the more authoritative advice inaccessible to less resourced users. To evaluate the extent of this problem, in Chapter 6 we conduct a large-scale measurement of the *quality*—comprehensibility, accuracy, and actionability—of one of the most prevalent forms of security education: online, text-based, security advice documents (including news articles). We collect 1,264 security- and privacy-related documents based on the results of user-generated search queries for security education materials and expert recommendations of security education websites. Manual annotation of these documents reveals 400 unique security behaviors: 196 identified in prior literature and 204 reported for the first time in this thesis. These behaviors are recommended to users 2,780 times across the 614 documents in our corpus that contained explicit advice imperatives.

An evaluation of the comprehensibility of these documents supports our hypotheses from Chapters 4 and 5 that much of online security advice is inaccessible to low-literacy users[3]: the average piece of security advice we collect is only partially comprehensible to a typical U.S. Internet user.

Comprehensibility is, however, only one aspect of quality. After recruiting 41 experts to evaluate the accuracy of this advice, we find that experts love advice: 89% of the advice imperatives we identify were evaluated as accurate (helpful in improving users' security).

Users, too, find the advice relatively palatable. We recruited a census-representative sample of nearly 300 U.S. Internet users to evaluate this advice. They rated the majority of the advice at most slightly time-consuming, difficult, or disruptive to implement and were somewhat or very confident they could implement the majority of it. When asked whether they have adopted this advice which they rate so highly, both experts and users report applying the majority of the recommended practices at least some of the time.

This analysis suggests a security education ecosystem in crisis: experts, unable to identify which behaviors are actually the most impactful for users' security have fallen in love with all of it, recommending hundreds of different practices that users work faithfully to implement, at least some of the time. The problem with security advice, it seems, may not be that much of it is bad, but rather, that there is too much of it. And, experts seem at a loss for how to prioritize it. As a result, this deluge of advice has left an army of half-armed guards protecting a multitude of half-closed safes, forgotten in the rush toward new secure practices.

[3] Literacy is strongly correlated with socioeconomic status.

Finally, in Chapters 7 and 8 we describe two methodological background studies that support the validity of the work presented in this thesis. In Chapter 7 we describe a study verifying the validity of survey data for asking questions regarding security and privacy experiences—a systematic comparison between between log data and survey data regarding an exemplar security behavior. We find that users are indeed able to accurately self-report broad security and privacy experiences in ways that generate accurate models of their behavior. This empirical validation of the methods used in Chapters 4 and 5 lends further support to the validity of our results about the relationship between security advice, socioeconomics, and security outcomes. In Chapter 8 we describe the development of a novel natural-language-processing tool—*Smart Cloze*—for automatically creating comprehension tests for domain-specific documents such as security advice. In this chapter we describe a set of experiments evaluating the validity of five different reading comprehension measures, including our *Smart Cloze* tool, on four domains of documents—Wikipedia articles, simple stories, security advice, and health advice. The findings of our evaluation provide a foundation for our advice comprehensibility measurement, as described in Chapter 6.

Chapter 2: Related Work

A significant body of prior work has focused on trying to understand why users do not behave securely online. In this section, I review related work that seeks to explain user security behavior through theoretical frameworks for security behavior change; examining the relationship between security behavior and user factors (e.g., personality, demographics); and understanding the security content to which users are exposed (e.g., warnings, news articles).

2.1 Security Behavior Frameworks

Cranor et al. propose a model for user behavior in digital security to help explain users' security choices. The Human in The Loop model [51] describes the process for communicating digital security information to users as the following set of steps: (1) attention switch, (2) attention maintenance, (3) comprehension, (4) knowledge acquisition, (5) knowledge retention, and (6) knowledge transfer. Relatedly, protection motivation theory (PMT) [195], a general framework from the psychological sciences, explains users' behavior in response to threats. PMT suggests that behavior depends on four factors: perceived threat severity, perceived likelihood of threat occurring, perceived efficacy of preventative behavior, and perceive ability to protect themselves. In Chapter 6 we draw on factors from these models (specifically, comprehension, perceived efficacy of preventative behavior, and perceived ability to protect) in developing metrics to evaluate the quality of security advice.

Finally, security decisions are sometimes framed in terms of traditional economic models of rationality and bounded rationality. Herley, for example, theoretically argues that it would be economically irrational for users to follow all of the security advice offered to them, as the time cost of doing so would be much larger than the risk [110]. Along the same lines, Beautement et al. propose the concept of a *compliance budget*—the limited time and resources users can spend on security behavior—over which users optimize their security behavior choices [26]. In my own work outside this thesis, I empirically validate a variation of these theoretical economic models of end-user security behavior, finding that users' 2FA behavior follows a boundedly rational model [187]. This work on economic tradeoffs in end-user security behavior highlights the importance of understanding the burden of different security behaviors to users, especially when considering how to educate them in a way that will align with their existing trade-off-based behavioral patterns. We return to this concept in Chapter 6 in which we include actionability—of which behavioral burden is a component—as one of our three metrics for evaluating the quality of security advice.

2.2 Factors Influencing Security Behavior

In addition to examining security behavior through the lens of general frameworks, several researchers have examined how specific user factors influence security behaviors. Das et al. demonstrated the importance of social influence; for example, showing users information about their Facebook friends' security behaviors made them more likely to adopt the same behaviors [56,57]. Relatedly, Rader et al. found that security stories from non-expert peers affect how users think about computer security and how they make security decisions like whether to click on a link [176]. Wash identified "folk models" of security, such as viewing hackers like digital graffiti artists, that influence users' perceptions of what is and is not dangerous [231]. In my work outside this thesis, I've found that security advice influences users' causal attributions for security incidents and plays a key role in informing their defensive decisions following an attack on their accounts [181].

Other researchers have considered how demographics affect security and privacy decision-making. Howe et al. note that socioeconomic status, and the corresponding belief that one's information may not be "important enough to hack," can affect security behaviors [119]. Wash and Rader investigated security beliefs and behaviors among a large, representative U.S. sample and found that more educated users tended to have more sophisticated beliefs but take fewer precautions [232]. Yet others have investigated how demographic and personality factors influence susceptibility to phishing [98,182,207]. More broadly, previous research on technology adoption has established the existence of a *digital divide*: an access, skill, and knowledge gap in digital literacy between lower- and higher-socioeconomic status (SES) demographics [103–105, 192, 215, 227]; but no security research to our knowledge examined this digital divide until the work I describe in Chapter 4.

2.3 Security Learning and Education

Security behaviors are not naturally ingrained: people learn security behaviors. In this thesis, I explore broadly from where users learn security behaviors, particularly focusing on their advice sources, the quality of the advice they receive, and ways to enhance advice.

Warnings. A large body of related work has examined the effectiveness of security warnings for training users in-the-moment on specific behaviors. More specifically, considerable prior work has analyzed how best to communicate risk, the impact of message readability, the use of metaphors, and interactive and adaptive messages for phishing and SSL warnings in browsers [21, 68, 101, 218, 235], banking security warnings [203], and security-warning habituation generally [35]. Behavioral nudging has also been explored, especially for password creation [70,87,126,225,232] and mobile app permissions [76, 125].

Advice. In addition to warnings, users receive many other types of information about security, including advice from professionals and information from news articles and privacy policies. Rader and Wash examined security advice topics, using topic modeling to analyze connections between user security decisions and the topics and words in three types of security advice [175], finding that news articles focus on the consequences of security-related attacks, web pages focus on the methodology of attacks, and the stories from peers focus on the people who conduct such attacks.

Researchers have also applied measures of readability to warnings and privacy policies as another measure of quality. Harbach et al. evaluated the readability of warning message descriptions using multiple traditional measures of text readability including the Flesch-Kincaid readability test and Gunning-Fog Index [100]. They evaluated the accuracy of these metrics using the Cloze procedure [219], which involves creating comprehension tests by removing every nth word in a given document and requiring the reader to "fill-in-the-blank" with the correct word, or computed metrics. Harbach et al. conclude that applying readability measures to warning messages is a promising approach to help developers and designers estimate the clarity of a warning for a specific audience, improving the warning message design process; yet the authors raise concerns about the accuracy of these traditional readability metrics for an adult population (as opposed to the grade-school readability measurement for which they were designed). Relatedly, McDonald et al. and Singh et al. measured the readability of the privacy policies, measuring readability via word counts, passive word proportions, the Flesch-Kincaid test and/or using the Cloze test [153, 212].

Finally, any attempt to improve the dissemination and adoption of security advice will require decisions about which advice is relevant and important and whether the advice itself can be easily comprehended by users. A relatively small body of prior work has analyzed the quality, topics, and comprehensibility of advice itself. Ion et al. surveyed more than 200 security experts to determine what behaviors they most often practice and/or strongly recommend [120], in order to provide context on what behaviors should be put in advice. The work we present in Chapter 6 extends this work, building on this initial taxonomy of security behaviors and evaluation to consider a larger number of security advice imperatives drawn directly from advice text.

Targeted Interventions. Finally, significant additional research has focused on targeted educational interventions for users, for example to educate users about phishing prevention [23, 136, 208]. Researchers have also used comics and stories to facilitate security education. Srikwan and Jakobsson developed a series of security cartoons to communicate digital security risk to users [197] and Kumaraguru et al. developed a set of PhishGuru comics, which they evaluated as part of an embedded training system that sent the comics in email to users following a simulated phishing attack [135]. Kumaraguru et al. found that users retained knowledge from the comics and found the training enjoyable. Relatedly, Zhang-Kennedy et al. showed the efficacy of interactive comics and infographics for improving users' updating and anti-virus behaviors [238, 239]. Finally, multiple studies have focused on how video games can improve security education and security behaviors, especially when embedded in corporate training [62, 208].

Chapter 3: How Users Learn Digital Security Behaviors

In this chapter[1] we present the results of a semi-structured interview study with 25 participants of varied demographics. The interviews were designed to provide a broad exploration of how users learn security behaviors. During a 60-minute interview, we asked participants questions designed to help them articulate their digital-security habits at home, as well as where they learned these strategies and why they chose to implement them, with the assumption that participants could in most cases accurately recall their habits and articulate reasons for those habits. We also addressed where participants learned security strategies and why they may reject certain strategies that they have heard about but choose not employ. We explicitly compared this information to the ways that participants learn and process physical-security advice, to determine whether mechanisms that inform physical-security advice-taking can be imported to the digital domain.

Further, we recruited participants in two groups: security-sensitive users who handle data governed by a security clearance or by HIPAA or FERPA regulations, and general users who do not. This allowed us to consider the effect that regular exposure to a data-security mindset has on the ways that users process security

[1] Published as [186].

advice in their personal (non-work) lives. Finally, we explored as a case study participants' reactions to 2FA, which has been identified as a highly effective but underutilized security tool in prior work [120].

3.1 Methods

We conducted semi-structured interviews in our laboratory between March and October 2015. To support generalizable and rigorous qualitative results, we conducted interviews until new themes stopped emerging (25 participants) [42]. Our subject pool is larger than the 12-20 interviews suggested by qualitative best-practices literature; as such, it can provide a strong basis for both future quantitative work and generalizable design recommendations [96].

The study was approved by the University of Maryland Institutional Review Board. Below, we discuss our recruitment process, interview procedure, details of our qualitative analysis, and limitations of our work.

3.1.1 Recruitment

We recruited participants from the Washington D.C. metro area via Craiglist postings and by sending emails to neighborhood listservs. We also distributed emails in public- and private-sector organizations with the help of known contacts in those organizations. In addition, we posted flyers in University of Maryland buildings and emailed university staff members. We collected demographic information including age, gender, income, job role, zip code, and education level from respondents in order to ensure a broad diversity of participants. Participants were compensated $25 for an approximately one-hour interview session.

3.1.2 Procedure

We asked participants to bring a device they use to connect to the Internet for personal use with them to their interview. Two researchers conducted all of the interviews, which took between 40 and 70 minutes. We used a semi-structured interview protocol, in which the interviewer primarily uses a standard list of questions but has discretion to ask follow-ups or skip questions that have already been covered [106]. Semi-structured interviews allow researchers to gather information about participants' practices, habits, and experiences as well as their opinions and attitudes.

During the interview, we asked questions about participants' digital- and physical-security habits as well as where they learned those habits. We also asked participants to "act out" their use of technology in a series of scenarios. We asked questions about participants' behaviors and advice sources for digital-security topics such as device security, including password protection and antivirus use; web browsing and emailing, including 2FA and phishing questions; and online banking and shopping, including questions about the participant's banking login process and payment methods. We asked similar questions regarding physical-security topics such as dwelling security, including questions about locking methods and alarm systems; transit (e.g. car and bike) security, with questions similar to those asked for dwelling security; and personal safety when walking alone, including questions about carrying weapons. We validated that our list of digital security topics broadly covered the same topics as those mentioned as high priority in prior work by Ion et al. [120].

On each of these topics, participants were first asked a general open-ended question regarding their security behaviors: for example, “How do you protect your devices?” and then asked sequentially more specific questions, for example: “Can you show me how you access the home screen on your smartphone?”, “Have you always had/not had a password on your smartphone?”, and “Are there other strategies you use for protecting your devices which you have not mentioned?”

Participants were subsequently asked a series of follow-up questions on each topic, such as “Why do you use this strategy?” ; “Have you ever had a negative experience with...?” ; and “Where or from whom did you learn this strategy?” . In addition to questions regarding specific security topics, participants were asked more generally about where, from whom, and why they accepted security advice, as well as about strategies they had considered but not adopted . Participants were also asked to compare digital- and physical-security advice in terms of usefulness and trustworthiness. Finally, participants were asked to briefly describe their current or most recent job. They were specifically asked if they handled sensitive data as part of their job, and if so, what kind. The full interview protocol is included in Appendix A.1.

3.1.3 Analysis

The interview data was analyzed using an iterative open-coding process [216]. Once the two interviewers completed the interviews, they transcribed 17 of the interviews. The remaining eight interviews were transcribed by an external transcription service. The interviewers then met in person to develop and iteratively update an initial set of codes for the data. Subsequently, they independently coded each interview, incrementally updating the codebook as necessary and re-coding previously coded interviews. This process was repeated until all interviews were coded. The codes of the two interviewers were then compared by computing the inter-coder percent agreement using the ReCal2 software package [84]. The inter-coder percent agreement for this study is 75%. This is a reasonable score for an exploratory semi-structured study, with a large number of codes, such as ours [142]. Further, after calculating this percent agreement score, the interviewers met to iterate on the codes until they reached 100% agreement on the final codes for each interview.

3.1.4 Signifying Prevalence

For each finding, we state the number of participants who expressed this sentiment, as an indication of prevalence. However, our results are not quantitative, and a participant failing to mention a particular item for which we coded does not imply they disagree with that code; rather the participant may have simply failed to mention it. As a result, we opted not to use statistical hypothesis tests for comparisons among participants. Our results are not necessarily statistically generalizable beyond our sample; however, they suggest many areas for future work and provide novel contributions to the body of work surrounding users' strategies for learning digital-security behaviors.

3.1.5 Limitations

Our study has several limitations common to qualitative research. While we asked participants to search their memory for answers to our questions, they may not have fully done so, or they may have forgotten some information. Further, we assume that participants are largely able to correctly identify which of their behaviors are security behaviors and why they practiced those behaviors. To mitigate satisficing [115], interviewers repeatedly prompted participants to give full answers to all questions. Participants may also have tired and provided less thorough answers toward the end of the interview, and those who were particularly concerned about the interviewer's perception of them may have altered their answers in order to not portray themselves as overly secure or insecure [115, 224]. Additionally, the age, gender and race of the interviewers may have introduced some bias into participants' responses. We recruited a diverse pool of participants to increase the odds that relevant ideas would be mentioned by at least one participant, despite these limitations.

3.2 Results

In this section we detail the results of our study. First, we will discuss our participants' demographics and security sensitivity. An overview of these demographics is shown in Table 3.1. Second, we will address the sources from which participants accept security advice and how these sources differ across genders and for physical and digital security. A summary of these sources is shown in Figure 3.1. Third, we will address the different reasons our participants gave for accepting and rejecting digital- and physical-security advice; some of the differences in these reasons were unanticipated. Fourth, we address differences between security-sensitive and general participants, which imply imply that exposure to digital-security information in the workplace may have effects on advice processing. Finally, we present a case study on 2FA, a behavior found by Ion et al. to have high security importance, but low adoption [120].

3.2.1 Participants

We recruited 158 potential participants and selected 47 to interview. We selected a balance of men and women, as well as a diversity of age, ethnicity, and education. Of the 47 participants selected for interviews, 25 attended their interview appointments.

Demographics for our 25 participants are shown in Table B.17. Fifty-six percent of our participants are female, slightly more female than the general U.S. population in 2014 (51%) [13]. Our sample is somewhat less Hispanic (8% vs. 17%)

ID	Gender	Age	Race	Educ.	Income	Sec. Type
P1	M	31-40	W	M.S.	\$90-\$125k	F
P2	F	22-30	A	B.S.	\$50-\$70k	—
P3	M	18-22	W	S.C.	\$90-\$125k	F
P4	F	51-60	W	Ph.D.	\$150k+	S
P5	F	22-30	B	M.S.	\$90-\$125k	F
P6	F	41-50	W	M.S.	\$30-\$50k	—
P7	F	31-40	H	M.S	\$70-\$90k	F
P8	F	31-40	B	M.S.	\$90-\$125k	—
P9	M	22-30	W	B.S.	\$50-\$70k	S
P10	M	22-30	B	B.S.	\$50-\$70k	S
P11	M	60+	W	P.	\$90-\$125k	C
P12	M	41-50	B	S.C.	\$0-\$30k	S
P13	F	31-40	A	M.S.	\$0-\$30k	—
P14	F	31-40	B	S.C.	\$90-\$125k	—
P15	F	41-50	B	Assoc.	\$50-\$70k	C
P16	F	31-40	H	H.S.	\$0-\$30k	—
P17	F	18-22	B	H.S.	\$0-\$30k	—
P18	M	18-22	B	H.S.	\$0-\$30k	—
P19	F	22-30	B	M.S.	\$50-\$70k	F
P20	F	60+	W	Ph.D.	\$150k+	—
P21	M	41-50	W	Ph.D.	\$150k+	C
P22	M	60+	W	S.C.	\$90-\$125k	—
P23	F	22-30	B	Assoc.	\$70-\$90k	H
P24	M	41-50	W	B.S.	\$30-\$50k	S
P25	M	18-22	B	Assoc.	\$70-\$90k	H

Table 3.1: Participant Demographics. The columns show: participant identifiers (coded by interview date order), gender, age, race (White, Black, Asian, and Hispanic), education, gross household income in 2014, and security sensitivity at work. The abbreviations in the education column stand for high school graduate, some college, Bachelors degree, Associates degree, Masters degree, Doctoral degree, and Professional degree (e.g. MBA, J.D.). The abbreviations F/H/S/C/—in the security type column stand for FERPA, HIPAA, and SSN data handling, the holding of a Security Clearance, and no work with sensitive data, respectively.

and less White (40% vs. 62%), but more Black (44% vs. 13%) than the U.S. population [13]. We had a proportional number of Asian participants (8%). However, the racial makeup of our sample more closely matched the racial proportions of the Washington D.C. metro area, which is 43% White (our sample: 40%), 46% Black (our sample: 44%), 10% Hispanic (our sample: 8%) and 4% Asian (our sample: 8%) [8]. Our participant sample is wealthier than the US population and our demographic area: 28% of our participants have a household income under $50,000, whereas 47% of households in the general US population and 40.1% of households in the D.C. area earn less than $50,000 per year [8, 10]. Our sample is, however, representative of the educational attainment in our demographic area: 88% of our participants hold a high school degree or higher, compared with 90.1% per the D.C. area census; and 60% of our participants hold a Bachelor's degree or higher, compared to 55% in the D.C. area [8].

3.2.2 How Security Behaviors Are Learned

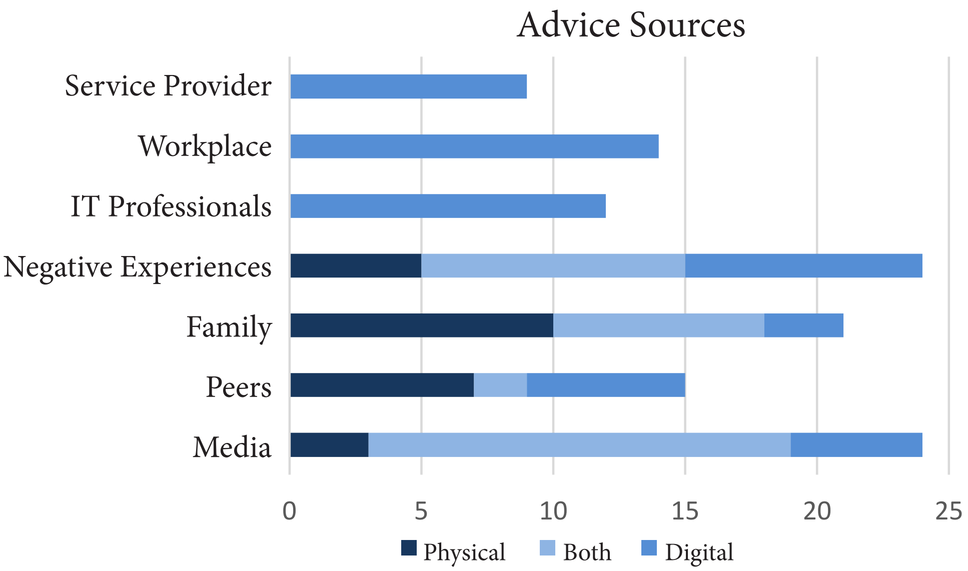


Figure 3.1: Number of participants who reported using each advice source for digital and physical security, respectively.

Participants reported implementing digital- and physical-security advice from a number of sources. While many sources were common to both digital and physical security (media, peers, family), in this section we emphasize advice sources unique to digital security, including IT professionals, the workplace, and providers of participants' digital services (e.g. Comcast). Next, we discuss a new source of security information: fictional portrayals of negative-security events through TV shows and movies. Our findings emphasize and expand prior findings on the importance of negative security stories for teaching digital security behaviors [176]. We then consider common sources—media, family members, and peers—in more detail. We examine which specific people and sources in this group our participants considered authoritative. Finally, we include an interpretive section discussing gender-based differences in advice sources.

Digital Only: IT Professionals. IT professionals are an information source strictly for digital-security methods (N=12). These professionals can be colleagues in a participant's work environment or friends of the participant. As we will discuss in Section 3.2.3, a participant's belief that a digital-security advice source is trustworthy is a primary factor in whether they choose to accept the advice; it seems that participants view IT professionals as especially trustworthy. "For personal [digital security advice], I might talk to one of the IT guys about that. I just talk to . . . the one I'm most friends with, I always try to get information: what's the best intervention, what do you think?" comments P15. Further, participants may use IT professionals to evaluate the trustworthiness of advice they have seen elsewhere. For example, P19 says that when she is looking for new digital-security advice, she will "talk to the IT guy at my office. I've talked to him a couple of times about my phone and whatever I hear or read." Although participants may receive useful advice from colleagues and friends who are IT professionals, we hypothesize that this advice may not be sufficient. For example, as P13 notes: "My friends who work in IT, they just tell you to change your password as often as possible."

Digital Only: Workplace. In addition to information users solicit from IT professionals, users also receive unsolicited security advice from their workplaces in the form of newsletters, IT emails, or required trainings. Fourteen participants cited receiving this type of advice. P4 says, for example, that she learned from work not to click links in emails that claim she needs to update her password. "We got an email from IT telling us that, never will there'll be an email from them that would require you to do that." Similarly, P8 pays attention to her security trainings at work: "They'll do yearly IT security training, which is not even necessarily for work, but just for life . . . they talk about things like not sending people money over Facebook . . . they also email out updates when things change. I do actually pay attention to those emails when they send them, like about privacy notice updating." Further, P2 says she "always reads the IT newsletter" put out by her workplace.

Digital Only: Service Provider. Another source of digital security information cited by nine participants is the corporations that provide a service to the participant (e.g. SunTrust Bank, Apple, Verizon). For example, P23 comments: "I usually call my carrier (Comcast) and they have security stuff for your Internet and they'll tell me what I can do."

Negative Experiences. As reported in Rader and Wash's work on security

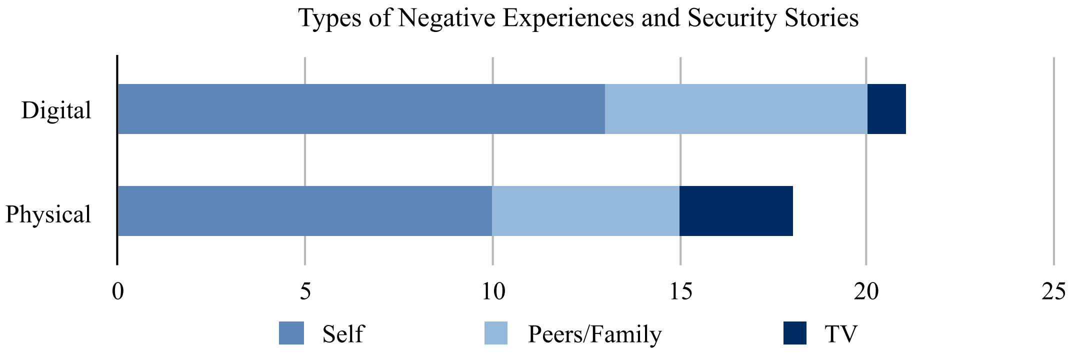


Figure 3.2: Distribution of types of negative experiences from which participants learned new security behaviors: personal events, stories told by peers, and stories in TV shows or movies.

stories, negative events described by peers or directly experienced by participants can be strong learning tools [176]. In our study, we found that 24 participants either had negative experiences themselves or were told stories of negative-security events by peers, which led to behavior changes. The distribution of the types of negative-security situations (events that happened to the participant, to the participant's friend, or that the participant heard about through TV) on which participants relied is shown in Figure 3.2. Our participant sample was smaller, yet broader, than that used in Rader and Wash's work, and our results thus confirm the generalizability of their findings beyond the college student population [176].

Participants tend not to learn from security stories told by others or from events that happen to themselves when they feel that they or the victim did all they could to prevent the event, when they feel that they or the victim placed themselves in harm's way, or when they cannot find a cause for the negative event. For example, P2 had a friend who was robbed, but did not change her own behavior "because I think she took all the precautions she reasonably could. She parked in a brightly lit area and a reasonably safe neighborhood...I don't think that there was much...[that she could] have changed." P24 and P9 have had friends who got viruses, but they did not do anything differently afterwards, because they felt that the friends were victimized due to their lack of technical expertise. Finally, P18 comments, "I actually think recently someone tried to log into my email from China and Google sent me an email and Google blocked it and said it looked strange and I said it was very strange," but he did not alter his behavior after this incident.

Although only four participants cited TV shows specifically, each strongly recalled stories of negative physical or digital security-related events happening to characters in those shows. They directly credited these shows with leading to a specific change in their behavior. For example, P12 put a password on his WiFi network after watching a tech show that showed "people going by houses and WiFi snooping and knocking on people's doors saying, 'Oh your WiFi is open, you need to protect it' . . . shows like that, [they] make you think." P14 had a similar experience: watching a movie motivated her to always check the back seats in her car for a lurking person. "People had mentioned that you should check your back seats before but I never paid attention to it until [this] movie," she says. Thus, it seems that TV shows or movies may serve as strong proxies for a negative experience that happens directly to the user or someone she knows. We hypothesize two reasons for this: (1) while participants often blamed themselves or their friends for personality or behavioral flaws that led to security problems, they were more likely to give relatable fictional characters or the unknown real victims shown on TV the benefit of the doubt; and (2) TV shows and movies are typically designed to be vivid, realistic, and believable, thus making participants feel that what is happening on the screen could happen to them, too.

Evaluating Authority in Common Advice Sources. Prior work has identified media, family, and peers as important sources of digital-security advice [56]. Our results confirm these findings, and offer additional insights into which media participants feel is most authoritative and how participants evaluate the expertise of their family and peers.

Almost all participants (N=24) reported receiving both digital- and physical-security information from media. Media included online articles, forums, television shows, news shows, the radio, magazines, and advertisements. Of the participants who cited media as an advice source for digital security, five participants cited a specific technology-oriented resource as authoritative or trustworthy: "Some of the blog[s] I read [are] by computer people, those are the most trustworthy. For example, I read Wired," says P20. In general, the technical sources cited by these participants were: CNet, Wired, Bruce Schneier's blog and Mashable. [1, 5–7].

Another common source of digital- and physical-security advice are family members (N=21) and peers (N=15). In describing why they chose to take security advice from their family members or friends, 11 participants said they consulted their peer or family member because they considered this person an expert. For example, P1 says he always asks his father-in-law for digital security information because his father-in-law is "a bit of a techie in his spare time. He's the one that I go to for advice and feedback, new stuff, articles, he'll send links. He knows the best of what's going on." Interestingly, however, expert status in our sample was not necessarily determined by education or job role (e.g. IT professional, police officer) but rather by participant's perceptions of the "tech-savviness" or physical-security expertise of their peer or family member.P3 says that he purchased anti-virus software at his father's direction. He says, he's "very tech-savvy and he'll say, 'You need to get this. This is important.' I don't question him because he's very much in the know." When asked what makes his father 'tech-savvy', P3 says "he's always loved computers and all that entails, but he doesn't work in technology." Further exploration of specific cues leveraged by users to assess the 'tech-savvy' or expertise of their friends, family, and the media could aid researchers in signaling advice-source trustworthiness, which is a primary motivator for users' acceptance of digital-security advice, as discussed further in 3.2.3.

Gender and Advice. Eighteen participants, evenly split between men and women, cited a man as a source of digital-security advice, while only three cited a woman. If this trend holds true among a larger population, it may be because men have historically been overrepresented in technology and computing fields and thus are considered to be more authoritative on that topic [79]. Alternatively, men may simply offer more unsolicited advice in the domain of digital security, or perhaps because women are still underrepresented in IT and computing fields there are fewer women who chose to offer digital-security advice [40].

On the other hand, 12 participants cited a woman as a source of physical-security advice, compared to three participants who cited men. Eight of these 12 participants who received physical-security advice from women were women themselves. Historically, women have had higher rates of crime victimization, perceive themselves to be at higher risk of victimization, and express greater fear of crime than do men [151]. It is probable that women are aware of this gendered difference in threat levels and perceptions, and thus find each other more relatable sources of advice.

3.2.3 Why Advice is Accepted

What leads users to accept advice from the sources mentioned above? In this section, we discuss participants' reasons for accepting security advice. We find that the trustworthiness of the advice source is the key metric for digital security. This finding may be explained by another of our findings: participants struggle to assess the plausibility and value of digital-security advice. In contrast, participants' relative confidence in their assessment of the plausibility of and necessity for physical-security advice leads them to cite their own evaluation of the advice's content as the primary assessment metric in the physical domain. We also in this section compare which advice, physical or digital, participants feel is more useful and/or more trustworthy.

Digital-Security Advice. Eleven participants used the trustworthiness of the advice source to determine whether to take digital-security advice.

In the case of media advice, participants must determine whether advice offered by an unknown author is trustworthy. Participants mentioned five heuristics that they use to measure the trustworthiness of a media advice source, including:

their knowledge and trust of the advice author, other users' reviews of the advice, how widespread the advice was on various media outlets, whether the content of the advice differed strongly from their current behavior, and the simplicity of the advice. All of these heuristics were equally prevalent in our data.

The first technique mentioned for evaluating media advice source trustworthiness was to assess the author or media outlet providing the advice: P20 notes that her acceptance of advice, “depends on the author and how the article is written.” P22 says he finds advice useful “If I would quote that source to someone else, like the Washington Post, [or another] reputable media outlet. If it's just some Matt Drudge on the Internet advising about computer security, I would just ignore that more quickly than I saw it.”

A second evaluation metric was other users' reviews of the advice. Two security-sensitive participants, one who holds an M.S. in digital security (P24) and another who handled FERPA data as an HR file clerk (P10), crowd-sourced their advice and software evaluation. P24 comments, “I evaluate howto videos and other advice channels via user comments.” Similarly, P10 says, “I look at reviews and the software and the website to decide whether to use the advice or download [software]. I look at whether it has a good reputation—whether it is popular with online reviewing.”

A third heuristic for advice evaluation was how widespread across different media outlets the advice became, with the implicit assumption that distribution outlets who reprinted a given piece of advice had evaluated the sources and information and found it to be valid. P25 comments that he trusts “news that's backed up by facts and is across multiple channels, because if it's not good, multiple places won't pick it up."

A fourth metric for evaluating a media advice-source trustworthiness was how much the content of the advice differed from the participant's current behavior: P5 says she took the advice because "it was the opposite of what I was doing, so it automatically made it seem as though it was more credible." P2 comments that she took the advice since "it made sense; I guess if [my password is] a bit longer, it's harder for [a malicious] computer to figure it out."

Finally, a fifth heuristic for media advice-source evaluation is the simplicity of the advice. P2 adds, "If it's just tips that you can implement in your everyday life, then the advice feels more trustworthy" and P16 wishes that advice "would have a better setup to say 'Here, this is what you have to do for step one, step two, step three.' . . . like from Google when they're saying that you can [add] privacy."

Participants may rely on the trustworthiness of the advice source because they are not confident in their own ability to evaluate the content of the advice. Indeed, P7 says, "physical security is related more to me and my body . . . it makes sense to me whereas with computer security, I'm securing myself from threats that I don't even know anything about...I know when somebody walks up with a gun that I should be worried." P12 also notes that the tangibility of physical security can make personal safety strategies more trustworthy and easier to implement, commenting, "you know, cyber security is great, but the people who are doing it are so smart that they can put back doors in it that you don't even know about, so sometimes, I don't even trust the advice...with physical security, I can touch that, or I know someone that I can relate to.”

That said, participants’ ability to accurately judge the trustworthiness of advice sources may vary. As an example of good advice, P9 learned to use incognito browsing from a friend, “incognito came out in college and a friend came over and needed to use gmail and just said look at this and logged himself into gmail and didn’t need to log me out and it was useful.” Similarly, P15 learned about security alarm systems “years ago, from a friend of mine who had a security alarm business.” However, P17 mentioned being told less credible information such as the following: “A lot of my friends don’t have iPhones because, this is the term they use, ‘iPhones are hot’. Like they attract all the attention to your phone, like anything you’re doing illegal it can get caught on your phone, ‘cause it’s like a hot box iPhone. It can be tracked in any type of way, stuff like that. I didn’t even know that, I was like whoaaaaa it can be tracked? If I had known that, I wouldn’t have gotten an iPhone, yeah.”

Physical-security advice. As participants are more confident in their ability to evaluate the plausibility of physical-security advice content, for physical security, the advice source is of lesser importance. Only three participants cite the trustworthiness of a physical-advice source as an important metric, and those participants also cited this metric for digital security. Instead, participants rely on their own assessments of physical-security advice to determine whether to implement new behaviors (N=7). On the subject of plausibility, P22 says about physical-security advice, “if it doesn’t pass the smell test, in other words if it just doesn’t seem plausible, then I dismiss it. If it's something that I recognize as making sense," then he will consider implementing it.

Digital vs. Physical Advice: Usefulness and Trust.

Figure 3.3 shows participants' assessments of the trustworthiness and usefulness of digital- and physical-security advice.

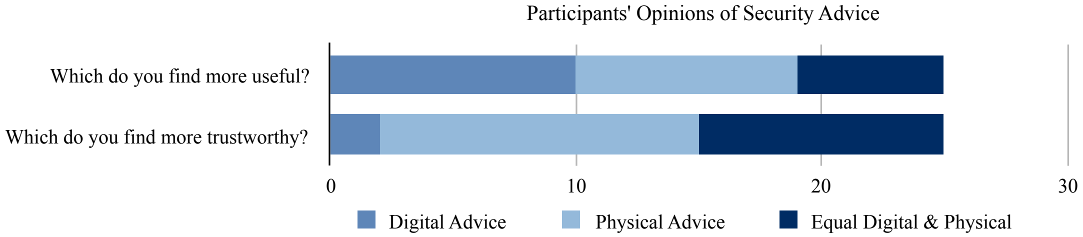


Figure 3.3: Participants' opinions regarding which security advice, digital or physical, is most useful.

Half of our participants (N=13) felt that physical-security advice was more trustworthy overall than digital-security advice. Only two participants felt that digital-security advice was more trustworthy than physical-security advice. The remaining 10 participants felt that digital- and physical-security advice was equally trustworthy. We suspect that this was largely because, as mentioned above, participants find physical-security advice easier to mentally evaluate (N=7). P9 comments that he would probably trust physical-security advice more than digital-security advice because: "there are a lot fewer variables. I trust it more because it's easier to evaluate if it's legitimate." Similarly, P23 says that she trusts physical-security advice more because it is "more hands on and visual, it's in your face a little bit more."

Relatedly, five participants trust physical-security advice more because they feel it is simpler and easier to implement than digital-security advice. “Physical-security advice is more trustworthy because it’s more common sense and they don’t typically require you to download and install something that would be trouble in itself,” comments P20.

Participants are more split on which advice, digital or physical, is more useful. Nine participants feel that physical advice is more useful, primarily for the same reasons they found physical advice more trustworthy: “I can see the relevance in the personal security whereas the computer security, again I am trusting that because I have a little icon on the right that it is doing its job. Do I know what it it’s doing? No.” says P7. Similarly, P3 comments that he finds physical-security advice more useful because: “Again, it’s my understanding. It just comes so much more naturally.”

On the other hand, the 10 participants who feel that digital advice is more useful noted that there are more techniques available for digital than physical security and that they feel a higher risk of digital threats. To the first point, P15 says: “digital-security advice is more useful—because with digital I can probably do more research, and there’s more to do there than the physical. Physical you can only do so much, I don’t care what I have on me, someone can overpower me.” With regard to feeling that there is more digital than physical security risk, P11 comments, “[I] find digital security more useful and more trustworthy because there is so much more research on it and it’s so much more pervasive.”

3.2.4 Why Advice is Rejected

While trustworthiness and plausibility are the two main reasons our participants choose to *accept* advice, there are a multitude of reasons for which they reject it. Inconvenience is often cited as a possible explanation for users rejecting digital-security advice [26, 102, 110], but it was not the most prevalent reason we discovered. Our participants related frustrations with advice content, such as the content being too marketing-oriented, or less surprisingly, too advanced. They also rejected digital-security advice when they believed that they were not at risk or felt that implementing security measures was not their job. Figure 3.4 summarizes the prevalence of these reasons for rejecting digital- and physical-security advice. Below, we provide further detail on these reasons, and compare and contrast participants' motivations for rejecting advice in each domain.

Too Much Marketing. Eight participants rejected digital- and physical-security advice because it appears to be more about selling a product than about providing advice: "I don't do anything with a price tag attached. I could be persuaded to do it if I had a serious problem. I did have my identity stolen one time but I was able to fix it, but I'm not one of these people who signs up for [identity theft protection] or something like that," says P22. Similarly, P16 wishes that physical-security advice could be more substantive and distributed primarily through mechanisms other than advertisements.

I'm Not At Risk. Eight participants rejected physical-security advice as unnec-

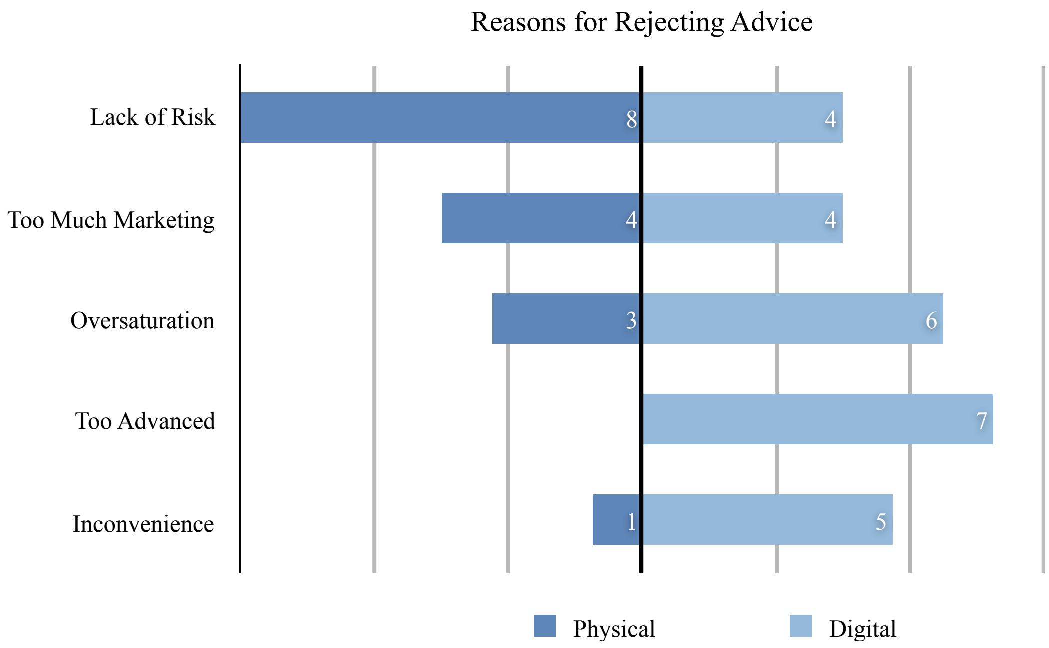


Figure 3.4: Distribution of reasons participants rejected digital- and physical-security advice.

essary due to their low risk profile. For example, P24 says: "[I've] heard about 24-7 monitoring and crap like that, I think it's overkill. If everyone [in my neighborhood] was driving fancy cars, maybe."

Four participants rejected digital-security advice for the same reason. P5 says he does not put a password on his phone because, "I just don't feel I have that much interesting stuff on there." P10 comments that she does not use or look for security tactics for her tablet, because "there's nothing personal on the tablet." Similarly, P3 does not take security advice for browsing because he is "not so concerned about browsing as opposed to personal financial information." The participants who cited these feelings for digital security were of varied incomes, and the overall incidence of feelings of "unimportance" around digital security was quite low. This is in contrast to prior work, which had proposed that many users, particularly those with lower incomes, might not execute security behaviors due to low valuation of their data [119]. One possible cause for this change is that as technology becomes more ubiquitous, users are becoming more aware of the value of their data. Overall, feelings that risk was low and therefore implementing a new behavior was unnecessary were more common for physical than digital security.

It's Not My Job. Eighteen participants rely on the companies whose software, hardware, or services they use to keep them safe. These participants do not seem to be making explicit cost-benefit calculations about particular personal behaviors being redundant to the services provided by these companies; rather, they simply assume that they are not responsible for the security of a given system because a corporation they trust is taking care of it. This motivation for rejecting security advice was unique to the digital-security domain. For example, P8 comments, "I had been banking with a bank that I wasn't happy with. Then I went to Bank of America, which was this big bank. I'm like, 'Oh, they're awesome so I don't have to worry about anything. I will be safe.'"

In addition to trusting corporations to take care of security for them, participants also rely on browser and device prompts (N=20), software defaults (N=20) and security requirements imposed by their services (e.g., your password must be 16 characters long) (N=14) to keep them safe. For example, many participants use a password or passcode to lock their phone because the phone prompted them to do so at set-up. P2 says, "When you boot up these phones now, they just give you the option." Relatedly, P4 says she only has passwords or passcodes on her Mac products because, “the Mac products prompt you to set up the security things...I never thought about it [for the Kindle]. I guess it wasn’t prompted...I would have to look up, how to do it on the Kindle.” In addition to prompts, participants rely on software defaults, such as those in anti-virus software, to provide security tactics: P17 comments, that she has a script and popup blocker because it “was through McAfee and it was automatic. . . . I’m not really technical savvy where I can block stuff and...go into my settings and know what I’m messing with.”

Other reasons for rejecting advice. Nine participants stated that they felt oversaturated and lacked the time to implement the advice they saw, even if they thought it was good advice. P7 says: “Part of it is just saturation. You get so much information from so many sources. I don’t even know sometimes what’s worth looking at.” Additionally, P6 notes that in general he often does not take security advice because he has “kind of reached a level of don’t care. It’s so obvious to me that I don’t know what I don’t know, that it’s frustrating to try to tease apart what would be helpful and what wouldn’t.”

The advice may also be too advanced (N=7), too inconvenient (N=6), or participants may feel that no matter what, they will be hacked (N=11). Even participants who are highly educated may reject digital-security advice for being too advanced (N=4). P9 holds a computer engineering degree and says he knows that HTTPS and SSL exist, but “I don’t even know what the acronyms mean, I know that some websites are more secure and others aren’t, and I don’t pay attention to it.” P8, who holds a master’s degree, also struggles to understand too-complex advice: she sometimes rejects advice, “Depending on the number of steps and the complexity of it because I’m not a IT person . . . it can be complex what they’re asking me to do.”

Finally, a few participants described reasoning that was less common but still interesting, with possible implications for design. One participant (P3) noted that he rejects advice because he see it in the wrong venue: “I see the information while on [public transit] to work and then by the end of the day, looking at a computer is the last thing I want to do.” We hypothesize that this factor may be important for many users, even though no other participants explicitly mentioned it. A few other participants reported rejecting what they perceived as good advice for others because they were already confident in their own behaviors (N=3). P25 notes that having others tell him how to be digitally secure is pointless, because: “I do what I do based on my own personal feelings and intellect, so I don’t find it useful, but for someone who didn’t know it would be useful. Never found any of the advice useful. I just have my own way of protecting what I do, so it’s like if someone’s telling you how to make a PB&J sandwich, and I’m like I know how to do it. But if they’re saying something drastic—don’t do this, this, and this—then I’ll look at it, but usually, no.”

3.2.5 Security-Sensitive vs. General Participants

In addition to differences between participants’ behavior in the physical- and digital-security domains, we also noted possible differences between participants in our sample who are and are not security-sensitive. We recruited security-sensitive participants to investigate how extra training in handling confidential or sensitive data at work would affect how participants process security advice in their personal lives. Below, we discuss some observed trends that appear to differentiate security-sensitive from general participants; given our qualitative data and limited sample size, these findings mainly serve to suggest directions for further exploration. The prevalence of these differences in our sample is summarized in Figure 3.5.

Two-Factor Authentication. Seven of 15 security-sensitive participants in our study had adopted 2FA, compared to eight of 10 general participants. Four of these security-sensitive participants cite privacy concerns as a reason for not using 2FA. Thus, we hypothesize that security-sensitive users may be less trusting that the service requesting 2FA can protect their personal information. Participants' motivations for accepting and rejecting 2FA are discussed in more detail in Section 3.2.6. We quantitatively explore this potential difference between the privacy concerns of security-sensitive and general users in Section 4.

Advice Evaluation. Nine of 15 security-sensitive participants cited the trustworthiness of the advice source as their key metric for choosing to take digital-security advice, compared to only two of 10 general participants. We suspect that security-sensitive users may be more discerning about advice because they have been trained to look critically at the digital information they come across. A primary component of workplace digital-security training is reminders not to trust unknown individuals [2, 4].

Workplace Digital-Security Advice. Thirteen out of 15 security-sensitive participants took advice from their workplace, contrasted with four of 10 regular participants. This is perhaps unsurprising given the workplace emphasis on digital-security and regular trainings that occur for security-sensitive users.

Beliefs About the Utility of Digital Security Advice. Eight of 15 security-sensitive participants in our sample believed that digital-security advice was more useful than physical security advice, compared to two of 10 general participants. We speculate this may be related to these participants being more frequently reminded to pay attention to digital security and data sensitivity.

Feelings of Inevitability. General participants in our sample expressed more feelings of inevitability ('no matter what, I will be hacked') than did security-sensitive participants. Six out of 10 general participants expressed these feelings, contrasted with three out of 15 security sensitive participants. We hypothesize that less formal training may contribute to general users having more feelings of powerlessness.

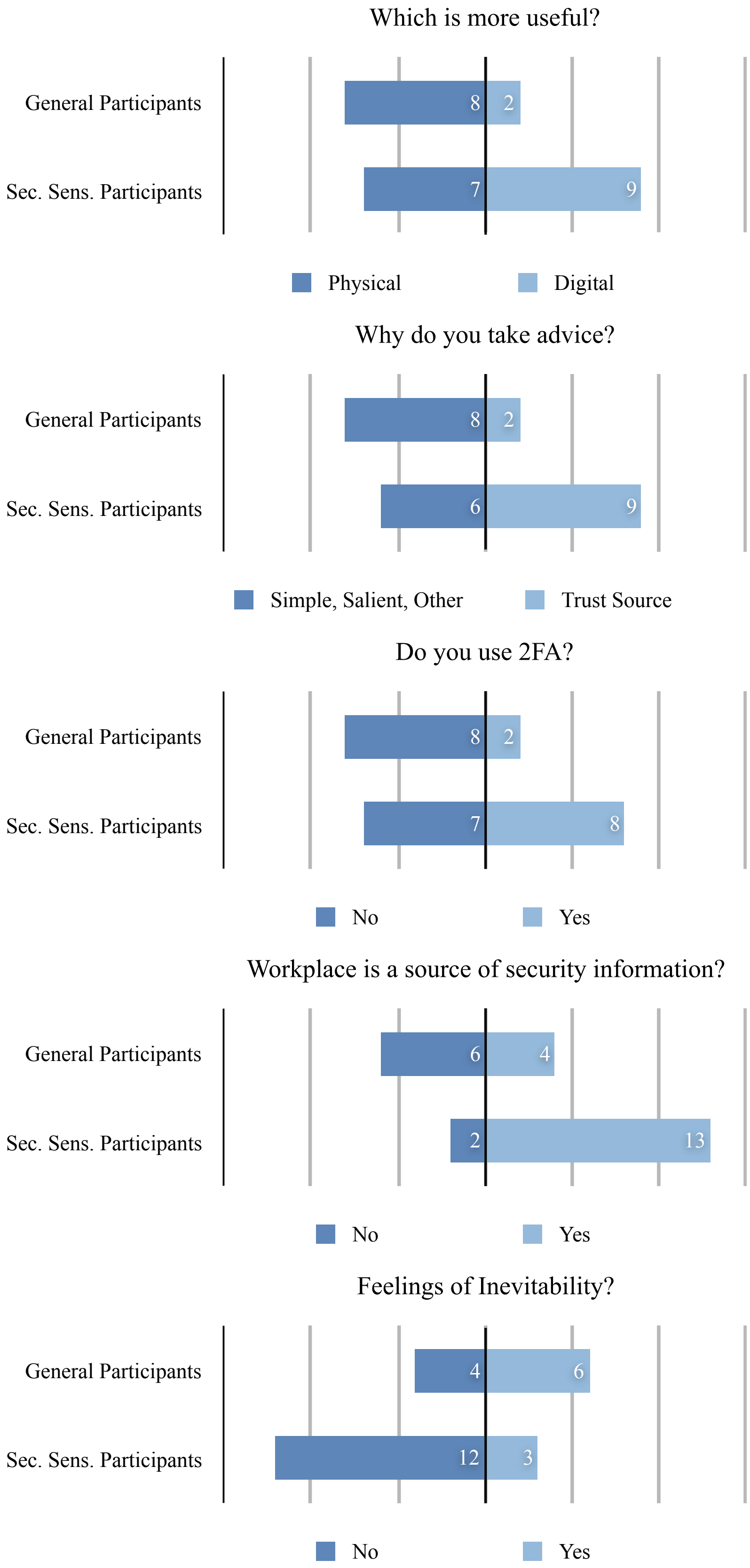


Figure 3.5: Security-sensitive participants in our sample tend to differ from general participants in their valuation of digital-security advice, their reasons for taking advice, their use of two-factor authentication, and some of their advice sources.

3.2.6 Case Study: Two-factor Authentication

Ion et al. report that use of 2FA is one of the top three security behaviors recommended by or used by security experts [120]. However, only 40% of the non-expert participants in that study reported using 2FA. Our results shed some light on the reasoning behind users' acceptance or rejection of this behavior.

How and Why I Use Two-Factor Authentication. Of the participants we interviewed, more than half reported using 2FA (N=14). In our interview questions about 2FA, we started by defining 2FA as "a service where you might put in your phone number and then be sent a verification code." Given this definition, all participants recognized 2FA and were able to substantively answer our interview questions on this topic. Of our 14 participants who had used 2FA, five used 2FA for some, but not all, services for which it is offered. These participants use 2FA for those services they feel are particularly important: P6 says, "I've got 2FA on one thing, and that is my insurance company. I did that because [of a negative experience at my workplace]. I figured that [my insurance] was one of the most important things, because...it covers every aspect in my life. I didn't want anyone to mess with that."

Alternately, participants may only use 2FA on services that strongly encourage or force them to do so: "I do that with Xbox Live, they force me to do that. I think Google, they want me to do that but I always say later," comments P12. [2] Similarly, P14 says: "Yes, at one time Verizon, because I have a Verizon email account, it

[2] Note that XBox Live does not require 2FA, but this participant may have misinterpreted the prompt screen as a requirement.

asked me to do [2FA], it takes a while but I've done it...it forced me to do it." Of the remaining nine participants who used 2FA, two did not understand what they were doing: P16 comments, "You mean when it asks to use by text or phone call? I do that, even though I hate doing it, because I'm trying to figure out what is the purpose, but it says the purpose is your safety and security."

Why I Don't Use Two-Factor Authentication. Eleven participants knew about but chose not to use 2FA. Five of these participants declined 2FA due to privacy concerns: specifically, they worried about giving out their personal phone number, about GPS tracking based on that phone number, and about the service providing 2FA's ability to keep their information secure. For example, P13 says: "No, [I want] nothing connected to the phone. So, the phone is directly connected to the email. I don't feel comfortable to let people in if it's connected to the email account." Similarly, P3 says: "I think I do have that [2FA] capacity. I think I've always declined Gmail enabling that access...Based on what I know about Gmail, it just seemed like giving up too much information to Google." With regard to protecting the information used for verification, P23 says: "Google has prompted but I've always ignored it because I think that someone will get ahold of it, I'm not saying they would, but I'm just always like, you know, yeah."

In addition to privacy concerns, two participants declined to use 2FA due to convenience concerns: "Two years ago, at the beginning of the summer, Google introduced 2FA, and this was an issue because I tried to log in and I didn't get cell service and I couldn't get the text message to log in, and that was the last time I tried to change anything," says P9. And two participants declined the service due to not understanding the purpose of the tool.

3.3 Discussion

The primary contribution of this chapter is a taxonomy of security advice sources and qualitative insight into why users chose to accept and reject advice. Our results underscore the importance of minimizing digital-security advice so that users do not feel overwhelmed. As mentioned in Section 3.2, several users felt oversaturated with advice and further, felt that they lacked the time to implement all the advice they were given. This is consistent with the compliance-budget model [26]. Given that we found that many users struggle to evaluate the plausibility and usefulness of digital-security advice, reducing the amount of advice they need to process will significantly reduce their cognitive load. This of course requires identifying a small set of recommendations that provide high value to users, which we seek to do in Chapter 6. While the amount of advice provided should be strictly limited, our findings suggest that critical advice can be made more effective in several ways. More than half of our participants felt that physical-security advice was more trustworthy, or more useful, than digital-security advice, we hypothesize this is due, in part, to the small number of simple instructions typically provided in physical-security advice.

Additionally, we found that users reject security advice for a number of somewhat surprising reasons, including containing too much marketing information and threatening users' sense of privacy. Thus, while participants trusted the companies from which they purchase their technology and services as a good source of trusted advice, it is important for organizations to avoid the perception of marketing so that users can easily recognize the credibility of their information. In subsequent related work, I and coauthors explored the potential for companies to serve as advice givers, at the time of purchase of a new computer [168], confirming the importance of avoiding marketing perception, and developing a framework for intervention between service providers and consumers.

More broadly, our findings indicate that users believe they lack the skills to evaluate the content of digital-security advice and must instead rely on their evaluation of the trustworthiness of the advice *source* when determining whether to accept the advice. For example, we found that participants rely on IT professionals, particularly those from their workplaces, as a source of credible digital-security advice, even for personal technology. Given that many IT professionals are already overloaded with requests, organizations may need to consider providing them with extra support and training for this potentially critical but under-acknowledged role.

Chapter 4: Quantifying Security Education

In this next chapter[1], I draw on the exploratory, qualitative analysis of users' security education mechanisms and advice sources described in the prior chapter as a foundation for a quantification of users' security educational mechanisms. Additionally, I specifically explore whether there exists a *digital divide* in security education. Previous research has established the existence of a *digital divide*: an access, skill, and knowledge gap in digital literacy between lower- and higher-socioeconomic status (SES) populations [103–105,192,215,227]. However, the bulk of research on this topic has not directly - and empirically - addressed security and privacy [34,119,122].

In this work, we conduct a census-representative survey of 526 U.S. residents, analysis of which supports statistically generalizable conclusions regarding user learning, beliefs, and behavior [25]. This survey queried respondents' behaviors, advice sources, reasoning, and beliefs across four digital-security domains identified as highly important by experts [120]: password strength, antivirus use, software updating, and 2FA. To enable comparisons between digital security and the more well-developed domain of physical security, we asked similar questions about the securing of exterior doors in respondents' homes.

[1] Published as [179].

4.1 Methods

We conducted a computer-administered, closed-answer survey of a census-balanced sample of 526 respondents in April 2016 via the Survey Sampling International panel. To ensure our survey instrument could produce generalizable and rigorous results, we pre-tested our questionnaire by conducting cognitive interviews and expert reviews. These methods are best practices in survey methodology for minimizing biases and improving validity in survey data collection [173].

This study was approved by the University of Maryland Institutional Review Board. Below, we discuss our survey development process, sampling procedure, details of our statistical analysis, and limitations of our work.

4.1.1 Survey Development

Our survey queries respondents': digital and physical security advice sources, reasoning for accepting or rejecting advice from these sources, beliefs about the purpose and value of different digital-security behaviors, and general opinions regarding the importance and utility of digital- and physical-security advice. In addition to asking standardized demographic questions regarding respondents' age, race, gender, education level, and income, we also asked whether respondents had ever held a government security clearance, and if not, whether they currently work with data governed by HIPAA (U.S. health-privacy regulation) or FERPA (U.S. student-privacy regulation). We refer to these participants collectively as *sensitive-data participants*. Further, we asked whether respondents held a degree in or worked

Gender	**Age**	**Race**	**Educ.**	**Income**
M	40-49 yrs	Asian	B.S.	$100-$125k
M	18-29 yrs	Hispanic	M.S.	$30-$50k
M	30-39 yrs	Black	Some College	$30-$50k
F	50-59 yrs	Black	High School	<$30k
F	40-49 yrs	White	B.S.	$50-$75k

Table 4.1: Cognitive Interview Demographics.

in the fields of computer science, computer engineering or IT. We also administered the six-item web-use skills index to assess the respondent's technical skill level [105]. We used this to explore how exposure to a security-sensitive mindset, educational background or work experience on computer science or IT, and technical skill, respectively, influence users' learning mechanisms and security behaviors.

After developing the initial set of questions, we conducted cognitive interviews with five demographically-diverse participants (see Table 4.1). Cognitive interviewing is a method of pre-testing questionnaires that provides insight into how respondents interpret and answer questions, so that errors can be corrected before deployment [173, 233].

During the interview, participants were instructed to "think aloud" as they answered each interview question via the Qualtrics survey interface. After answering each survey item, they were asked one of the following questions: "Was that question hard to answer?"; "How did answering that question make you feel?"; "Was there an answer choice missing or one that you would have preferred?" Participants frequently volunteered information about how they felt or missing answer choices, even when unprompted. The results of these interviews were used to iteratively revise and re-write certain questions until they were clearly understood by respondents. No participants reported finding the questionnaire stressful nor the questions uncomfortable [61]. The cognitive interviews were 20 to 30 minutes in length.

After the third cognitive interview was complete, three experts reviewed our survey instrument to evaluate question wording, ordering, and bias: our university's statistical and survey methodology consultant and two human-computer-interaction researchers with survey expertise. Expert reviewing is another best practice typically used to identify sensitive questions, questions that may need additional clarification, and problems with question ordering or potential biases [173]. We updated a number of our questions following the expert review, and then completed cognitive interviews until no additional questionnaire problems emerged.

The final survey was administered via the Qualtrics web interface. Each question was required, and a "Prefer not to answer" choice was offered for any questions identified as sensitive by the researchers or the expert reviewers. Additionally, sensitive questions and questions that may have had social-desirability bias (in which the respondent may feel socio-cultural pressure to respond in an "acceptable" manner) were rewritten to reassure respondents that all answers were acceptable, according to best practices [224]. For example, a question asking whether or not respondents used antivirus software was phrased as follows: "There are different reasons that people decide to use or not to use antivirus software on their personal devices. Which of the following best describes you:...". A list of the final survey questions are provided as supplementary material.

The order in which the questions about each of the four digital security behaviors were presented was randomized to prevent order bias [133]. The physical-behavior question was not included in the randomization, as the results of the cognitive interviews showed that respondents needed to be prepared for the topic switch to physical security and found it cognitively challenging to switch between topics. In order to improve the quality of the data collected, a commonly-used attention check question was included: “Please select Unhappy as your answer choice to the following question. This question is designed to check that you are paying attention” [214]. Finally, demographic questions were placed at the end of the questionnaire to minimize sensitivity and bias, as per expert recommendations and best practices [201]. The full questionnaire is available in Appendix A.2.

4.1.2 Sampling

We wanted to examine a representative sample of Internet-using adults in the United States. To this end, we followed the American Association for Public Opinion Research guidelines and used sample quotas to obtain obtain a census-balanced sample of US adults for our survey [25]. We contracted Survey Sampling International (SSI) to recruit respondents who matched the US-census sample quotas on the metrics of age, race, income and gender. SSI administered our Qualtrics survey to these respondents through their platform, and compensated respondents according to their agreement with SSI. Respondents were provided with benefits such as gift cards, airline frequent flyer miles, and donations to charities of their choice.

By using this large and representative sample, we can make statistically significant and broadly generalizable conclusions about user behavior, beliefs, and practices [25]. In comparison, the majority of work on user security behavior and learning is drawn from convenience samples on platforms such as Amazon Mechanical Turk and also from small qualitative lab studies [69, 120]. Prior work, including my own, comparing Mechanical Turk samples with the general population has shown these samples to have important limitations when generalizing to the Internet-using population both with regard to demographics and security and privacy attitudes and experiences [121, 123, 184, 196]. As such, our work provides a more robust picture of user behavior at a national scale than has prior work.

Our sample is nearly representative of the demographics of the United States with regard to age, education, gender, and race. Our sample is very slightly wealthier than the general population, potentially due to lack of Internet or device access among those earning less than $30,000. Additionally, we had a 5% higher incidence of Caucasian respondents and a 5% lower incidence of Hispanic participants than in the general population. This may be due to the fact that we used a single "select all that apply" race and ethnicity question which offered both Hispanic (ethnicity) and White (race) as answer choices to the same question. Table 4.2 compares the demographics of our sample to the 2014 United States Census [11].

To assess the knowledge and skills of our respondents, we administered the extensively validated six-item web-use skills index, which measures Internet skills on a scale from 1 (low) to 5 (high) [105]. The mean for our participants is 3.75 (SD=0.99). This is slightly higher than the mean of 3.37 that would be anticipated from Hargittai and Hsieh's work developing this scale. However, our result still seems

Metric	**Sample**	**Census**
Male	49%	49%
Female	50%	51%
Caucasian	69%	64%
Hispanic	11%	16%
African American	12%	12%
Other	8%	8%
Some HS	3%	8%
Completed HS	23%	28%
Completed Some College	25%	18%
Associates Degree	10%	9%
College Degree	26%	26%
Master's	10%	7%
Doctoral	4%	4%
18-29 years	22%	23%
30-39 years	20%	17%
40-49 years	19%	17%
50-59 years	16%	18%
60-69 years	15%	14%
70+ years	8%	11%
< \$30k	26%	32%
\$30k-\$50k	19%	19%
\$50k-\$75k	17%	18%
\$75k-\$100k	13%	11%
\$100k-\$150k	14%	12%
\$150k+	9%	8%

Table 4.2: Demographics of participants in our sample. Some percentages may not add to 100% due to item non-response. Census statistics from the American Community Survey [11].

reasonable, as Hargittai and Hsieh collected their data in 2010 and the Internet skill level of the population has almost certainly increased in the past six years. Additionally, thirty percent of our respondents were 'sensitive-data' respondents and 19% of our respondents held a degree in or worked in the fields of computer science (CS), computer engineering, or IT.

4.1.3 Statistical Analysis

In addition to presenting descriptive statistics regarding the prevalence of respondents' behaviors, advice sources, reasoning, and beliefs, we compare Likert-scale factors between participant sub-groups (e.g. beliefs between sensitive-data and general participants) using the Mann-Whitney U test [145]. We also construct several logistic-regression models in an effort to describe the relationship between respondent security behavior and informal learning. Logistic regression is a well-known statistical method for modeling binary outcomes [118]. In order to avoid over-fitting these models, we used the standard backward-elimination technique, removing one factor from the model at a time, until we minimize the Akaike Information Criterion (AIC) [20,237]. For each model, we present the outcome variable, included and eliminated factors, log-adjusted regression coefficients (*odds ratios*), 95% confidence intervals, and p-values.

We use Pearson's X^2 test to assess independence among categorical variables, such as between respondents' security behavior and their sources of computer security information [86]. For comparisons across many categories, we use omnibus tests; if the omnibus tests are significant, we then apply pairwise tests selected a priori to compare individual categories. One limitation of this method is that our data contains repeated measurements of the same respondent, as every respondent answered questions about multiple advice sources and security behaviors. Pearson's X^2 test does not take into account this repeated measurement, meaning it is possible reported test statistics are overstated; nonetheless we believe X^2 is the most appropriate test for these analyses. We interpret the results with this limitation in mind.

4.1.4 Limitations

Our representative sample provides for robust, broadly generalizable results. It is currently not possible to obtain a purely probabilistic sample via an Internet survey [25], as such we cannot precisely state the prevalence of user behaviors and advice sources in the entire U.S. population. Nonetheless, our work provides a strong foundation for understanding national behaviors and trends.

As with any survey, some respondents may have selected the first answer that seemed to satisfactorily answer the question, without thinking deeply about their own beliefs [132]. To mitigate this, we included an attention check question to screen out inattentive participants and kept our questionnaire to 10 minutes in length, following expert recommendations for minimizing respondent fatigue.

It is also possible that respondents mis-reported answers in an effort to answer in a socially desirable manner. However, we focus primarily on asking respondents to recall their advice sources, about which significant social desirability bias seems unlikely; additionally, our questionnaire-testing procedure revealed no evidence of social desirability bias. Further, while we asked participants to search their memory for answers to our questions, they may not have fully done so, or they may have forgotten some information. We also assume that participants are largely able to correctly identify which of their behaviors are security behaviors and why they practiced those behaviors. Finally, it is possible that our survey questions do not accurately assess the constructs we sought to measure. To mitigate these errors, we extensively pre-tested the questionnaire. While we made every effort to eliminate errors and biases, as with any survey, there may have still been lingering measurement errors.

4.2 Results

Overall, we find that 53% of respondents report making stronger passwords for some accounts than for others and 84% of respondents report using antivirus software. Although there has been no prior work requesting users to self-report their antivirus use, our findings agree with prior industry work by McAfee, which analyzed log data to determine that 88% of computers had antivirus software installed [211].

The majority of our respondents updated their software: 37% reported updating all of their software immediately and 41% reported updating their software a little while after learning of updates. Only 5% of respondents reported rarely or never updating their software, while 17% of respondents reported updating some but not all software. This self-report data contradicts the findings of Nappa et al., who used Symantec logs to find that at most 14% of vulnerabilities were repaired when an exploit was released [159]. Furthermore, our findings are also higher than those reported by Ion et al. who reported that 25% of experts and 9% of non-experts in their survey reported installing updates “immediately” [120]. These discrepancies could have several possible explanations: different sample (Nappa et al. measured machines rather than people; Ion et al. used Amazon Mechanical Turk while we used a census-representative sample), social desirability (although we believe our wording was more neutral), an increase in user updating behavior since the Ion et al. survey in 2014, differences in how different respondents interpret "immediately" updating their software, and/or the explicit "Updates are installed automatically" option in used in Ion et al's survey but not ours.

Finally, of our respondents, 25% used 2FA on all of the devices or services that offered it; 45% used 2FA on some, but not all services; and 28% never used 2FA (2% NR). The proportion of our respondents that use 2FA is higher than the rates cited in prior work by Ion et al. [120]. This may reflect an increase in 2FA adoption since the Ion et al. survey was conducted in 2014; we also defined 2FA for all respondents, which may have prevented some under-reporting from participants who did not recognize the term. We asked the 236 (45% of total) respondents in our sample who used 2FA on some but not all services why they used 2FA for those services. The majority (62%) said they used 2FA on only some services because they were required to do so by those platforms. Twenty-eight percent of these 236 participants said that they used 2FA for the services that were more important to them. Very few participants (8%) said that they activated 2FA on only some services because it was easier to do so on those particular services. Of those who did not use 2FA for any services (149, 28% of total), 64% had never seen information about nor had been prompted to use this security strategy.

4.2.1 Beliefs about Behavior Purpose

To assess respondents' beliefs about the purpose of security behaviors, we asked respondents what they thought the "primary reason" was for updating software and for using 2FA. We did not ask these questions for passwords and antivirus use, as cognitive interview results showed that these behaviors had a single, intuitive answer, and asking a question deemed "obvious" by all respondents caused negative survey sentiment.

Security was not believed to be the primary driver for completing updates. The highest proportion of respondents (40%) believed that the primary purpose was to "ensure the software is free of bugs and crashes less often." Twenty-nine percent of respondents selected "to increase the security of the software software" as the primary purpose, and 30% believed the purpose was to get the "latest and greatest software features" (1% NR). However, for 2FA, security was cited as the primary purpose by the majority, 67% of respondents. The remaining 21% of respondents believed that 2FA was used to ensure that they could regain access to their account, and 10% believed 2FA was to enable the website to contact them (2% NR).

4.2.2 Beliefs about Security Importance

In addition to asking respondents where they learned digital (and physical) security behaviors, we asked them to compare the usefulness and trustworthiness of digital and physical security advice using two 5-point Likert scales anchored on "Digital is a lot more useful (trustworthy) than physical" and "Physical is a lot more useful (trustworthy) than digital". Our prior qualitative work suggests that respondents who handle sensitive data value digital-security advice more highly [186]. In our sample, we found that sensitive-data respondents were significantly more likely to find digital-security advice more trustworthy than their non-sensitive peers (MWU *p<0.01*), while they were not quite significantly more likely to find digital-security advice more useful (MWU *p=0.08*) . We hypothesize receiving digital-security information and emphasis in the workplace leads to higher levels of trust of security information in general, and that implementing security practices in their routine workplace activities leads to sensitive-data respondents regarding digital-security advice as more useful.

4.2.3 How Users Learn Security Behaviors

For each behavior that a respondent reported completing, we asked how they learned the behavior or what instigated them to do the behavior. See Figure 4.1 for a summary of respondents' advice sources. Note that participants were allowed to select multiple options; as a result, percentages may add to more than 100%.

The majority of respondents (80%) cited device- or software-based prompts or requirements as a reason for doing at least one digital-security behavior. These prompts included password meters, update reminders, or invitations to use 2FA. Further, 53% of respondents cited being required to use a behavior or automatic behaviors, such as automatic software updates, as a reason for using a behavior.

Media and family/friends were the most prevalent sources of digital-security advice. This finding confirms the results of prior, less representative studies [88,186]. For the majority of respondents, online, print, or TV news articles were at least one of the types of media advice they saw (67.5%). Online forums served as a digital-security advice source for 40% of respondents, and fictional narratives and advertisements accounted for 25% of media advice, each. Additionally, the majority (60%) of advice from a family members and friend was given by a person with a background in CS or IT. This confirms prior qualitative work, indicating that respondents may feel that these individuals are experts or individuals with these backgrounds may volunteer more unsolicited advice [172].

As also noted in prior qualitative work, negative experiences appear to be a key source of digital-security advice [176, 186]. Our work confirms this finding, although at a lower level of prevalence than may have been expected: 28% of our respondents learned to practice a behavior due to a negative experience or a story told about a negative experience.

We were surprised to find that digital-service providers, such as TimeWarner or Bank of America, were a source of advice for 33% of respondents, as this source of advice is little discussed in prior research. Respondents also reported receiving a significant amount of advice from their workplace (29.5%). Of those who received advice from work, over 50% received advice from an IT newsletter or a friend or colleague who worked in the IT department. The remainder received advice from formal security training or from colleagues who did not have an IT background.

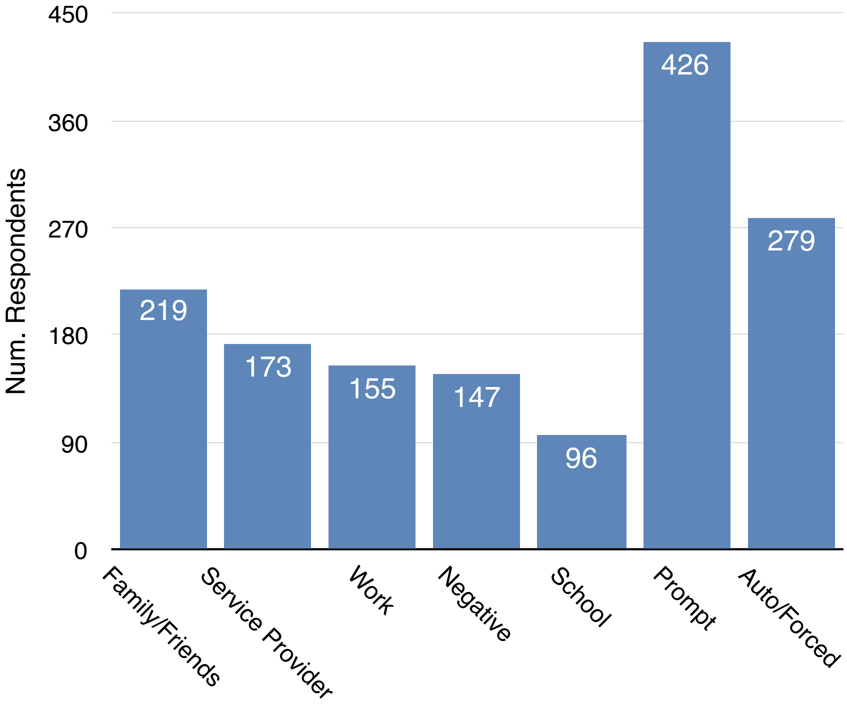


Figure 4.1: Prevalence of security education channels.

4.2.4 A Digital Divide in Who Takes Which Advice

How do these beliefs about the usefulness and trustworthiness of digital-security advice, as well as demographic and knowledge factors, impact from where users take advice? Here, we present the results of binary logistic regression models for each source from which respondents reported learning behaviors. These models provide insight into the audience of each advice source. Whether or not a respondent reported a given advice source at least once was used as the outcome variable in our models. The input factors considered in each model are listed in Table 4.3. Included in these factors are two interaction factors—between sensitive-data and beliefs about the usefulness and trustworthiness, respectively, of digital- vs. physical-security advice—which were informed by our prior qualitative work [186]. Below, we describe and interpret the significant factors included in each model. Because we did not conduct a controlled experiment, these results do not imply causality. The final regression results for each model after backward elimination, including nonsignificant factors, are shown in Table 4.4.

Factor	**Description**	**Baseline**
Gender	Male, female or other.	Female
Age	18-39 years, 40-59 years, and over 60 years.	18-39 yrs
Income	<$50,000, $50,000-$100,000, and >$100,000	<$50,000
Race	Black, Hispanic, White, and Other.	White
Education	Less than Bachelor's degree, Bachelor's degree, Graduate degree.	<B.S.
CS Background	Whether or not the respondent reported working in or holding a degree in CS or IT.	N/A
Sensitive Data	Whether or not the respondent reported working with HIPAA, FERPA, social security/credit card data, or holding an active or prior clearance.	None
Internet Skill	Level of Internet skill as measured by the six-item general web-use skills index [105].	N/A
Belief: Useful	Response to whether digital-security advice was **a lot** more useful than physical-security advice, somewhat more useful, equal, or that physical-security advice was somewhat or a lot more useful than digital. On a five-point Likert scale.	N/A
Belief: Trust	Response to whether digital-security advice was **a lot** more trustworthy than physical-security advice, somewhat more trustworthy, equal, or that physical-security advice was somewhat or a lot more trustworthy than digital. On a five-point Likert scale.	N/A
Sens. Data & Useful	Interaction between the Sensitive Data and Belief: Useful factors described above.	N/A
Sens. Data & Trust	Interaction between the Sensitive Data and Belief: Trust factors described above.	N/A

Table 4.3: Factors used in regression models. Categorical factors are represented by binary variable sets and individually compared to the baseline; a numerical value, centered on the middle value, was used for Likert factors.

As described in Methods, we used backward elimination, minimizing AIC, to reach our final model. The final model for media advice included Internet skill, exposure to sensitive data, age, income, and belief about the usefulness of digital security as factors. We find that respondents with higher Internet skill are 32% more likely to use media as an advice source, potentially because media is increasingly being distributed online rather than through print, TV or radio.

The final model for work advice included Internet skill, exposure to sensitive data, age, income, education, and belief factors. Those who work with sensitive data are 4.5× more likely to cite their workplace as an advice source than those

Source	**Factor**	**OR**	**CI**	***p*-value**
Media	Internet Skill	1.32	[1.09, 1.6]	< 0.01*
	40-59yrs	0.59	[0.39, 0.91]	0.02*
	Over 60yrs	0.87	[0.54, 1.41]	0.58
	$50k - $100k	0.85	[0.55, 1.31]	0.45
	> $100k	1.82	[1.12, 2.97]	0.02*
	Sensitive Data	1.50	[0.98, 2.29]	0.06
	Belief: Useful	1.09	[0.86, 1.38]	0.47
	Sens. Data & Useful	1.31	[0.94, 1.83]	0.11
Work	$50k - $100k	1.93	[1.16, 3.21]	0.01*
	> $100k	2.91	[1.63, 5.18]	< 0.01*
	Sensitive Data	4.53	[2.65, 7.74]	< 0.01*
	Internet Skill	1.41	[1.11, 1.8]	< 0.01*
	Sens. Data & Useful	1.58	[1.04, 2.41]	0.03*
	Belief: Useful	0.69	[0.48, 0.99]	0.04*
	40-59yrs	0.86	[0.53, 1.39]	0.54
	Over 60yrs	0.53	[0.29, 0.97]	0.04*
	Bachelors degree	1.53	[0.92, 2.53]	0.1
	Graduate degree	1.83	[0.97, 3.47]	0.06
Negative Experience	Internet Skill	1.31	[1.06, 1.63]	0.01*
	Sensitive Data	1.50	[1, 2.23]	0.05*
	40-59yrs	0.55	[0.35, 0.87]	0.01*
	Over 60yrs	0.61	[0.36, 1.04]	0.07
School	CS Background	6.22	[3.46, 11.17]	< 0.01*
	40-59yrs	0.26	[0.14, 0.47]	< 0.01*
	Over 60yrs	0.11	[0.04, 0.27]	< 0.01*
	$50k - $100k	0.57	[0.3, 1.09]	0.09
	> $100k	1.99	[1.06, 3.74]	0.03*
	Belief: Useful	1.14	[0.8, 1.64]	0.46
	Sensitive Data	0.99	[0.55, 1.78]	0.97
	Belief: Trust	1.23	[0.95, 1.6]	0.12
	Sens. Data & Useful	0.71	[0.46, 1.11]	0.13
Prompt	CS Background	0.20	[0.12, 0.35]	< 0.01*
	Belief: Useful	0.80	[0.65, 0.97]	0.02*
	$50k - $100k	0.48	[0.28, 0.83]	< 0.01*
	> $100k	0.64	[0.35, 1.19]	0.16
	Internet Skill	1.26	[0.99, 1.6]	0.07
Automatic	Belief: Useful	0.79	[0.67, 0.92]	< 0.01*
	CS Background	0.59	[0.37, 0.94]	0.03*
	Bachelors degree	0.60	[0.39, 0.92]	0.02*
	Graduate degree	0.68	[0.39, 1.18]	0.17
	40-59yrs	1.00	[0.66, 1.5]	1
	Over 60yrs	1.74	[1.08, 2.79]	0.02*
Family & Friends	40-59yrs	0.40	[0.26, 0.61]	< 0.01*
	Over 60yrs	0.51	[0.32, 0.83]	< 0.01*
	Black	0.66	[0.37, 1.21]	0.18
	Hispanic	0.47	[0.25, 0.89]	0.02*
	Other	1.56	[0.81, 2.99]	0.18
	CS Background	1.37	[0.85, 2.2]	0.2
	Internet Skill	0.87	[0.72, 1.05]	0.15
Service Provider	Sensitive Data	1.81	[1.21, 2.7]	< 0.01*
	Male	1.57	[1.07, 2.31]	0.02*
	Internet Skill	1.24	[1.01, 1.53]	0.04*
	40-59yrs	1.40	[0.9, 2.19]	0.14
	Over 60yrs	2.10	[1.27, 3.48]	< 0.01*
	Belief: Useful	1.14	[0.96, 1.35]	0.13
	CS Background	0.65	[0.39, 1.09]	0.11

Table 4.4: Regression results for advice source models. OR is the odds ratio between the given factor and the baseline; CI is the 95% confidence interval; statistically significant factors (p¡0.05) are denoted with *.

who are not exposed to sensitive data. This confirms results of prior work: those who have clearances and/or handle sensitive data may receive benefits that improve their security through workplace training [186]. Those who had higher Internet skill were 41% more likely to cite the workplace as a source of advice. Further, those who cited their workplace as a source of advice were more likely to believe that digital security was more useful than physical security. Thus, there is a need for increased digital equity and improved security interventions for users who do not have the opportunity to receive workplace training.

The final model to describe those who cited negative experiences, or stories of these experiences included Internet skill and sensitive data. Those respondents with higher Internet skill levels were 31% more likely to cite a negative experience as an advice source and those who handled sensitive data were 50% more likely. Those with higher Internet skill or those who handle sensitive data may be more likely spend more time online, and thus may be more likely to have and learn from a personal negative experience. Furthermore, more skilled users, and those who are exposed to sensitive data, may be more likely to recognize a negative experience when it occurs and identify the underlying cause of that experience than less skilled or experienced users.

The final model for advice from school included age, CS background, income, sensitive data, and beliefs. Those with a CS background were approximately 6× more likely than those without this educational background to cite school as an advice source, and those who were over 60 were only 11% as likely as younger respondents to cite school as an advice source. This is likely due to the fact that computers were not used in schools until relatively recently, and those who work in the field of CS or IT and/or hold a degree in this field most likely received digital-security advice as they were obtaining their degrees or training.

Perhaps because they have already learned digital-security behaviors from school, before they ever see a prompt, those with a background in CS were only 20% as likely as those without this background to cite a device prompt as a way that they learned about a particular security behavior. That said, those who used prompts as a source of security advice were more likely to believe that security advice was important—perhaps because this belief encouraged them to heed the security prompts. The final model for device prompts also controlled for Internet skill and income.

The model for device automation—that is, whether respondents reported learning about a security behavior because it was automated or required—included CS background, belief about the usefulness of digital-security advice, age, and education. Those with a background in CS or IT were only 59% as likely as those without this background to learn from security requirements or automations. Similarly to device prompts, this may reflect that respondents with technical education already know about security behaviors before encountering them as requirements. Additionally, those who used a behavior because it was automated or required were more likely to believe that digital security advice was important—a belief in the importance of security may inspire users to utilize prompts and automation.

The final model for advice from family and friends included age, ethnicity, CS background, and the respondent's Internet skill. There appears to be a relationship between age and taking advice from family and friends. Although there is no discernible pattern between the coefficients for ages 40-59 and over 60, there appears to be a significant difference between respondents who are 18-39 years old and those 40 and over. Additionally, respondents who are Hispanic were only 47% as likely to report taking advice from family and friends as White respondents. These findings indicate potential differences in how respondents in different age and ethnic groups chose to solicit advice.

Finally, the model for service provider advice included age, Internet skill, exposure to sensitive data, CS background, gender, and beliefs about the usefulness of digital-security advice. Respondents who are male were 57% more likely to report receiving advice from service providers, than Female respondents. Additionally, respondents with higher Internet skill levels were 24% more likely to report taking advice from a service provider, and those who are exposed to sensitive data were 81% more likely to have taken advice from this source.

4.2.5 Why Users Accept and Reject Advice

In addition to examining the factors that affect which users take advice from which sources, we wanted to understand why the users of each advice source choose to accept and reject advice and behaviors. Prior work has identified a number of different reasons that users accept and reject security advice [186]. In order to evaluate these findings and determine their prevalence, we asked respondents why they chose to practice (or not practice) a behavior based on the advice that they received. The answer choices that we provided were drawn from our prior qualitative work and feedback gained during our cognitive interview sessions [186]. We detail the results below.

To assess respondents' reasons for accepting advice, we asked, "Which of the following best describes why the information you received made you decide to do this behavior?" We then presented four answer choices. The first two "I trusted the person or source of the information." and "The information made sense to me." were drawn from our prior work [186]. The other two choices, "The information increased my fear of a negative event." and "Other." (with a write in option), were added based on the cognitive interviews and expert reviews to ensure that we captured all possible respondent answers.

From our prior work, we hypothesized that respondents would be more likely to accept physical-security advice based on their evaluation of the advice content, while they would accept digital-security advice based on their trust of the source. As shown in Figure 4.2, we found that trusting the source was most popular for antivirus (53%) and updating (51%), while trusting the content was most popular for 2FA (52%), passwords (57%), and door locking (58%). This appears to confirm our hypothesis for some digital-security behaviors, but not others.

To investigate further, we ran an omnibus X^2 across all advice sources and behaviors (X^2 = 58.96, p = 4.79e−12), followed by planned pairwise comparisons of each digital behavior to door locking. Our results strongly support that most digital behaviors are different from the physical behavior, especially antivirus (X^2 = 41.15, p = 1.41e−10) and updating (X^2 = 25.49, p = 4.45e−7). 2FA was also significantly different from physical, but to a lesser degree ($X^2 = 6.78$, $p = 0.0083$). We hypothesize that passwords are not significantly different from physical ($p > 0.05$) because participants have been exposed to enough passwords advice to feel comfortable evaluating its content directly.

Finally, we validate that the two choices presented in our prior qualitative work [186] are near-exhaustive of the reasons that users accept advice: for each of the behavior questions, only a small portion of respondents (μ= 5%) reported that increased fear of a negative event caused them to take advice; an average of 2% of our respondents selected "Other."

Similar to advice acceptance, we drew the answer choices for our question regarding why respondents chose *not* to practice a behavior, even after seeing information recommending that behavior, from multiple prior studies [176, 186, 231, 232] and from the results of our survey pre-testing. We only asked these questions regarding antivirus, updating, and 2FA; as it is rarely, if ever, an option to not use passwords. Nearly half of our respondents (43%, 225) rejected at least one of these three behaviors. Among these respondents, we found that inconvenience (28%) and advice that contained too much marketing material (17%) were the two most common reasons for advice rejection, across all behaviors. We also found that a *lack* of negative experience was the third most common reason (13%) for rejecting a behavior. Although believing that one's data has no value [119], difficulty understanding advice [19], and being 'careful' on the Internet [231, 232] have been offered as reasons for rejection in prior work, these reasons were all cited by less than 10% of our respondents.

The reasons respondents selected for rejecting advice varied by behavior. For antivirus software, “They were trying to sell me something” was most often cited as a reason for rejecting advice related to antivirus software (33%). Other reasons for rejection included having had no prior negative experience (13%), feeling that they were “careful” when using their computer and the Internet (15%), and finding antivirus software too difficult to use (11%). Inconvenience, which included both “it was inconvenient” and “I did not have time”, was the most common reason (50%) selected by respondents for why they did not complete software updates. All other reasons provided were cited by fewer than 10% of respondents and are thus not reported here. Inconvenience was also the most common reason given by respondents for not using 2FA (41%). Our prior work also suggested that privacy concerns may inhibit advice taking, especially for 2FA, where users may be reluctant to share a phone number. We found that while privacy concerns were not very prevalent in general, they were somewhat more prevalent for 2FA (15%) [186]. See Figure 4.3 for more detail on respondents’ reasoning for rejecting security advice.

4.2.6 Advice Sources and Behavior

We next examine which advice sources were most commonly associated with which security behaviors (Figure 4.4). We find that media was the primary source of advice for both passwords (28%) and 2FA (26%), while family and friends accounted for a larger portion of the advice about antivirus behavior (28%) and software updates (24%). Service providers (21%) were also a primary advice source for 2FA.

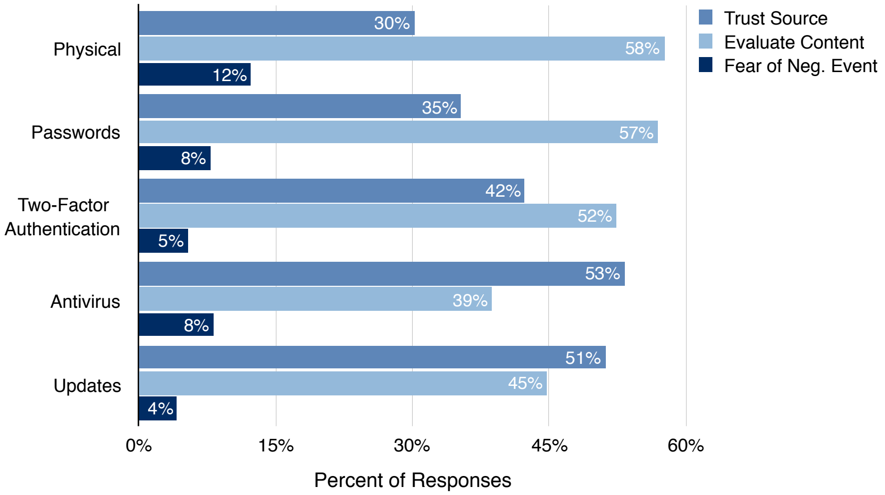


Figure 4.2: Reasons for accepting digital-security advice. Percentage per behavior.

Finally, learning a behavior via a negative experience was most common for antivirus use (19%). An omnibus X^2 test showed that these differences among behaviors are significant ($X^2 = 59.05$, $p = 3.67e{-7}$).

4.3 Discussion

The primary finding of this chapter is shedding light on digital inequity in security advice sources. Users with lower socioeconomic status tend to be part of a *knowledge gap*: they have diminished access to digital media and more difficulty finding reputable and useful information on the web [103, 104, 192, 215, 227]. Our work expands these findings to digital security: we find evidence that users with higher levels of Internet skills—demonstrated by prior work to be wealthier and somewhat more secure [105, 126]—use different advice sources. In particular, lower-skill users rely more on prompts, the advice of family and friends, and service providers than higher-skill users do. These differences, combined with discrepancies

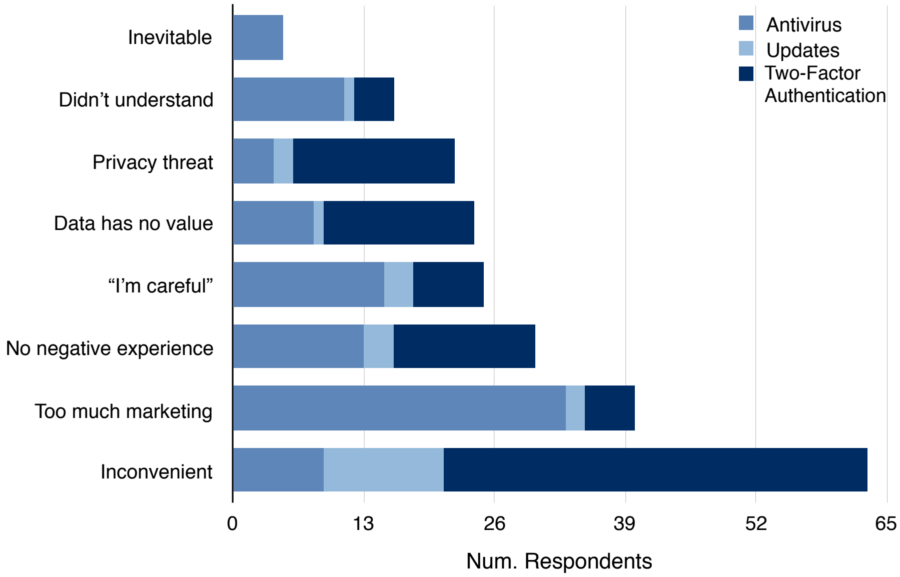


Figure 4.3: Reasons for rejecting digital-security advice. Total per behavior, multiple responses possible. This question was not asked for passwords, as not using them is rarely, if ever, an option.

in skills and resources, may lead already disadvantaged users to be disproportionately victimized [34, 122]. Indeed, while we could not control for all confounding factors, we found that users with lower incomes were less likely to update their software (X^2=28.03, $p \leq 0.001$), to use 2FA (X^2=15.60, $p = 0.004$), and slightly less likely to use stronger passwords for sensitive accounts (X^2=9.60, $p = 0.048$). Although prior work has suggested that differences in security behaviors may be caused by lower-SES users not highly valuing their data [119], in our sample this was not the case.

Our respondents were 41% more likely to cite their workplace as an advice source if they had higher Internet skill, and 4.5× more likely if they held a job that we categorized as security-sensitive. While the workplace may be a valuable source of advice for those who have access to it and the skills to understand this training,

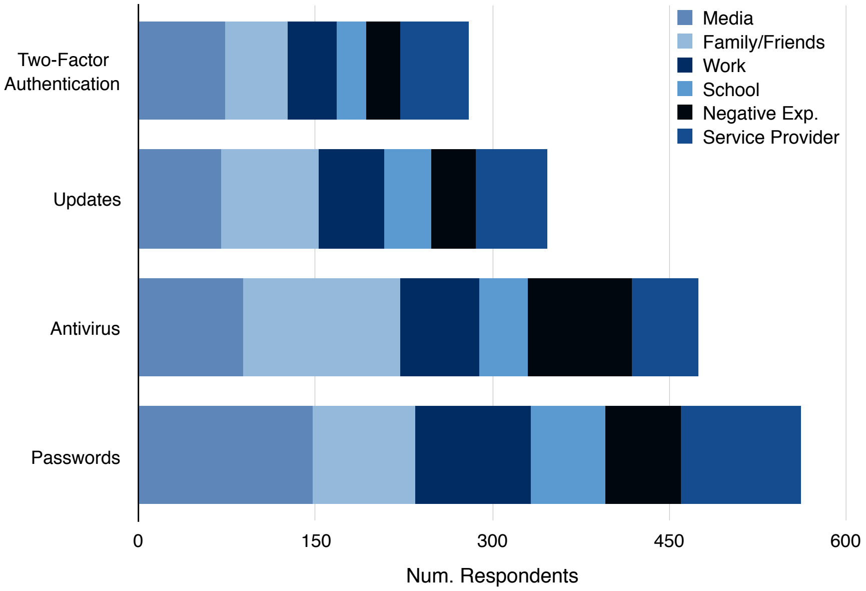


Figure 4.4: Education method by behavior.

such resources may not be available for low-SES users; furthermore, security is often forgotten in digital literacy interventions that do exist in the workplace. For example, the Kesla+ project aimed at increasing the digital skills of low-skill office workers in the workplace included no training on digital security [127]. Thus, we advocate piloting and evaluating digital literacy programs which include or focus entirely on digital security. Additionally, future work should include analyzing and improving the grade-level readability and clarity of security advice to avoid widening the digital security gap.

We also found that participants with higher levels of Internet skill were more likely to have learned from a negative experience, either their own or someone else's. We hypothesize that lower-skill users are less likely to recognize the causes of a negative experience and therefore learn from it. Because our results indicate that users who have not learned from negative experiences are more likely to reject advice, this inequity may put lower-skill users at additional risk. Of course, we want to minimize all users' direct exposure to negative experiences; instead we recommend amplifying stories of others' negative experiences. Future work could examine how to effectively simulate negative experiences, for example by using short, relatable stories that clearly demonstrate how to prevent the problem.

Chapter 5: How Security Education Sources Impact Security & Privacy Outcomes

Now that we have identified a digital divide in security education, this begs the question: does this inequity in advice sources lead to a divide in security outcomes? In this chapter[1], I explore this question: do sources of security education—explored in the last two chapters—actually relate to users' security outcomes? To do so, I draw on a large probabilistic, random digital dial survey of people in the United States.[2] The thoroughly pretested survey queried respondents' security and privacy experiences, including becoming the victim of a scam, having your identity stolen, having an email or social media account compromised, losing a job or other opportunity as a result of something posted online, and having someone post something about you online without consent; other questions examined respondents' advice sources, available Internet resources, and demographics. The relationships identified in this analysis, and the prevalence of the experiences reported by respondents, are accurate within 2.7% of their true values in the entire U.S. population.

A commonly raised question regarding survey data is whether people are able to accurately report their experiences around digital security. To address this con-

[1] Published as [183].

[2] This dataset was awarded to Elissa Redmiles through a Data Access Grant from Data&Society.

cern, I include in Chapter 7 a comparison of the validity of survey data to log data in a security context. This work provides support for the validity of self report data for asking particular security questions, including broad questions about e.g., negative experiences such as those used in the work described in this chapter.

5.1 Methods

In this work, we modeled the results of a 3,000-respondent telephone survey using binary logistic regression. Our Institutional Review Board (IRB) determined that our analysis of existing data did not constitute human subjects research. Below, we discuss the dataset and survey development process, sampling procedure, details of our statistical analysis, and limitations of our work.

Dataset. The dataset was collected by Princeton Survey Research Associates International (PSRAI) for Data&Society via a computer-assisted-telephone-interview (CATI), random digit dial (RDD) census-representative survey of 3,000 respondents from November 18 to December 23, 2015. We received this dataset through a Data Grant from Data&Society (funded by the Digital Trust Foundation).[3]

The survey was developed by a senior researcher at Data&Society with the intent of releasing the data for analysis regarding the impact of SES on security and privacy. She assembled the survey both by authoring and pre-testing new items and by leveraging a number of pre-tested questions from surveys conducted by Pew

[3] The survey development and deployment portion of this study was approved by Chesapeake IRB [14].

and Reason-Rupe [12, 15–17]. The survey asks questions regarding respondents' security and privacy experiences including their advice sources, their prior negative experiences, and the resources available to them, as well as standard demographic questions. The order in which the questions were asked was randomized to prevent order bias [133]. Additionally, demographic questions were placed at the end of the questionnaire to minimize sensitivity and bias, as per expert recommendations and best practices [201].

Prior to deployment, the questionnaire was pretested with a small number of respondents. These interviews were monitored by PSRAI and conducted by experienced interviewers to ensure that respondents understood the questions.

The survey was administered via CATI by professionally trained interviewers in both English and Spanish. Calls were made throughout the day, on multiple days to both landline and cell phones to maximize the chance of connecting with different respondents. Every person in the United States had a non-zero chance of being selected for the survey.[4] This was a probabilistic survey, the dataset was weighted to be representative of the U.S. population, and the findings we report are accurate within 2.7% of the true prevalence in the population. A full outline of the survey items, weighting methodology, and analysis code can be found **[[at** `go.umd.edu/2124`**.]]**

Our unweighted sample was nearly representative of the U.S. population with respect to gender, age, education, geographic region, number of adults in the house-

[4] Those who did not have a telephone were contacted via mail and, if interested, were provided with a phone to use for the survey.

hold, population density, household phone usage, and race/ethnicity. The weighted sample is fully representative of the population, such that the 95% confidence interval for this survey is 2.7 points. This confidence interval is calculated based on the survey design effect, which represents the loss in statistical efficiency that results from a disproportionate sample design and systematic non-response. Table B.17 compares a subset of the demographics of our weighted and unweighted sample to the 2013 American Community Survey [11]. Further, the prevalence of negative experiences in our data is in line with prior work. [5]

Metric	**Unweighted**	**Weighted**	**Census**
Male	52.4%	48.7%	48.2%
Female	47.6%	51.3%	51.8%
Caucasian	58.1%	62.8%	65.8%
Hispanic	18.6%	15.6%	15%
African American	14.0%	11.8%	11.5%
Other	6.7%	7.4%	7.6%
<H.S.	12.8%	12.6%	13.3%
H.S. grad	27.4%	27.8%	28.0%
Some college	24.0%	30.0%	31.0%
B.S. or above	34.6%	28.7%	27.7%
18-29 years	16.3%	20.1%	20.9%
30-49 years	24.6%	32.6%	34.7%
50-64 years	28.8%	25.4%	26.0%
65+ years	27.0%	18.6%	18.4%
<$20k	20%	*NA*	32%
$20k-$40k	21%	*NA*	19%
$40k-$75k	18%	*NA*	18%
$75k-$100k	10%	*NA*	11%
$100k-$150k	8%	*NA*	12%
$150k+	7%	*NA*	8%

Table 5.1: Sample demographics, percentages may not add to 100% due to non-response. Income was the unweighted variable of interest.

[5] See go.umd.edu/2124 for a comparison with Pew 2013 data [9].

Analysis. We built two sets of binary logistic regression models in our analysis, in order to identify independent and covariate relationships between advice sources, SES, and security outcomes. We use the survey R library to incorporate the survey weights [143]. The first set of models was used to predict the odds of an individual reporting having experienced a security or privacy outcome: one model included respondents' advice sources, another included socioeconomic status (SES), and the third combined both. The second set of models predicted the likelihood that respondents with different SES used particular advice sources. We chose a simple grouped model rather than five individual models for ease of interpretability.

To reduce the chance of overfitting our data, we deliberately chose parsimonious models with input factors based on Chapter 3. To further prevent over fitting, we performed 5-fold cross validation in line with commonly used classification and regression practices [108]. We calculated the Akaike Information Criterion (AIC) [20] across five folds for each model, and we found that the AIC values for each fold were within an average of 3% of each other. For each model, we present the outcome variable, including factors, log-adjusted regression coefficients (*odds ratios*), 95% confidence intervals (moderated by the survey design effect [128]), and *p*-values.

Limitations. Self-reported surveys have several common limitations, chiefly related to under- and over-reporting, which may be caused by satisficing (selecting the first satisfactory answer without thinking deeply) [115], recall bias (misremembering experiences), desirability bias (selecting a socially desirable rather than honest answer), and the potential for questions to be misinterpreted. These were mitigated by using thorough question-development and pre-testing processes and by interviewers reminding respondents to answer thoroughly and honestly. The survey was brief, minimizing respondent fatigue.

This survey measures only whether respondents have ever used certain advice sources or had certain negative experiences. As a result, we cannot determine how often a particular advice source was consulted or how many negative experiences a respondent had; nor can we determine whether an advice source was consulted before or after any negative event. Thus, we report our findings together with several hypothetical explanations and suggest that future work should investigate these relationships further. In addition, we did not conduct a controlled experiment, and thus these results should not be interpreted as implying causality.

5.2 Results

All Internet-using respondents were asked questions regarding negative security and privacy incidents that they had experienced, such as "Have you ever had important personal information stolen, such as your Social Security Number, your credit card, or bank account information?" We find that 49% of all respondents in the weighted data reported at least one of the negative experiences shown in Figure 1. To determine how these reported incidents relate to respondents' SES, advice sources, and resources, we utilized binary logistic regression models to predict a participant's likelihood of reporting one or more of these experiences. We created three models, detailed below. The results of all three models are presented in Table 5.2.

Our first model evaluates whether the likelihood of reporting at least one

Model	**Factor**	**OR**	**CI**	***p*-value**
SES & Resources Only	H.S. or Less	0.40	[0.2, 0.81]	< 0.01*
	H.S. to B.S.	0.65	[0.44, 0.97]	0.03*
	$< \$20K$	1.09	[0.69, 1.74]	0.71
	$20-$40K	1.20	[0.79, 1.84]	0.39
	R: Cell only	0.74	[0.52, 1.06]	0.11
	R: Home Internet	1.94	[0.87, 4.32]	0.1
Advice Only	A: Friend	1.85	[1.25, 2.73]	< 0.01*
	A: Website	1.92	[1.15, 3.21]	0.01*
	A: Coworker	1.59	[0.98, 2.58]	0.06
	A: Gov. Website	1.59	[0.87, 2.88]	0.13
	A: Librarian	1.73	[0.75, 4.02]	0.2
	A: Teacher	0.95	[0.45, 2.01]	0.9
Advice & SES	A: Friend	1.84	[1.24, 2.72]	< 0.01*
	A: Website	1.76	[1.06, 2.94]	0.03*
	A: Coworker	1.53	[0.95, 2.46]	0.08
	A: Gov. Website	1.52	[0.85, 2.74]	0.16
	A: Librarian	1.88	[0.82, 4.31]	0.14
	A: Teacher	0.92	[0.44, 1.96]	0.83
	<$20K	1.09	[0.68, 1.76]	0.72
	$20-$40K	1.19	[0.77, 1.83]	0.44
	H.S. or Less	0.53	[0.25, 1.09]	0.09
	S.C.	0.75	[0.5, 1.13]	0.17
	R: Mostly cell	0.77	[0.53, 1.12]	0.17
	R: Home Internet	1.73	[0.77, 3.88]	0.18

Table 5.2: Regression results for three different models of reporting at least one negative experience (binary). ‘A’ and ‘R’ indicated boolean advice sources and resources, respectively. “Mostly cell” indicates primary Internet access via mobile, and “Home Internet” means Internet at home. Baseline for the categorical household income factor is >$40K; baseline for education is a bachelor’s or above. OR is the odds ratio between the given factor and the baseline; CI is the 95% confidence interval.

negative incident is significantly related to advice sources. We find that respondents who take advice from friends and websites are 85% and 92% more likely to report at least one negative experience, respectively. Of those who took advice from friends, 49% reported a negative experience, compared to 25% of those who took advice from a co-worker, 21% from a non-governmental website, 14% from a government website, and 8% from a teacher or librarian. This may indicate that respondents more often

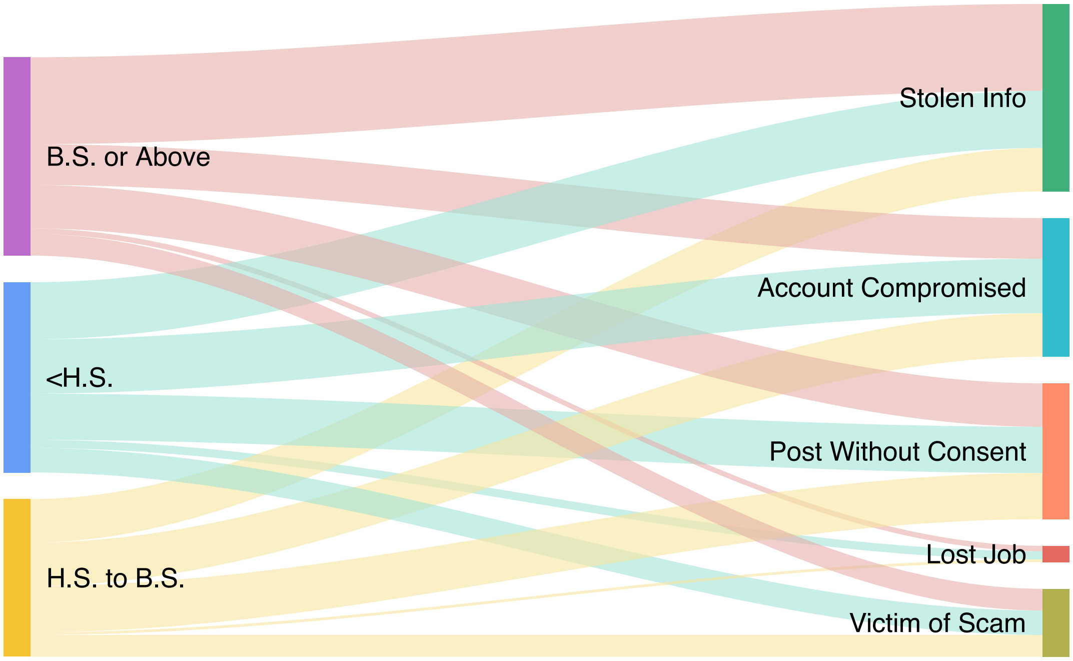


Figure 5.1: Prevalence of negative security outcomes by education level. Interactive diagram: `jsfiddle.net/5orqbkp4/3/`.

seek advice from certain sources after a negative experience, that librarians and teachers give particularly good advice, or that respondents are receiving detrimental or difficult to interpret advice from friends, coworkers, and websites. .

Our second model evaluates whether SES alone relates to respondents' reported security and privacy incidents, we modeled these incidents as a function only of SES factors. In this model, we find that education is the only factor significantly related to a respondent's likelihood of reporting a negative experience (income was not correlated). Surprisingly, we find that those with lower levels of education—high school diploma or less education and less than a bachelor's—are 60% and 35% *less* likely, respectively, to report at least one of the five negative experiences (Table 5.2). While 53% of those in the weighted dataset who hold a bachelor's or above reported a negative experience, only 47% of those who had less than a bachelor's

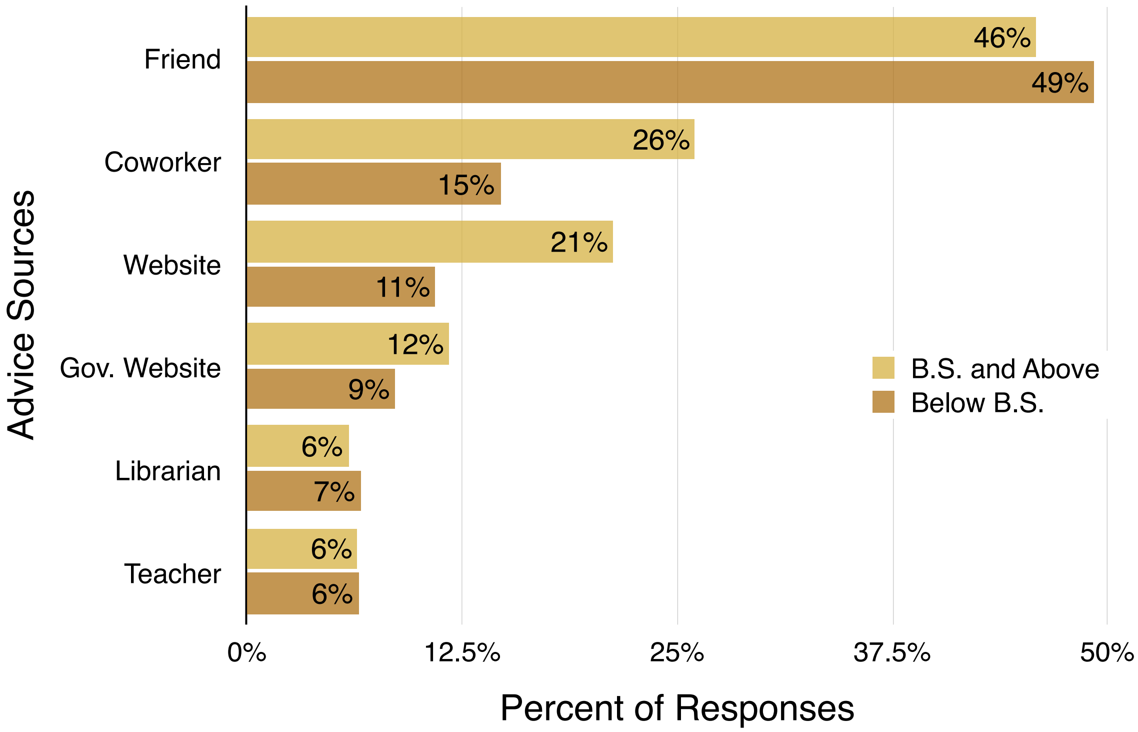


Figure 5.2: Respondents' advice sources by education (weighted).

reported such an incident. Figure 1 illustrates that negative experiences were unevenly distributed across educational groups; 32% of those holding a bachelor's or above reported having information stolen, compared to 20% of those with less than a bachelor's. There are several potential explanations for this finding, which should be explored in future work: less-educated users may be targeted less frequently for scams or identity theft, they may have more difficulty recognizing or recalling negative events, or they may have protective skills or resources not measured in this survey.

Finally, we wanted to understand whether advice and SES were both related to the security and privacy incidents that users report, or whether if we controlled for both variables, only one would remain significant. We therefore constructed a third model containing both advice and SES as explanatory factors. We find that only advice sources are significant factors. Using a likelihood ratio test [204], we find that this combined model has a goodness of fit significantly better than the SES-only model (X^2=45.09, $p < 0.001$, $df = 1164$) and not significantly different from the advice-only model (X^2=7.33 $p = 0.29$, $df = 1164$). This suggests that users' negative experiences relate to their advice sources, regardless of SES.

We do however find a second-level effect of SES. While SES alone does not relate to security incidents, SES does relate to advice sources, which in turn, relate to incidents. Figure 5.2 provides an overview of respondents' reported advice sources organized by education. We find that users who have a high school diploma or less education are 99% less likely to report a coworker as an advice source, and those who hold less than a bachelor's degree, but who completed high school, are 51% less likely. Similarly, those who held a high school diploma were 50% less likely to report coworkers and those with under a high school education were 73% less likely to report using government websites. Perhaps surprisingly, there was no significant difference in the SES or resources of respondents who reported taking advice from librarians, friends, and teachers. Overall, these results confirm our prior findings.

We hypothesize that these findings relate to less-educated users having different job roles, possessing relatively fewer Internet skills [105], and distrusting websites that provide general advice without a clear source [186]. We also hypothesize that advice from websites may be more difficult to read and interpret than advice from other sources. Of note, there was no relationship between available Internet resources and advice sources, implying that accessibility of advice related to devices and Internet access may not be a problem.

5.3 Discussion

This chapter illustrates a clear relationship between respondents' security and privacy experiences and advice outcomes and shows evidence of a second-level digital divide: with SES affecting respondents' advice sources, which in turn relate to outcomes. The precise direction of this relationship between advice sources and security and privacy outcomes is unclear: do people receive bad advice that leads to worse experiences, or do they wait to seek advice until after a negative experience? We hypothesize some of both.

In either case, however, this finding confirms that the current advice ecosystem is not working, and should be reevaluated. In the next chapter, I explore what makes bad advice bad—outdated or incorrect content, poor presentation, a lack of readability, belief in the talisman of useless advice [112], or some combination—and look for ways to remove it or replace it with better advice.

Chapter 6: Quality of the Security Advice Ecosystem

In this chapter we conduct a large-scale measurement of the quality of the security advice ecosystem. We break quality into three components: comprehensibility, accuracy, and actionability. We explore the relationship between these components, topics of advice, and advice-givers – looking for patterns in the creation of good, and bad, advice to better understand the flaws in the existing security education ecosystem and identify remedies.

6.1 Corpus of Security Advice

In this section we describe the collection and annotation of the corpus of security advice we measure and provide a descriptive overview of the corpus.

6.1.1 Collecting the Corpus

We used two approaches to collect text-based security advice: (1) We collected search queries for security advice from crowdworkers and scraped the top 20 articles surfaced by Google in response to their queries, and (2) We collected a list of authoritative security-advice sources from computer security experts and librarians and scraped articles recommended by those resources.

User Search Query Generation. We recruited 50 participants from Amazon Mechanical Turk to write search queries for security advice. To obtain a broad range of queries, we used two different surveys. The first survey asked participants to list three digital security topics they would be interested in learning more about, then write five search queries for each topic. Participants in the second survey were shown the title and top two paragraphs of a security-related news article[1], then asked if they were interested in learning more about digital security topics related to the article. If the participant answered yes, they were prompted to provide three associated search queries. Participants who answered no were asked to read additional articles until they reported interest; if no interest was reported after six articles, the survey ended without creating queries. Twenty-five people participated in each survey and were compensated $0.25 (first survey) or $0.50 (second survey). Participants completed these tasks in four minutes or less, and our protocol was approved by the University of Maryland IRB.

From these surveys, we collected 140 security-advice search queries. After manual cleaning by the researchers (removing duplicates and off-topic queries), 110 queries remained. Examples of these queries include, "how safe is my information online?," "how to block all windows traffic manually?," and "common malware."

We then aggregated the top 20 Google search results for each query using parameterized GET requests, yielding a cumulative URL index that preserved rank order of search results. We then used the Diffbot API [3] to parse and sanitize HTML body elements within each identified site, merging all such elements to create one

[1] See Appendix A.6 for the list of articles shown.

text file per site. Our collection was conducted in September 2017.

In total, the resulting search corpus includes 990 documents. Examples of advice in this corpus include Apple and Facebook help pages, news articles from Guardian, the New York Times, and other media sources, and advice or sales material from McAfee, Avast, or Norton.

Expert Advice List. We also wanted to represent the types of articles users might be referred to if they asked an authority figure for advice. To do so, we collected a corpus of online security advice recommended by experts. We asked 10 people for a list of websites from which they personally get security advice or which they would recommend to others. These included five people holding or pursuing a Ph.D. in computer security, two employees of our university's IT department who have security-related job responsibilities, and three librarians from our university and local libraries.

Two researchers visited each recommended website and collected URLs for the referenced advice articles. Manual collection was required, as many of these expert sites required hovering, clicking images, and traversing multiple levels of sub-pages to surface relevant advice. (An initial attempt to use an automated crawl of all URLs one link deep from each page missed more than 90% of the provided advice.) As with the search corpus, we then used the Diffbot API to parse and sanitize body elements.

The resulting expert corpus includes 894 documents. Exemplar pieces of advice in this corpus include U.S. CERT pages, FBI articles, and articles from Bruce

Schneier's blog. Only five documents of advice were in both the expert and search corpi: an article from the FTC on malware, a veracrypt page, an article on passwords from security-in-a-box, an article from `safetynetkids.org`, and an article from `axantum.com`.

Corpus Cleaning. We followed a two-step, manual cleaning process for our corpus. First, to ensure that all of the documents in our corpus actually pertained to online security and privacy we recruited CrowdFlower crowd workers to review all of the documents and answer the following Yes/No question: "Is this article primarily about online security, privacy, or safety?". We retained all documents in our corpus for which three of three workers answered 'Yes'. For documents for which two of the three initial workers answered 'Yes', we recruited an additional two workers to review the document. We retained all documents for which four of five of the workers who evaluated the document answered 'Yes'. After this cleaning, 1,264 documents were retained in our advice corpus.

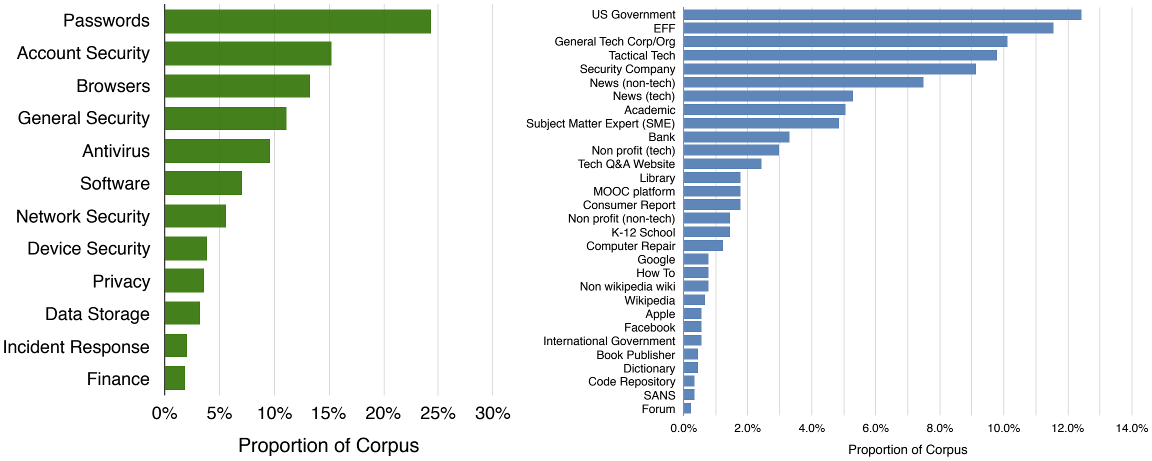
Extracting and Evaluating Advice Imperatives. In order to evaluate the actionability and accuracy of our security advice corpus, we needed to identify the individual advice imperatives (e.g., "Add ! at the end of your password") contained in the documents so that we could evaluate them. Two members of the research team manually annotated each of these 1,264 documents to extract the advice imperatives.

We constructed an initial taxonomy of advice imperatives by drawing on the limited set of prior work that had identified user security behaviors [29,39,120,161]. We manually reviewed each article and made a list of all described behaviors. In addition, we reached out to the article authors to see if they would share their original data, in order to find any behaviors that may have been mentioned by a small number of participants, but not reported in the papers. The authors of [120] shared their codebook with us. After merging duplicate behaviors between articles, this process gave us an initial list of 196 individual pieces of security advice.

We used this taxonomy as a starting point for annotating our security advice corpus. To ensure validity and consistency of annotation, two researchers double-annotated 165 (13.1%) of the advice documents.

The researchers annotating the corpus reached a Krippendorff alpha agreement of 0.69 (96.36% agreement) across the 12 high level code categories, which is classified as substantial agreement [137]. Given this substantial agreement, the large time burden of double annotating all 1264 documents, the researchers proceeded to independently code the remaining documents. To evaluate the consistency of our independent annotations, we compute the intraclass correlation (ICC): a commonly used statistical metric [210] for assessing the consistency of measurements such as test results or ratings. We find that both annotators had an ICC above 0.75 (0.823 for annotator 1 and 0.850 for annotator 2), indicating "good" consistency in their annotations [129].

At the end of the annotation process, the researchers reviewed each other's taxonomies to eliminate redundancies. Ultimately, our analysis identified 400 unique advice imperatives: 206 newly identified in our work, 170 identified in prior literature and also found in our corpus, and 26 from the research literature that did not appear in any of our documents. The full list of advice can be explored here:



(a) Distribution of topics.

(b) Distribution of domains/domain categories.

Figure 6.1: Distribution of topics and domain categories across the security advice corpus.

`securityadvice.cs.umd.edu`. This tool includes the evaluation data for accuracy and actionability (for the 374 pieces of advice that we found in or corpus), a list of the source documents, and examples of how the advice was stated in the source documents, and for each piece of advice.

As part of this analysis process, we also identified two categories of irrelevant documents present in our corpus - 229 documents that were advertisements for security or privacy products and 421 documents that had no actionable advice such as definitions, news reports, specific help pages for various software, or organizational descriptions that were security-related but contained no specific imperatives. This left a final corpus of 614 documents containing security advice.

6.1.2 Corpus Descriptives

Here, we provide an overview of the advice and documents in our corpus.

Twelve Topics of Security Advice. Our manual coding of our corpus of 614 documents resulted in 12 high level categories of advice related to:

- Account Security: advice concerned with avoiding account compromise (this does *not* include mechanisms for authentication e.g., passwords, 2FA). For example, documents about identifying compromise on your social media account, avoiding spam in your email account, and not signing up for "unnecessary" accounts.

- Antivirus: advice concerned with antivirus software and avoiding viruses and malware. For example, imperatives to use antivirus software, to not be lulled into a false sense of security from using antivirus software, and to run virus scans on new devices.

- Browsers: advice concerned with browser-based security and privacy. For example, advice on clearing browser history, to only download things you are looking for, to verify website signatures and certificates, and advice about VPNs.

- Data Storage: advice concerned with how to safely store data. For example, advice regarding keeping sensitive information on removable storage media, advice regarding backups and SSDs, and advice regarding encrypting data.

- Device Security: advice concerned with securing physical computing devices. For example, documents regarding covering your webcam, keeping your devices with you, or locking your smartphone.

- Finance: advice concerned with security and privacy of digital financial information. For example, to do online banking on certain devices, to use secure payment methods, and to type banking links manually.

- General Security: advice concerned with general security and privacy awareness. For example, advice to seek out expert help, to avoid overconfidence online, and to use parental controls for children.

- Incident Response: advice concerned with what to do following a security or privacy incident. For example, advice to cancel or change accounts, to report suspicious incidents to IT/support, and to document the incident.

- Network Security: advice related to securing home networks, routers, Bluetooth, and firewalls. For example, advice to use a password to protect your wifi, to change your router name from the default, and to turn off Bluetooth.

- Passwords: advice related to passwords (including advice to use 2FA). For example, advice to use strong passwords (including specific imperatives regarding how to construct such a password), to use unique passwords, and advice on how to store passwords.

- Privacy: advice related to privacy. For example, advice to use Tor, read privacy policies, and act anonymously online.

- Software: advice related to securing software on your devices. For example, advice to update applications, only install trusted software, and to remove unnecessary programs.

Figure 6.2 shows the distribution of topics across our corpus. In total, we extracted 374 unique advice imperatives, which occurred 2780 times across the 614 documents in our corpus. An average of 5.79 specific pieces of advice were contained in each document.

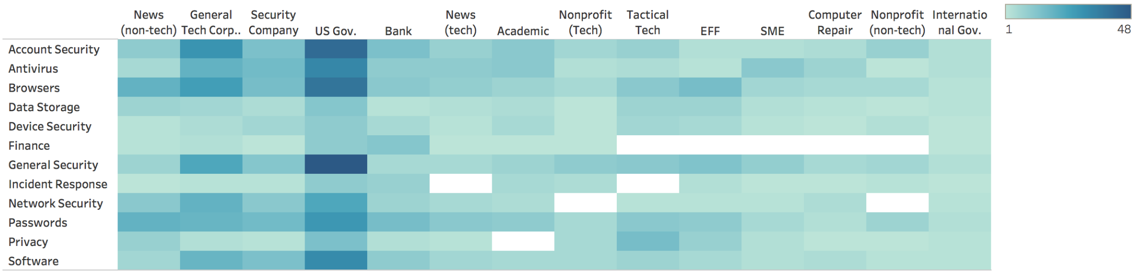


Figure 6.2: Heatmap of **number of documents** from each domain about a given topic.

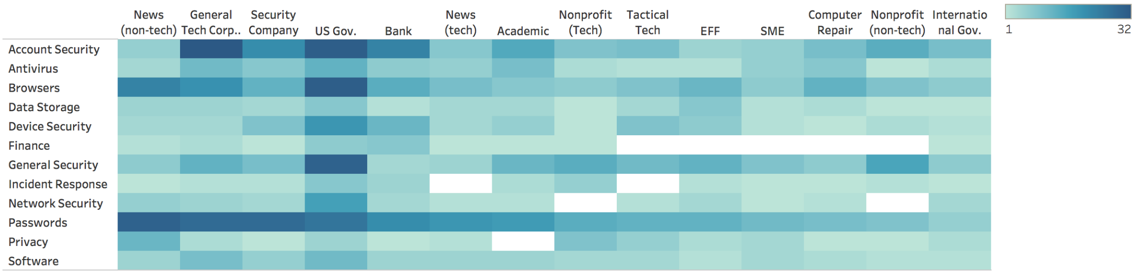


Figure 6.3: Heatmap of **number of unique advice imperatives** extracted from each domain about a given topic

Origins of Security Advice. We identified 476 unique web domains in our corpus. Through manual analysis, we further grouped these domains into broader categories (including retaining certain, relatively high-frequency domain owners of interest such as 'Google' and 'EFF').[2]

To provide context on the coverage of topics by different domains, Figure 6.2

[2] In the remainder of this chapter, we use the term "domain" to refer to these domain categories for brevity.

presents a heat map of the number of documents containing advice about each topic from each domain and Figure 6.3 presents a heat map of the number of unique advice imperatives extracted about each topic from each domain.

6.2 Comprehensibility

This section addresses the comprehensibility of the documents in the advice corpus.

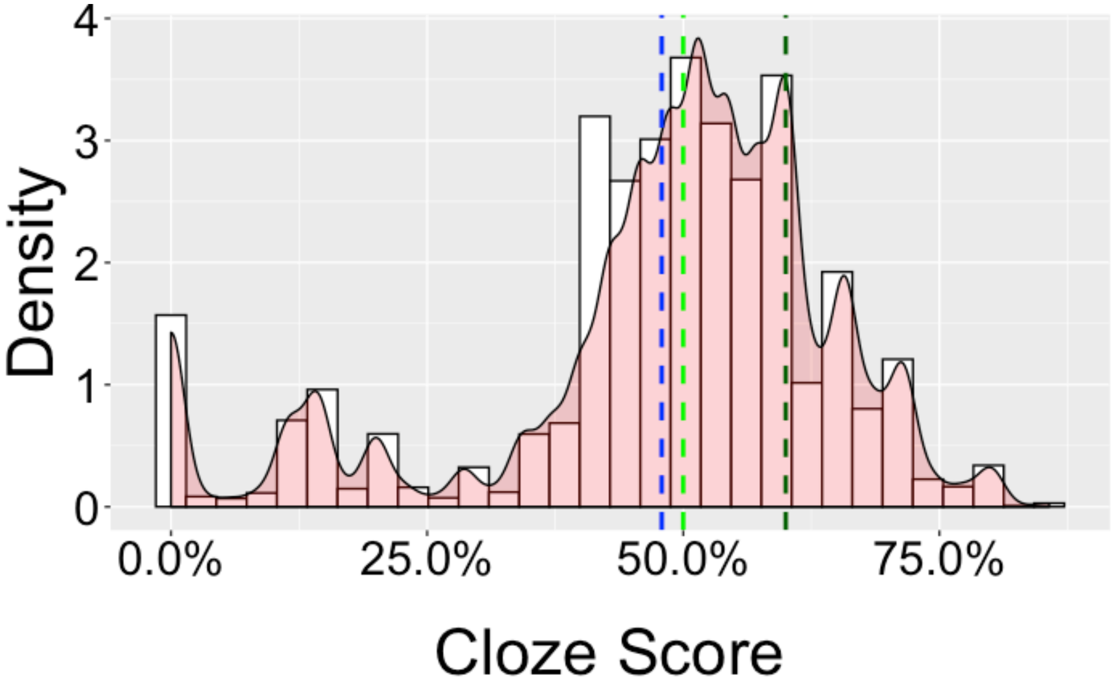


Figure 6.4: Security advice corpus Cloze scores. Higher scores indicate more comprehensible documents, light green shading and dotted line indicates mean Cloze score >50% which signifies partial comprehensibility, brighter green shading and line indicates mean score >60% which signifies full comprehensibility. Dashed blue line indicates corpus mean.

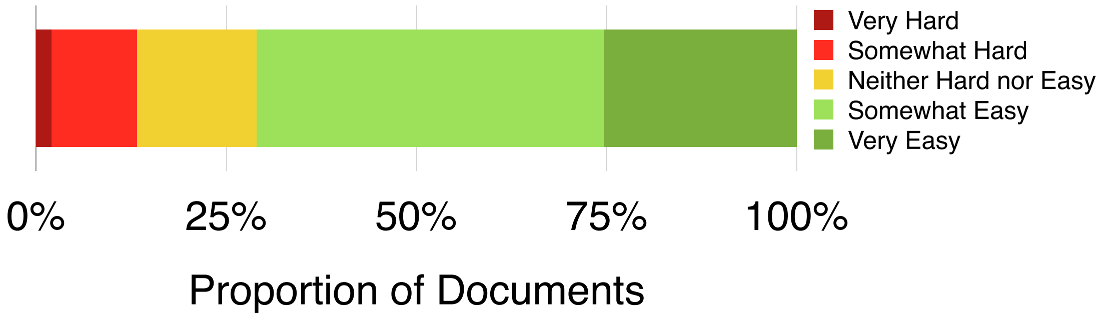


Figure 6.5: Distribution of perceived ease ratings across the security advice corpus.

6.2.1 Measuring Comprehensibility

A variety of metrics are available for assessing the comprehensibility of texts: human-expert-written comprehension questions, automatically generated comprehension tests, and computed metrics requiring no human input [31, 82, 93, 219].

Given the scale of our dataset, it is not possible to construct human-expert written comprehension questions for all 614 documents. Thus, as our results from Chapter 8 recommend, we opted for the next-best option: automatically generated comprehension tests and ease-perception assessments.

Specifically, we use the Cloze procedure [219], which involves creating comprehension tests by removing every nth word in a given document and requiring the reader to "fill-in-the-blank" with the correct word. The Cloze procedure was validated as a scalable method of comprehension assessment through comparison with expert-written comprehension questions for grade-school texts [30, 109, 164, 178]. We use the *Smart Cloze* variant of the Cloze procedure [180], which is one of many multiple-choice Cloze test variants [36, 90, 92, 117, 158, 160, 170]. In Chapter 8, we validate *Smart Cloze* on multiple types of documents, including security-related documents, and show that *Smart Cloze* reduces participant burden in domain-specific contexts while retaining validity. Cloze scores are computed as a percent out of 100 based on the number of blanks correctly completed.

We also measure perceived reading ease to augment our Cloze measurements. Perceived ease is an importantly different component of comprehensibility—as described in more detail in Chapter 8—as users may not even continue to read documents perceived at first glance as too difficult to read. To measure perceived ease, we use a single item [190, 200]: "How easy is this document to read?" with 5-point Likert-item response choices of "Very Easy" (2), "Somewhat Easy" (1), "Neither Easy nor Hard" (0), "Somewhat Hard" (-1), and "Very Hard" (-2). We compute perceived ease per document as the median of the three ease scores given to the document by the three survey respondents who evaluated it.

Recruitment. We recruited three people to take each type of test (hereafter referred to as Cloze tests and ease tests) for each of our documents; each test consisted of four randomly selected documents, three from our corpus and one attention check document designed to ensure that the readers were paying attention to the task and not just filling in blanks at random, per best practices in web survey design [63]. We recruited test takers using the Cint census-representative survey panel (n=638 for Cloze and n=635 for ease); our respondents were representative of the U.S. within 5% on age, gender, race, education and income. Participants were compensated in accordance with their agreement with Cint. The respondents who took our Cloze tests had 'excellent' reliability [129], with an ICC of 0.989. Those who took our ease tests achieved 'good' reliability (ICC = 0.757).

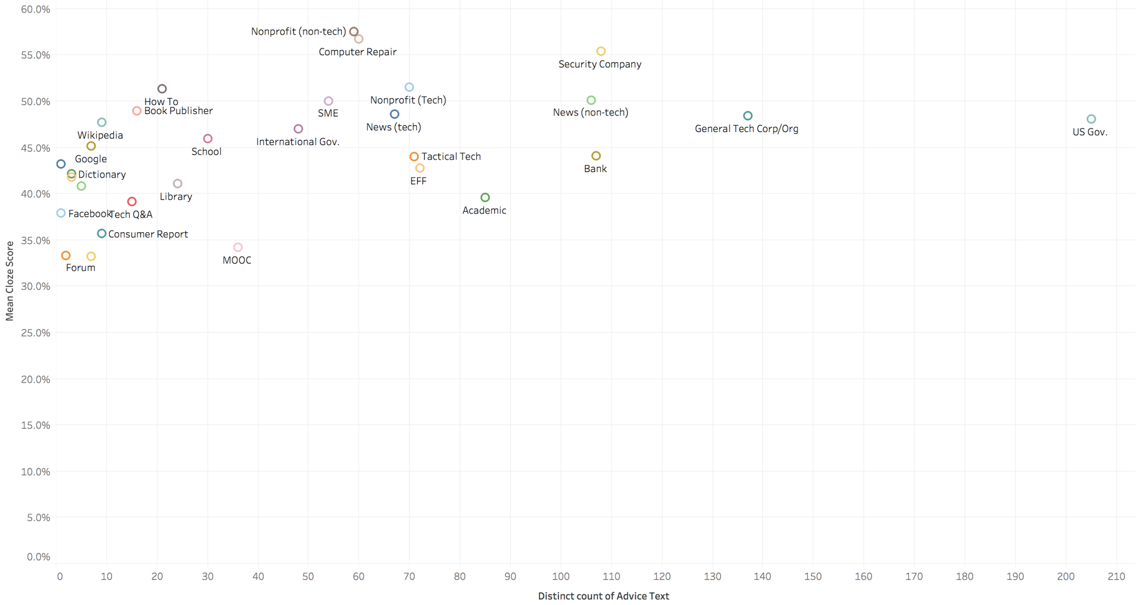


Figure 6.6: Cloze scores by domain and number of unique advice pieces.

6.2.2 Overall Comprehensibility

Wide variance in comprehensibility of security advice, with general population having at best partial comprehension of the average document. Overall, we find that the documents in our corpus have an average Cloze score of 47.9%, median Cloze score of 51.4%, and a standard deviation of 17.9%. A Cloze score below 50% indicates low comprehension, 50-60% indicates some, but not complete comprehension, and 60% indicates sufficient comprehension [71]. A little over half of the documents (55%) had a mean score of at least 50% and a little over a quarter (27%) had a score at or above 60%.[3]

[3] As Cloze scores are continuous, we report the Cloze score of a document as the mean of the Cloze scores of the three test takers.

Nearly three quarters of documents perceived as somewhat easy to read. Ease perceptions were, in general, more rosy than the results of the comprehensibility tests. People perceived the average document in our corpus as "somewhat" easy to read. Further, only 13.3% of documents in our corpus were perceived as "somewhat" or "very" hard to read with an additional 15.6% perceived as 'neither hard nor easy to read'. There was, however, very high variance in ease perceptions: the standard deviation across documents, 1.05, is a little over a full step on the Likert scale.

Figure 6.4 summarizes the overall comprehensibility and Figure 6.5 the perceived ease of the corpus. Figure 6.6 summarizes the comprehensibility (Cloze scores) by domain and by number of unique advice imperatives obtained from documents in that domain.

Security document comprehensibility correlates with age of acquisition, concreteness, polysemy, and hypernymy of context words. To understand, linguistically, why some documents were easier to comprehend than others, we considered a set of psycholinguistic factors that have been found to correlate with the difficulty of texts [53, 213].[4]

To evaluate the effect of these factors, we constructed a series of mixed-effect linear models evaluating the effect of seven different psycholinguistic factors [5], which

[4] We computed these factors for our documents using Cohmetrix [93]. While many computer-security-specific words contained in our documents are not included in the dictionaries Cohmetrix uses, the context words in the advice documents were covered. It is typical that not all words in a document will be covered by linguistic databases [53, 213].

[5] To evaluate the significance of each factor we constructed models without a given factor and conducted log likelihood tests against the null model. Full results of these models are in Appendix B.3.

controlled for the effect of having multiple test takers evaluate multiple documents. We find that the following factors have a significant effect on the comprehensibility of security documents:

- *age of acquisition*: the average age at which the words in the document are acquired by "typical" children.
- *concreteness*: the abstractness of the words in the document.
- *polysemy*: the number of senses that can be meant by a particular word; this can be indicative of text ambiguity, because the words can be interpreted in many ways.
- *hypernymy*: the specificity of a word.

Documents that contain words with a lower age of acquisition, higher concreteness, lower polysemy (fewer possible meanings for a given word), and lower hypernymy (more specificity) correlate with higher Cloze scores.

6.2.3 Comprehensibility by Topic

Documents containing advice about account security, browsers, data storage, device security, and finance are the most comprehensible. There is a significant difference in the Cloze scores for security documents containing advice about different topics ($p < 0.001$, ANOVA; all pairwise tests remain significant after Holm Bonferonni correction[6]), with the mean Cloze scores by topic varying by

[6] We omit the p-value table for the pairwise comparisons as all p-values were below 0.05, even after correction.

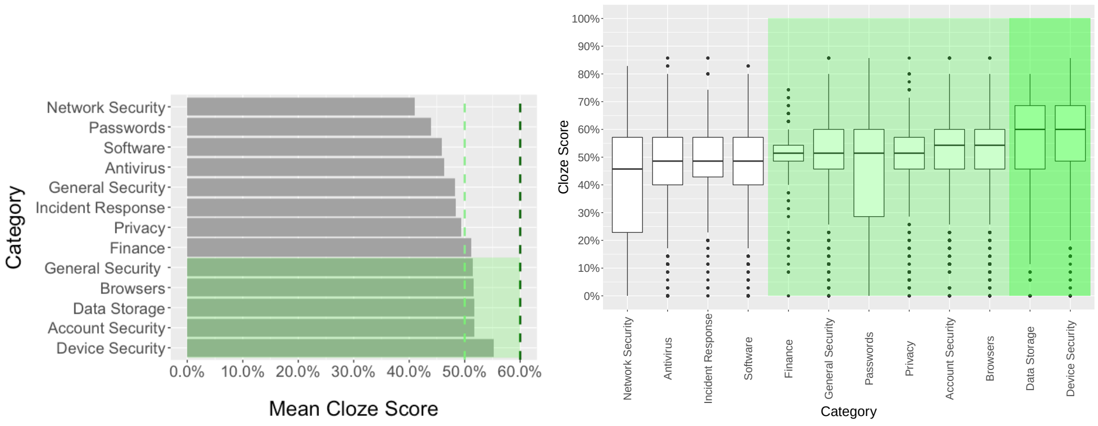


Figure 6.7: Mean (left) and median (right) Cloze scores by topic across the security advice corpus. Light green shading indicates mean Cloze score of 50% or above and brightergreen shading indicates a median score of 60% or above.

Metric	**Mean**	**Median**	**S.D.**
Device Security	54.3%	60.0%	20.0%
Account Security	51.8%	54.3%	14.3%
Data Storage	51.8%	60.0%	21.5%
Browsers	51.7%	54.3%	14.9%
Finance	51.1%	51.4%	6.22%
Privacy	49.5%	51.4%	16.0%
Incident Response	48.4%	48.6%	11.3%
General Security	48.2%	51.4%	16.1%
Antivirus	46.4%	48.6%	17.4%
Software	45.9%	48.6%	18.3%
Passwords	43.6%	51.4%	23.2%
Network Security	41.2%	45.7%	19.7%

Table 6.1: Cloze score summary statistics by topic.

13.1%. Figure 6.7 and Table 6.1 summarize the differences in topic Cloze scores. Documents that contained advice about account security, browsers, data storage, device security, and finance achieved at least partial mean comprehension (Cloze scores above 50%). Additionally, finance-related documents had particularly low variance in scores, with a standard deviation of 6.22%.

The remaining topics all had mean Cloze scores under 50%, indicating that the majority of test takers struggled to comprehend the average text on these topics. Password and network security related documents had particularly low mean scores, and very wide score spreads. We hypothesize this may be the case because there were the highest number of documents containing advice about passwords, meaning these documents covered a large breadth of the corpus (as evidenced by the fact that documents containing advice about passwords had the highest SD in Cloze scores across the corpus: 23.2%). On the other hand, documents containing advice about network security may have been difficult to comprehend perhaps because network security is a particularly technical topic, and documents containing advice about network security may thus contain more technical jargon.

There was not a significant difference in reading ease perceptions for documents containing advice about different topics ($p = 0.999$, Kruskal-Wallis test[7]).

6.2.4 Comprehensibilty by Domain

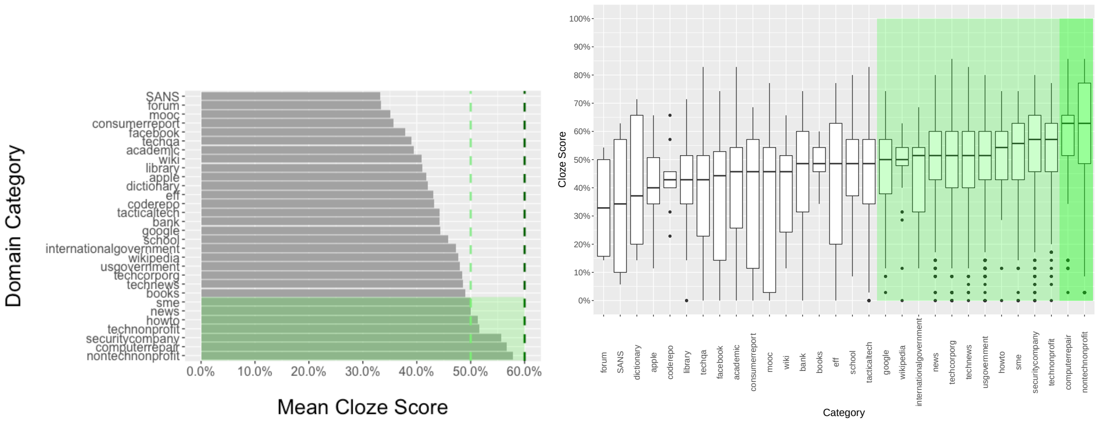


Figure 6.8: Mean (left) and median (right) Cloze scores by domain across the security advice corpus. Light green shading indicates mean Cloze score of 50% or above and brightergreen shading indicates a median score of 60% or above.

Seven classes of advice givers provide particularly comprehensible security advice: general news channels, subject matter experts, non-

[7] We use a Kruskal-Wallis rather than ANOVA test for the ease scores as the ease scores were not normally distrubuted, while the Cloze scores had a near-normal distribution as determined by a qqplot.

profits (both technology and non-technology focused), as well as security and computer repair companies. We also wanted to understand whether some advice-givers provided more readable advice than others. To answer this question, we examined Cloze scores grouped by domain (Figure 6.8) summarizes these results. The Cloze scores of the domains were significantly different: $p < 0.001$, ANOVA (all pairwise tests remain significant after Holm Bonferonni correction[8]). We see that 7 of the 30 groups of domains we considered had mean Cloze scores above 50%: SMEs, general news outlets, how-to websites, both non-tech-focused and tech-focused non-profit organizations, security companies, and computer repair companies.

Within particular categories, we see that some organizations perform better than others (see Figures 6.9-6.12).

Government Organizations. Among U.S. government organizations, `ic3.gov`, `whitehouse.gov`, `ftc.gov`, and `dhs.gov` all have mean Cloze scores above 50%; while the remaining domains perform worse (Figure 6.9). We had only five non-U.S. government domains in our dataset, three of which (`csir.co.za`, `staysmartonline.gov.au`, and `connectsmart.gov.nz`) had mean scores of 50% or above.

Child-Focused Organizations. Encouragingly, documents from non-profit organizations (both technology focused and not) that were aimed toward children (e.g., childline.org.uk, netsmartz.org, safetynetkids.org.uk) appear to be among the most readable (Figure 6.10). That said, content collected from school websites was not particularly readable, with mean Cloze scores below 50%, suggesting that schools

[8] We omit the p-value table for the pairwise comparisons as all p-values were below 0.05, even after correction.

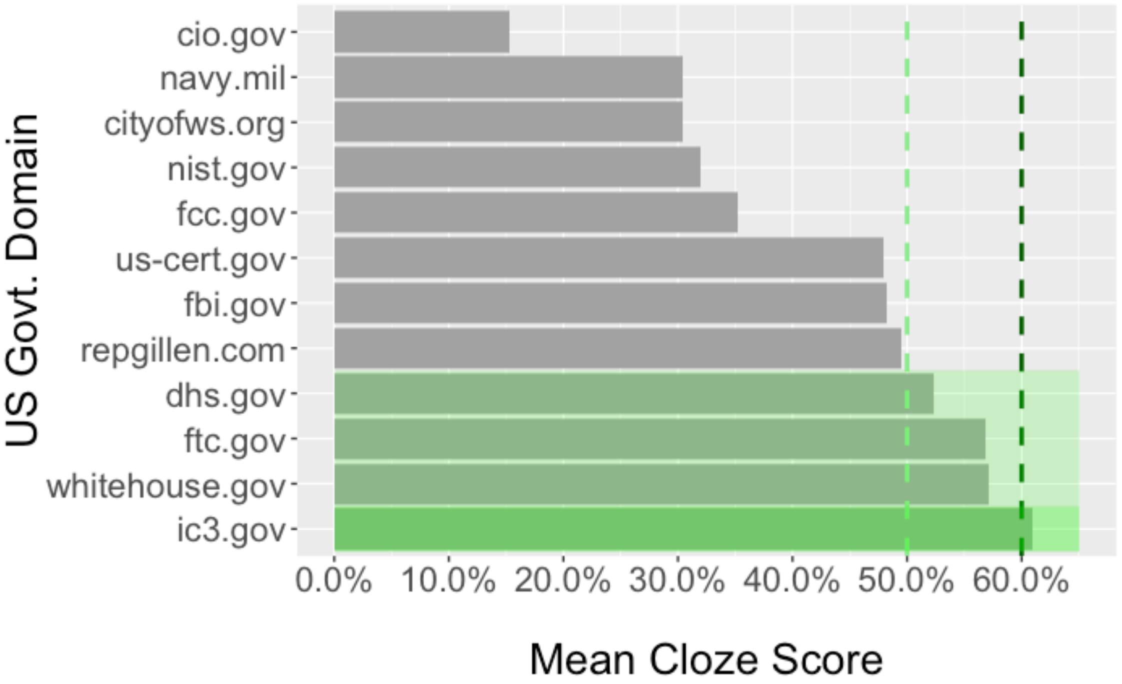


Figure 6.9: Mean Cloze scores for U.S. Government domains. Light green shading indicates mean Cloze score of 50% or above and brightergreen shading indicates a median score of 60% or above.

may be better off obtaining content from child-focused nonprofit organizations rather than writing their own.

Technical Non-profits. Documents from more technical non-profit organizations had wider variance. Documents from the Tor Project, GNU, and `techsoup.org` all had mean Cloze scores above 50%. On the other hand, documents from nine other technical non-profits, including Mozilla, Chromuium, Ubuntu, and organizations focused specifically on helping non-experts e.g., librarians (libraryfreedomproject.org) had mean Cloze scores well below 50%. Documents from the EFF and Tactical Tech-sponsored organizations also had mean Cloze scores below 50%. This is important to note, as documents from these two organizations alone make up 21% of our corpus.

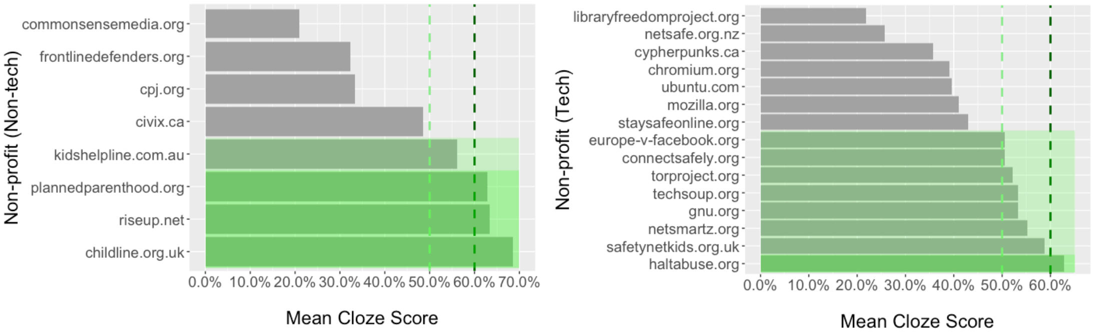


Figure 6.10: Mean Cloze scores for non-proft domains; non-tech nonprofits (left) and tech-focused nonprofits (right). Light green shading indicates mean Cloze score of 50% or above and brightergreen shading indicates a median score of 60% or above.

Subject Matter Experts. Blogs from individual SMEs exhibited similar score patterns to documents from technical non-profits (Figure 6.11). SME content had an average Cloze score of 50.9%, with 6 of 26 subject matter expert publications (`malwaretruth.com`, `tamingthebeast.net`, `macexpertguide.com`, `Internet-online-privacy.com`, `cknow.com`, and `thatoneprivacysite.net`) achieving Cloze scores in the full comprehension range (above 60%).

Corporations. Security-focused companies and those offering computer repair services both scored very high on comprehensibility. Even more so than for SMEs, the material these companies present is designed to bring in new clients, and in the case of computer repair services especially, those clients are typically lay users who have recently experienced a computer problem. Thus, creating readable materials directly affects these companies' bottom lines; they appear to be responding to this need by providing readable content. This is in contrast to non-security-focused organizations, including those frequently under fire for privacy and security issues such as Google (mean Cloze = 45.1%), Facebook (mean Cloze = 37.9%), and Apple (mean Cloze = 41.7%).

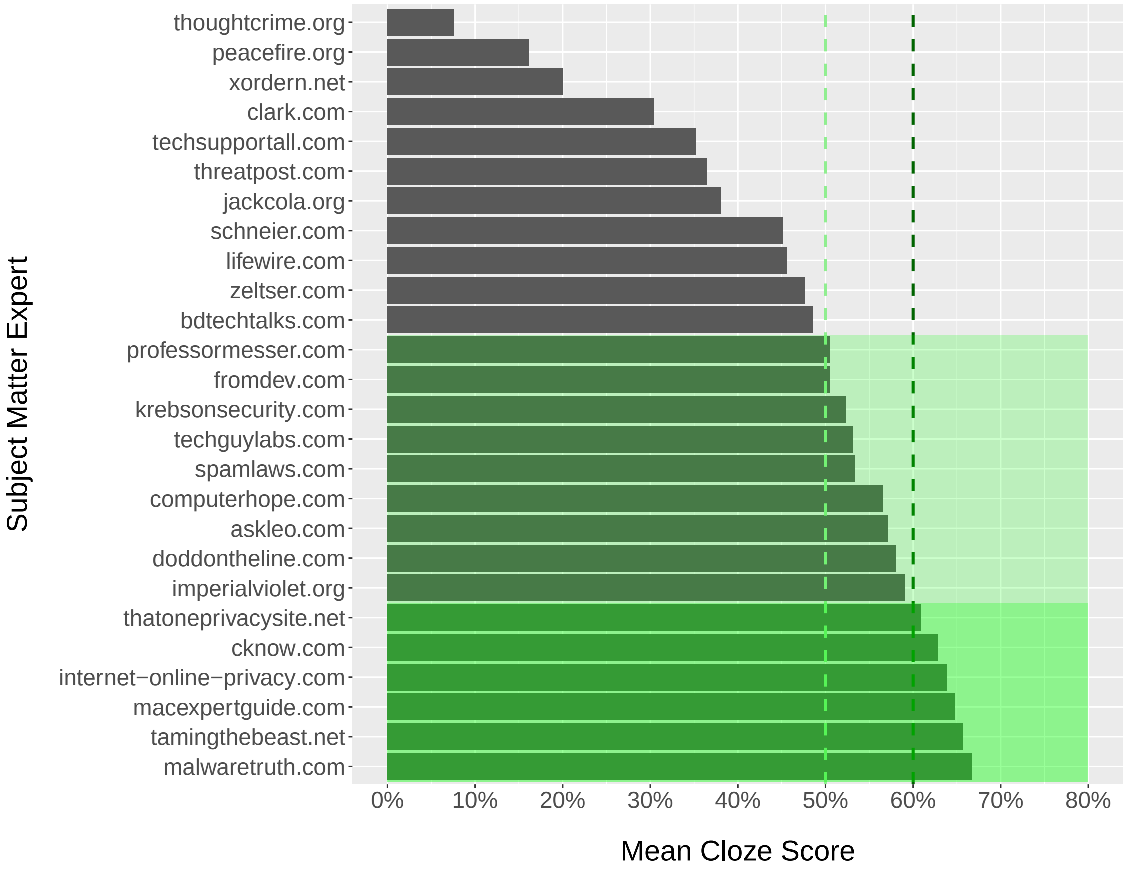


Figure 6.11: Mean Cloze scores for domains run by subject matter experts (SMEs). Light green shading indicates mean Cloze score of 50% or above and brightergreen shading indicates a median score of 60% or above.

News Organizations. Non-tech-focused news publications had a mean Cloze score of exactly 50%, with `billmoyers.com`, `medium.com`, `brighthub.com`, `lifehacker.com`, two local news sites (`desmoinesregister.com` and `deccanchronicle.com`), `usatoday.com`, `huffingtonpost.com`, `lovetoknow.com`, `boardingarea.com`, and `thehindu.com` performing particularly well, with mean Cloze scores above 60% (Figure 6.12).

Tech-focused news publications achieved significantly lower comprehension ($p < 0.001$, ANOVA), with a mean Cloze score of 48.6%. Of tech-focused pub-

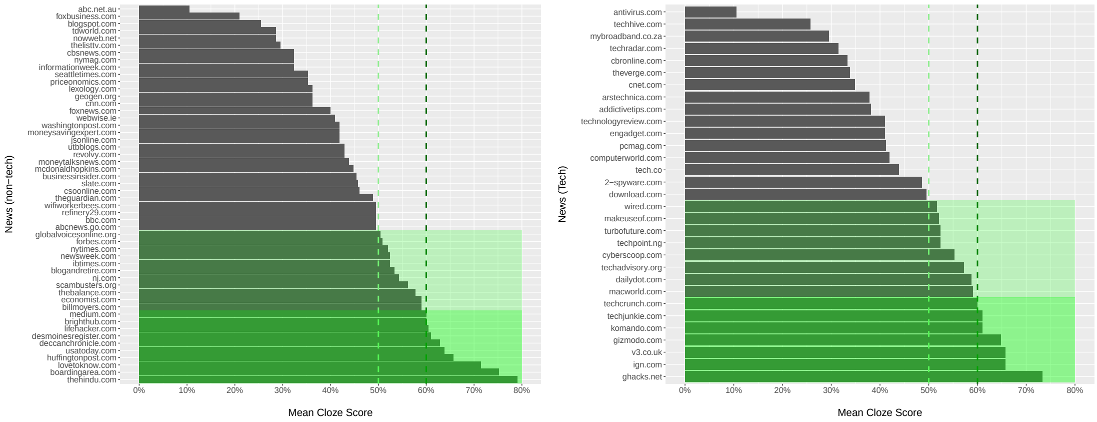


Figure 6.12: Mean Cloze scores for news producer domains; general news producers (left) and tech-focused news outlets (right). Light green shading indicates mean Cloze score of 50% or above and brightergreen shading indicates a median score of 60% or above.

lications, 15 of 31 achieved partial comprehension (Cloze scores > 50%), and 7 (techcrunch.com, techjunkie.com, komando.com, gizmodo.com, v3.co.uk, ign.com, and ghacks.net) achieved scores in the full comprehension range (> 60%).

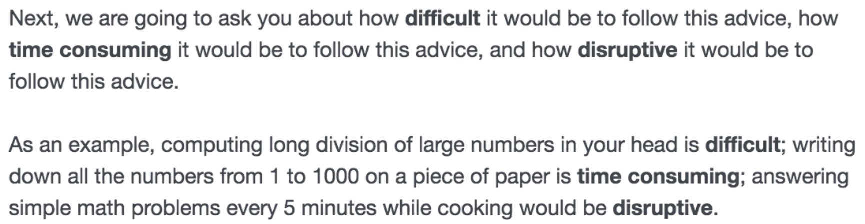
Low Comprehension Platforms. Finally, 7 of the 30 advice givers we examined provided particularly difficult to read advice (mean Cloze scores under 40%): SANS (sans.org), security forums such as (malwaretips.com and wilderssecurity.com), MOOC platforms such as (lynda.com and khanacademy.org), consumer reports such as (consumerreports.org and av-comparatives.org), Facebook, Technical Q&A websites like stackoverflow.com and stackexchange.com, and academic publications such as those hosted by usenix.org and ieee-security.org.

There was not a significant difference in ease perceptions for security documents containing advice from different domains ($p = 0.999$, Kruskal-Wallis test).

6.3 Actionability

In this section, we describe our approach to measuring the actionability of the advice in our corpus and report the results of this measurement.

6.3.1 Measuring Actionability



Next, we are going to ask you about how **difficult** it would be to follow this advice, how **time consuming** it would be to follow this advice, and how **disruptive** it would be to follow this advice.

As an example, computing long division of large numbers in your head is **difficult**; writing down all the numbers from 1 to 1000 on a piece of paper is **time consuming**; answering simple math problems every 5 minutes while cooking would be **disruptive**.

Figure 6.13: How the different types of actionability were explained to survey respondents in the actionability evaluation questionnaire.

As experts do not always have an accurate sense of users' burden from following advice, we use general U.S. population evaluations to assess advice actionability: how hard it would be for a typical user to put into practice.

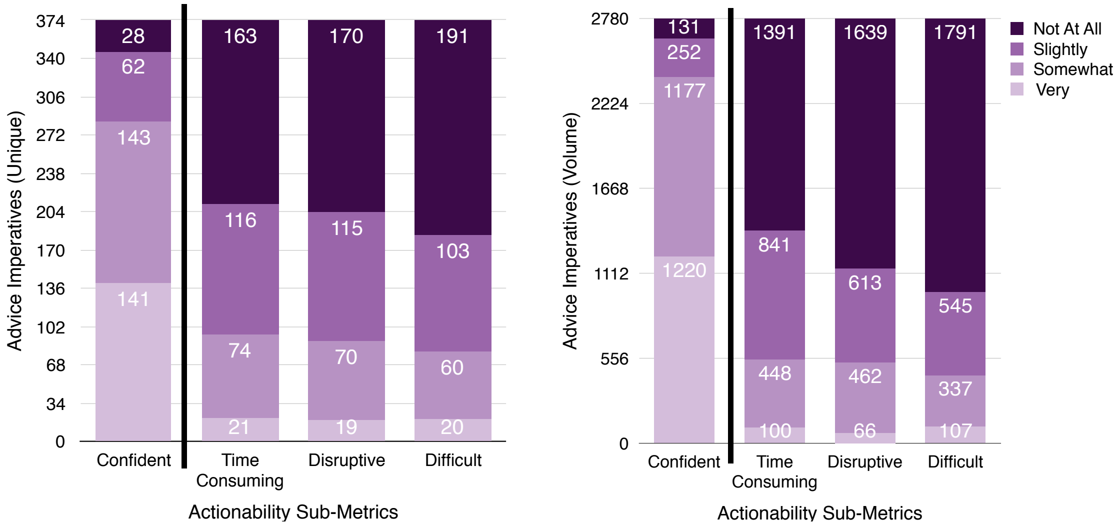
Evaluation Questionnaire. The actionability questionnaire evaluated advice along four sub-metrics:

- *Confidence:* How confident the user was that they could implement this advice.
- *Time Consumption:* How time consuming the respondent thought it would be to implement this piece of advice.
- *Disruption:* How disruptive the user thought it would be to implement this advice.
- *Difficulty:* How difficult the user thought it would be to implement this advice.

on a 4-point likert scale from 'Not at All' to "very". As an example for the distinction between time consumption, disruption, and difficulty, we provided respondents with the description shown in Figure 6.13. Each survey had respondents evaluate five advice imperatives. The full questionnaire is included in the Appendix A.8.

We broke actionability into these four sub-metrics in alignment with prior work on security behavior. The confidence sub-metric is drawn from Protection Motivation Theory [195], which identifies perceived ability to protect oneself as a key component of protective behavior implementation, and from the Human in the Loop model [51], which identifies knowledge acquisition—knowing what to do with information—as a key component of security behavior change. The time consumption and disruption sub-metrics are created to align with the "cost" of the behavior, which has been found to be an important decision-making factor in economic frameworks of secure behavior [26, 110, 187, 188]. Finally, the difficulty sub-metric is used to align with the capabilities component of the Human in the Loop model [51].

Recruitment. We again recruited using the Cint panel (n=313), which provided a sample matched to the U.S. population within 5% on age, gender, education, race and income. Participants were compensated in accordance with their agreement with Cint. Respondents had 'good' reliability in evaluating these sub-metrics, with $ICC = 0.896$ for confidence evaluation, $ICC = 0.854$ for time-consumption evaluation, and $ICC = 0.868$ for both disruption and difficulty evaluation.



(a) Actionability of unique advice imperatives based on user ratings on the four actionability sub-metrics: confidence, time consumption, disruption, and difficulty.

(b) Actionability of advice based on user ratings by volume across all 2780 pieces of advice in the corpus.

Figure 6.14: Actionability of unique advice imperatives (left) and of all advice imperatives in the corpus (right).

6.3.2 Overall Actionability

Overall, the advice in our corpus had a median confidence rating of "somewhat" confident – people were somewhat confident that they could implement the advice – a median time consumption rating of "slightly" time consuming, a median disruptive rating of "slightly" disruptive, and a median difficulty rating of "not at all" difficult. The distribution of actionability ratings across advice is shown in Figure 6.14a.

People were at least "somewhat" confident about implementing majority of advice and felt it was at most "slightly" time consuming, disruptive, or difficult to implement. People were very (37.7%) or somewhat (38.2%) confident about implementing three quarters of the advice, with only 7.49% of advice receiving a median confidence rating of “not at all” confident. Further, nearly half of the advice was considered “not at all” time consuming (43.6%), “not at all” disruptive (45.5%), and just over half was considered not at all difficult (51.1%).

49 pieces of advice were identified as unactionable on at least one sub-metric. We define unactionable advice as advice that people were not at all confident about implementing or advice that was rated as very time consuming, very difficult, or very disruptive to implement. Only 21, 19, and 20 pieces of advice were rated “very” time consuming, disruptive, and difficult, respectively. 28 pieces of advice received a median confidence rating of “not at all” confident. In sum, these 49 pieces of advice made up 13.1% of the 374 pieces of advice we evaluated. About half of this advice (24 pieces) was rated as very unactionable on one of four submetrics, the remainder was rated very unactionable on multiple submetrics.

Of these 49 pieces of advice, 41 were rated as accurate by a majority of experts who evaluated the advice, 6 were rated as useless, one was rated as harmful(“Not change passwords unless they become compromised”) and one was a piece of advice on which the experts reached no consensus (“Lock your SIM card in your smartphone”). The median risk reduction estimated by the experts for the 41 pieces of advice rated unactionable was 32.5%, slightly below the median risk reduction for all advice. Table 6.2 lists the unactionable advice and the accuracy metrics for this advice.

Over 80% of the 2780 pieces of advice in our corpus were rated as “somewhat” or “very” actionable across actionability submetrics. Examining the advice in our corpus by volume (2780 pieces of advice total), the 49

Advice	Not Confident	Very Time Consuming	Very Disruptive	Very Difficult	Accuracy	Risk Reduced
Apply the highest level of security that’s practical	✗	✗		✗	All Accurate	50%
Be wary of emails from trusted institutions	✗				All Accurate	25%
Beware of free VPN programs		✗		✗	All Accurate	30%
Change your MAC address	✗				Majority Accurate	32.5%
Change your username regularly		✗	✗	✗	Majority Useless	NA
Consider opening a credit card for online use only	✗				All Useless	NA
Cover your camera			✗		Majority Accurate	30%
Create a network demilitarization zone (DMZ)	✗				Majority Accurate	27.5%
Create keyboard patterns to help remember passwords		✗	✗	✗	Majority Useless	NA
Create separate networks for devices	✗	✗	✗	✗	Majority Accurate	40%
Disable automatic download of email attachments		✗			All Accurate	40%
Disable Autorun to prevent malicious code from running	✗	✗			All Accurate	50%
Disconnect from the Internet	✗				All Accurate	25%
Do online banking on a separate computer				✗	All Accurate	32.5%
Encourage others to use Tor			✗	✗	Majority Accurate	25%
Encrypt cloud data	✗			✗	Majority Accurate	45%
Encrypt your hard drive	✗		✗	✗	All Accurate	5%
Isolate IoT devices on their own network	✗	✗	✗	✗	Majority Accurate	20%
Keep sensitive information on removable storage media		✗			Majority Accurate	22.5%
Leave unsafe websites		✗	✗		Majority Accurate	22.5%
Limit personal info being collected about you online	✗				Majority Accurate	15%
Lock your SIM card in your smartphone	✗	✗	✗	✗	No Consensus	NA
Not blindly trust HTTPS	✗				Majority Accurate	20%
Not change passwords unless they become compromised	✗				All Harmful	-30%
Not identify yourself to websites	✗				Majority Accurate	30%
Not let computers or browsers remember passwords	✗				Majority Accurate	45%
Not overwrite SSDs	✗	✗	✗	✗	All Accurate	45%
Not send executable programs with macros			✗	✗	All Accurate	20%
Not store data if you don’t need to				✗	All Accurate	40%
Not use credit or debit cards online	✗	✗		✗	Majority Useless	NA
Not use encryption when sending e-mail to a listserv	✗	✗	✗	✗	Majority Useless	NA
Not use extensions or plugins	✗				Majority Accurate	35%
Not use Facebook	✗			✗	Majority Accurate	30%
Not use your real name online			✗		All Accurate	30%
Not write down passwords				✗	Majority Accurate	50%
Remove unsafe devices from the network		✗	✗		All Accurate	50%
Run a virus scan on new devices	✗				All Accurate	35%
Set up auto-lock timers for your smartphone		✗	✗		All Accurate	30%
Turn off Bluetooth	✗				All Accurate	40%
Understand who to trust online	✗		✗		All Accurate	20%
Unmount encrypted disks		✗			All Accurate	50%
Use a password manager		✗			All Accurate	50%
Use an air gap			✗		Majority Accurate	50%
Use an unbranded smartphone		✗			All Useless	NA
Use different computers for work and home use	✗				All Accurate	50%
Use encryption	✗		✗	✗	All Accurate	50%
Use incognito mode		✗	✗	✗	Majority Accurate	45%
Use single sign-on SSO	✗				All Accurate	10%
Use unique passwords		✗			All Accurate	50%

Table 6.2: List of the most unactionable advice based on user ratings of confidence, time consumption, disruption, and difficulty. The first four columns indicate whether the advice was rated with a median rating of “not at all” confident, “very” time consuming, disruptive, and/or difficult. The fifth column indicates the expert rating for accuracy: all accurate (all experts rated the advice as accurate), majority accurate (majority of experts rated this advice as accurate), etc. The final column provides the median risk reduction estimated by the experts for this piece of advice; advice rated as useless by a majority of experts or on which the experts did not reach consensus does not have a risk reduction reported, harmful advice has negative risk reduction.

pieces of most unactionable advice made up only 3.6% of the advice in the corpus by volume despite making up 13.1% of the unique advice imperatives (Figure 6.14 shows this contrast). Our respondents were “very” or “somewhat” confident they could implement 86% of the advice. Further, respondents found 81.3%, 81.0%, and

84.0% of the advice at most "slightly" time consuming, disruptive, and difficult to implement, respectively (Figure 6.14b).

In over half the documents, there is at least one piece of advice that people rated as "somewhat" or "very" time consuming, disruptive, or difficult to implement. Despite the infrequent repetition of unactionable advice, this advice was spread widely across the documents in the corpus. We find that 20.5% of documents contained at least one piece of advice that people were "not at all" confident they could implement, while 16.5% of documents contained advice that people rated as "very" time consuming to implement, 11.3% of documents contained advice people rated as "very" disruptive to implement, and 18.6% of documents contained advice that people rated as "very" difficult to implement.

6.3.3 Actionability by Topic

Each of the four sub-metrics of actionability differ significantly by topic, with $p = 0.040$ for confidence, $p < 0.001$ for time consumption, $p = 0.042$ for disruption, and $p = 0.022$ for difficulty (Kruskal-Wallis tests). These differences are summarized in Figure 6.15. The pairwise comparison tables, including which specific topic-differences were significant can be found in Appendix B.10.

Overall, people were confident they could implement at least 50% of advice on all of the topics. They also rated over 50% of the advice on all topics except data storage as at most slightly time consuming, disruptive, or difficult.

People are most confident about account security, antivirus, and

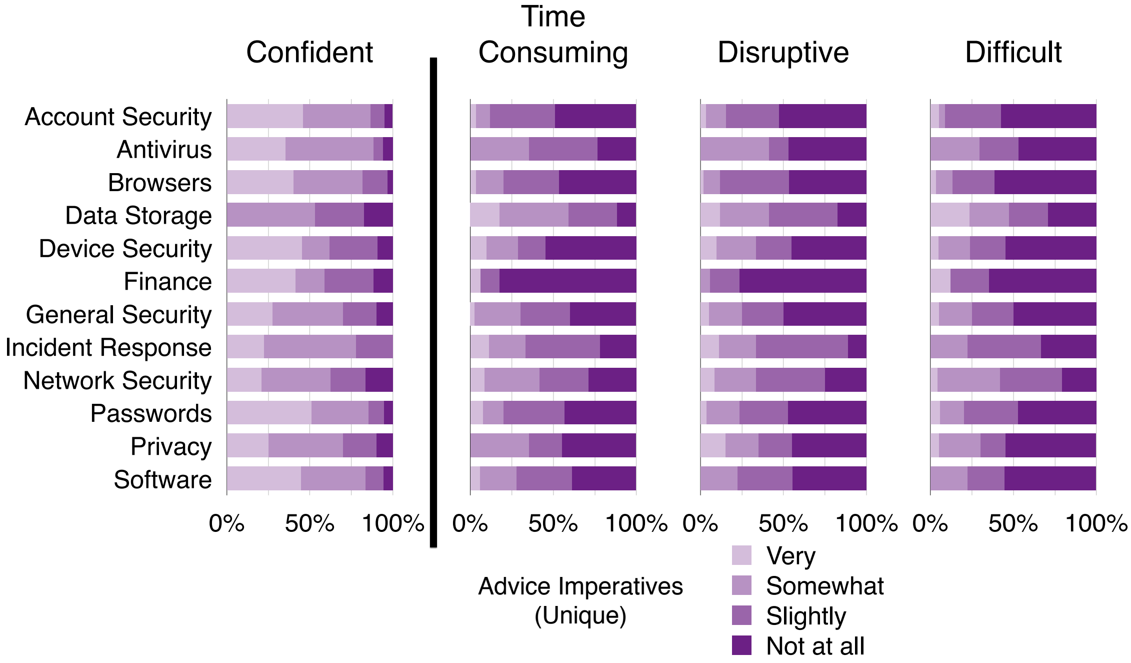


Figure 6.15: Actionability of advice by topic.

passwords; least confident about data storage and network security. People were at least "somewhat" confident that they could implement more than 80% of the advice related to account security, antivirus, and passwords, perhaps because these are among the most common security advice topics (Figure 6.2). On the other hand, people were at least "somewhat" confident they could implement barely half of the advice about data storage and network security. There was no advice about data storage about which people were "very" confident and network security had the lowest proportion of advice about which people were "very" confident (20.8%). These two topics differ significantly in confidence ratings from the remaining categories, as do the five high performing categories aforementioned (Table B.13).

People evaluated advice related to data storage and network security as the most time consuming, and advice about finance and account

security as the least. The vast majority of advice (94.1%) about finance was rated as “slightly” or “not at all” time consuming. Account security advice was similarly perceived as not time consuming, with 88.1% of account security advice rated as at most “slightly” time consuming. These two topics differed significantly in their time consumption ratings from the rest of the topics (Table B.14).

On the other hand, 58.8% of the advice about data storage was rated as at least “somewhat” time consuming to implement. Although not as time consuming as data storage, network security had the next most time consuming advice, with 41.7% of the advice about network security being rated as “somewhat” or “very” time consuming. It is additionally interesting to note that the time consumption of privacy advice was quite split: near equal proportions of privacy advice were rated as at least “somewhat” time consuming (35%) and “not at all” time consuming (45%).

People rated advice related to data storage the most disruptive, and advice about finance the least. Finance had the lowest proportion of advice that was rated disruptive to implement, with no advice that was “very” disruptive and only 5.9% advice that was “somewhat” disruptive. Data storage advice was once again given the lowest actionability rating, with 41.2% of data storage advice being rated at least “somewhat” disruptive to implement. Additionally, we observe that, while advice about browsers, account security, and privacy do not statistically differ (Table B.15) in their proportions of disruptive advice, the shape of the distribution of disruptive advice is different for privacy. Similar to the split in ratings for the time consumption of privacy advice, we observe that the highest proportion of advice that was considered “very” disruptive to implement was related to privacy (15%), while on the other hand, 45% of the advice about privacy was considered “not at all” disruptive.

People rated advice about data storage and network security as the most difficult, advice about account security, browsers, finance, and passwords as the least. Finally, data storage also had the most difficult advice, with 47.1% rated as at least “somewhat” difficult to implement. Network security advice was also similarly (Table B.16) quite difficult to implement, with 41.7% of the advice about network security being rated as at least “somewhat” difficult to implement. On the other hand, more than 80% of the advice about account security, finance, browsers, and passwords was rated as at most “slightly” difficult to implement. It is interesting to note that while the majority of advice about browsers is rated “not at all” difficult to implement, advice about browsers was rated as relatively disruptive (the majority of advice about browsers was rated as at least “slightly” disruptive).

Proportion of unactionable advice differs significantly by topic. Finally, we also see a significant difference in the proportion of unactionable advice (advice that received a median rating of “very” for difficulty, disruption, or time consumption or a median rating of “not at all” for confidence; see list in Table 6.2) between topics ($p = 0.037$, Kruskal-Wallis test). These results are summarized in Figure 6.16. Overall, data storage had the highest proportion of unactionable advice (35.3%), followed by privacy (25.0%), network security (20.8%), finance (17.6%), device security (16.7%), passwords (14.5%), incident response (11.1%), general security (10.0%), account security (8.5%), browsers (6.7%), antivirus (5.9%), and software

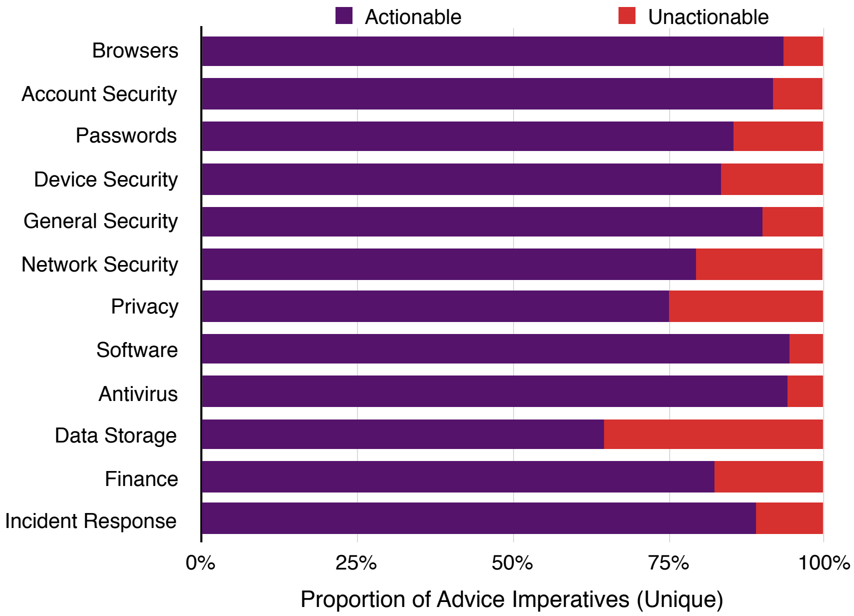


Figure 6.16: Proportion of unactionable of advice by topic. Unactionable advice is advice that received a median rating of "very" for difficulty, disruption, or time consumption or a median rating of "not at all" for confidence; see list in Table 6.2) by topic.

(5.6%).

6.3.4 Actionability by Domain

Actionability did not differ significantly by domain (Kruskal-Wallis tests for confidence: $p = 0.906$, time consumption: $p = 0.852$, disruption: $p = 0.334$, and difficulty: $p = 0.873$).[9] Figure 6.17 summarizes the distribution of advice actionability by domain.

[9] There was also not a significant difference when considering just unactionable advice by domain ($p = 0.684$).

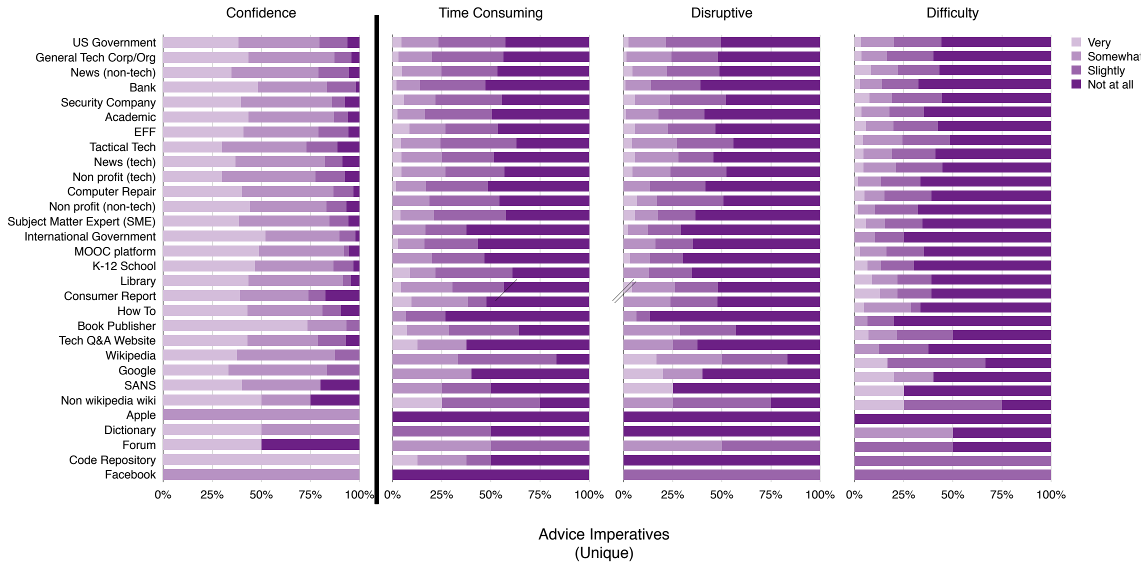
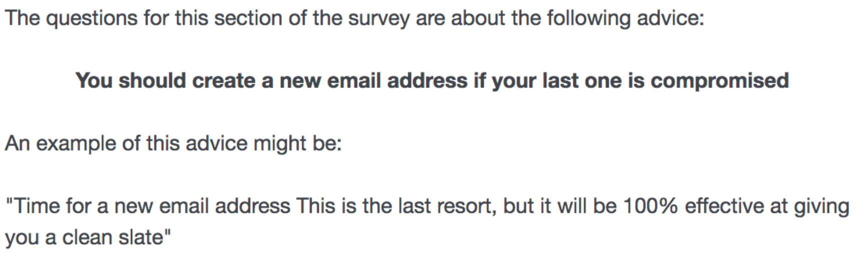


Figure 6.17: Actionability of advice by domain.

6.4 Accuracy

In this section we describe our approach to measuring the accuracy of the 374 advice imperatives we identified in our corpus and report the results of this measurement.

6.4.1 Measuring Accuracy



The questions for this section of the survey are about the following advice:

You should create a new email address if your last one is compromised

An example of this advice might be:

"Time for a new email address This is the last resort, but it will be 100% effective at giving you a clean slate"

Figure 6.18: Example of how a piece of advice with example drawn from our security corpus would be shown to experts and users in the accuracy and actionability evaluation questionnaires.

We also use human-generated data to measure the accuracy of the advice imperatives. We asked experts to answer an evaluation questionnaire for each piece of security advice. As with actionability, each advice imperative was evaluated by three users.

Evaluation Questionnaire. The accuracy questionnaire evaluated, for each advice imperative:

- *perceived accuracy:* whether the expert believed that a typical end user following this advice would lead to an improvement, no effect, or harm to the users' security.
- For advice that would reduce security risk:
  - *risk reduction:* how much the expert estimated risk would be reduced (numerically, on a scale from 0% to more than 50%) if the advice was followed.
  - *priority:* how highly the expert would prioritize recommending this piece of advice to users (number 1 behavior, in the top 3, in the top 5, in the top 10, would recommend but not in the top 10, or would not recommend).
  - *longevity:* how long the expert thought the advice would remain accurate (less than 1 year, 2-5 years, 5-10 years, more than 10 years).
- For advice the expert thought would increase security risk:
  - *risk increase:* how much the expert estimated risk would be increased if the advice was followed.

Each survey contained 10 pieces of randomly selected security advice. Each piece of security advice was accompanied by an example that was randomly selected from the sentences annotated with that advice code (Figure 6.18). The full questionnaire is included in Appendix A.7.

Recruitment. I recruited experts by tweeting from my personal Twitter account, asking well-known security Twitter accounts to retweet our call for experts, and leveraging my and my collaborators personal networks. We also posted in multiple professional LinkedIn groups and contacted authors of security blogs. All recruited individuals completed a screening questionnaire that assessed their security credentials, including what security certifications they held, whether they had ever participated in a CTF, what security blogs or publications they read, whether they had ever had to write a program that required them to consider security implications, whether they had ever penetration tested a system, and their current job title. We also asked them to upload their resume or link to their personal website so that we could verify their credentials.

We considered anyone who had done two or more of: participating in a CTF, penetration testing a system, writing programs that required them to consider security implications OR who held security certifications (including computer security professors) to be an expert. Ultimately, 41 qualified experts evaluated our security advice. The majority of our experts were practical experts; only three were academics.

Experts were paid $1 for each piece of advice they evaluated. Advice was evaluated in batches of 10; experts were allowed to complete as many batchess as they desired and were able to skip previously-evaluated pieces of advice. On average, experts evaluated 38 pieces of advice. We find that the experts achieved ‘good’ [129] reliability in evaluating the advice for accuracy, with an ICC of 0.876.

6.4.2 Overall Accuracy

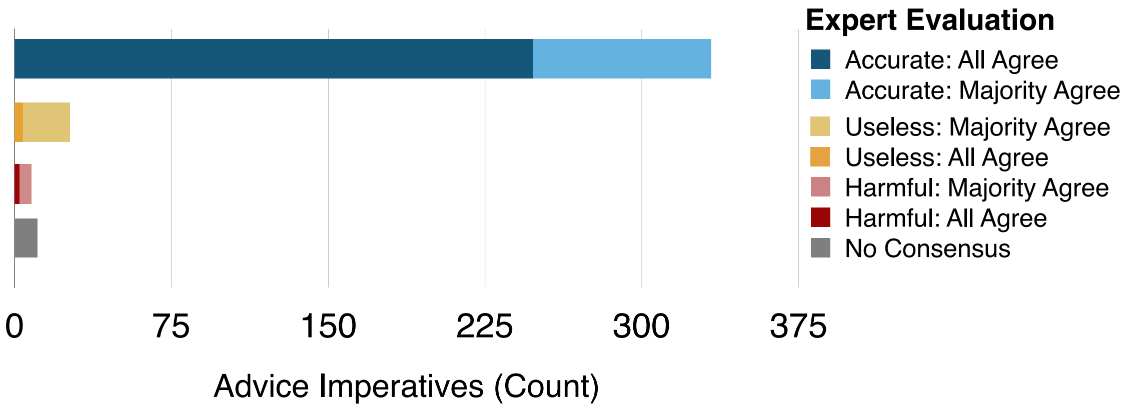


Figure 6.19: Accuracy of advice imperatives based on expert evaluations.

Almost all advice labelled as accurate, almost none as harmful. Overall, 248 imperatives (66.3%) were rated as accurate by all three experts who evaluated it, 85 imperatives (22.7%) were identified as accurate by two of three experts, and 31 imperatives (8.29%) were identified as accurate by one expert. Experts reported that following these imperatives would lead to a median 37.5% reduction in users’ security risk.

Four imperatives (1.1%) were classified as useless by all three experts:

- You should consider opening a credit card for online use only

- You should file taxes early (to avoid identity theft)

- You should let your children teach you about the Internet too

- You should use an unbranded smartphone

Twenty-two additional imperatives (5.9%) were classified as useless by two of three experts, and 77 (20.6%) were classified as useless by one expert. Appendix B.4 lists all 26 pieces of advice identified as useless by the majority (2 of 3) of experts.

Two imperatives (0.50%) were identified as harmful by all three experts: "You should not change your passwords unless they become compromised" and "You should write down passwords on paper." Five additional pieces of advice (1.3%) were identified as harmful by 2 of 3 experts:

- You should base passwords on upcoming events

- You should create a new email address if your last one is compromised

- You should use tracking applications (to monitor your online activity)

- You should use different personas online

- You should store passwords

The seven imperatives that a majority of experts rated as harmful were rated as having a median increase to users' security risk of 10%. 31 additional imperatives (8.23%) were identified as harmful by one expert. The full list of 38 imperatives identified as harmful by at least one expert is in Appendix B.5).

Finally, these eight pieces of advice had no consensus (each expert gave a different evaluation):

- You should feel comfortable making weak passwords for sites that don't keep personal information
- You should install firmware on mobile devices
- You should lock your SIM card in your smartphone
- You should not change browser security settings
- You should not open attachments from unknown senders
- You should not use a password manager
- You should protect your computer from power surges
- You should transfer sensitive files to network shares

Figure 6.19 summarizes these results.

Of the 2780 instances of advice imperatives in our corpus, only 3% are perceived by experts as harmful or useless. By volume across our corpus, which contained 2780 pieces of advice, 95.8% of the advice given out to users was considered accurate by a majority of experts (77.7% was considered accurate by all experts). 2.48% of advice was considered useless by a majority of experts, and 0.51% was considered harmful.Thus, even though 8.82% of advice was perceived as useless or harmful by a majority of experts, these imperatives make up only 3.00% of the advice in our corpus by volume[10] ($p < 0.001$, Kruskal-Wallis proportion test). Figures 6.20 and 6.21 illustrate this difference.

[10] When we say "by volume" in the remainder of this chapter we mean the proportion out of the 2780 instances of advice imperatives.

All documents in our corpus contain at least one piece of advice that a majority of experts rated as accurate. When examining the accuracy of advice at a document level, we find that all documents contained advice rated as accurate by a majority of experts. 82.7% of documents contained exclusively advice evaluated as accurate; on average, these documents contained 4.65 pieces of advice.[11] In contrast, documents that contained at least one piece of useless, harmful, or no-consensus advice contained an average of 11.19 pieces of advice, of which an average of 9.78 pieces (83.3%) of that advice was perceived as accurate.

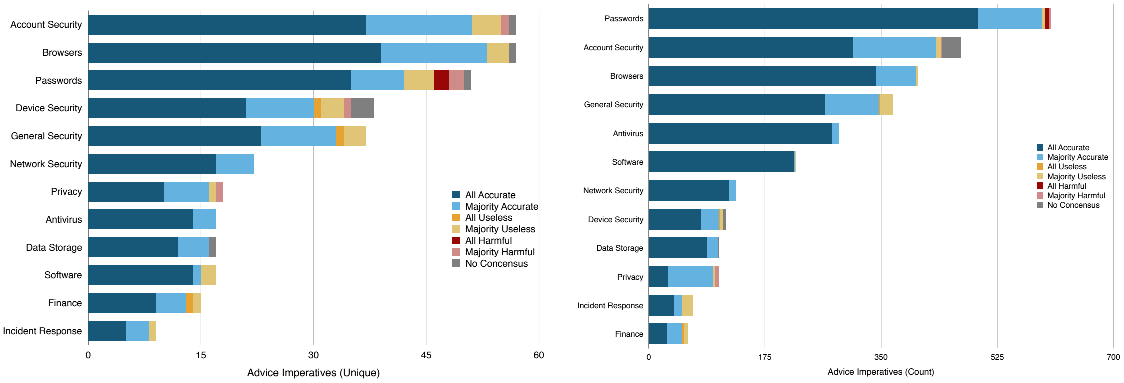
10.2% of documents in the corpus contained at least one of the 26 pieces of advice our experts identified as useless. On average, these documents contained 1.40 pieces of useless advice. 6.10% of the corpus contained one of the eight pieces of advice about which there was no consensus, these documents contained an average of 1.16 pieces of no consensus advice. Finally, 2.76% of the corpus contained harmful advice; on average, 1.07 pieces of it.

6.4.3 Accuracy by Topic and by Domain

Number of unique advice imperatives rated as accurate does not vary significantly by topic or domain.

Perhaps unsurprisingly given that the majority of advice is considered accurate by a majority of experts, accuracy does not vary significantly by topic ($p = 0.245$, Kruskal-Wallis test) or by domain ($p = 0.958$, Kruskal-Wallis test). Figures 6.20 and 6.21 provide an overview of advice accuracy across topics and domains. There

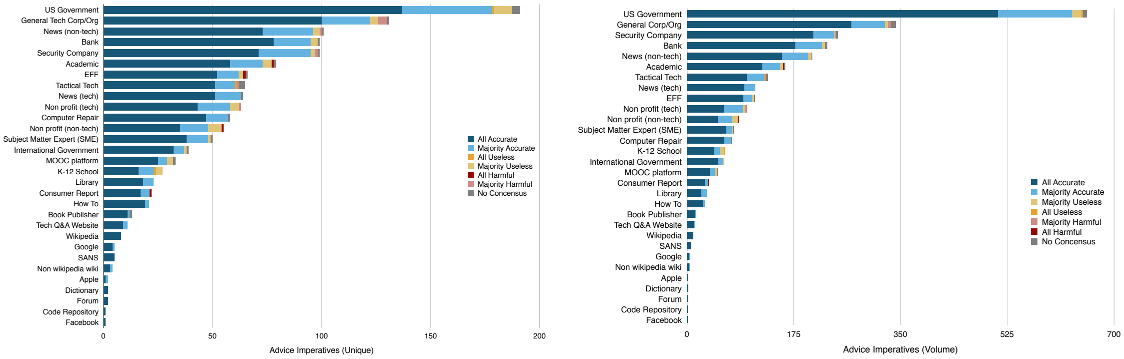
[11] As mentioned in Section 6.1 documents contained 5.79 pieces of advice on average.



(a) Accuracy distribution of unique advice imperatives in the corpus by topic.

(b) Accuracy distribution of all advice in the corpus by topic.

Figure 6.20: Accuracy distribution of unique advice imperatives (left) and accuracy distribution of all advice imperatives in the corpus (right) by topic.



(a) Accuracy distribution of unique advice imperatives in the corpus by domain.

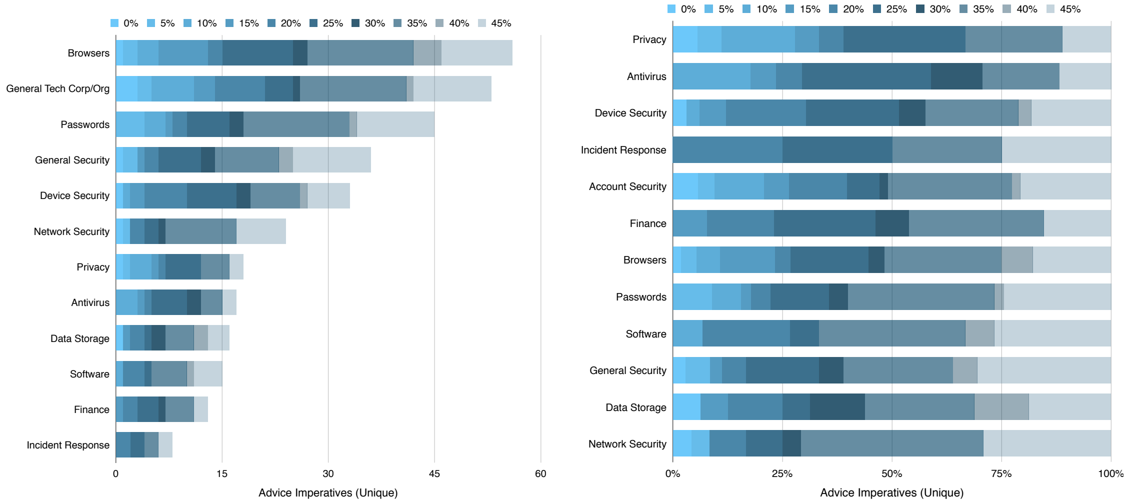
(b) Accuracy distribution of all advice in the corpus by domain.

Figure 6.21: Accuracy distribution of unique advice imperatives (left) and accuracy distribution of all advice imperatives in the corpus (right) by domain.

is also not a significant difference in the median risk reduction of the accurate advice given about different domains ($p = 0.210$, Kruskal-Wallis test) nor topics ($p = 0.312$, Kruskal-Wallis test).

6.4.4 Advice Priority

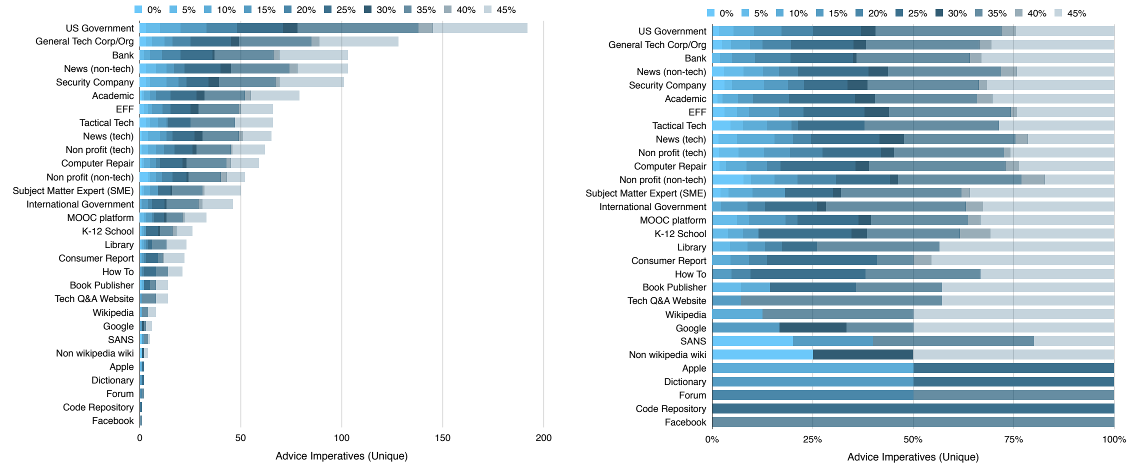
Only 25 pieces of advice given top priority—number 1 or top 3 recommended behaviors—by experts. While experts perceive the vast majority



(a) Expert-estimated risk reduction by count of advice in each topic.

(b) Expert-estimated risk reduction by proportion of advice in each topic.

Figure 6.22: Expert-estimated risk reduction of accurate imperatives by topic.



(a) Expert-estimated risk reduction by count of advice extracted from each domain.

(b) Expert-estimated risk reduction by proportion of advice extracted from each domain.

Figure 6.23: Expert-estimated risk reduction of accurate imperatives by domain.

of advice as accurate, they were somewhat more discerning when considering which behaviors to recommend to users. Twenty five pieces of advice (6.67%) received a median priority rating of “top 3 behaviors I would recommend” (see Appendix B.6 for the full list). Over a third (31.6%, 118 pieces) of all advice was rated as being in the “top 5” behaviors experts would recommend and, of the remaining 231 pieces of advice, 187 pieces (half of all advice) were rated as being among the “top 10” behaviors experts would recommend.

We also inferred priority across our data by computing the ranking of all pieces of advice using matrix factorization. Matrix factorization is a commonly used technique from information retrieval [130, 209] that is used to determine a full ranking of items when individual users have provided ratings on only a small portion of the space of items (in our case advice).[12]

This approach led to the following pieces of advice being identified as the top 10:

- You should use unique passwords for different accounts
- You should update devices
- You should use anti-malware software
- You should scan attachments you open for viruses
- You should use different passwords for different accounts/devices
- You should use unique passwords
- You should encourage others to use strong passwords
- You should not tell anyone your passwords, even IT
- You should use end-to-end encryption for communication

[12] We performed a matrix factorization with $1,000$ iterations, $alpha = 0.0002$, and $beta = 0.02$.

- You should remember your passwords

And the following advice being in the bottom strata (priority less than 9):

- You should not use banking apps or websites
- You should create keyboard patterns to help with remembering passwords
- You should lock your SIM card in your smartphone
- You should create multiple accounts
- You should disable and/or limit caching
- You should write down passwords on paper
- You should file taxes early (to avoid identity theft)
- You should install firmware on mobile devices
- You should use an unbranded smartphone
- You should create a new email address if your last one is compromised

The full advice ranking and computed rankings are included in Appendix B.7.

Users struggle to discern between advice, little correlation between expert and user priority rankings. As a comparison point, we also analyzed users' ratings of the priority of the same 374 pieces of advice. Users identified 13.5% of the advice as being the number one behavior they should follow, and an additional 25.6% of behaviors as being among the top 3 behaviors they should follow. This high proportion of advice labeled as being in the top 3 suggests that users have an

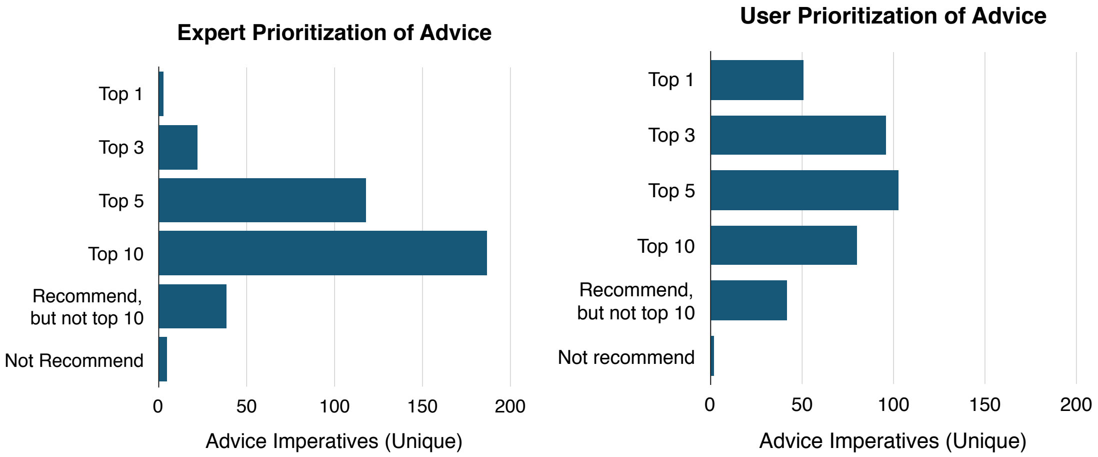


Figure 6.24: Priority with which experts would recommend the advice in our corpus (left) and priority with which users would implement the advice in our corpus (right).

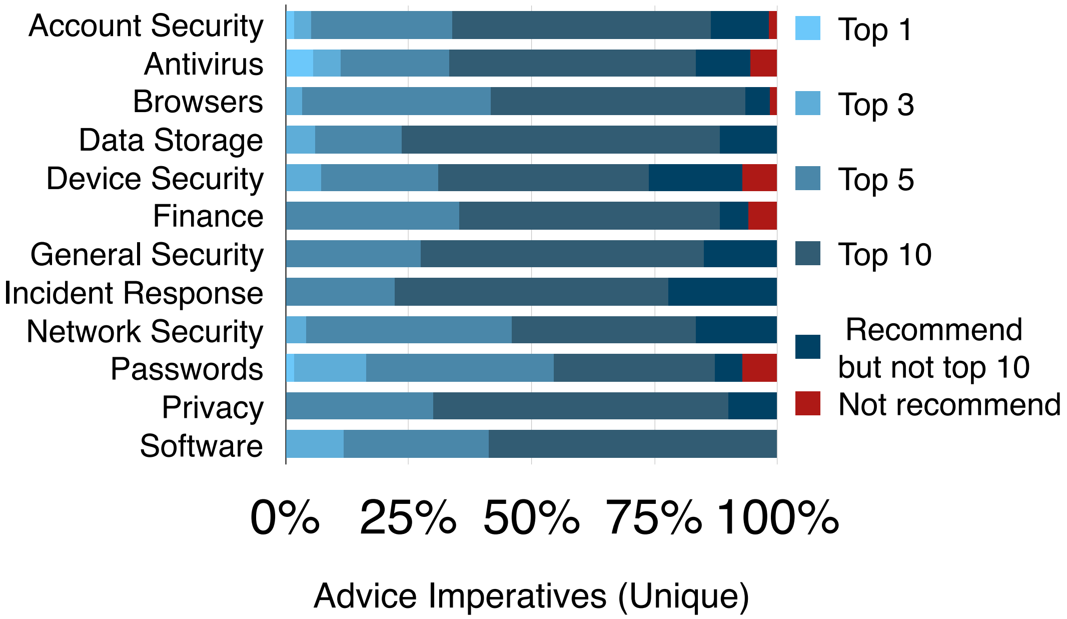


Figure 6.25: Priority with which experts would recommend the advice in our corpus by topic.

even more difficult time differentiating the importance of advice than do experts, as illustrated in Figure 6.24.

Experts' and users' priority rankings (as computed with the matrix factorization technique and parameters described above) do correlate ($p < 0.001$), albeit not very strongly ($tau = 0.175$).[13] This result aligns with the findings of Ion et al. who considered a smaller sample (20 pieces) of advice citeion2015no. The full prioritization of advice computed using matrix factorization based on users' rankings is provided in Appendix B.8.

Passwords, antivirus, finance, network, and software security advice given highest priority by experts. Priority differs significantly by topic ($p = 0.044$, Kruskal-Wallis test) but not by domain ($p = 0.089$); Table B.12 in Appendix B.9 shows the pairwise comparisons between the topics.[14] Advice about passwords is given especially high priority, along with advice about antivirus, finance, network security, and software security. Figure 6.25 summarizes these results.

6.4.5 Advice Longevity

Finally, among the advice in the corpus that was rated as accurate by a majority of experts, the median perception of longevity for the advice was that the advice would remain useful for improving people's security for the next 5-10 years. No advice was perceived as remaining useful for less than one year and only 38 pieces of advice (11.4%) were perceived as remaining useful for at most the next 2-5 years (see Appendix B.12 for a list of this advice). Over half of the advice (178 pieces, 53.8%) was perceived as remaining useful for 5-10 years and just over a third (115 pieces, 34.7%) was perceived as remaining useful for the next few decades.

[13] We use Kendall's Tau rank correlation as this coefficient is better designed for noisy data such as ours than is Spearman's Rho rank correlation.

[14] Post-hoc Mann-Whitney comparisons are corrected for multiple testing with the Holm Bonforonni correction.

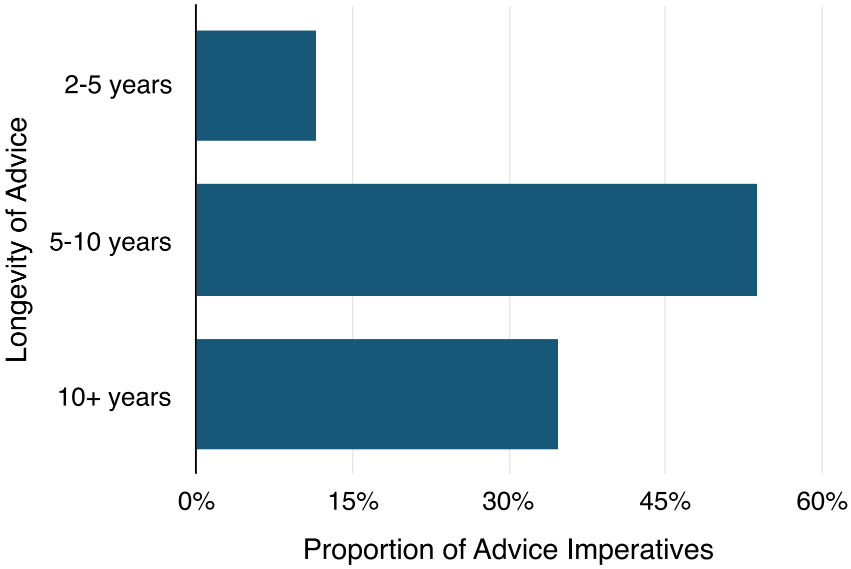


Figure 6.26: Experts perceptions of how long the accurate advice would remain useful.

Longevity did not differ significantly by topic ($p = 0.826$, Kruskal-Wallis test) or by domain ($p = 0.567$). Figure 6.26 summarizes these results.

6.5 Adoption

We also asked both our expert and general population advice-evaluators to report whether or not they themselves followed ("at least some of the time"[15]) the advice they were evaluating. The question assessing adoption was included in each of the respective surveys, with cushioning language to reassure respondents that their answer would not affect their survey compensation or qualification in any way (see Appendix A.7 and Appendix A.8 for full questionnaires). The advice adoption results are summarized in Figure 6.27.

[15] After many rounds of cognitive testing, we settled on asking whether advice was ever followed, as this was the most intuitive question for experts and general users alike across all of our advice.

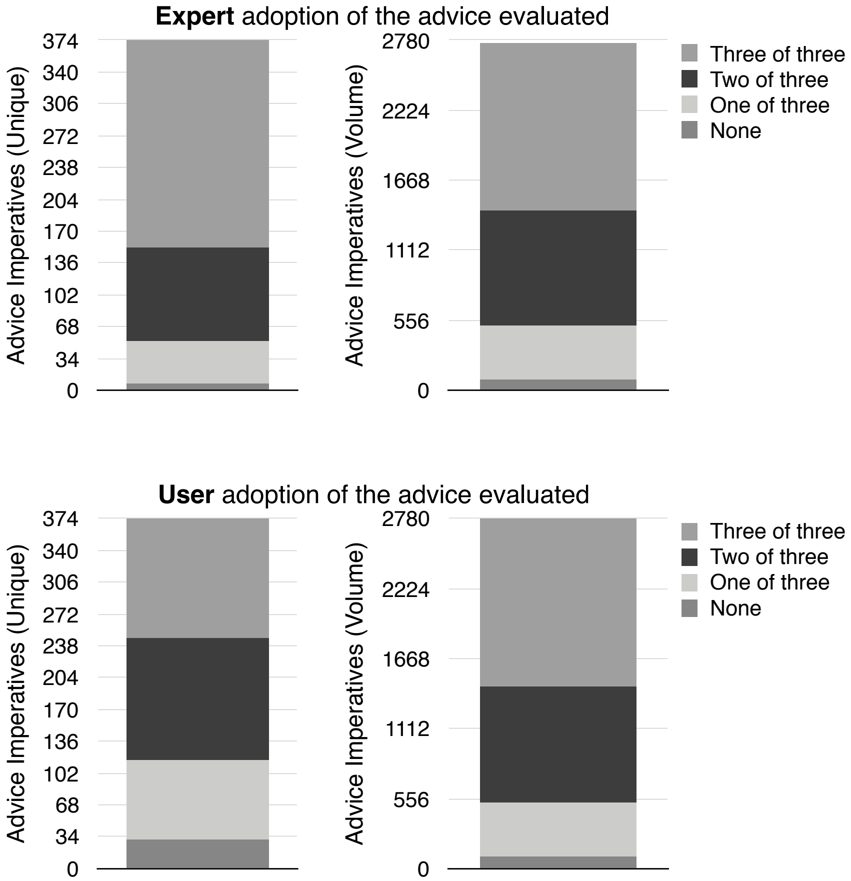


Figure 6.27: Adoption of unique advice imperatives (left) and all advice in our corpus by volume (right) by experts and general population respondents.

6.5.1 Experts

For 59.6% of the advice (224 imperatives), all three experts claimed to follow that advice at least some of the time. Another quarter of the advice (26.3%, 99 imperatives) was reported as adopted by two of three experts who evaluated it. On the other hand, only six imperatives (1.60%) were followed by none of the experts who evaluated the advice; a list of this unfollowed advice is included in Appendix B.11.

By volume, less than 1% of advice in the overall corpus was followed by no experts and only 5.6% of the advice was followed by at most one expert. The remaining advice (94.0%) was followed by the majority of experts who evaluated it. At a document level, 90.5% of the documents have at least one piece of advice followed by all the experts who evaluated it and only 2.1% of the documents contain one piece of advice followed by no experts.

6.5.2 General Population

Nearly 70% of the advice was reported as adopted, at least some of the time, by the majority of people who evaluated it. All the people who evaluated 34.4% (129 imperatives) of the advice claimed to follow it at least some of the time. Another 34.7% of the advice was reported as adopted by two of three people who evaluated it. Only 31 imperatives (8.27% of advice) were followed by no people (31 imperatives); a list of this unfollowed advice is included in Appendix B.11.

By volume, only 3.2% of the advice in the corpus was reported as unfollowed by all people who evaluated it and 15.4.% was followed by only one person who evaluated it. The remaining 81.4% was followed by the majority of people who evaluated it (48.0% was followed by all three evaluators). Examining the documents themselves, we find that 13.5% of the documents contain at least one piece of advice not followed by a respondent while 72.4% of the documents contain at least one piece of advice that all three respondents reported following.

6.5.3 Harmful and Useless Advice

Of the 40 pieces of advice that were rated harmful or useless by the majority of experts who evaluated that advice or about which the experts reached no consensus, 23 were practiced by one expert who evaluated that advice, 11 were practiced by two of three experts who evaluated the advice, and one ("Create pronounceable passwords") was followed by all three experts who evaluated it as useless. On the other hand, 14 of these 40 pieces of advice were followed by one respondent who rated that advice, 8 were followed by two of the three respondents who rated that advice, and 11 were followed by all three respondents who rated that advice.

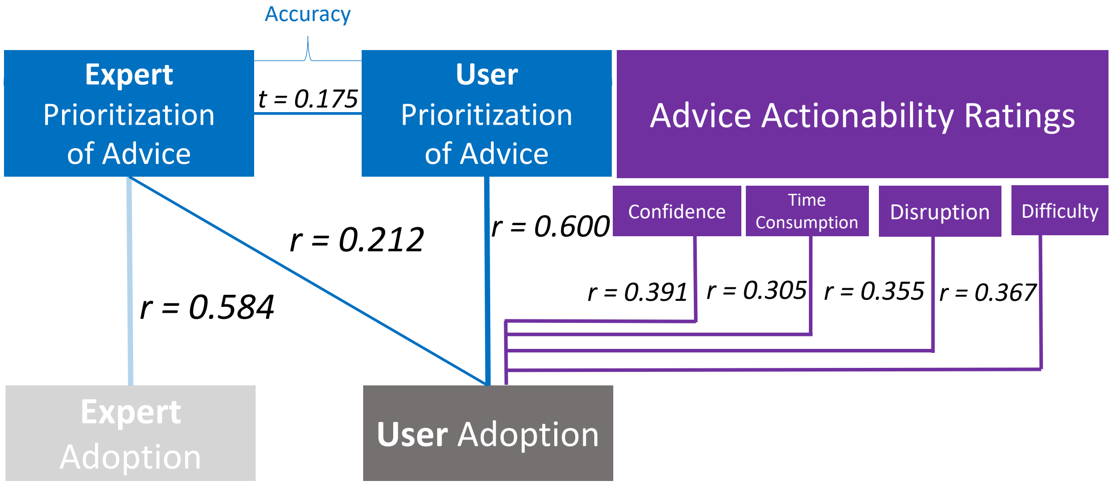


Figure 6.28: This figure summarizes the relationships (correlations) between security advice adoption and our actionability and accuracy (specifically, priority rankings of advice) measurements.

6.5.4 Actionability and Adoption

There is a significant correlation between actionability and adopting advice. Advice with higher confidence ratings is significantly more likely to have a higher adoption rate ($r = 0.391$, $p < 0.001$). Similarly, less time consuming ($r = 0.305$, $p < 0.001$), disruptive ($r = 0.355$, $p < 0.001$), and difficult ($r = 0.367$, $p < 0.001$), advice is more likely to be adopted.

Actionability correlates significantly with adoption by experts only for user-rated confidence ($r = 0.110$,$p = 0.034$) and difficulty ($r = 0.124$, $p = 0.017$), and does so with much smaller effect size.

6.5.5 Priority and Adoption

. Finally, expert and user priority rankings also correlate with advice adoption. The number of users who report adopting a behavior strongly correlates with users priority ranking of that behavior ($p < 0.001$, $r = 0.600$). Experts' behavior, too follows their priority rankings: the correlation between the expert-priority ranking for a piece of advice and the number of experts who adopt that advice is nearly identical ($r = 0.584$, $p < 0.001$). The discrepancy between user and expert ratings is underscored by the difference in correlation strength: user adoption correlates far more weakly with the expert-priority ranking of a piece of advice ($p < 0.001$, $r = 0.212$), underscoring the importance of experts filtering down to a much smaller set of key advice, across which users can then prioritize. Figure 6.28 summarizes the correlations between adoption, priority, and actionability.

Topic	**Prevalence**		**Comprehension**		**Actionability**			
	Unique	Volume	Cloze	Ease	Confident	Time Consuming	Disruptive	Difficu
Account Security	60	15.1%	51.8%	0.773	3.15	1.78	1.78	1.69
Antivirus	17	9.53%	46.4%	0.893	3.10	2.10	2.06	1.98
Browsers	59	13.1%	51.7%	0.843	3.04	1.84	1.80	1.72
Data Storage	17	3.20%	51.8%	0.706	2.51	2.63	2.31	2.47
Device Security	41	3.81%	54.3%	0.729	3.03	1.90	1.98	1.87
Finance	17	1.87%	51.1%	0.839	2.82	1.59	1.67	1.75
General Security	40	11.0%	48.2%	0.980	2.83	2.03	1.94	2.00
Incident Response	9	1.98%	48.4%	1.21	2.81	2.15	2.26	1.78
Network Security	24	5.50%	41.2%	0.720	2.68	2.17	2.08	2.19
Passwords	54	24.3%	43.6%	0.818	3.13	1.98	1.99	1.92
Privacy	20	3.52%	49.5%	0.884	2.80	2.12	2.12	1.98
Software	17	6.98%	45.9%	0.658	3.06	2.09	2.04	1.87

Table 6.3: This table summarizes our advice-quality findings by topic. Unique and volume describe the number of unique advice imperatives about that topic and the proportion of all 2780 pieces of advice in our corpus that are about that topic, respectively. Cloze is the mean Smart Cloze score and ease is the mean of the median ease scores for documents containing advice about that topic, respectively. Ease score of 0 is “Neither hard nor difficult,” an ease score of 1 is “Somewhat easy,” and a score of 2 is “Very easy.” The actionability metrics are the means of the median actionability sub-metrics for the advice about the given topic: where 1 is “not at all”, 2 is “slightly”, 3 is “somewhat” and 4 is “very” (e.g., a mean score of 2.27 for confidence indicates that, on average, participants were between slightly and somewhat confident about implementing advice on this topic). Effective is the proportion of advice on that topic that was rated accurate by the majority of experts and risk is the mean of the median risk ratings for the advice about that topic.

6.6 Discussion

The work in this chapter provides a comprehensive taxonomy of 400 end-user security behaviors, including 204 behaviors not previously catalogued in the literature. The full list of behaviors can be explored here: `https://advicequality.github.io`. Table 6.3 summarizes our quality findings across the 12 topics of security advice we identify.

We systematically evaluate the comprehensibility, perceived efficacy, and perceived actionability of our corpus. Overall, we find that while the average document in our corpus is *perceived* at first glance as 'somewhat easy' to read, it has only low to partial comprehensibility to the general public. On average, documents contained between five and six pieces of security advice, and all documents contained at least one piece of advice perceived by experts as useful in reducing people's security risk. General news channels, nonprofit organizations (both technology and non-technology focused), subject matter experts, and security and computer repair companies provided the most comprehensible documents.

There was, however, variance within these categories. Let us consider as an example the documents from the U.S. government, which provided the most advice in our corpus (205 unique imperatives and 2112 documents). While documents from the U.S. government had a mean Cloze score of 47%, those from `ic3.gov`, `whitehouse.gov`, `ftc.gov`, and `dhs.gov` performed significantly better (achieving at least partial comprehensibility) than those from other government providers. Similarly, despite the high comprehensibility of documents from some technical nonprofit organizations such as Tor, documents from the EFF and Tactical Tech – which provided nearly 20% of advice in our corpus – had mean Cloze scores below 50%. While it is not necessarily problematic for more technical content such as that from academic security publications and security forums to be incomprehensible to the average person, low readability of content from organizations such as the Library Freedom Project, MOOCs, Facebook Help pages, and Technical Q&A websites may make it difficult for non-experts to stay secure.

We establish axes of security advice quality. First, we find that our metrics correlate with (reported) adoption, lending support for the importance of the advice quality factors we have operationalized. We find that all four of our actionability sub-metrics correlate with reported behavior adoption by users. Additionally, we find that priority ranking — one of our metrics of efficacy — strongly correlates with reported adoption, as well. General users' and experts' priority rankings strongly correlate with their respective reported adoption.

Second, our results support that our quality metrics are discriminant: that is, they measure different components of advice quality. For example, while network security was least readable and also had low actionability, data storage did quite well on readability while rating consistently low on actionability (lowest confidence, most disruptive, most difficult). Similarly, documents containing advice about software security and antivirus were among the more difficult to read, but not considered high in implementation difficulty, indicating that the readability of the document in which the advice occurs is different from the actionability of the advice it contains.

On the other hand, experts' perceived assessments of accuracy are not as discriminant as we had hoped: experts evaluated 89% of the advice in our corpus as accurate, with median risk reduction reported as 37.5%, therefore failing to distinguish between more and less helpful advice. Further, because new attacks are common and proving security (as opposed to vulnerability) is difficult, recommended security behaviors tend to accumulate and are rarely deprecated [113]. These results point to a need for direct measurement of the efficacy of security behaviors, in order to distinguish the most and least useful. Further, future work may also seek to explore mechanisms for tailoring advice to users' threat models and personas: the advice needed to protect someone engaging in activism or who is at risk of being doxxed may not be appropriate for a low-exposure user.

We expand on the results of prior work. About half (196) of the behaviors we identify were referenced in prior literature addressing security advice. Our findings support the results of prior work by Ion et al. [120,189], recently replicated by Busse et al. [38], who asked experts to provide their top recommendations of security advice and rate the quality of 20 pieces of advice. Two of the three behaviors given "number one" priority by our experts overlap with the top three behaviors suggested by experts in both papers: "update system" and "use unique passwords." The third-most-important behavior identified by both papers "use two-factor auth", is rated as a "top 3" priority by our experts and ranked #25 out of 374 across all of our advice. Both Ion et al. and the Busse et al. replication also found that experts rated most of the 20 pieces of advice evaluated in their work as "good," perhaps foreshadowing the results of our work. Further, both papers found that users' and experts' reports about the most important security advice differed; our experts' and general users' rankings of the priority of the security advice only weakly correlate, confirming this finding across a much larger set of advice.

We establish that the primary problem with security advice is that there is too much of it. The vast majority of advice we evaluated was perceived by experts as accurate. Even the small portion of advice that experts identify as harmful, they view as relatively low-risk (median risk increase of 10%). While experts were somewhat more discerning when prioritizing advice they would recommend, they still identified 25 pieces of advice as being in the top three. This may be due to unfalsifiability — experts being unable to identify whether a piece of advice is actually useful, or prove when it is not — or due to different goals and focuses between experts: an expert focused on account compromise may emphasize different information than a privacy guru, leading to lack of overlap in expert advice [189]. In either case, this overload of advice is leaving users struggling to implement hundreds of imperatives. Further, experts appear slow to update their advice: for example, two pieces of advice in our study that were perceived as harmful were related to changing and storing passwords, despite this advice having been updated and disproven in the most recent NIST standards [94].

Further, for users to even find this plethora of advice, they may need to filter out significant non-instructive information returned as a result of Internet searches or on the websites recommended by experts. Only 50% of documents we collected based on expert recommendations and Google search results contained actual security advice (as opposed to very general information or advertisements for security products).

We identify security topics that need better advice. While the primary outcome of this work is that we need less advice, we do note that a few topics of advice performed consistently worse than others across our evaluations. Advice about data storage topics (e.g., "Encrypt your hard drive," "Regularly back up your data," "Make sure to overwrite files you want to delete") scored poorly in actionability across our metrics. This raises questions about whether we should be giving this advice to end users in the first place, and if so, how these technical concepts can better be expressed in an actionable way. Network-security advice performed nearly as poorly, especially on user ratings of confidence, time consumption and difficulty.

This is perhaps even more concerning, as the advice on network security is far more general (e.g., “Use a password to protect your WiFi,” “Secure your router,” “Avoid using open Wi-Fi networks for business, banking, shopping etc.”).

Privacy advice was more of a mixed bag. While a quarter of the advice about privacy was rated as un-actionable, a significant proportion of the other privacy advice scored quite high on actionability. Experts were less positive toward any privacy advice, with no advice about privacy being rated among the top 3 practices experts would recommend. As privacy becomes increasingly important, and prominent in users’ awareness, there appears to be significant room for improvement.

In sum, our results suggest that security advice is struggling from a crisis of unfalsifiability [112]—experts are unable to identify the most critical advice and prioritize it—thus leaving users to hopelessly attempt to follow hundreds of different behavioral imperatives [111]. We establish that the prioritization of advice is strongly correlated to self-reported adoption of that advice, underscoring the criticality prioritization. While issues of comprehensibility and actionability are present, the glaring problem appears to be with the overwhelming quantity of—even reasonably well constructed—advice.

Chapter 7: Methods: Comparing the Validity of Security Self-Report and Log-Data

A commonly raised question regarding survey data is whether people are able to accurately report their experiences around digital security. To address this concern, in this chapter[1] I describe a study that I conducted with my coauthors in order to compare the validity of survey data to log data in a security context [188]. This work provides support for the validity of self report data for asking particular security questions, including broad questions about e.g., negative experiences such as those considered in Chapter 5.

In prior work in security, results from user studies, although valuable, have not always translated to the real world: Fahl et al. found that password creation studies only somewhat reflect users' actual choices [74], and researchers from Google found that the best TLS warning messages identified by surveys did not always pan out in A/B field tests [21]. There are a number of possible reasons for such discrepancies, including: (1) despite the best efforts of the research teams, the user studies may not have been designed most optimally to elicit accurate reports; (2) the user studies may have not been conducted with a sample that effectively represents the actual

[1] Published as [188].

user population; (3) people may not know themselves well enough to accurately report on their in-the-wild behavior; or (4) the environment of user studies may simply not be effective for answering certain types of questions.

Other fields face similar challenges. For example, public health researchers who wish to measure and understand risky behaviors—- e.g., heavy drinking, unprotected sex, smoking—- often use surveys to measure the frequency of these behaviors and identify correlated factors to target with interventions [133, 147]. To enable good outcomes from these surveys, survey methodology researchers have painstakingly investigated how different survey designs and samples affect responses, and how these responses reflect real-world behavior [48, 77, 131, 134, 141, 224]. They discovered that cognitive biases, such as difficulty predicting behavior for hypothetical future situations, or reluctance to report socially undesirable practices, affect survey results [133, 173]. To compensate for these biases, researchers developed new methods and best practices that can be used to obtain more accurate measurements [50, 131, 224].

It is not clear whether these compensatory approaches will translate to the digital security and privacy domain. For example, best practices from warning design literature did not automatically translate to better security-warning comprehension [75]. Prior work comparing survey samples also suggests that using online samples to ask questions about online security and privacy has unique biases that must be accounted for [123, 185]. Research is therefore needed to understand how user study data deviates from real-world observations, in order to understand how to best mitigate and correct these biases. So far there is unfortunately little such work comparing user study results to the real world [21, 74, 152].

The work presented in this chapter takes an important step toward more thoroughly measuring biases between digital security surveys and real-world security practices. To explore the validity of self-report survey data in the security context we systematically compare real-world measurement data to survey results, focusing on an exemplar, well-studied [24, 64, 67, 73, 91, 100, 120, 148–150, 155, 156, 159, 165, 191, 198, 205, 223, 228, 229, 232, 236] security behavior: software updating. We align field measurements about specific software updates (n=517,932) with survey results in which participants respond to the update messages that were used when those versions were released (n=2,092). Specifically, we compare the results of log data of user behavior in response to 11 different software updating messages collected using the WINE system [66], to responses to a survey asking respondents to self-report their intended behavior and reasoning for updating in response to the same messages. To better understand self-report biases and sample source effects, we tested two different framings for our survey questions and collected responses from two different sources (n=2,092: 1,751 responses from Amazon Mechanical Turk and 455 from a demographically census-representative web-panel sample of U.S. Internet users).

Our ultimate goal was to examine whether insights about our exemplar security behavior (software updating) derived from survey data match well with real-world results, and whether any deviation we observe wass sufficiently systematic to be corrected in a straightforward manner. To this end, we quantified differences in reported and measured patching delay in response to the same update messages.

We also examined whether features previously identified by prior work as important to update decisions—- text of update message, length of update message, prior negative experiences with updates, and whether a restart is required—- produced similar effects in both the survey and measurement data.

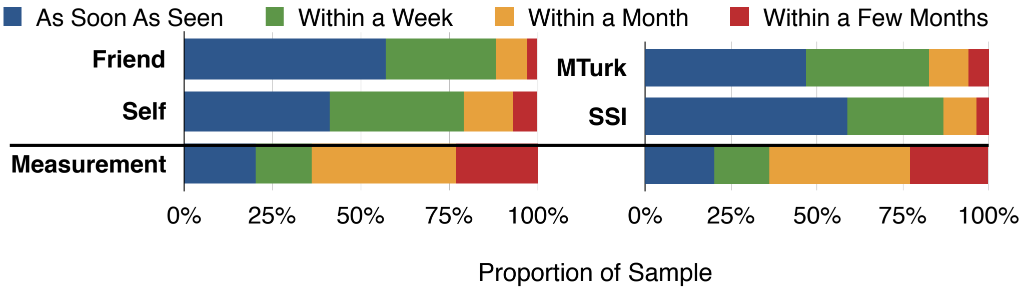


Figure 7.1: Comparison of self-reported update speeds by framing condition (left, full survey dataset) and survey source (right, per source over both framing conditions) to the measurement data.

For the most part, we observed systematic, consistent differences between the measurement results and the survey results. For speed of updating, survey respondents tended to report faster update speeds than we observe in reality, and survey framing matters: respondents asked to make a recommendation to a friend advised updating immediately, respondents reporting on their own behavior said they would update within one week, and measurement data indicates that in reality most users updated within a few weeks (Figure 7.4).

We also observed that surveys replicated the effects of high-level, user-specific factors—such as typical behavior, perception of risk, and negative experiences—identified in prior work, albeit sometimes with systematic differences in effect size. For example, in both survey and measurement data, past tendency to update is significantly correlated with speed of applying a new update; however, survey data shows a medium effect size, while measurement data shows only a small effect. On the other hand, we find that survey data does not well represent factors that require careful reading of update messages, such as the length of update message or whether they mention needing to restart. This may reflect that respondents are not reading carefully, that they are not accurately assessing which features drive their real-world decisions, or that they are not yet aware of the causes of their behavior.

Overall, our results suggest that security-related surveys are relatively accurate at representing perceptive real-world effects, but survey questions about specific message features appear not to work well for proxying real-world effects.

7.1 Research Questions and Datasets

To understand biases in self report data about digital security behavior, we conduct an in-depth comparison of empirical observations of host-machine updating behavior collected using the WINE system [66] (n=517,932) to survey data eliciting self-report responses (n=2,092) to the same update messages.

In particular, we address the following research questions:

RQ1: How well do self-reported security-behavioral intentions correlate with observed field data?

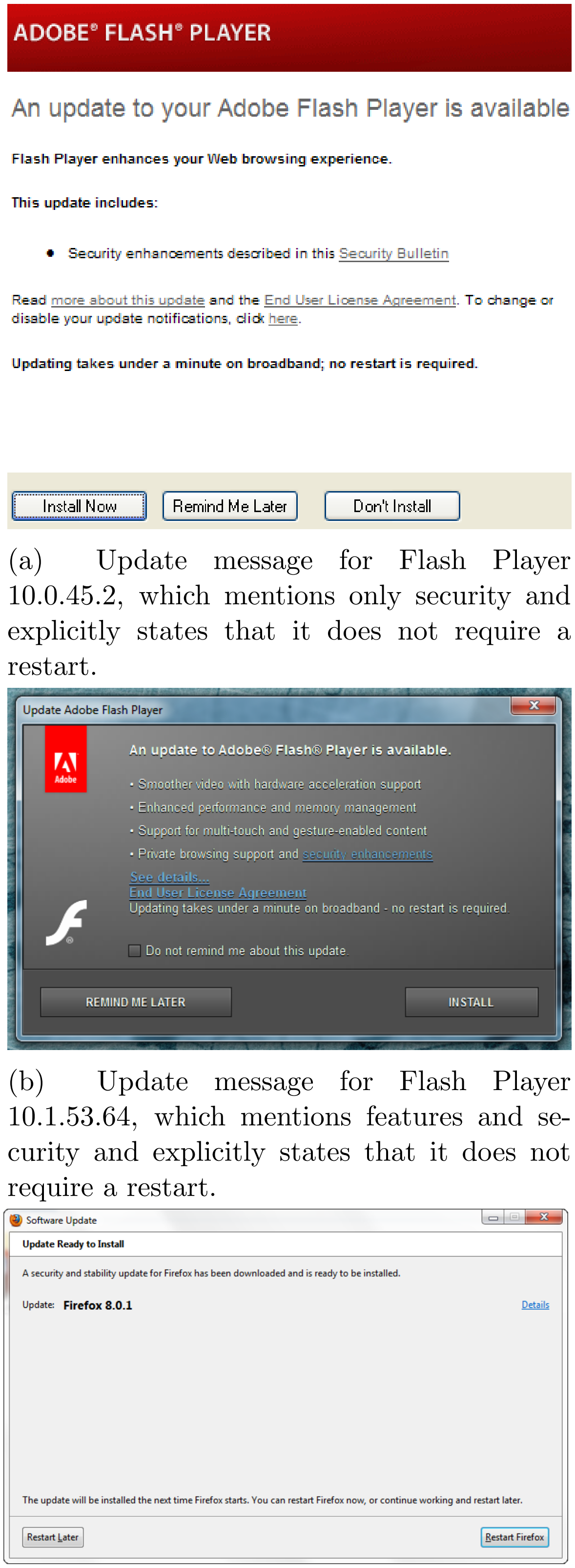
RQ2: How does framing the question in terms of the respondent's own behavior, as compared to behavior recommended for a friend, affect this correlation?

RQ3: How does sample source (i.e., demographic representativeness) affect these correlations?

RQ4: How does the correlation between self-reports and measurement data differ for research questions relating to general perceptions and behaviors, as compared to research questions related to the update messages, which require respondents to carefully read specific, displayed information?

In this section, we connect these research questions to our data sources: the update messages for which we analyze behavioral and self-reported responses, the field measurement data we obtained, and the survey data we collected.

In order to compare the measurement and survey data, we want to contrast self-reported responses to a given update message to observed behavior when encountering the same message. To this end, we needed to find images for update messages in our field data.(See details of measurement data in Section 7.1.1 below.) Because neither our measurement dataset nor application release notes archive the images that were displayed to users when various updates became available, we instead searched for update messages by performing Google image searches and asking IT staff at two universities for any saved screenshots related to updates. In the end, we were able to obtain 11 messages for which we had patching records in our measurement dataset: six Adobe Flash messages, one Firefox message, two Adobe Reader messages, and two Opera messages. All messages were for updates released between 2009 and 2012 (in Section 7.3 we evaluate and discuss potential time confounds). Figure 7.2 shows three messages; Appendix B.14 shows the remainder.



An update to your Adobe Flash Player is available

Flash Player enhances your Web browsing experience.

This update includes:

- Security enhancements described in this Security Bulletin

Read more about this update and the End User License Agreement. To change or disable your update notifications, click here.

Updating takes under a minute on broadband; no restart is required.

(a) Update message for Flash Player 10.0.45.2, which mentions only security and explicitly states that it does not require a restart.

An update to Adobe® Flash® Player is available.

- Smoother video with hardware acceleration support
- Enhanced performance and memory management
- Support for multi-touch and gesture-enabled content
- Private browsing support and security enhancements

See details...
End User License Agreement
Updating takes under a minute on broadband - no restart is required.

□ Do not remind me about this update.

(b) Update message for Flash Player 10.1.53.64, which mentions features and security and explicitly states that it does not require a restart.

Update Ready to Install

A security and stability update for Firefox has been downloaded and is ready to be installed.

Update: **Firefox 8.0.1** Details

The update will be installed the next time Firefox starts. You can restart Firefox now, or continue working and restart later.

(c) Update message for Firefox 8.0.1.4341, which states that a restart is required to install the update, and mentions both stability and security.

Figure 7.2: Examples of Update Messages

7.1.1 Measurement Data

We use patch deployment data sets from the Worldwide Intelligence Network Environment (WINE) [66], a platform for accessing Symantec field data for cybersecurity. WINE collects data from machines that have installed home (as opposed to corporate) versions of Symantec security products, and is designed to ensure that the available data is a representative sample of data collected by Symantec [66]. Symantec makes measurement data collected using WINE from 2008 to 2014 available to researchers.

Our dataset includes records of the timestamp when specific files first appear on a given machine. We use data from Nappa et al. [159] to map software version updates to specific file hashes, allowing us to identify when a particular software patch was installed. We can therefore calculate updating speed as the time interval between patch release time and installation time, for a given patch version and machine.

We also use WINE log data to measure features of individual hosts, such as their history of update responses, history of crashes for particular applications and for the entire system, and whether or not specific applications are installed.

Sampling Measurement Data To obtain an appropriate subset of the measurement data, we selected only hosts for which we have a record that one of the 11 update versions we target was eventually installed. We then remove any machines whose local time is visibly incorrect: in particular, where the patch time is one day or more earlier than the actual patch release date. We retain only U.S. users, for ease of survey sample matching and reliability to findings from prior self-report work (nearly all of which were conducted with U.S. respondents).

Finally, we note that one machine can have multiple records in the data, if more than one of our eleven targeted updates was applied on the same machine. These repeated measures would complicate statistical analysis, particularly because we only have multiple records for a minority of hosts, so it would be difficult to account for them using standard methods. Instead, we randomly select only one of the available records for each host where multiple events were available. This random selection is performed last, after all other filtering steps, which selects 517,932 out of 730,270 update events that correspond to our 11 messages.

7.1.2 Survey Data

To compare with the measurement data, we collected self-report data about users' intended updating behavior using a between-subjects survey. Each survey began by showing the respondent exactly one of our 11 update messages; respondents were only shown update messages for an application which they reported either using or having on their device within the past 5 years. Section 7.3 provides more details on the demographic comparability of the survey and measurement samples.

The respondent then answered questions about how quickly they would apply the indicated update (RQ1) and then the reasoning behind their decision (RQ4). Appendix A.9 shows the questionnaire.

Framing (RQ2). RQ2 addresses one possible source of potential discrepancy between survey results and real-world phenomena: social-desirability bias—which from respondents' beliefs about the proper or expected answers to survey questions—and personalization biases that arise from respondents' having difficulty accurately assessing their own behavior [80]. To investigate this, respondents were randomly assigned to one of two framing conditions: *self*, where they answer questions about their own intended behavior, or *friend*, in which they answer questions about what behavior they would suggest to a friend.

To measure self-reported updating speed, *self* respondents were told to "Imagine that you see the message below appear on your computer," and the update message image was displayed. They were then asked (on the same survey page) whether they would intend to update this application, with the following answer choices: "Yes, the first time I saw this message," "Yes, within a week of seeing this message," "Yes, within a few weeks of seeing this message," "Yes, within a few months of seeing this message," "No," and "I don't know."

In contrast, *friend* respondents were told to"Imagine that a friend or relative sees the message below on their computer and asks you for advice," and the update message image was displayed. These respondents were then asked (on the same survey page) how soon, if at all, they would recommend that their friend updated their application.

We hypothesized that asking about friends would provide respondents with a more neutral, less personal scenario. Asking about friends is a well-known tactic in behavioral economics and survey methodology for obtaining such normative judgements, and has previously been applied in human-centered security research [41,80,125,162].

Recruitment (RQ3). To address RQ3, we collected responses to our survey using two sampling platforms: Amazon Mechanical Turk (MTurk) and Survey Sampling International (SSI).

Respondents from MTurk were invited to take a survey about online behavior, and were paid $0.50 for completing the brief (¡5 min) survey. MTurk is known to produce demographically biased survey samples [121,167,196]; however, it is the most commonly used sampling platform in security research. In line with findings from prior work about response validity, we recruited only Turkers with 95% approval ratings [169].

Respondents recruited through SSI were sampled such that the demographic makeup of the respondent pool closely matched the demographics of the U.S. with regard to age, education, gender, race, and income (demographics for our SSI sample are shown in Appendix B.15). Such census-representative samples are expected to provide results more generalizable to the U.S. population [49]. SSI respondents took an identical survey to that shown to the MTurk respondents and were paid according to their agreement with SSI (compensation often takes the form of charity donations, airline miles, or cash).

We obtain a final survey sample of 2,092 respondents who use antivirus software and Windows computers (we refer to this dataset as the "full" survey dataset), which consists of 1,751 from Amazon Mechanical Turk (the MTurk dataset) and 455 from SSI (the SSI dataset).

Validity. To ensure that our survey was representative of surveys in the field, we drew our survey questions from prior work related to software updates [120, 149, 228, 229, 232], in some cases with slight modifications to specific questions. As described below, we selected and modified these pre-existing questions as needed to most closely match measurements available in the WINE data.

To maximize construct validity and ensure that our survey was easy for respondents to interpret, we conducted six cognitive interviews [173, 233] with a demographically diverse set of respondents. In these interviews we asked respondents to "think aloud" as they answered the survey questions and probed them on areas of uncertainty. We updated the survey after each interview and continued conducting interviews until areas of uncertainty stopped emerging.

7.2 Experimental Approach

Using the datasets described above, we developed experimental approaches to answer each of our research questions.

For all analyses, we use the updating speed measurement data and the main updating speed survey question defined in Section 7.1.2 above. We exclude any respondents who report that they would not install an update (n=138 in MTurk, n=64 in SSI) or that they do not know (n=45 in MTurk; n=19 in SSI), because we are unable to identify a parallel population in the measurement data.

Throughout our analysis, we apply Holm-Bonferroni correction as appropriate to account for multiple-testing effects [116].

7.2.1 RQ1—3: Comparing measurement and survey data

Our primary goal, encapsulated in RQ1, was to understand how well self-reported survey data can proxy for field measurements when considering users' security behavior. More specifically, we wanted to know whether, even if self-report data is not entirely accurate, it deviates systematically enough that it can still provide a useful understanding of end-user behavior. In the process, we compare across framing conditions (RQ2) and across sample sources (RQ3).

The answers to this updating speed question are thus treated as a 4-point Likert measurement. To align the survey answer choices with the measurement data we bin the measurement results to match the Likert responses: as soon as I see the message is equivalent to updating within 3 days, within a week is equivalent to updating between 3 and 7 days after the patch appears, within a few weeks is equivalent to updating between 7 and 30 days, and within a few months is equivalent to patching in 31 days or more.

To compare the update speeds observed in the measurement data and reported in the survey data, we use a X^2 proportion tests—which are robust to sample size differences—to compare updating speeds in the measurement and survey data, both over the full survey dataset and both conditions (RQ1), the full dataset by condition (RQ2), and by sample (RQ3). For the per condition and per sample comparisons, if the omnibus (e.g., friend vs. self vs. measurement) is found to be significant, we conduct planned pair-wise comparisons: RQ2: friend vs. measurement and self vs. measurement on the full dataset; RQ3: MTurk vs. measurement and SSI vs.

measurement, and a replication of the RQ2 analysis on the separated MTurk and SSI datasets, respectively.

7.2.2 RQ4: Comparing Question Types via Factors That Affect Updating

RQ4 investigates how the relation between self-report and measurement data is affected by the type of construct being measured.Within our exemplar context of software updates, we identified two types of constructs: general constructs, such as how often the respondent typically updates, or how often the respondent's computer typically crashes, and detailed constructs, such as self-reporting in the presence of a subtle experimental manipulation, such as the differences in the text of the update messages we tested.

For this investigation, we examine features that have been found in prior work to be relevant to update speeds and decision-making, and that were obtainable in our datasets:

- the **application** being updated;
- the **cost** of installing the update, in terms of whether it requires a restart;
- whether the update mentions **only security** (as opposed to other features) [2];
- the **length** of the message;
- the **risk** associated with the update, typically informed by the user's prior **negative experiences** with updating and stability;

[2] All of the messages we collected mentioned security, thus we compare the effect of mentioning *only* security to mentioning both security and other enhancements, as prior work suggests that user may be wary of additional enhancements [228].

	Feature	Measurement	Survey	Prior Work
Detailed Constructs	Application	Source application.	Same as measurement.	[149,150,223, 228]
	Update Cost	Whether the update mentions requiring a restart.	Same as measurement.	[150]
	Security-Only	Message mentions security but not features or stability.	Same as measurement.	[73, 100, 149, 223]
	Message Length	Number of words in message.	Same as measurement	[223]
General Constructs	Update Risk	Negative experiences characterized by two different features: average number of application and system crashes per week over past one year and the average change in crashes for the application and the overall system before and after the past updates within one year.	Responses to four survey questions about experiences with application and system crashes in general and related to updates of this application.	[95, 149, 150, 229]
	Tendency to Update	Mean updating speed for prior patches from the same application.	Responses to the following survey question: "In general, how quickly do you install updates for applications on your computer or for your computer itself (e.g., the computer operating system)?"	[191,198]

Table 7.1: Summary of the factors considered in our models, how they were operationalized in each dataset, and from what related work they were drawn.

- and the user's prior history of updating speed, which we refer to as **tendency to update**.

Table 7.1 summarizes how we instantiate these factors in each dataset, as well as which related work supports their inclusion.

The first several features—- application being updated, whether a restart is required, whether security is the only feature mentioned, and message length—- are determined by the update message under consideration. Table 7.2 summarizes the update messages we collected according to these features. Messages were considered to be security-only if they mentioned that the patch addressed security issues and

Version	Application	Release date	Security Only	Requires Restart	No. of Words	Risk Metrics Available
10.0.22.87	Flash	2/24/2009	✓		57	
10.0.45.2	Flash	2/11/2010	✓		57	
10.1.53.64	Flash	6/3/2010			48	
10.2.152.26	Flash	2/8/2011			55	✓
10.3.181.14	Flash	5/12/2011			50	✓
11.0.1.152	Flash	10/4/2011			57	✓
9.3.2.163	Reader	4/13/2010		✓	35	
9.5.1.283	Reader	4/10/2012			23	✓
10.61.3484.0	Opera	8/9/2010		✓	80	
11.64.1403.0	Opera	5/10/2012		✓	80	✓
8.0.1.4341	Firefox	11/22/2011		✓	45	✓

Table 7.2: Summary of update messages.

made no mention of features or stability. For example, Figure 7.2a shows a security-only message, while the message in Figure 7.2b mentions both security and features. Message "cost" was characterized by whether the message mentioned requiring a restart (e.g., Figure 7.2a states that it requires no restart, while Figure 7.2c states that a restart is required). If restart is not mentioned in the message, then we consider it as "not required" since users are likely unaware of restart. Finally, message length was characterized as the number of words in the message.

We consider the first four features to be "detailed constructs," especially security only, restart, and the length of the message, which require respondents to be paying close attention to the displayed messages. The last two features: risk informed by prior experiences and general tendency to update, are "general constructs."

In order to isolate the effects of the detailed constructs as much as possible we identified sets of messages to compare:

- Application effects: we use the full dataset to compare effects among the four applications
- Cost effects: we compared the two Adobe Reader messages to each other, as

one message mentioned a restart requirement and the other did not. (This is the only pair of messages with this within-application variation). The Reader messages were otherwise quite similar (same description of security and stability enhancements, same application), although the number of words in the messages did vary.

- Effect of message mentioning only security: we compared the six Flash messages to each other. Two of the six messages mention only security, while the other messages mention additional enhancements. Additionally, one of the security messages is the same length as a message that mentions security and features, allowing us to include message length in our model and control for this factor. All mention that you do not need to restart.
- Message length: we also use the Flash messages, as they have the largest variation in length and are similar on all other features, as just described.

The remaining features—- update risk and tendency to update—- are user-specific, and thus were inferred from measurement results and survey responses. For these two features, we compare messages within applications, to control for potential application effects, and between applications, to control for covariance with other features.

7.2.2.1 Inferring Features from Measurement Data

Risk Metrics. We characterize update risk in terms of a user's prior experience with overall stability, as well as specifically how updates affect stability.

To measure this in the measurement data, we use WINE's *binary stability* dataset, which records both system crashes and application crash/hang events.

We define four risk metrics:

- Average weekly frequency of system crashes and hangs during the year before the user installs the target patch.
- Average weekly frequency of crashes and hangs for the target application during the year before the user installs the target patch.
- Average change in the number of system crashes and hangs between the week before and the week after a new patch was installed. Averaged over all updates of the target application installed in the year prior to installing the target patch. If the average is positive, we consider this an increase in system crashes post-update.
- Average change in the number of crashes and hangs between the week before and the week after a new patch was installed. Averaged over all updates of the target application installed in the year prior to installing the targeted patch. if the average is positive, we consider this an increase in application crashes post-update.

The former two metrics are used to capture the overall crash tendency of the system or application, while the latter two are used to capture the user's past negative experience in system/application crashes when they update the applications.

For ease of analysis, we center and normalize the raw crash counts. This data was only collected starting in 2011. Thus, we are only able to obtain stability features for the 5 update messages, 30,623 users, as indicated in Table 7.2.

General Tendency to Update. We define general tendency to update as the average update speed for all versions of a given application prior to the targeted update. Let V_N be the selected version, such that $\{V_1, V_2, \cdots, V_{N-1}\}$ are the prior versions. $D(v, m)$ is the speed of updating version v for machine m. Then the tendency to update for machine m is calculated as $\frac{1}{N}\sum_{n=1}^{N} D(V_n, m)$.

7.2.2.2 Inferring Features from Survey Data

Risk Metrics. To assess perceived prior negative experience with updating—- specifically around crashing risk—- we asked respondents a series of four questions. The first two were "Over the past year, how frequently do you feel like [application for which patch message is shown] has frozen (e.g., hang) or crashed?" and "Over the past year, how frequently do you feel like any application on your computer or your computer itself crashed?" Both questions provide answer choices on a four point scale: "Less than once a week", "At least once a week but not more than three times a week," "At least three times a week but not more than five times a week," and "Five times a week or more."

We also asked, "Over the past year, have you noticed that updating [application for which patch message is shown] changes how frequently it freezes (e.g., hangs) or crashes? and "Over the past year, have you noticed that updating [application] changes how frequently any application on your computer or your computer itself crashes?" These questions had the following answer choices: "Yes, my computer crashes more after I update," "Yes, my computer crashes less after I update," and "No, updating [application] has no impact on how frequently my computer crashes."

General Tendency to Update. We assessed tendency to update by asking respondents "In general, how quickly do you install updates for applications on your computer or for your computer itself (e.g., the computer operating system)?" with answer choices: "As soon as I see the update prompt," "Within a week of seeing the prompt," "Within a few weeks of seeing the prompt," "Within a few months of seeing the prompt," "I don't install updates that appear on my computer," and "I don't know." This question was constructed to be similar to a question asked by Wash and Rader [232].

7.2.2.3 Statistical Modeling to Compare Effects of Relevant Factors

To compare the effect of factors suggested by prior work as related to people's updating behavior between the survey and measurement data, we construct ordinal logistic regression models, which accommodate Likert outcome variables such as our measure of update speed [163].

We construct one set of models to examine the detailed constructs; these models include all survey and measurement data for the messages being considered. We also construct a second set of models to examine the risk metrics, as these metrics were only available in the measurement data for five of our 11 messages. We refer to these as the *detailed* and *risk* models respectively.

To best isolate the effects of the individual constructs, we use a hierarchical modeling approach. We construct a baseline model and then add feature sets so we can examine their impact in isolation [234]. For both detailed and risk model sets, our baseline models contain a single feature: general tendency to update.[3] We then add sets of features to examine the constructs of interest. Specifically, for the detailed constructs, we construct the following models:

- Across All Applications (Construct of Interest: Application)
  - Baseline: General Tendency (ordinal DV, four-point scale)
  - General Tendency and Application (categorical DV, Flash is the baseline)
- Reader (Construct of Interest: Cost)
  - Baseline: General Tendency
  - General Tendency and Cost (boolean DV, whether the message mentioned a Restart)
- Flash (Constructs of Interest: Length, Security)
  - Baseline: General Tendency
  - General Tendency and Length (continuous DV, number of words)
  - General Tendency and Security-Only (boolean DV, whether the message mentioned anything other than security)

[3] To further address RQ3 and control for sample effects, we also include survey sample source as a factor in the survey models.

  - General Tendency, Length, Security-Only: constructed to control for covariance between length and security-only

For the risk model set, we construct models across all applications. The baseline model for each consists of general tendency to update, and for the survey data, sample source. The risk model for each consists of the four risk factors: frequency of system and application crashes (ordinal DVs) and existence of an increase in system crashes and application crashes post-update (boolean DVs), and controls for general tendency to update and application.

To ensure model validity, we performed backward AIC selection on the baseline model in each case (retaining the baseline factors in all cases). For each model we report the log-adjusted regression coefficients, known colloquially as *odds ratios* (O.R.s), and indicate significance (p-values < 0.05). To further examine RQ3, we include the sample source (MTurk or SSI) as a factor in all of our survey regression models.

7.3 Dataset Comparability and Limitations

We next discuss threats to validity related to our datasets and experimental approach.

Sampling The majority of WINE hosts are located in the United States. For consistency, we sample only U.S. survey respondents and include only U.S. WINE hosts in our analysis. Additionally, we recruited only survey respondents who use

Windows devices, as all WINE hosts are Windows. Finally, we conduct our modeling using only those survey respondents who reported using antivirus software, in order to closely mirror the measurement data (eliminates 416 respondents).

Differences in Timing One crucial confounding factor in our analysis is the difference in time between when the measurement and survey data were collected. The measurement data available from Symantec was collected from 2009 to 2013, while the survey data was collected in 2018. We attempt to quantify the importance of this time delay by investigating how time affects each dataset.

To understand how updating frequency in the real world has changed over time, we tested the effect of time in measurement data. The effect is significant, but of small size (X^2=72412, $p < 0.001$, V=0.181). Additionally, although time is a significant factor, the effects are not in a consistent direction for each application (Figure 7.3): Opera is updated significantly faster in 2012 than in 2010, while Reader is updated slower in 2012 than in 2010; Flash is updated slower in 2010 than in 2009 and faster in 2011 than in 2010. Given the inconsistencies in these time biases, we do not suspect that time will create systematic biases in our results.

To evaluate whether self-reports about updating frequency have changed over time, we compared our results with the oldest work with comparable data [232]. Wash and Rader conducted a census-representative survey of 2000 people, in which they asked respondents to report their general updating frequency, also on a five-point Likert scale. Using a Mann-Whitney U test, standard for Likert scale data [146], we find no significant difference in updating frequencies between their results and

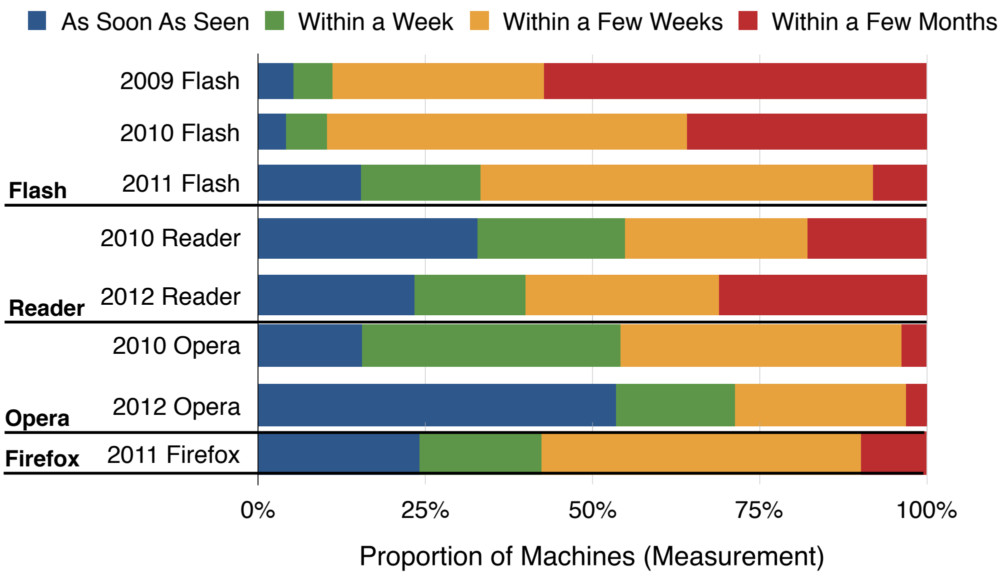


Figure 7.3: Measurement update speed by year and application.

our survey.

Thus, while time confounds are possible,we hypothesize that they are unlikely to be so significant as to invalidate our results. Taking into account that real-world data of the size and quality provided by WINE is rarely available, we argue that our analysis can provide many valuable insights despite this potential confound.

Machines vs. Users The measurement data measures machines, while our survey data measures users. For our analysis, we assume that there exists a one-to-one mapping in the measurement data between machine and user, but it is of course possible that one user manages multiple machines. Although we cannot determine how many of these cases may exist, we believe the effect of this should be relatively minimal given the large size of our dataset. Additionally, it is possible that some hosts in the measurement data are not personal computers, but rather corporate-managed machines. However, machines managed by large organizations typically use an enterprise Symantec product and therefore are not recorded by WINE. The percentage of corporate managed machines *not* using the enterprise software is anticipated to be quite low [159].

Self-Report Biases As is typical of survey studies, self-report methodologies have a number of biases and limitations. For example, social-desirability bias, where people report what they think will make them seem most responsible or socially desirable [133]. However, it is important to note that in this study, we wanted specifically to compare the survey results, which are inherently biased in some ways, with the measurement data, which is inherently biased in other ways. We apply best practices for extensively pre-testing our survey, randomizing answer choices, and placing demographic questions last. Biases which are not mitigated by these steps are therefore a key aspect of our results.

Generalizability Finally, our work has three potential threats to generalizability. First, we sample only antivirus users. However, as antivirus users are estimated to make up at least 83% of the online population (see Chapter 4) and it is unlikely to be able to draw a truly random sample of log data, we consider this population to cover the population of Internet users relatively well. Second, we examine only software updating behavior. As such, we can indeed only hypothesize about similar bias effects in other security behaviors. We opt to provide detailed, in-depth analysis of a single behavior rather than more cursory analysis of multiple behaviors; this follows the approach of nearly all prior work in survey methodology, which tends to consider one behavior (e.g., smoking) at a time to enable thorough analysis. Third and finally, automatic updates have been growing in adoption since the time when our measurement data was collected. However, automatic updates may still offer users a choice to delay and require user-controlled application restarts. Thus, users still must make time-related software update choices, even if they may not have the option to chose *whether* to update.

7.4 Results

Below, we detail our findings by research question.

7.4.1 RQ1—3: Speed, Framing, and Sampling

	Comparison	X^2	**p-value**
RQ1	Measurement vs. Survey	103630	< 0.001
RQ2	Omnibus: Measurement vs. S: Self vs. S:Friend	103730	< 0.001
	Measurement vs. S: Friend	103310	< 0.001
	Measurement vs. S: Self	102850	< 0.001

Table 7.3: X^2 tests comparing the speed of updating reported in the surveys (S) with the speed of updating observed in the measurement data (WINE).

We start by examining self-report biases in estimating update speed.

RQ1: Updating Speed. To obtain an overall comparison between survey and measurement data, we compare the full survey dataset (which consists of responses from both the MTurk and SSI survey samples, across both framing conditions) with the measurement data. We find a significant difference ($X^2 = 103630$, $p < 0.001$)

between the combined survey responses and the measurement data: the median update speed in the survey data is “Within a week” (Likert value 2), while the median speed in the measurement data is “Within a few weeks” (Likert value 3).

RQ2: Survey Framing. To examine the effect of the survey framing, we separately compare the *friend* and *self* conditions (described in Section 7.2.1) to each other and to the measurement data. (This comparison also combines both sample sources.) We find significant and consistent differences in outcomes between our two survey framings (Table 7.3): median update speed in the Friend condition is “Immediately” (Likert value 1), compared to a median of “Within a Week” (Likert value 2) in the Self condition and “Within a Few Weeks” (Likert value 3) in the measurement data.

RQ3: Sample Comparison We also compare update speeds by survey sample. We find a significant difference between update speeds reported in the MTurk sample and those reported in the SSI sample ($X^2 = 1256.3$, $p < 0.001$). SSI respondents report a median update speed of “Immediately” (Likert value 1) compared to MTurk respondents who report a median speed of “Within a Week” (Likert value 2). Finally, the effect of the survey framing on the survey results for both samples is significant (MTurk: X^2=40.19, $p < 0.001$; SSI: X^2=16.5, $p = 0.009$).

Summary: systematic over-reporting of update speed in surveys; survey framing matters Figure 7.4 summarizes the results of our comparison of updating speeds reported in the two different survey framing conditions (friend vs. self) and samples (MTurk vs. SSI) against the measurement data. Overall, we find that survey re-

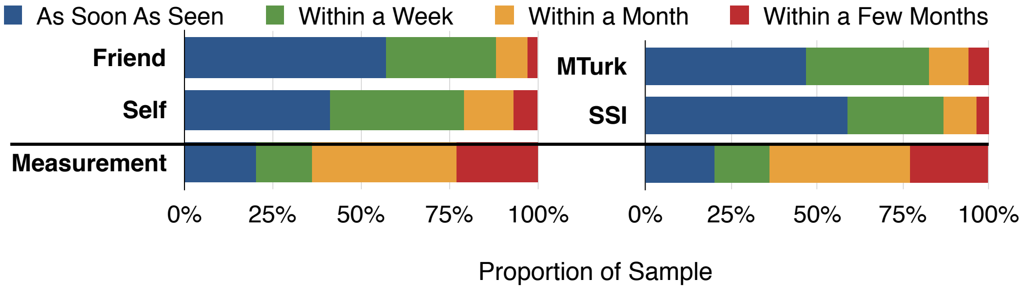


Figure 7.4: Comparison of self-reported update speeds by framing condition (left, full survey dataset) and survey source (right, per source over both framing conditions) to the measurement data.

spondents systematically report faster update speeds compared to the measurement data, and this bias is affected by survey framing. Finally, we observe reporting speed differences between the two survey samples: Perhaps surprisingly, the responses of the MTurk participants are somewhat closer to the measurement data than are those of the census-representative participants.

7.4.2 RQ4: Factors Affecting Update Speed

Next we examined the impact of various factors that prior work suggests may affect update speeds. To do so, we construct hierarchical regression models on both the survey and measurement datasets to compare variables of interest while controlling for other potentially relevant factors, as described in Section 7.2.2. In interest of brevity, we summarize the results here, and include in Appendix B.16.1 tables of regression results for all models constructed.

We detail our results by factor: general tendency to update, crash risk, and then the four message features. Finally, we review sample effects related to these factors (RQ3).

7.4.2.1 General Tendency to Update

In regression models for both the measurement and survey data, we find a significant relationship between general tendency to update and update speed for all applications. People who typically update more quickly, or report typically updating more quickly, are also more likely to report updating (or actually update) a given application faster. This is illustrated in Figure 7.5). This significant relationship holds in every model we test, for survey and measurement, both for the full dataset and for individual applications. However, the effect is larger in the survey data than in the measurement data: the odds ratios (O.R.s) for the survey models average 5.85 (SD=0.834), compared to 1.55 (SD=0.220) for the measurement data.

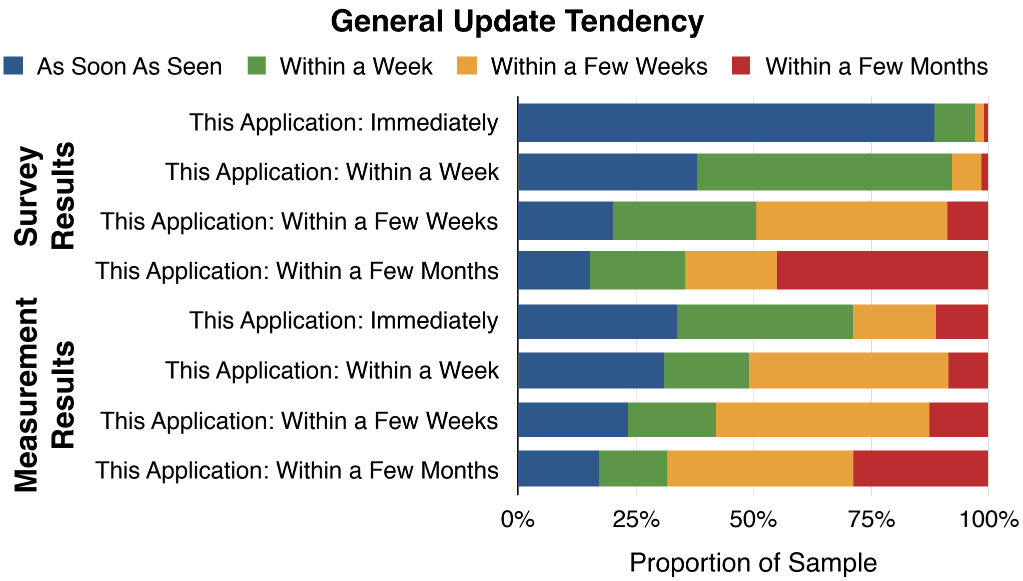


Figure 7.5: General tendency to update vs. update speed for a specific message in the survey (top) and measurement (bottom) data.

Summary: General tendency to update is significant in both datasets, but the effect is larger for survey data In sum, we observe that we would draw similar conclusions about general tendency to update being an important covariate from either the survey or the measurement data, but the effect sizes in the survey data are consistently larger than those in the measurement data.

7.4.2.2 Risk

We consider four risk metrics: average frequency of system and application crashes, and increases in system and application crashes after updating. In the measurement data, we observe mixed results regarding the relationship of these risk metrics to updating speed, finding a lack of consistency in which risk metrics, if any, are related to updating behavior; especially when controlling for other covariates. The relationship between prior negative experiences and updating speed was previously unstudied in measurement data.

In regression models controlling for general tendency to update and for the application being updated, we find in the measurement data that more frequent system crashes are associated with slower updating speeds (O.R.=1.03, $p = 0.005$), while increased crashes after prior updates are associated with faster updating speeds (O.R.=0.89, $p = 0.026$). These effects are fairly small. In contrast, in the survey data, none of the risk metrics show a significant relationship to updating speed.

To see if the discrepancy in results may have been caused by issues of respondent quality, we reconstruct our survey regression models using a smaller dataset of only "high-quality" survey responses. We borrow this approach from Fahl et al., who found that user study data more closely matched real-world data when filtering out low-quality responses [74]. In our context, we define low-quality responses as those who gave nonsensical answers: those who cited lack of restart as a reason to install an update message, but who saw a message did in fact require a restart (and reciprocally, those who cited needing to restart as a reason not to update, but who saw a message that did not require a restart) and those who cited like or dislike of features as a reason for installing, or not installing, but who in fact saw a message that mentioned only security (see Appendix B.16.2 for more detail). Examining the regression models built on this "filtered" survey dataset (n=981), we find significant effects, in the same directions as in the measurement data, albeit with larger O.R.s: perceived average number of system crashes (O.R. = 1.76, $p = 0.044$) and perceived change in crashes of the given application (O.R. = 0.53, $p = 0.440$) are related to self-reported update speed.

Summary: Risk effects replicated in survey data after filtering In sum, we observe a small but significant relationship between update speed in response to a particular message and crash risk factors in the measurement data. After filtering for respondent quality, we observe a similar effect in the survey data.

7.4.2.3 Message Features

We compare the effects of four features related to the message text: the *application* being updated, the *cost* of installing the update (whether it requires a restart), the *length* of the update message, and whether the message mentions *only security* or also other features or stability enhancements.

Application To examine the effect of the application on our results, we construct models over the full dataset, with application as a covariate. We find that the application is significantly related to the speed of updating in both the survey and the measurement data. The regression results for the measurement data show that Flash is updated more slowly as compared to Firefox (O.R.=0.66, $p < 0.001$) and Adobe Reader (O.R.=0.63, $p < 0.001$), and much more slowly than Opera (O.R.=0.29, $p < 0.001$). In the survey data, the overall effect is slightly smaller, but still significant: Firefox and Reader are have faster reported update speeds than Flash (O.R. = 0.82, $p = 0.048$; O.R.=0.81, $p = 0.007$). The survey model shows no significant result for Opera, however.

Cost: Reader To examine the effect of mentioning a restart requirement (implicitly suggesting a time or effort "cost" to the user) in update messages, we compare two Adobe Reader messages. We find that the message that mentions a required restart is updated more slowly in the measurement data than the message that does not mention such a cost (O.R. = 0.53, $p < 0.001$ in a regression model controlling for general tendency). In the survey results, this effect is not mirrored.

Length: Flash We compare the six Flash update messages to examine the impact of message length. In the update data, message length has a significant, albeit small effect on update speed: the length of the update message is significant both in the model that controls only for general tendency (O.R.=0.98, $p < 0.001$) and the model that also controls for mentioning security only (O.R.=0.93, $p < 0.001$); there are no significant effects in the survey data.[4]

Mentions Only Security: Flash Finally, the measurement data shows that users who saw one of the Flash messages that only mentioned security vs. mentioning security and features or stability improvements updated faster, even when controlling for the user's typical update frequency (O.R. = 3.33, $p < 0.001$) and typical update frequency as well as message length (O.R. =4.54, $p < 0.001$). The survey data does not mirror this effect.

Filtering Respondents and Internal Consistency We reconstructed each of the above models for message features using only the filtered subset of high-quality respondents (as described in Section 7.4.2.2 above. This approach did not produce any improvements in matching significant effects seen in measurement data.

To further investigate, we examined the *internal consistency* of the survey responses: how well users' responses about why they would (not) choose to install or recommend an update matched the actual properties of the messages they saw. Appendix B.16.3 details this answer-choice consistency mapping and results in table format.

We find that for the most part, reasons for updating that mentioned specific message properties were unrelated to the actual properties of the assigned message. Specifically, self-reports about update motivation related to a new version having features the user would want were not related to whether the update message men-

[4] We could not control for the other message feature, restart, because no Flash messages mentioned a restart requirement.

tioned features in addition to security (X^2=4.72, p = 0.067). Similarly, reports about not wanting to update because of the new version having features the user would *not* want were also not related to whether the update message mentioned features in addition to security (X^2=0.050, p = 0.823) Reports of not wanting to update because of needing to restart or because of time constraints (e.g., costs) were not related to the update message mentioning a restart (X^2=0.917, p = 0.384). On the other hand, participants who reported wanting to install or recommend an update because it looked fast or did not require a restart were more likely to have seen a message that did not mention a restart (X^2=6.39, p = 0.024). Figure 7.6 summarizes these results.

The application being updated, however, seems to be more salient than other message properties. Reporting that you would update because the given application was important (X^2=38.2, $p < 0.001$), or would not update (X^2=11.8, p = 0.019) because it was unimportant both varied significantly based on the queried application.

RQ3: Survey Sample Effects We note that all survey regression models controlled for sample source. When looking at the full dataset, the baseline model shows no effect from sample source, but controlling for application type shows that MTurk respondents updated significantly more slowly (OR=1.30, $p < 0.001$) than SSI respondents. This effect is also seen in the Flash-only models.

Summary: Survey respondents inattentive to most message features Overall, we observe small but significant effects in the measurement data for all message-related factors. However, we only observe application-related effects—- not more detailed message-related effects—- in the survey data. Internal consistency checks suggest that this may relate to survey respondents not noticing these specific details in the update messages.

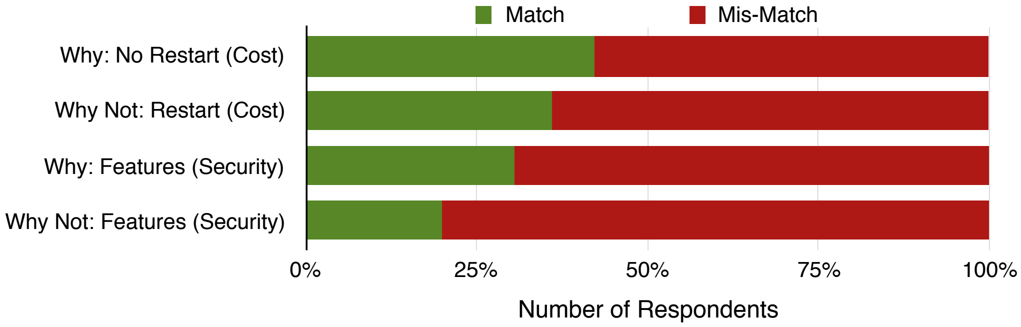


Figure 7.6: Comparison of the internal consistency of survey responses related to two of the three message features.

Chapter 8: Methods: Evaluating the Validity of Readability Measures for Domain-Specific Adult Texts

A variety of readability metrics are available for assessing comprehensibility of texts: human-expert-written comprehension questions, automatically generated readability tests, and computed metrics requiring no human/agent input [31, 82, 93, 219]. Human-written comprehension questions are the gold standard for measuring readability [65, 199], but developing such questions is costly and difficult to scale. As such, prior work has explored various automated, scalable approaches to generating comprehension questions. The first is automatic reading test generation, typically using the Cloze procedure. The Cloze procedure was validated as a scalable method of comprehension assessment through comparison with expert-written comprehension questions for grade-school texts [30, 109, 164, 178].

Recently, researchers have explored approaches to adjusting the construction of Cloze tests: selecting particular key sentences or parts of speech to use as blanks, often to assess retention of factual knowledge or awareness of vocabulary [43, 90, 92, 138, 140], and multiple-choice Cloze tests in which test-takers select from a set of distractors rather than filling in an open blank, which avoid potential scoring issues with typos and equally-correct synonyms [36, 90, 92, 117, 158, 160, 170]. In this chapter[1], I present the novel *Smart Cloze* tool that I and my collaborators developed, which builds on this prior work by choosing distractors from a domain-specific, rather than general dictionary, answering the call from Collins-Thompson's recent review of readability measures [45] for more domain-specific tool options.

The second alternative are readability metrics that take no reader input. The original form of these were readability formulae, the most popular of which is the Flesch reading ease score (FRES) [81, 82, 221], which assumes that longer sentences and words—- which often co-occur with complex syntax—- indicate greater reading difficulty [54, 78]. More recently linguistic feature-based [46, 83, 154] and machine learning approaches have also been used to predict the readability of text [45,59,124].

Despite being used frequently in computational contexts, the majority of these readability assessments were developed for grade-school texts and were validated with grade-school readers. Online texts—such as the security advice texts we seek to measure in this chapter—differ from grade-school texts in that they are targeted toward adult readers, which may lead to differences in text structure (e.g., bullet points), word abstraction [93], and domain-specificity (e.g., medical advice, digital security advice). Such differences may affect the accuracy of computed metrics and automatically generated readability tests, which are increasingly used to scale readability measurements in the digital world [27, 72, 85]. Yet, the validity of readability assessment techniques has rarely been re-evaluated for online contexts.

In this chapter, we we make three contributions: **First**, we evaluate the most commonly used methods for measuring readability in terms of *content validity* (the

[1] Published as [180].

degree to which different measures relate to theoretically-grounded linguistic components like text cohesion or syntactic complexity), *convergent validity* (the degree to which these measures correspond to each other), *redundancy* (the degree to which one measure is subsumed by another), and *score precision* (the shape of the distribution of score from a given measure, and how well it distinguishes among documents). **Second**, we identify a need for domain-specific automatically generated readability tests. To address this, we develop and evaluate a novel technique for automatically generating readability tests specifically for domain-specific texts: *Smart Cloze*. We find that Smart Cloze offers some benefits for domain-specific applications compared to existing measures. **Third**, we contribute two open-science resources: our open-source Smart Cloze tool, as well as a *Digital Readability* evaluation corpus of 100 documents, including 300 comprehension questions written by human experts, that we use in our evaluation.

8.1 Evaluation Corpus

In our evaluation, we compare readability scores from five sources: human-written comprehension questions; automatically generated readability tests including both traditional Cloze and our Smart Cloze domain-specific variant; annotator perceived ease [190, 200], which has been used to evaluate readability metrics in the past; and the Flesch Reading Ease Score (FRES) [82]. We compared these metrics across our *Digital Readability* evaluation corpus. Here we describe our corpus, how we generated each of the readability metrics, and how we conducted our validity analysis.

We draw our *Digital Readability* evaluation corpus from four source corpora: simple stories created by crowdworkers, from [193], wikipedia articles, from [206], health information documents, from [157], and security and privacy advice documents, from the corpus described in Chapter 6. The last two corpora—health and security advice—are domain-specific: focused on a singular domain and often containing jargon or topics not typically encountered in daily life.

1. Story corpus. We drew our crowd-worker-created stories from the MCTest [193] dataset which consists of 500 simple stories created by Amazon Mechanical Turk crowdworkers and validated manually for quality.

2. Wikipedia corpus. We drew our Wikipedia articles from a corpus of 20,000 Wikipedia articles scraped from Wikipedia and cleaned for quality [206]. We selected Wikipedia articles as a baseline of adult texts against which to compare the domain-specific texts. Wikipedia articles have a mean FRES similar to our domain-specific texts (mean FRES for the wikipedia sample = 47.9; for the health documents = 53.7; and for the security documents = 48.7), suggesting that, at least by one measure, the texts should be similar in readability.

3. Health corpus. We drew health articles from the 500-document Health Text Readability Corpus [157]. This corpus includes consumer health information documents made available for public use by the CDC, NIH, American Heart Association, American Diabetes Association, and the National Library of Medicine's Medline Plus resource. Worksheets, posters, infographics, and

websites are not included. More than half (N=293) of the documents were found in ”Easy to Read” collections; that is, the document has been designated by its source agency as appropriate for adults who read at or below a 7th-8th grade reading level.

4. Security corpus. We used the security advice corpus described in Chapter 6.

Final evaluation corpus. To ensure comparability of results, we used a standardized subsampling procedure to select 25 documents from each corpus. To ensure that our evaluation captured some variance in documents, we subsampled by length. We first remove the shortest and longest 5% of documents, then we then divide the documents into five bins by length, based on how many standard deviations the length of a given document is from the mean length for that corpus. We manually reviewed all selected documents to ensure that they were on-topic and appropriately clean.[2]

8.2 Readability Metrics

We created three **comprehension questions** for each of the documents in our evaluation corpus: one True/False question and two multiple choice questions with four answer options each, per comprehension question best practices [22, 58]. Domain-specific questions were written by three co-authors who were domain experts in digital security or in health; the general questions were written by two

[2] You can find the documents in our corpus, the 300 comprehension questions, and the code for generating traditional Cloze and Smart Cloze tests at: `https://github.com/SP2-MC2/Readability-Resources`.

other co-authors. All 300 comprehension questions were reviewed and edited by a paid comprehension question specialist, who had experience writing and evaluating comprehension questions for the SAT, Discovery Science, and similar organizations; the specialist spent more than 10 hours editing and refining the questions.

We selected the FRES as our **computed measure**, as it is the most-used by number of citations, and anecdotally, by wide-spread application. We computed the FRES for each document using the Python `textstat` package [3].

For our **annotator perception of ease** measurement, we use a single-item question "How easy is this document to read?" with 5-point Likert-item response choices ranging from "Very Easy" to "Very Hard."

Finally, for our **automatically generated readability tests** we used both the traditional Cloze Procedure and our Smart Cloze procedure. Prior work suggests that the frequency of blanks does not significantly affect results [219]. We select set $n = 5$, up to a maximum of 35 target words, for both our traditional Cloze implementation and our Smart Cloze tests, as was done in the original Cloze implementation [220].

Smart Cloze tool. Prior work to improve Cloze tests offered a multiple-choice variant of the traditional Cloze procedure in which distractors (incorrect answer choices) are randomly drawn from a general dictionary containing other words with the same part of speech. While such multiple-choice variants offer improvements in test-taker time, they are potentially inappropriate for domain-specific applications. For example, replacing the word "encryption" in a cybersecurity text

[3] `https://pypi.org/project/textstat/`

with “dog” creates a very easy test. As such, we implemented a novel approach that we call Smart Cloze: we construct a domain-specific dictionary from the same corpus for which we are generating tests and draw distractors from it. The goal is to offer relevant alternatives such as “antivirus” and “key” as distractors for “encryption.”

To construct a Smart Cloze test for some document d selected from a domain-specific corpus c, our tool follows the following procedure. First, we bin all of the words in c by part of speech (tagged using Spacy[4]) to create a domain-specific dictionary. We then construct a similar part-of-speech-tagged *document-specific dictionary* using only the words in d. Third, we identify *target words* in d to be replaced by multiple-choice questions. Fourth, we generate distractors for each target. We randomly select up to 14 potential distractors with the same part of speech as the target word from each of the domain-specific and document-specific dictionaries. We then process these distractors in random order, optimizing to obtain two from each dictionary, until we have found four satisfactory distractors.

We measure whether a potential distractor is satisfactory by examining how probable it is that the distractor might substitute for the target word within d. To do this, we first look up the bigram probabilities of the target word (w_c) with its preceding (w_{c-1}) and following (w_{c+1}) words in Google’s n-gram corpus. This gives us a baseline for how probable the correct answer is. We then look up bigram probabilities of the potential distractor (say w_d) in combination with the same preceding (w_{c-1}) and following (w_{c+1}) words. Satisfactory distractors have both preceding-distractor and distractor-following bigram probabilities within two orders

[4] https://spacy.io

of magnitude of those for the correct target word. [5] More precisely, a distractor w_d will be accepted if:

$$[P(w_d|w_{c+1}) \geq P(w_c|w_{c+1})] \wedge [P(w_{c-1}|w_d) >= P(w_{c-1}|w_c)]$$

If we do not find four satisfactory distractors (by this definition) within the candidate 28, we instead select the potential distractors with the highest bigram probabilities until we obtain the desired four distractors. Finally, to avoid very small lists of distractor options for certain part of speech (e.g., `TO` only contains to'), we merge parts of speech with small wordlists with larger, related parts of speech until enough unique distractors can be found.

8.3 Validity Evaluation

To evaluate the validity of these readability metrics and compare them, we needed readers to answer the comprehension questions, Cloze tests, and ease question for our documents. We recruited U.S. Amazon Mechanical Turk workers (MTurkers) with a 95% approval rating or above to complete these tasks. Each worker completed one randomly selected readability measure for four documents, including one randomly selected from each of the four corpora. MTurkers were compensated with $1.50 for completing the task. We recruited at least five distinct MTurkers for each type of measure and each document (n=841).

We compare our five readability metrics by examining their construct valid-

[5] We selected two orders of magnitude heuristically to narrow the search space for faster computation while obtaining an appropriate difficulty for the test. Future work could explore alternative heuristics in more detail.

ity [52]: the degree to which it appears that the measures are accurately measuring readability. To do so, we examine:

- *Content validity:* the degree to which the measures relate to concepts that have been theorized to be relevant to readability; and
- *Convergent validity:* the degree to which related measures (e.g., multiple measures of the same construct) are correlated.

We also explore three factors that are relevant to selecting an appropriate readability measure:

- *Redundancy:* the degree to which any measure is fully, and redundantly, covered by another measure;
- *Score precision:* the precision with which the measure distinguishes between different documents; and
- *Participant burden:* the cost of the measure to the participant (and the researcher) in time to complete.

To assess **content validity**, we examine the degree to which five core linguistic components (narrativity, syntactic simplicity, word concreteness, referential cohesion, and deep cohesion) theorized to be related to readability [93] can *explain the variance* in the measure scores. We measure these components using the Cohmetrix tool [93]. We construct linear regression models, in which the mean measure score for a document is the outcome variable and the input variables are the five linguistic components.

As we wish to understand *which* components are related to which measures, we seek to ensure that we construct a model of best fit. To do so, we perform feature selection via stepwise backward selection, minimizing AIC [37]. We further measure applicability to domain-specific texts by including the source corpora of the document as a sixth covariate in the regression model. We set Wikipedia as the baseline for corpora source, as it represents a broad set of non-domain-specific documents with similar FRES to the domain-specific documents.

To assess **convergent validity**, we compute the Pearson correlation between the scores for each readability method in our evaluation dataset. We report the ρ value (strength of the correlation) for correlations significant at $\alpha < 0.05$; Holm-Bonferonni [18] correction is applied to account for multiple testing.

We also assess **redundancy**, which is not strictly a property of convergent validity, but is relevant when comparing multiple measures that attempt to assess the same construct. Demonstrating that two related measures are correlated establishes convergent validity, but if they are perfectly correlated, then it is unlikely both are needed [174]. For this analysis, we construct linear regression models in which the mean score from a given measure for a given document is the outcome variable and the input variables are the three other types of measures (note that we do not include both Cloze measures in any model, but instead construct separate, three-variable models, each with FRES, comprehension questions, ease, and one of the Cloze measures). We consider the degree of redundancy to be the proportion of variance in measure scores explained by the other measures (that is, the R^2 value of this regression model).

To assess **score precision**, we examine the shape of the distribution of scores for a given measure. Per best practice for observing distributions, we do so both through visual inspection and by measuring kurtosis (a statistical measure of the 'tailness' of a distribution) [60].

Finally, we assess **participant burden** in terms of time to complete the task (which also proxies for researcher cost). We compare time by bootstrapping confidence intervals for the mean time for completion of a readability assessment for a given document. Non-overlapping confidence intervals indicate a significant difference in completion time.

8.3.1 Limitations

Our work is subject to four primary limitations. First, automatic selection of distractors means that there may be differences in the difficulty of different distractors (or variances in difficulty of tests generated by the method when used repeatedly). Based on a manual review of the Cloze tests we conducted before deployment, we did not find trivial distractors to be highly prevalent, given the breadth of words available in each dictionary. However, future work may wish to explore methods for measuring and ensuring consistency in distractor difficulty. Second, MTurk respondents are known to be more educated than the general population, and thus the results of our work may not generalize to low-literacy populations, second-language learners, and others [121,185]. Third, while we attempted to cover a relatively broad space of online documents, other types of documents (e.g., news articles, Facebook posts) may perform differently. Finally, it is possible that MTurkers were inattentive to our tasks, limiting the validity of our data. We mitigate this possibility by restricting our sample to workers with 95% approval rates on past tasks, as shown in prior work to ensure participant attention to surveys as well as gold-standard 'test' questions [169].

8.4 Results

	Linguistic Components (Content Validity)					Additional Considerations			
	Narrativity	Syntactic Simplicity	Word Concreteness	Referential Cohesion	Deep Cohesion	Burden (Mean Time)	Mean Score	Score Precision (Distribution Trend)	Domain Sensitivity
Comprehension	✓	✓				2.86 min	75.7%	exponential	
Traditional Cloze	✓			✓		5.05 min	34.1%	normal	
Smart Cloze	✓	✓		✓		4.55 min	52.4%	normal	✓
Ease			✓			1.67 min	67.1%	uniform	✓
FRES	✓	✓	✓			—	61.0%	uniform	✓

Table 8.1: Summary of our results on content validity (significant relationships between readability measure and linguistic components theorized to explain comprehension) and other considerations for selecting a readability measure (time for participants to complete a test for a given measure on an average document, average score achieved across documents, trend in the shape of the distribution of scores achieved with a measure, and whether the measure exhibits variation by document domain.

In this section we summarize our results for content validity (including domain sensitivity of measurements), convergent validity, redundancy, score precision, and participant burden (Table 8.1).

8.4.1 Content Validity

We find that comprehension question scores are significantly related to the narrativity ($p = 0.003$) and syntactic complexity ($p = 0.035$) of the document, while performance on comprehension questions is not significantly related to the other three linguistic factors we examined (word concreteness, referential cohesion, deep cohesion) or to type of document (source corpus).

Traditional and Smart Cloze scores are significantly related to the narrativity

(Traditional: $p < 0.001$; Smart: $p = 0.040$) and referential cohesion (Traditional: $p = 0.035$; Smart: $p = 0.008$) of the document. Smart Cloze scores were significantly related to the syntactic complexity ($p = 0.005$) of the document; traditional Cloze scores were not significantly related to syntactic complexity. Finally, neither type of Cloze score was significantly related to deep cohesion or to word concreteness. Smart Cloze scores vary significantly by document domain, while traditional Cloze scores do not. Specifically, Smart Cloze scores are significantly higher for domain-specific documents: those from the health ($p < 0.001$) and security (0.031) source corpora, than for Wikipedia documents. We hypothesize that this is the case because the topics of domain-specific documents are narrower — there are fewer reasonable options for any given blank space — than in the Wikipedia documents, resulting in easier multiple-choice questions. (Anecdotal observation of the generated questions seems to align with this theory.)

Ease perceptions are significantly related only to word concreteness ($p = 0.015$) and document domain: stories ($p = 0.027$) and security ($p = 0.015$) documents are perceived as significantly easier to read than Wikipedia articles. The relationship between ease perceptions and concreteness (and lack of relationship with the other linguistic features we examined) is worth remark. Concreteness of words appears to be easy for readers to assess with a quick glance at an article. This assessment, and their overall perception of ease, may in turn determine whether readers are willing to further read a document they encounter "in the wild," at which point other readability factors may become more relevant. We therefore hypothesize that ease and other measures may complement each other. Finally, FRES scores are significantly related to narrativity ($p < 0.001$), word concreteness ($p < 0.001$), and syntactic complexity ($p < 0.001$); but not to either referential or deep cohesion. Perhaps unsurprisingly, FRES scores were significantly higher for stories than for Wikipedia ($p < 0.001$). FRES scores were also higher for security than for Wikipedia ($p = 0.015$), but the health and Wikipedia documents in our sample did not differ in FRES.

While the regression models we constructed explained a significant portion of the variance in scores for ease [6] ($R^2 = 0.504$), FRES ($R^2 = 0.758$), Smart Cloze (R^2= 0.389) and traditional Cloze ($R^2 = 0.334$), these factors explained much less of the variance for comprehension question scores ($R^2 = 0.132$).

8.4.2 Convergent Validity

To examine **convergent validity**, we examine the correlation between scores from different measures (Figure 8.1). Comprehension question scores have the least correlation with scores from the other methods: no correlation with traditional Cloze or ease ratings, and small correlation with FRES ($\rho = 0.22$) and Smart Cloze ($\rho = 0.23$).

This low correlation between comprehension questions and the other methods of measuring readability, together with the low explanation of variance noted above, suggest that comprehension questions assess a combination of the readability of the text and the reader's cognitive abilities, different from the other metrics, which

[6] This result closely parallels prior work, which predicted perceived ease of Wall Street Journal articles using discourse, vocabulary and length, resulting in an R^2 of 0.503 [171].

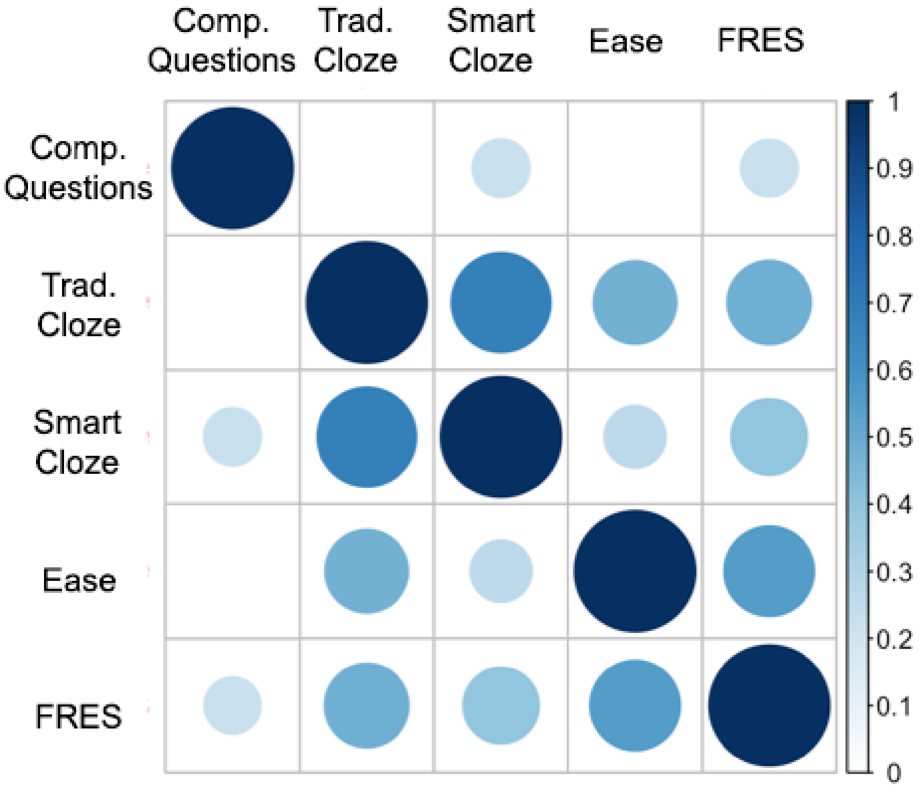


Figure 8.1: Correlation matrix showing the convergent validity of the measures. That is, the correlation between readability measurement methods. Non-significant correlations ($p > 0.05$) are not shown.

may be more specific to just the text itself [199]. Traditional Cloze, on the other hand, correlates relatively well with all other methods. Perhaps unsurprisingly, there is high correlation ($\rho = 0.71$) between traditional and Smart Cloze scores. Traditional Cloze also correlates well with ease ($\rho = 0.47$) and FRES ($\rho = 0.48$). Smart Cloze correlates less with ease than does traditional Cloze (ease: $\rho = 0.264$, FRES: $\rho = 0.44$). Finally, ease and FRES correlate relatively strongly with each other ($\rho = 0.56$).

8.4.3 Redundancy

By constructing regression models with the mean score from a given measure on a given document as the outcome variable, and the other measures as the input variables, we find that 4.02% of the variance in the comprehension question scores can be explained by ease perception, FRES, and traditional Cloze (7.92% with Smart Cloze). 20.1% of the variance in traditional Cloze is explained by the other measures, while 22.1% of the variance in Smart Cloze is explained by these measures. 36.0% of the variance in ease perception is explained by mean comprehension question scores, FRES, and traditional Cloze (31.8% with Smart Cloze), while 35.8% of the variance in FRES measurements is explained by scores on comprehension questions, ease perception, and traditional Cloze (37.8% Smart Cloze). Thus, none of the measures are redundant, as the variance in no measure is fully (or even more than 50%) explained by the others.

8.4.4 Score Precision

Researchers selecting a readability measurement method may also wish to consider the **score precision**: that is, are you trying to find a few bad outliers in a corpus of highly readable documents, or are you expecting a relatively normal distribution of document quality? Figure 8.2 shows the score distributions by method across all documents and for each document type.

Across domains, the Cloze tests provide the most normal distributions of scores

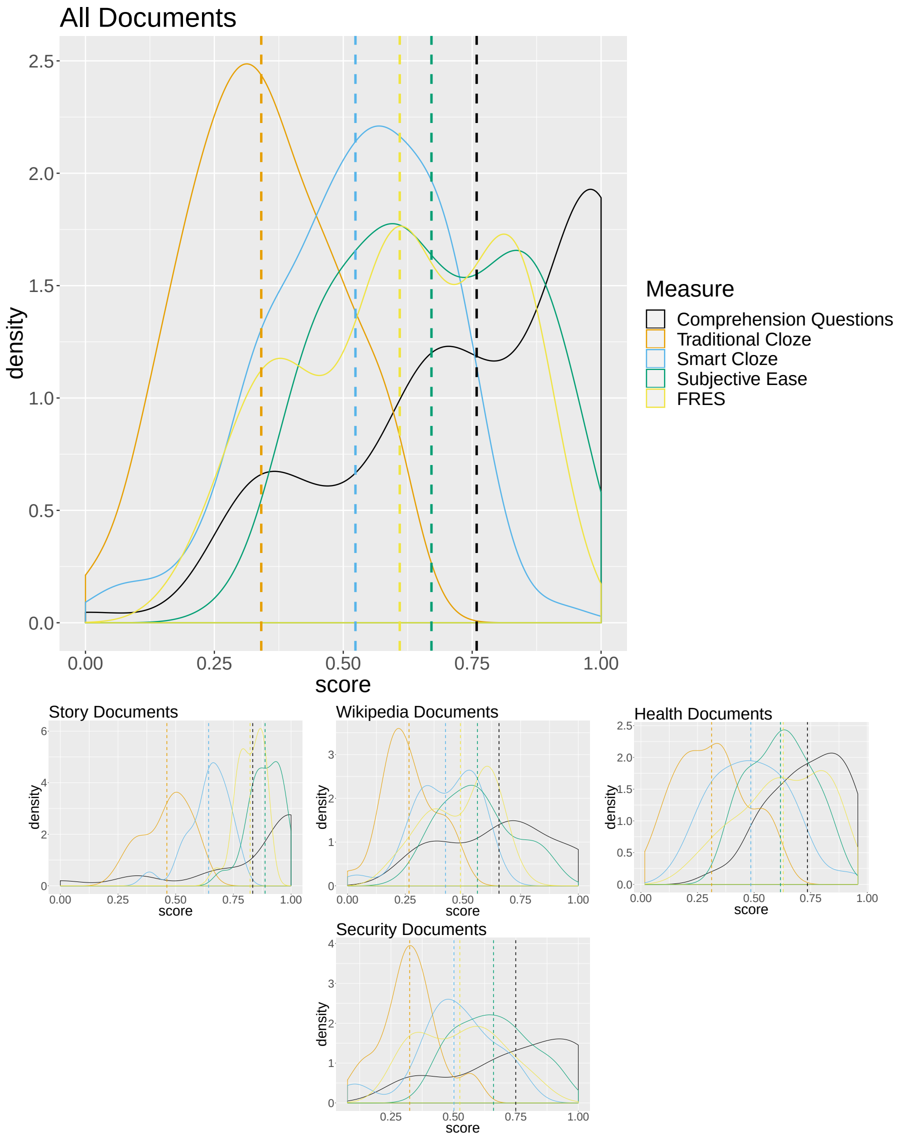


Figure 8.2: Score distributions by method, across all corpora (top) and by corpus (bottom).

(average traditional Cloze kurtosis = 2.34, average Smart Cloze kurtosis = 3.08)[7]. Cloze scores are thus useful in cases where the relative readability of documents is of interest and where you hypothesize that a normal distribution of readability may be appropriate. The distribution of traditional Cloze scores is transposed left, with a mean of 0.341 (95% confidence interval: [0.329, 0.353]), while the Smart Cloze

[7] The kurtosis of a normal distribution is 3; the kurtosis of a unifom distribution is 1.

distribution is centered, with a mean of 0.524 (95% confidence interval: [0.510, 0.537]). Traditional Cloze scores may thus need to be scaled (considered relative to each other rather than as absolute values) to account for this observed ceiling effect.

Ease ratings and FRES, on the other hand, have a more platykurtic distribution (ease: average kurtosis 1.91; FRES: average kurtosis 1.94; fully uniform or platykurtic distribution is 1). A platykurtic distribution has fewer outliers than a normal distribution. Thus, these methods may be more useful in corpora where you expect few readability outliers. Further, ease ratings and FRES both have means higher than 0.5: ease has a mean across domains of 0.671 (95% CI: [0.657, 0.685]) and FRES has a mean of 0.610 (95% CI: [0.594, 0.625]). Given these relatively high means, these methods may also need to be scaled, or may be most useful in cases where you anticipate that an average document in your corpus will be fairly readable. Comprehension questions provide a similarly platykurtic distribution (average kurtosis: 2.06), but with a very high mean (0.757, 95% CI:[0.739, 0.778]).

8.4.5 Participant Burden

Finally, research is often constrained by resources, including time and budget, and ethically we must be mindful of the burden we impose on our participants. Ease perception (one question) is the fastest test for a worker to complete, with participants spending an average of 1.67 minutes (95% CI: [1.56, 1.78]) per document. Comprehension questions (three questions) took a significantly longer period of time, averaging 2.86 minutes ([2.64, 3.12]) per document, followed by Smart Cloze with an average of 4.55 minutes ([4.08, 4.60]) per document. Finally, traditional Cloze took significantly longer than Smart Cloze, averaging 5.05 minutes per document ([4.72, 5.42]).

8.5 Selecting a Readability Metric for Online Document Evaluation at Scale

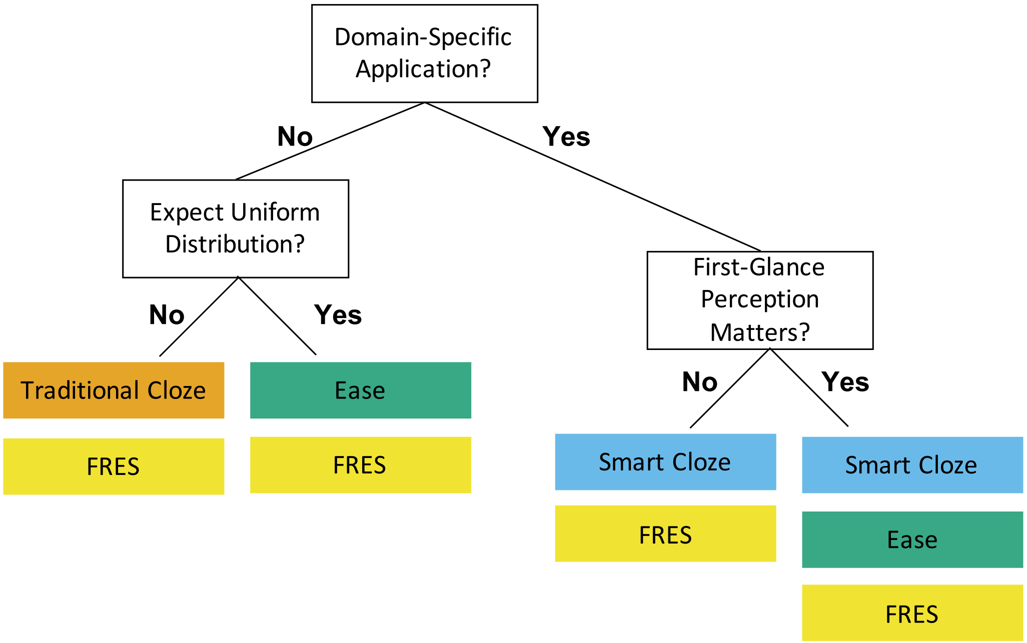

Domain-Specific Application?
No
Yes
Expect Uniform Distribution?
No
Yes
Traditional Cloze
FRES
Ease
FRES
First-Glance Perception Matters?
No
Yes
Smart Cloze
FRES
Smart Cloze
Ease
FRES

Figure 8.3: Flow chart for selecting readability measures.

In sum, no single readability metric outperformed all the others. Each metric offers different benefits and tradeoffs, and human-written comprehension questions differ the most from the other metrics. We summarize the relevant considerations for selecting a readability metric in Figure 8.3 and encourage the use of multiple metrics in cases where creating comprehension questions is not scalable.

We find that comprehension questions and Smart Cloze both relate significantly to syntactic complexity, perhaps because they require selection among different possible answer choices. Traditional and Smart Cloze relate to referential cohesion, which makes logical sense, as filling-in-the-blank questions require context from prior sentences. Finally, ease and FRES relate to word concreteness, potentially providing relevant assessments of "first glance" readability reactions. The readability metrics examined also exhibit convergent validity, with the three traditional methods (traditional Cloze, subjective ease, and FRES exhibiting the strongest correlation in scores. Finally, the measures are not redundant: a significant portion of the variance in each remains unexplained by the others.

These different methods offer different levels of precision: the Cloze methods trend toward normal distributions with low (traditional) and centered (Smart) means. On the other hand, ease and FRES assessments are more uniformly distributed, with higher means (near 60 and 70%, respectively). Further, Smart Cloze, FRES, and ease measurements all significantly co-varied with document type: Smart Cloze scores were significantly higher for the domain-specific documents (health, security) than for Wikipedia articles, while FRES and ease scores were significantly higher for the story and security documents than for Wikipedia.

While it may be tempting to exclusively use linguistic features because they are cheap and easy to obtain, we find that for the five linguistic factors we explored in this work, these factors explain only 30-50% of the variance in the reader-input readability metrics. Future work may wish to explore additional linguistic factors [99, 124, 222, 226], beyond those covered in this work. In the mean time, our results suggest that when possible, researchers should still consider augmenting these factors with a human-input method. The Smart Cloze tool we propose offers improvements in participant burden, especially for domain-specific documents: scores are higher on average than for traditional Cloze, and tests are 30 seconds faster on average (54 seconds faster for domain-specific documents). However, Smart Cloze is less correlated with perceived ease than traditional Cloze, possibly because the multiple choice option makes the test easier to complete, lessening the chance that participants will "give up." Thus, Smart Cloze is best used in cases where cursory or first glance assessment of readability is less relevant, or in combination with an ease assessment.

Chapter 9: Discussion

The work in this thesis provides insight into people's sources of security education, how those sources relate to their security outcomes, and how the quality - and quantity - of advice available in the security education ecosystem has left users to fend for themselves.

The results from Chapters 3, 4, and 5 showed that individual users were receiving security education from a wide variety of sources. Our results from Chapter 6 suggest that users are receiving an overwhelming amount of advice from one of the most prevalent of these sources: websites. If the other sources from which they receive security education offer even a fraction of the 374 unique pieces of advice we identified from websites, users must be filtering through hundreds of pieces of advice. Even if those advice items are perceived by experts as accurate, the scale of this task is daunting.

Our results from Chapter 6 support that it is not primarily the quality of security education itself that is the problem, but rather the quantity and lack of prioritization of the information provided. While it is possible that some sources of security education such as friends and coworkers provide exclusively different advice from the 374 imperatives we identified and evaluated in Chapter 6, this seems unlikely. Thus, the evaluation of advice in this thesis should provide a reasonably representative picture of the security education users receive through various channels. While we do identify issues of comprehensibility and some of actionability (see Section 6.6 for a more detailed summary), which may be especially pronounced for text-based advice, more critically, we find that experts love security advice: too much so.

9.1 How People Evaluate Advice

These results provide additional context to explain our findings in Chapters 3 and Chapter 4 regarding advice evaluation. In Chapter 3 we found that many participants only accepted advice that they had seen repeated across multiple channels. Given the large quantity of advice available, this appears to be a prioritization heuristic, as well as a trust heuristic. In Chapters 3 and 4 we also saw that negative experiences were one of the primary ways that people learned security behaviors. This implies that negative experiences, too, may serve to help users prioritize important advice. In Chapter 6 we find that this prioritization of advice is critically important, as users' prioritization of advice strongly correlates with self-reported adoption.

Additionally, in Chapter 4 we saw that different behaviors had different advice-evaluation heuristics, with advice about some topics (antivirus, software updating) being evaluated chiefly based on trust of the advice source and advice about others (passwords) being evaluated based on the content itself. These differences align well with the results from Chapter 6 that users were very confident about their ability to implement advice about passwords and a relatively unconfident regarding advice about antivirus and software security (which included software updating). Finally, we also find in Chapter 4 that one of the primary reasons users reject security education is because it is trying to promote a product. In Chapter 6 we found that advertisements for security products made up nearly 20% of the advice we collected based on expert recommendations and user-generated search queries, underscoring users' need for this evaluation heuristic.

9.2 Expert Evaluations and Unfalsifiability

Experts, too, must evaluate advice. As we show in Chapter 6 they struggle to do so, marking the vast majority of advice as accurate, and report nearly 40% of advice as being amoung the top 5 behaviors they would recommend to users. This is likely due to issues of unfalsifiability [112]: without measurement of impact on actual security, or proven harm, everything appears at least slightly useful as a talisman against potential harms. Yet, by asking users to hang hundreds of talismans, each chipping away a little bit at their compliance budgets [26], experts are instead creating a hole-filled blanket of protection, exposed across a myriad of attack surfaces.

9.3 Relationship Between Advice and Outcomes

This issue of expert evaluation suggests reason for concern with users' other sources of security education, as well. In addition to websites, advice from friends and coworkers (particularly expert friends and IT professionals in the workplace) are the most prevalent sources of advice we identified in Chapters 3-5. IT professionals and expert friends are likely to exhibit the same struggles with falsifiability and prioritization as the experts in the study we presented in Chapter 6.

Further, we hypothesize that non-expert friends may be similarly ineffective at prioritizing advice, given the results in Chapter 6, which showed that asking the general population to prioritize advice resulted in even less clearly identified top advice than asking experts (see Appendix B.13). Similarly, the results from Chapter 7 suggest that people are overly optimistic when recommending security behaviors to others.

Our measurement results thus support, and refine, one of our hypotheses from Chapter 5: that the correlation between receiving advice from friends, coworkers, and websites and being more likely to experience a security incident was due to a problem with the advice from those sources. While we initially hypothesized that this may have been due to users receiving bad or difficult to interpret advice from the sources, or due to them turning to these sources *after* an incident, our subsequent results from Chapter 6 suggest that users are receiving too much advice from these sources, rather than bad advice. This flooding of advice may leave users not much better off than when they looked for advice in the first place. On the other hand, our results from Chapter 5 lead us to hypothesize that teachers and librarians may be better at paring down advice to a few key points, making these advice givers better at helping those who turn to them.

9.4 Digital Divide in Security Advice

In Chapters 4 and 5 we found differential use of advice channels based on Internet skill and socioeconomic status. Specifically, in Chapter 4 we found that low-SES and low-skill users were less likely to rely on information from digital media (such as that we evaluate in Chapter 6) and more likely to rely on advice from friends, family, and service providers. In Chapter 5 we confirm, on a fully weighted, U.S. representative sample, this divide in advice sources. This divide, in part, prompted the examination of the comprehensibility of advice in Chapter 6. In this analysis, we found that the average security advice document has low to partial comprehensibility to the average U.S. user, making most of the documents nearly incomprehensible to those with lower literacy and/or English fluency [32, 33, 166, 217, 230], skills strongly correlated with socioeconomic status [55, 107]. Further, those from low-SES backgrounds tend to have far more limited free time and higher cognitive burden [97, 144], making it even more difficult for these users to sort through the over-abundance of advice provided online than it is for others. In combination, these issues of comprehensibility and overload likely explain the significant gap in use of online advice by lower and higher SES users.

9.5 Moving Forward with Security Education

Overall, the work presented in this thesis suggests that experts have created a crisis of security advice through their own indecisiveness: Unable to narrow down a multitude of relatively actionable, but half-heartedly followed security behaviors to a key, critical set that are actually necessary for keeping users safe. The U.S. Government, alone, offers 205 unique pieces of advice to end users and this information is a primary source of advice for more than 10% of users (Chapter 5), who are consequently flooded with authoritative advice. Non-technical news media, such as CNN and Forbes, also offer more than 100 unique pieces of advice to users, while the EFF and Tactical Tech each offer over 60 pieces of advice, respectively.

That the 41 experts we recruited in Chapter 6—many of whom were senior security professionals at reputed security organizations—considered nearly all of this advice accurate and helpful in reducing security risk suggests that we need a new approach. In line with arguments by Herley and van Oorschot calling for a science of security [114], our work calls for a science of security behavior.

Future work must aim toward *measuring* the direct impact of following secure practices. For example, comparing the effect of different practices on frequency of account and device compromise. Such experiments will enable falsifiability of the claim that all behavior (advice) enhances security [112]. Further, such measurement can enable us to identify the most impactful behaviors such that the remaining multitude can be archived and de-prioritized.

The data collected in Chapter 6 can provide a starting point for such measurement. As it is likely infeasible to evaluate 374 different behaviors, this data can be used to prioritize evaluation of the most prevalent advice, and that given highest priority by experts and users. Additional future work may also wish to consider identifying which of the advice sources catalogued in this thesis *are* good at prioritizing advice. For example, results in Chapter 5 suggest that teachers and librarians may be such sources, as, potentially, are non-technology focused non-profit organizations, which we found in Chapter 6 provided a small amount of high quality advice. Further, the findings in Chapter 6 highlight a few key security topics (network security, privacy, and incident response) that may be in need of an advice overhaul.

Finally, the work presented here suggests that security behavior is not, chiefly, a matter of education. Rather, staying safe online appears to require a baseline level of security education—which the adoption results from Chapter 6 suggest that many U.S. users have, albeit in excess—after which a few, measurement-validated behavior recommendations tailored to account for users' threat models (see Chapter 3 and my work outside this thesis [181, 187]) and compliance budgets [26] may be most effective.

Chapter 10: Conclusion

Security advice is everywhere, and there is a plethora of it. In the large-scale evaluation of the quality of one of the most popular sources of security advice—online articles—I find 374 unique security behaviors being recommended to users 2,780 times across 614 documents (Chapter 6).

As I show in Chapter 3 and 4 users struggle to determine which advice to accept, and which (most) to reject in an effort to stay safe online. This struggle is not restricted to users. In Chapter 6 I find that experts love advice, too much so. Experts fail to distinguish between helpful and useless advice, considering nearly 90% of the 374 unique advice imperatives "accurate" and beneficial to users' security. Thus, despite—or in fact, because of—the plethora of advice available to users, the probabilistic, census-representative telephone survey that I analyze in Chapter 5 reveals that more than 50% of Americans have had at least one major security incident, such as having their information stolen [183].

In this thesis, I establish that this burden does not fall equally (Chapter 4): I find empirical evidence of a digital divide in security education in which lower-skill and less-educated Americans encounter and apply authoritative advice less frequently than their more-skilled, more-educated neighbors. In Chapter 5 I show that security advice is strongly correlated with security outcomes: less authoritative advice correlates with more security incidents. Further, in Chapter 6 I show that the majority of security advice is only partially comprehensible to the average American, leaving most people—and especially those with lower socioeconomic status and literacy skills—behind. Taken together, these results suggest a security-education ecosystem in crisis: unable to help most Americans, and especially those with the least resources, effectively defend themselves from digital crime.

Appendix A: Study Instruments

A.1 Security Education Interview Protocol

Employment

- Could you tell me a little bit about what you do?
- Do you handle sensitive or private data as part of your job?
  - Could you tell me a little bit more about that data?

Digital Security

Device Protection

- How many devices do you use to access the Internet for personal use?
  - Do you have a smartphone? Tablet? Multiple computers?
  - What type or brand of smartphone or computer (e.g. Windows/Mac/Linux) do you use?
- Can you show me how you access your devices?
  - When was the last time you changed this password?
- Are there any other tactics you use to protect your devices?

- Do you use antivirus software?
    - How often do you run the software?
    - Did you install it or did it come with your computer?
    - Why do you use it?
- Why do you use these strategies for protecting your [phone/computer/devices]?

  For each strategy, ask:

    - When did you start using this strategy?
    - How do you feel that this strategy works to protect you?
    - Why did you choose to use this strategy over using a different one?
    - What are you most worried about?
    - Have you ever had a negative experience?
    - Do you know anyone who has had a negative experience?
    - Are there ever times when you do not choose to use this strategy?
    - Where or from whom did you learn this strategy?
- Are there strategies you have considered or heard about but do not use?
- Is there a password on your wireless Internet at home?
    - Did you set up this password?
    - When was the last time you changed this password?
    - Were you prompted to do so?

- Is there a password on your router?

- Are there any other tactics you use to protect your wireless Internet?

- Why do you use these strategies for protecting your wireless Internet? *For each strategy, ask:*

  - When did you start using this strategy?

  - How do you feel that this strategy works to protect you?

  - Why did you choose to use this strategy over using a different one?

  - What are you most worried about?

  - Have you ever had a negative experience?

  - Do you know anyone who has had a negative experience?

  - Are there ever times when you do not choose to use this strategy?

  - Where or from whom did you learn this strategy?

- Are there strategies you have considered or heard about but do not use?

- How secure do you feel your devices and your wireless Internet are?

Internet Activities

Browsing and Emailing

- Do you browse the Internet?

- Do you access your email via a web browser (e.g. Safari/Firefox/Chrome/Internet Explorer)?

- Do you shop online or bank online?

- Do you do all of these activities on all of your devices?

- Scenario: Let's imagine that you have a family member (parent/spouse/sibling/child) with whom you share a computer. You are searching for a surprise birthday gift, lets say a necklace, for this person, and you are using the Internet to research potential gifts. Can you show me what you would do to start this project?

- In general, how do you stay secure when browsing the Internet or checking your email?

  - When was the last time you changed your email password?

    - Were you prompted to do so?

  - Do you use two-factor authentication?

    - Two-factor authentication is a service where you might put in your phone number and then be sent a verification code.

  - Do you use the privacy settings when browsing?

  - Do you ever use incognito browsing or private browsing?

  - Do you use a script, popup, or cookie blocker?

  - How do you treat emails from unknown individuals?

  - Are there any particular precautions you take when downloading from the Internet?

- Are there any other tactics you use when browsing the Internet/accessing your email via the Internet?
- Why do you use these strategies for staying secure while browsing the Internet or accessing your email? *For each strategy, ask:*
  - When did you start using this strategy?
  - How do you feel that this strategy works to protect you?
  - Why did you choose to use this strategy over using a different one?
  - What are you most worried about?
  - Have you ever had a negative experience?
  - Do you know anyone who has had a negative experience?
  - Are there ever times when you do not choose to use this strategy?
  - Where or from whom did you learn this strategy?
- Are there strategies you have considered or heard about but do not use?
- How secure do you feel you are when browsing the Internet and accessing your email?

Online Shopping/Banking

- Narration: Can you please walk me through what you would do to login to your banking website? Now please pretend you are exiting the website as if you had just completed your banking business.

- How often do you change your password for online banking or shopping accounts?
- Are there any other tactics you use when shopping online or doing online banking?
  - Do you always use the same credit card?
  - Do you use paypal?
  - Do you use a single use credit card number?
- Why do you use these strategies for staying secure while online shopping or online banking? *For each strategy, ask:*
  - When did you start using this strategy?
  - How do you feel that this strategy works to protect you?
  - Why did you choose to use this strategy over using a different one?
  - What are you most worried about?
  - Have you ever had a negative experience?
  - Do you know anyone who has had a negative experience?
  - Are there ever times when you do not choose to use this strategy?
  - Where or from whom did you learn this strategy?
- Are there strategies you have considered or heard about but do not use?
- How secure do you feel you are when online shopping and online banking?

General Advice

- Do you store your passwords anywhere?
    - Where do you store them?
    - In what format do you store them?
    - Is it password protected or locked?
    - Why did you start doing this?
    - When did you start doing this?
- Do you ever look for new information or talk to someone about tactics such as [what they mention above for security]?
    - Where do you look for this information and with whom do you talk?
- Do you often see news pieces, ads, or articles on TV, in the newspaper, or online with tips or advice about how to protect yourself online?
    - How do you feel about the information provided?
    - Are there strategies you have learned from these sources?
- What other sources do you consult when seeking security advice?
- Do you see any security advice that you do not take?
    - Why do you not take it?
- Do you feel that you have the ability to make yourself more digitally secure?

- Whom or what would you say has most influenced your overall approach to computer security, and in what way?

Physical Security

Dwelling Security

- Do you live in a house or an apartment?
  - Do you own your dwelling?
  - Do you live alone, with a partner, family, or with roommates?
- Can you walk me through what you do as you leave your dwelling?
  - Are there one or two locks?
  - Is it a hard lock or an electronic lock?
  - Is that something that came with the building or something you installed?
    - Why did you install the locks?
- Can you walk me through what you do when you prepare to go to bed in the evening and when you return from your day of work?
- Are there any other strategies, which you have not mentioned, that you use to secure your dwelling?
  - Light timers?
  - Security system?
  - Security system or guard dog signs?

- Is there anything that led you to buy or rent in the location you did?
- Why do you use these strategies for securing your dwelling? *For each strategy, ask:*
  - When did you start using this strategy?
  - How do you feel that this strategy works to protect you?
  - Why did you choose to use this strategy over using a different one?
  - What are you most worried about?
  - Have you ever had a negative experience?
  - Do you know anyone who has had a negative experience?
  - Are there ever times when you do not choose to use this strategy?
  - Is this strategy something that is important to you, or something you feel is more important to other members of your household who share the dwelling?
  - Why would you say that it is more important to [you/other]?
  - Where or from whom did you learn this strategy?
- Are there strategies you have considered or heard about but do not use?
- How secure do you feel that you are when you are at home?
- How secure do you feel that your belongings are when you are not home?

Transit Security

Car (if applicable)

- What is your primary method of transportation?
- Do you own or lease your car?
- Where is it typically parked?
- Can you walk me through what you do when you get out of your car, once it is parked?
  - What do you do if you have to store items in the car?
- Are there any other strategies, which you have not mentioned, that you use to protect your vehicle?
- Why do you use these strategies for protecting your vehicle? *For each strategy, ask:*
  - When did you start using this strategy?
  - How do you feel that this strategy works to protect you?
  - Why did you choose to use this strategy over using a different one?
  - What are you most worried about?
  - Have you ever had a negative experience?
  - Do you know anyone who has had a negative experience?
  - Are there ever times when you do not choose to use this strategy?
  - Is this strategy something that is important to you, or something you feel is more important to people with whom you share the car (if applicable)?

    - Why would you say that it is more important to [you/other]?
    - Where or from whom did you learn this strategy?

- Are there strategies you have considered or heard about but do not use?
- How secure do you feel that your car is when it is parked?
- How secure do you feel the belongings you have in your car are, when the car is parked?

Bicycle (if applicable)

- Do you own or rent or bikeshare your bicycle?
- Where is it typically stored?
- Can you walk me through what you do when you get off your bicycle once it is parked somewhere?
    - What type of lock do you use?
    - To what object do you lock the bike?
    - Where do you affix the lock?
- Are there any other strategies, which you have not mentioned, that you use to protect your bike?
- Why do you use these strategies for securing your bike? *For each strategy, ask:*
    - When did you start using this strategy?

    - What are you most worried about?
    - Have you ever had a negative experience?
    - Do you know anyone who has had a negative experience?
    - Are there ever times when you do not choose to use this strategy?
    - Is this strategy something that is important to you, or something you feel is more important to people with whom you share the bike?
        - Why would you say that it is more important to [you/other]?
    - Where or from whom did you learn this strategy?
- Are there strategies you have considered or heard about but do not use?
- How secure do you feel that your bike is when it is unattended?

Personal Security (walking)

- Where do you tend to walk?
    - Do you walk more than 10 minutes a day?
- Are there any particular approaches you take, or items you carry, when walking alone?
- Have you had any martial arts/self defense training?
    - Why did you undergo this training? Who administered the training?
- Why do you use these strategies? *For each strategy, ask:*
    - When did you start using this strategy?

  - How do you feel that this strategy works to protect you?
  - Why did you choose to use this strategy over using a different one?
  - What are you most worried about?
  - Have you ever had a negative experience?
  - Do you know anyone who has had a negative experience?
  - Are there ever times when you do not choose to use this strategy?
  - Where or from whom did you learn this strategy?

- Are there strategies you have considered or heard about but do not use?
- How secure do you feel you are when walking?

General Advice

- Do you ever look for new information or talk to someone about tactics such as for protection your [dwelling, vehicle/bike, self, other members of your family]?
  - Where do you look for this information and with whom do you talk?
- Do you often see news pieces, ads, or articles on TV, in the newspaper, or online with tips/advice, social media posts, chain emails on how to protect your [dwelling, vehicle/bike, self, other members of your family]?
  - How do you feel about the information provided?
  - Are there strategies you have considered or heard about but do not use?
- What other sources do you consult when seeking physical security advice?

- Do you feel that you have the ability to make yourself more physically secure?
- Whom or what would you say has most influenced your overall approach to physical security, and in what way?
- Would you say that you see more advice about digital security or about physical security?
- Which security advice, digital or physical, do you find more trustworthy?
- Which more useful?

A.2 Security Education Questionnaire

- Which of the following personal computing devices do you use at least once a week? This does not include devices used in the workplace.

  1. A desktop computer.
  2. A laptop computer.
  3. A smartphone (e.g. iPhone 5S, Samsung Galaxy S 5).
  4. A feature phone (i.e. a non-touch screen mobile phone).
  5. A tablet (e.g. iPad, Nexus 7 Tablet).
  6. An eReader (e.g. Kindle, Nook).

- The questions in the next section will address your behavior on the computer. Please answer all of the following questions with regard to your behavior at home, not your behavior in the workplace.

- Which of the following best describes your password(s)?

  1. My password(s) for certain accounts (such as banking accounts) are stronger than my passwords for other accounts.
  2. All of my passwords are relatively equal in strength.
  3. Prefer not to answer.

- *For each of the behaviors these "where did you learn" questions were asked, the example below is for passwords.*

  – Where did you learn about making strong passwords? [Multiple Choice]

    1. I read or saw a piece of online, TV, or print media about making strong passwords. For example, a news article, TV show, blog post, or advertisement.
    2. I received information from my workplace regarding making strong passwords.
    3. I received information about making strong passwords while taking a course in school.
    4. I received information from a family member or friend regarding making strong passwords.
    5. I had a negative experience or someone told me about their negative experience, such as having my account hacked, that made me decide to make strong passwords.
    6. I received information from a service provider regarding making strong

passwords. Examples of service providers include TimeWarner, Verizon, and Bank of America.

7. A website or application in which I was making a password displayed a password meter (e.g. measure of password strength).

- If 1 to “where did you learn”: Which of the following best describes the media in which you read or saw information about making strong passwords.

    1. I watched or read a news article about making strong passwords.
    2. I watched a TV show or movie, or read a book, in which a character either made or didn’t make strong passwords.
    3. I was reading an online forum in which making strong passwords was discussed.
    4. I watched or read an advertisement that recommended making strong passwords.

- if 2 to “where did you learn”: Which of the following best describes how you received information about making strong passwords at your workplace?

    1. I received information about making strong passwords from a friend or colleague who is an IT professional.
    2. I received information about making strong passwords from a colleague who is NOT an IT professional.
    3. I received information about making strong passwords through an

IT email or announcement.

4. I received information about making strong passwords through a workplace email or announcement that was NOT from the IT department.
5. I received information about making strong passwords from a workplace course or training.

– If 3 to "where did you learn": Which of the following best describes the professional and/or educational background of the family member or friend who provided you with information regarding making strong passwords.

  1. This family member or friend has an education in, or works in, the field of computer science, computer engineering or IT.
  2. This family member or friend does NOT have an education in, nor do they work in, the field of computer science, computer engineering or IT.

– Which of the following best describes why the information you received made you decide to make strong passwords:

  1. I trusted the person or source of the information.
  2. The information made sense to me.
  3. The information increased my fear of a negative event.
  4. Other: [text entry]

– *This "who benefits" question is asked for each behavior.* In your opinion, who benefits or would benefit from you making strong passwords: you or the website that stores your information?

1. Only I benefit.
2. Mostly I benefit, but the website benefits somewhat.
3. I and the website benefit equally.
4. Mostly the website benefits, but I benefit somewhat.
5. Only the website benefits.
6. Prefer not to answer.

- There are different reasons that people decide to use or not to use anti-virus software on their personal devices. Which of the following best describes you:

  1. I use anti-virus software, such as Norton AntiVirus or Avast, on my computer or mobile phone.
  2. I do not use anti-virus software on my computer or mobile phone.
  3. Prefer not to answer.

- *If the respondent said they used antivirus software, the above block of "where did you learn" questions is asked.*

- *If the respondent said they did not use antivirus software, the following block of "why not" questions is asked.*

  – Which of the following most closely matches your experiences?

    1. I tried using anti-virus software, but then I stopped.

2. I have never tried using anti-virus software.
3. Prefer not to answer.

- *If “I have never tried using anti-virus software”:*
  - Which of the following have you experienced? [Multiple Choice]
    1. My computer, mobile phone, a website, or application, prompted me to use anti-virus software.
    2. I have seen information or been told about anti-virus software.
    3. I have never seen information about anti-virus software and I have never been prompted to use it.
  - *If “My computer...prompted me to use antivirus software”:* After your computer, a website, or an application prompted you to use anti-virus software, what made you decide NOT to use it. [Multiple Choice]
    1. It would be inconvenient.
    2. I did not have time to install anti-virus software.
    3. They were trying to sell me something.
    4. It violated my privacy.
    5. I do not have anything valuable on my computer or mobile phone, so I do not need anti-virus software.
    6. I think I will be hacked anyway, so why bother.
    7. It was too hard to do or I did not understand how to use anti-virus software.

8. I use an Apple (i.e. Macintosh) computer or mobile phone so I do not need anti-virus software.
9. I use a Linux computer so I do not need anti-virus software.
10. Someone else owns the device and I do not decide what software it runs.
11. I have never had a negative experience, so I have not felt the need to use anti-virus
12. I'm careful with my devices, so I do not need to use anti-virus.

∗ *If "I have seen information or been told about antivirus software":* After seeing information about anti-virus software, what made you decide NOT to use it? [Multiple Choice]

1. It would be inconvenient.
2. I did not have time to install anti-virus software.
3. They were trying to sell me something.
4. It violated my privacy.
5. I do not have anything valuable on my computer or mobile phone, so I do not need anti-virus software.
6. I think I will be hacked anyway, so why bother.
7. It was too hard to do or I did not understand how to use anti-virus software.
8. I use an Apple (i.e. Macintosh) computer or mobile phone so I do not need anti-virus software.

            9. I use a Linux computer so I do not need anti-virus software.
            10. Someone else owns the device and I do not decide what software it runs.
            11. I have never had a negative experience, so I have not felt the need to use anti-virus
            12. I’m careful with my devices, so I do not need to use anti-virus.

    - *If “I tried using anti-virus software, but then I stopped.*
        - What made you decide to stop using anti-virus software. [Multiple Choice]
            1. Using anti-virus software was inconvenient.
            2. I do not have anything valuable on my computer or mobile phone, so I do not need anti-virus software.
            3. I started to use an Apple (i.e. Macintosh) computer or mobile phone so I no longer needed anti-virus software.
            4. I started to use a Linux computer so I no longer needed anti-virus software.
            5. Someone else owns the device and I do not decide what software it runs.
            6. I saw or read information that I should not be using anti-virus software.

- *The above questions are then asked for software updating and 2FA. We’ve provided the initial behavior question for each of these sections here:*

– Which of the following best describes your behavior when software updates are available for your personal device? Here are three examples of updates, and how they might appear on your device.

1. I install the updates as soon as I learn of them.
2. I install the updates a little while after I learn of them.
3. I rarely or never install the updates.
4. I install some, but not all updates.
5. Prefer not to answer.

– People have many different reasons for using or not using two-factor authentication. Which of the following answer choices best describes you?

What is two-factor authentication?

- Two-factor authentication is also known as two-step verification.

- Two-factor authentication uses not only a password and username but also an additional verification code, such as a 4-digit code texted to your phone.

- When setting up two-factor authentication, a website might ask for your phone number so that they can text you this code in the future.

1. I use two-factor authentication on all of the websites and apps that offer it.
2. I use two-factor authentication on some but not all of the websites and apps that offer it.

    3. I never use two-factor authentication.
    4. Prefer not to answer.

- Please select the answer choice that says ”Very unhappy”. This question is designed to check whether you are paying attention.

  1. Very happy
  2. Somewhat happy
  3. Neutral
  4. Somewhat unhappy
  5. Very unhappy

- The questions in the next section will address your behavior off the computer.

  – There are many different reasons why people choose to or not to lock their home?s exterior door(s). Which of the following best describes you:

    1. I always lock my home?s exterior door(s).
    2. I sometimes lock my home?s exterior door(s).
    3. I never lock my home?s exterior door(s).
    4. Prefer not to answer.

  – *The same questions as for the other behaviors are asked.*

- With which of the following statements about the usefulness of physical- and digital-security advice do you most agree?

  What is physical-security advice?

Some examples of physical-security advice include information about how to protect your house and information about how to stay safe when walking alone.

What is digital-security advice?

Some examples of digital-security advice include information about anti-virus software or information about making strong passwords.

1. Digital is a lot more useful than physical
2. Digital is somewhat more useful than physical.
3. They are about the same.
4. Physical is somewhat more useful than digital.
5. Physical is a lot more useful than digital.

- With which of the following statements about the trustworthiness of physical- and digital-security advice do you most agree?

  1. Digital is a lot more trustworthy than physical
  2. Digital is somewhat more trustworthy than physical.
  3. They are about the same.
  4. Physical is somewhat more trustworthy than digital.
  5. Physical is a lot more trustworthy than digital.

- *We then asked the 5-Item Web Use Skills Index [105].*

- The next section is the final section in the questionnaire. This section contains demographic questions.

  - Please specify the gender with which you most closely identify.

    1. Female
    2. Male
    3. Other: [text entry]
    4. Prefer not to answer.

  - Please specify your age.

    1. 18-29
    2. 30-39
    3. 40-49
    4. 50-59
    5. 60-69
    6. Over 70
    7. Prefer not to answer.

  - Please specify your ethnicity.

    1. Hispanic or Latino
    2. Black or African American
    3. White
    4. American Indian or Alaska Native
    5. Asian, Native Hawaiian, or Pacific Islander

6. Other: [text entry]
7. Prefer not to answer.

- Please specify the highest degree or level of school you have completed.

  1. Some high school credit, no diploma or equivalent
  2. High school graduate, diploma or the equivalent (for example: GED)
  3. Some college credit, no degree
  4. Trade/technical/vocational training
  5. Associate degree
  6. Bachelor?s degree
  7. Master?s degree
  8. Professional degree
  9. Doctorate degree
  10. Prefer not to answer.

- Please select the response option that best describes your current employment status.

  1. Working for payment or profit
  2. Unemployed
  3. Looking after home/family
  4. A student
  5. Retired
  6. Unable to work due to permanent sickness or disability

7. Prefer not to answer.

- Which of the following best describes your educational background or job field?

  1. I have an education in, or work in, the field of computer science, computer engineering or IT.
  2. I DO NOT I have an education in, nor do I work in, the field of computer science, computer engineering or IT.
  3. Prefer not to answer.

- Please select the statement that best describes your government security clearance status. A security clearance is defined as a clearance to view privileged information, which is given by any government agency in the United States or abroad. For example, a United States Top Secret (TS) clearance or a United Kingdom Security Check (SC) clearance.

  1. I currently have an active security clearance.
  2. I have previously held a security clearance, but currently do not have an active security clearance.
  3. I have never held a security clearance.
  4. Prefer not to answer.

- *If "I have never held a security clearance":*

  1. Health care patient information controlled under the United States Health Insurance Portability and Accountability Act (HIPAA) regulations.

2. Student information controlled under the United States Family Educational Rights and Privacy Act (FERPA) regulations.
3. Other sensitive data such as social security numbers or credit card information.
4. None of the above.
5. Prefer not to answer.

- Please specify the range which most closely matches your total, pre-tax, household income in 2015

    1. Less than $29,999
    2. $30,000-$49,999
    3. $50,000-$74,999
    4. $75,000-$99,999
    5. $100,000-$124,999
    6. $125,000-$149,999
    7. $150,000-$199,999
    8. $200,000 or more
    9. Prefer not to answer.

A.3 Digital Divide Questionnaire

A.4 Survey Questions Used in Our Analysis

Princeton Survey Research Associates International
for
Data & Society Research Institute

Privacy and Security Experiences of Low-Socioeconomic Status Populations

Final Questionnaire
11/18/2015

Total n=3,000 U.S. adults age 18+ with oversample of low-SES adults
n=1,050 landline
n=1,950 cell phone
Pretest: November 11, 2015
Field Dates: November 18-December 22, 2015 (tentative)
Job#: 35017

START TIMING MODULE
LANDLINE INTRO:
Hello, my name is _____ and I'm calling for Princeton Survey Research. We are conducting a telephone opinion survey about some important issues today and would like to include your household. This is NOT a sales call.

May I please speak with the YOUNGEST **[RANDOMIZE:** (MALE / FEMALE)**]**, age 18 or older, who is now at home? **[IF NO MALE/FEMALE, ASK:** May I please speak with the YOUNGEST (FEMALE / MALE), age 18 or older, who is now at home?**]**
GO TO MAIN INTERVIEW

CELL PHONE INTRO:
Hello, my name is _____ and I'm calling for Princeton Survey Research. We are conducting a telephone opinion survey about some important issues today and would like to include you. I know I am calling you on a cell phone. This is NOT a sales call.

[IF R SAYS DRIVING/UNABLE TO TAKE CALL: Thank you. We will try you another time...**]**

VOICEMAIL MESSAGE [LEAVE ONLY ONCE -- THE FIRST TIME A CALL GOES TO VOICEMAIL:] I am calling for Princeton Survey Research. We are conducting a national opinion survey of cell phone users. This is NOT a sales call. We will try to reach you again.

CELL PHONE SCREENING INTERVIEW:
S1 Are you under 18 years old, OR are you 18 or older?

1 Under 18
2 18 or older
9 Don't know/Refused

IF S1=2, CONTINUE WITH MAIN INTERVIEW
IF S1=1, THANK AND TERMINATE – RECORD AS AGE INELIGIBLE: This survey is limited to adults age 18 and over. I won't take any more of your time...

IF S1=9, THANK AND TERMINATE – RECORD AS SCREENING REFUSAL: This survey is limited to adults age 18 and over. I won't take any more of your time...

READ TO ALL CELL PHONE RESPONDENTS
INTRODUCTION TO MAIN INTERVIEW: If you are now driving a car or doing any activity requiring your full attention, I need to call you back later. The first question is...

RESOURCES

ASK ALL INTERNET USERS (EMINUSE=1 OR INTMOB=1):
HOME3NW Do you ever use the internet at HOME? {Modified PIAL Libraries Survey March 2015}

1	Yes
2	No
8	**(VOL.)** Don't know
9	**(VOL.)** Refused

ASK SMARTPHONE OWNERS WHO USE THE INTERNET (SMART1=1 AND [EMINUSE=1 OR INTMOB=1]):
Q4 Overall, when you use the internet, do you do that mostly using your cell phone or mostly using some other device like a desktop, laptop or tablet computer? {PIAL Trend}

1	Mostly on cell phone
2	Mostly on something else
3	**(VOL.)** Both equally
4	**(VOL.)** Depends
8	**(VOL.)** Don't know
9	**(VOL.)** Refused

OUTCOMES

[READ TO ALL:] Now on a different subject...

ASK ALL:

Q16 As far as you know, have you ever had any of these experiences? **[INSERT ITEMS; RANDOMIZE]**. Have you ever had this experience, or not, or are you not sure? How about **[INSERT NEXT ITEM]**? **[READ FOR FIRST ITEM, THEN AS NECESSARY:** Have you ever had this experience, or not, or are you not sure?**]** {PIAL July 11-14, 2013}

a. Had important personal information stolen such as your Social Security Number, your credit card, or bank account information
b. Had medical or health information stolen
c. Had inaccurate information show up in your credit report

ASK ITEMS d-j OF ALL INTERNET USERS (EMINUSE=1 OR INTMOB=1):

d. Had an email or social networking account of yours compromised or taken over without your permission by someone else
e. Had difficulty paying off a loan or cash advance that you signed up for online
f. Been the victim of an online scam and lost money
g. Experienced persistent and unwanted contact from someone online
h. Lost a job opportunity or educational opportunity because of something that was posted online
i. Experienced trouble in a relationship or friendship because of something that was posted online
j. Had someone post something about you online that you didn't want shared

CATEGORIES

1	Yes
2	No
8	Not sure/Don't know
9	**(VOL.)** Refused

ADVICE SOURCES

ASK ALL INTERNET USERS (EMINUSE=1 OR INTMOB=1):

Q17 Next... Have you ever turned to any of the following people or places for advice about how to protect your personal information online? (First,/Next,) **[INSERT ITEM; RANDOMIZE; ITEM 'SOMEONE OR SOMETHING ELSE' ALWAYS LAST]**? **[READ IF NECESSARY:** Have you ever turned there for advice about how to protect your personal information online?**]** PIAL Modified Teens and Privacy Management Survey July 2012

a. A friend or peer
b. A family member
c. A co-worker
d. A librarian or resources at your library
e. A government website
f. A website run by a private organization
g. A teacher
h. Someone or something else? **(SPECIFY)**

CATEGORIES

1 Yes
2 No
3 **(VOL.)** Doesn't apply
8 **(VOL.)** Don't know
9 **(VOL.)** Refused

SOCIOECONOMIC STATUS

ASK ALL:

EDUC2 What is the highest level of school you have completed or the highest degree you have received? **[DO NOT READ] [INTERVIEWER NOTE: Enter code 3-HS grad if R completed training that did NOT count toward a degree]** {EDUC2}

1 Less than high school (Grades 1-8 or no formal schooling)
2 High school incomplete (Grades 9-11 or Grade 12 with NO diploma)
3 High school graduate (Grade 12 with diploma or GED certificate)
4 Some college, no degree (includes some community college)
5 Two year associate degree from a college or university
6 Four year college or university degree/Bachelor's degree (e.g., BS, BA, AB)
7 Some postgraduate or professional schooling, no postgraduate degree
8 Postgraduate or professional degree, including master's, doctorate, medical or law degree (e.g., MA, MS, PhD, MD, JD)
98 Don't know
99 Refused

[MAKE FULL NOTE AVAILABLE FOR INTERVIEWERS: Enter code 3-HS graduate if R completed vocational, business, technical, or training courses after high school that did NOT count toward an associate degree from a college, community college or university (e.g., training for a certificate or an apprenticeship)**]**

ASK ALL:

INC Last year -- that is in **[IF INTERVIEWED IN 2015:** 2014 **/ IF INTERVIEWED IN 2016:** 2015**]** -- what was your total family income from all sources, before taxes? Just stop me when I get to the right category... **[READ]** {INC}

1	Less than $10,000
2	10 to under $20,000
3	20 to under $30,000
4	30 to under $40,000
5	40 to under $50,000
6	50 to under $75,000
7	75 to under $100,000
8	100 to under $150,000, OR
9	$150,000 or more?
98	**(VOL.)** Don't know
99	**(VOL.)** Refused

ASK IF DK OR REFUSED INCOME (INC=98,99):

INC1. It's important for us to have some information about household finances to make sure our survey is accurate. Keeping in mind that this is a completely confidential survey, can you please tell me if your total family income BEFORE taxes last year was **[READ]** {INC1}

1	Under $40,000, OR
2	$40,000 or more?
8	**(VOL.)** Don't know
9	**(VOL.)** Refused

A.5 PSRAI Weighting Information

Privacy and Security Experiences of Low-Socioeconomic Status Populations

Survey Methodology Report

Princeton Survey Research Associates International for
Data & Society Research Institute
January 2016

WEIGHTING AND ANALYSIS

Weighting is generally used in survey analysis to adjust for effects of the sample design and to compensate for patterns of nonresponse that might bias results. The weighting was accomplished in multiple stages to account for the disproportionately-stratified samples, the overlapping landline and cell sample frames, household composition, and differential non-response associated with sample demographics.

The first stage of weighting corrected for different probabilities of selection associated with the number of adults in each household and each respondent's telephone usage patterns.[1] This weighting also adjusts for the overlapping landline and cell sample frames and the relative sizes of each frame and each sample. Since we employed a disproportionately-stratified sample design, the first-stage weight was computed separately for each stratum in each sample frame.

The first-stage weight for the i^{th} case from stratum h can be expressed as:

$$WT_{hi} = \left[\left(\frac{S_{LLh}}{F_{LLh}} \times \frac{1}{AD_{hi}} \times LL_{hi}\right) + \left(\frac{S_{CPh}}{F_{CPh}} \times CP_{hi}\right) - \left(\frac{S_{LLh}}{F_{LLh}} \times \frac{1}{AD_{hi}} \times LL_{hi} \times \frac{S_{CPh}}{F_{CPh}} \times CP_{hi}\right)\right]^{-1}$$

Where S_{LLh} = the size of the landline sample in stratum h
F_{LLh} = the size of the landline sample frame in stratum h
S_{CPh} = the size of the cell sample in stratum h
F_{CPh} = the size of the cell sample frame in stratum h
AD_{hi} = Number of adults in household i of stratum h
LL_{hi}=1 if respondent i of stratum h has a landline phone, otherwise LL_{hi}=0.
CP_{hi}=1 if respondent i of stratum h has a cell phone, otherwise CP_{hi}=0.

[1] i.e., whether respondents have only a landline telephone, only a cell phone, or both kinds of telephone.

This first-stage weight was used as an input weight for the demographic raking. The data was first divided into three groups – African-Americans, Hispanics and others. Each group was raked separately to population parameters for sex, age, education, region and number of adults in the household.

After the raking by race/ethnicity, the combined dataset was raked to total adult population parameters for sex, age, education, region, number of adults in the household, household telephone usage, population density and race/ethnicity.

The telephone usage parameter was derived from an analysis of recently available National Health Interview Survey data[2]. The population density parameter is county-based and was derived from Census 2010 data. All other weighting parameters were derived from an analysis of the 2013 American Community Survey 1-year PUMS file.

Each stage of weighting incorporated previous weighting adjustments. Raking was accomplished using SPSSINC RAKE, an SPSS extension module that simultaneously balances the distributions of all variables using the GENLOG procedure. The rakings correct for differential non-response that is related to particular demographic characteristics of the sample. The weight ensures that the demographic characteristics of the sample closely approximate the demographic characteristics of the target population. Table 1 compares weighted and unweighted total sample demographics to population parameters.

[2] Blumberg SJ, Luke JV. Wireless substitution: Early release of estimates from the National Health Interview Survey, July-December, 2014. National Center for Health Statistics. June 2015.

Table 1. Sample Demographics

	Parameters	Unweighted	Weighted
Sex			
Male	48.2%	52.4%	48.7%
Female	51.8%	47.6%	51.3%
Age			
18-29	20.9%	16.3%	20.1%
30-49	34.7%	24.6%	32.6%
50-64	26.0%	28.8%	25.4%
65+	18.4%	27.0%	18.6%
Education			
LT HS	13.3%	12.8%	12.6%
HS graduate	28.0%	27.4%	27.8%
Some college	31.0%	24.0%	30.0%
College graduate	27.7%	34.6%	28.7%
Region			
Northeast	18.0%	13.7%	17.3%
Midwest	21.3%	13.6%	20.3%
South	37.3%	46.2%	38.5%
West	23.4%	26.6%	23.8%
# of adults in HH			
1	16.5%	27.4%	17.5%
2	51.9%	48.9%	51.3%
3+	31.6%	23.8%	31.2%
HH phone use			
LLO	7.4%	4.7%	6.4%
Dual	44.8%	55.5%	45.2%
CPO	47.8%	39.8%	48.4%
Population Density			
1-Lowest	19.9%	30.3%	20.7%
2	20.0%	18.6%	20.0%
3	20.1%	14.6%	19.9%
4	20.0%	13.3%	19.4%
5-Highest	20.0%	23.3%	20.0%

Continued…

Table 1. Sample Demographics (continued)

	Parameters	Unweighted	Weighted
Race/ethnicity			
White, not Hispanic	65.8%	58.1%	62.8%
Black, not Hispanic	11.5%	14.0%	11.8%
Hispanic, native born	7.5%	8.9%	7.8%
Hispanic, foreign born	7.5%	9.7%	7.8%
Other, not Hispanic	7.6%	6.7%	7.4%

EFFECTS OF SAMPLE DESIGN ON STATISTICAL INFERENCE

Specialized sampling designs and post-data collection statistical adjustments require analysis procedures that reflect departures from simple random sampling. PSRAI calculates the effects of these design features so that an appropriate adjustment can be incorporated into tests of statistical significance when using these data. The so-called "design effect" or *deff* represents the loss in statistical efficiency that results from a disproportionate sample design and systematic non-response. PSRAI calculates the composite design effect for a sample of size *n*, with each case having a weight, w_i as:

$$deff = \frac{n\sum_{i=1}^{n} w_i^{2}}{\left(\sum_{i=1}^{n} w_i\right)^2}$$

In a wide range of situations, the adjusted standard error of a statistic should be calculated by multiplying the usual formula by the square root of the design effect (√*deff*). Thus, the formula for computing the 95% confidence interval around a percentage is:

$$\hat{p} \pm \left(\sqrt{deff} \times 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

where $\hat{p}$ is the sample estimate and *n* is the unweighted number of sample cases in the group being considered.

The survey's margin of error is the largest 95% confidence interval for any estimated proportion based on the total sample — one around 50%. For example, the margin of error for the total sample is ±2.7 percentage points. This means that in 95 out every 100 samples using the same methodology, estimated proportions based on the entire sample will be no more than 2.7 percentage points away from their true values in the population. It is important to remember that sampling fluctuations are only one possible source of error in a survey estimate. Other sources, such as measurement error, may contribute additional error of greater or lesser magnitude.

A.6 Articles Used to Prompt Search Query Generation

- https://www.zdnet.com/article/previously-unseen-malware-behind-cyberattack-ag

- https://mobile.wnd.com/2017/03/operating-system-movie-computer-virus-stored-

- https://www.pbs.org/newshour/show/ransomware-attack-takes-down-la-hospital-f

- https://www.mysanantonio.com/business/local/article/Computer-hackers-steal-Sa
php

- https://www.marketwatch.com/story/your-childs-teddy-bear-may-now-be-hacked-2

- https://www.wired.com/2017/03/Internet-bots-fight-theyre-human/

A.7 Advice Quality Expert Accuracy Evaluation Questionnaire

See Figure 6.18 for an example of how the advice for which these questions were asked was displayed to the respondent.

For each piece of advice:

1. People have many different practices when it comes to online privacy and security. Do you currently follow this advice? Your answer will have no bearing on your payment for this study.

   (a) Yes (at least some of the time)

   (b) No (never)

   (c) Not applicable

2. Please select the option that best matches your opinion.

    (a) Following this advice would IMPROVE someone's digital security or privacy at least a little bit (e.g., this advice is beneficial)

    (b) Following this advice would HARM someone's digital security or privacy at least a little bit (e.g., this advice is harmful)

    (c) Following this advice would have ABSOLUTELY NO EFFECT on someone's digital security or privacy (e.g., this advice is useless)

3. *[If they answered 1 to Q2]*:

    - How much would you estimate that following this advice would IMPROVE the typical end user's digital security or privacy (e.g., DECREASE security/privacy risk)?

        (a) 0% decrease in risk

        (b) 5% decrease in risk

        (c) 10% decrease in risk

        (d) 15% decrease in risk

        (e) 20% decrease in risk

        (f) 25% decrease in risk

        (g) 30% decrease in risk

        (h) 40% decrease in risk

        (i) 50%+ decrease in risk

- For how long do you think this advice will remain useful for improving people's security?
  (a) For the next year (0-1 years)
  (b) For the next few years (2-5 years)
  (c) For the next five to ten years (5-10 years)
  (d) For the next few decades (10+ years)
  (e) Other: [text entry]
  (f) I don't know

4. *[If they answered 2 to Q2]*: How much would you estimate that following this advice would HARM the typical end user's digital security or privacy (e.g., INCREASE security/privacy risk)?
   (a) 0% increase in risk
   (b) 5% increase in risk
   (c) 10% increase in risk
   (d) 15% increase in risk
   (e) 20% increase in risk
   (f) 25% increase in risk
   (g) 30% increase in risk
   (h) 40% increase in risk
   (i) 50%+ increase in risk

5. How important do you think this advice is to recommend to the typical end user for personal computer or mobile device use?

    (a) The number 1 behavior I would recommend

    (b) In the top 3 behaviors I would recommend

    (c) In the top 5 behaviors I would recommend

    (d) In the top 10 behaviors I would recommend

    (e) I would recommend it, but it's not in the top 10 behaviors I would recommend

    (f) I would not recommend this advice

A.8 Advice Quality User Actionability Evaluation Questionnaire

See Figure 6.18 for an example of how the advice for which these questions were asked was displayed to the respondent.

1. People have many different practices when it comes to online privacy and security. Do you currently follow this advice? Your answer will have no bearing on your payment for this study.

    (a) Yes (at least some of the time)

    (b) No (never)

    (c) Not applicable

2. How **difficult** do you think it would be for you to follow this advice on your personal/non-work computer or mobile device?

(a) Very difficult

(b) Somewhat difficult

(c) Slightly difficult

(d) Not at all difficult

3. How **time consuming** do you think it would be for you to follow this advice on your personal/non-work computer or mobile device?

    (a) Very time consuming

    (b) Somewhat time consuming

    (c) Slightly time consuming

    (d) Not at all time consuming

4. How **disruptive** do you think it would be for you to follow this advice on your personal/non-work computer or mobile device?

    (a) Very disruptive

    (b) Somewhat disruptive

    (c) Slightly disruptive

    (d) Not at all disruptive

5. How **confident** do you feel that you could implement this advice?

    (a) Very confident

    (b) Confident

(c) Slightly confident

(d) Not at all confident

A.9 Software Update Questionnaire

- Condition: Self
  - Q1S: Imagine that you see the message below appear on your computer. [image of update message] Would you install the update?
    - Yes, the first time I saw this message.
    - Yes, within a week of seeing this message.
    - Yes, within a few weeks of seeing this message.
    - Yes, within a few months of seeing this message.
    - No.
    - I don't know.
  - Q2S: What would make you want to install this update? [multiple selection, optional]
    - I always install updates *(Mapping: General Tendency)*
    - I trust this software company *(Mapping: Application)*
    - The features seem like something I would want *(Mapping: Features / Security-Only)*
    - I wasn't satisfied with the current version
    - The current version was broken

    - It was a security related update
    - I use this software frequently, so keeping it updated is important *(Mapping: Application)*
    - Previous updates that I have installed for this software made the software or my computer crash less *(Mapping: Risk)*
    - I don't have to restart to install this update *(Mapping: Cost)*
    - It seemed like it wouldn't take very long to complete this update *(Mapping: Cost)*
    - Other: [text entry]

  - Q3S: What would make you not want to install this update? [multiple selection, optional]
    - I rarely install updates *(Mapping: General Tendency)*
    - I wouldn't have time *(Mapping: Cost)*
    - I wouldn't want to restart *(Mapping: Cost)*
    - I wouldn't want to lose stuff while updating *(Mapping: Risk)*
    - It looked like it would be disruptive
    - This update didn't seem important
    - The update was not related to security
    - I do not use this software frequently, so keeping it updated is not important *(Mapping: Application)*
    - I wouldn't want the features it would add *(Mapping: Features / Security-Only)*

        - I'm satisfied with the current version
        - The update might make the application harder to use *(Mapping: Risk)*
        - I don't trust this software company *(Mapping: Application)*
        - Too many updates for this software
        - The software or my computer crashed more after I have updated in the past *(Mapping: Risk)*
        - I have had trouble updating this application in the past *(Mapping: Risk)*
        - I would worry about compatibility issues *(Mapping: Risk)*
        - I wouldn't want to lose stuff while updating *(Mapping: Risk)*
        - Other: [text entry]

- Condition: Friend
    - Q1F: Imagine that a friend or relative sees the message below on their computer and calls you for advice. What would you tell them?
        - Install the update immediately.
        - Install the update sometime this week.
        - Install the update within a few weeks.
        - Install the update within a few months.
        - Don't install the update
        - I don't know

- Q2F: What would make you tell your friend to install this update? [multiple selection, optional]
  - I always install updates
  - I trust this software company
  - The features seem like something they would want
  - They weren't satisfied with the current version
  - The current version was broken
  - It was a security related update
  - They use this software frequently, so keeping it updated is important
  - Previous updates that they have installed for this software made the software or their computer crash less
  - They don't have to restart to install this update
  - It seemed like it wouldn't take very long to complete this update
  - Other: [text entry]
- Q3F: What would make you not recommend that your friend install this update? [multiple selection, optional]
  - I don't install updates
  - They wouldn't have time
  - They wouldn't want to restart
  - They wouldn't want to lose stuff while updating
  - It looked like it would be disruptive

        - ∗ This update didn’t seem important
        - ∗ The update was not related to security
        - ∗ They do not use this software frequently, so keeping it updated is not important
        - ∗ They wouldn’t want the features it would add
        - ∗ They are satisfied with the current version
        - ∗ The update might make the application harder to use
        - ∗ I don’t trust this software company
        - ∗ Too many updates for this software
        - ∗ The software or their computer crashed more after they have updated in the past
        - ∗ They have had trouble updating this application in the past
        - ∗ I would worry about compatibility issues
        - ∗ They wouldn’t want to lose stuff while updating
        - ∗ Other: [text entry]

The order of [Q4-7], Q8, and Q9 was randomized.

- Q4: Over the past year, how frequently do you feel like [application] has frozen (e.g., hung) or crashed?
    - Less than once a week
    - At least once a week but not more than three times a week
    - At least three times a week but not more than five times a week

  - Five times a week or more

- Q5: Over the past year, have you noticed that updating [application] changes how frequently it freezes (e.g., hangs) or crashes?
  - Yes, it crashes more after I update.
  - Yes, it crashes less after I update.
  - No, updating [application] has no impact on how frequently it crashes.
- Q6: Over the past year, how frequently do you feel like any application on your computer or your computer itself crashed?
  - Less than once a week
  - At least once a week but not more than three times a week
  - At least three times a week but not more than five times a week
  - Five times a week or more
- Q7: Over the past year, have you noticed that updating [application] changes how frequently any application on your computer or your computer itself crashes?
  - Yes, my computer crashes more after I update.
  - Yes, my computer crashes less after I update.
  - No, updating [application] has no impact on how frequently my computer crashes.

- Q8: In general, how quickly do you install updates for applications on your computer or for your computer itself (e.g., the computer operating system)?
  - As soon as I see the update prompt.
  - Within a week of seeing the prompt.
  - Within a few weeks of seeing the prompt.
  - Within a few months of seeing the prompt.
  - I don't install updates that appear on my computer.
  - I don't know.
- Q9: Do you use any of the following software on your home or work computer? [Multiple answer]
  - A Norton software product (for example, Norton AntiVirus, Norton Family Premier, Norton Mobile Security, Norton Small Business)
  - A Symantec software product (for example, Symantec AntiVirus, Symantec Endpoint Protection)
  - Another anti-virus software (for example, McAfee Antivirus Plus, Kaspersky AntiVirus, Bitdefender Antivirus Plus)
  - None of the above
  - I Don't Know

Appendix B: Additional Data

B.1 Prevalence of Security and Privacy Outcomes Surveyed in Chapter 5

Table 1 presents a comparison of the prevalences of negative security and privacy outcomes in our sample overall in comparison with a 1,002 respondent Pew Research Center survey conducted in 2013, which asked the same questions [177].

For additional context, Table 2 compares the prevalences of security and privacy outcomes by income group, and Table 3 compares the prevalences of security and privacy outcomes by education group.

Outcome	**Overall Sample**	**Pew**
Stolen Info.	25%	10%
Account compromised	18%	21%
Scam Victim	7%	6%
Lost Job	2%	1%
Posted Without Permission	18%	N/A
At Least One Neg. Experience	49%	N/A

Table B.1: Comparison of outcome prevalence in our sample vs. Pew Research Center 2013 Trendline

Outcome	**Overall Sample**	**<$20K**	$20-$40K	**>$40K**
Stolen Info.	25%	20%	20%	29%
Account compromised	18%	19%	15%	24%
Scam Victim	7%	13%	10%	8%
Lost Job	2%	5%	2%	1%
Posted Without Permission	18%	19%	19%	15%
At Least One Neg. Experience	49%	52%	46%	50%

Table B.2: Comparison of outcome prevalence by income.

Outcome	**Overall Sample**	**< H.S.**	**H.S. - B.S.**	**B.S. or Above**
Stolen Info.	25%	16%	21%	32%
Account compromised	18%	15%	16%	20%
Scam Victim	7%	8%	9%	8%
Lost Job	2%	1%	2%	3%
Posted Without Permission	18%	17%	16%	17%
At Least One Neg. Experience	49%	47%	47%	53%

Table B.3: Comparison of outcome prevalence by education.

B.2 Regression Tables for Chapter 5 Advice Models

Below we present the regression results from modeling whether a user reported advice from a: coworker (Table 4), website (Table 3 in main paper), government website (Table 5), librarian (Table 6), teacher (Table 7), or friend (Table 8), as a function of their SES factors.

B.3 Regression Results for Psycholinguistic Features

We modeled Cloze scores in a series of mixed effect models as a function of the following psycholinguistic features:

Factor	**OR**	**CI**	***p*-value**
<H.S.	0.43	[0.15, 1.26]	0.13
H.S. to B.S.	0.50	[0.32, 0.77]	¡ 0.01*
<$20K	0.85	[0.46, 1.58]	0.61
$20-$40K	1.07	[0.63, 1.81]	0.8
R: Cell only	0.84	[0.55, 1.3]	0.44
R: Home Internet	1.75	[0.55, 5.62]	0.34

Table B.4: Regression results for coworker advice source model. Factors starting with 'R' are boolean resource factors, while the baseline for the categorical income factor is income <$40K and the baseline for education is a B.S. or above.

Factor	**OR**	**CI**	***p*-value**
<H.S.	0.23	[0.07, 0.74]	0.01*
H.S. to B.S.	0.75	[0.43, 1.31]	0.31
<$20K	1.42	[0.71, 2.83]	0.32
$20-$40K	1.59	[0.85, 2.96]	0.15
R: Cell only	0.61	[0.36, 1.03]	0.06
R: Home Internet	1.18	[0.31, 4.48]	0.8

Table B.5: Regression results for government website advice source model.

Factor	**OR**	**CI**	***p*-value**
<H.S.	0.60	[0.1, 3.78]	0.59
H.S. to B.S.	1.21	[0.52, 2.83]	0.66
<$20K	1.77	[0.63, 4.94]	0.28
$20-$40K	1.08	[0.46, 2.51]	0.87
R: Cell only	1.27	[0.61, 2.63]	0.52
R: Home Internet	0.59	[0.1, 3.51]	0.56

Table B.6: Regression results for librarian advice source model.

Factor	**OR**	**CI**	***p*-value**
<H.S.	0.44	[0.11, 1.71]	0.23
H.S. to B.S.	1.04	[0.48, 2.24]	0.93
<$20K	1.52	[0.69, 3.36]	0.3
$20-$40K	1.34	[0.61, 2.95]	0.47
R: Cell only	0.94	[0.5, 1.78]	0.85
R: Home Internet	4.04	[0.52, 31.22]	0.18

Table B.7: Regression results for teacher advice source model.

Factor	**OR**	**CI**	***p*-value**
<H.S.	0.57	[0.27, 1.21]	0.15
H.S. to B.S.	0.74	[0.5, 1.09]	0.12
<$20K	1.02	[0.63, 1.64]	0.95
$20-$40K	1.23	[0.8, 1.88]	0.36
R: Cell only	0.94	[0.65, 1.35]	0.73
R: Home Internet	2.16	[0.88, 5.32]	0.09

Table B.8: Regression results for friend advice source model.

- *age of acquisition*: the average age at which the words in the document are acquired by "typical" children. Word ratings are drawn from the MRC psycholinguistic database [47], which scores 1903 words; higher scores indicate words learned later.

- *familiarity*: how familiar the words are to a "typical" reader. Word ratings are drawn from MRC, which scores 3488 words; higher scores are given to more familiar words.

- *concreteness*: how non-abstract (concrete) the words in the document are. Word ratings are drawn from MRC, which scores 4293 words.

- *meaningfulness*: how broadly applicable a word is (e.g., people has high meaningfulness vs. abbess which has low meaningfulness). Word ratings are drawn from MRC, which scores 2627 words; higher score indicates a more meaningful word.

- *imagability*: how easy it is to construct a mental image of the words used in the document. Word ratings are drawn from MRC, which scores 4825 words; a higher score indicates a more imagable word (e.g., hammer vs. dogma).

Factor	O.R.	95%CI	p-value
ageofacq	0.996	[0.995,0.997]	< 0.001
familiarity	0.997	[0.995,0.999]	0.340
concreteness	0.997	[0.996,0.999]	0.035
meaningfulness	1.00	[0.999,1.00]	0.187
imagability	0.993	[0.990,0.996]	0.231
polysemy	0.994	[0.991,0.998]	0.003
hypernymy	0.458	[0.416,0.504]	< 0.001

Table B.9: Regression statistics for psycholinguistic Cloze analysis.

- *polysemy*: the number of senses that can be meant by a particular word; this can be indicative of text ambiguity, because the words can be interpreted in many ways. Word ratings are determined via WordNet synsets (i.e., groups of related lexical items); higher polysemy means a word has more possible senses.

- *hypernymy*: how specific a word is, lower scores indicate more specificity (lower on a WordNet tree, with fewer subordinates).

In Table B.9 we summarize the odds ratios and significance. Significance is determined by comparing models built with all factors but this factor to the null model using log likelihood tests.

B.4 Useless Advice

The pieces of advice listed below were rated as useless for improving security by the majority of experts (at least two of three).

All advice is of the form “You should...”:

- consider opening a credit card for online use only [all experts agree]

- file taxes early[all experts agree]

- let your children teach you about the Internet too [all experts agree]
- use an unbranded smartphone [all experts agree]
- ask people to remove your personal information and photos
- be aware of your online reputation
- bring proof-of-purchase for computer equipment when traveling
- carry laptops in something other than laptop cases
- change your respondentname regularly
- contact police or authority figures in case of a cyberattack or cyberbullying
- create keyboard patterns to help with remembering passwords
- create pronounceable passwords
- disable and/or limit caching
- encourage the positive sides of the Internet with children and friends
- install software in phases
- keep the computer in a common room in your house if you have children
- not meet up with people you've met online
- not use credit or debit cards online
- not use encryption when sending e-mail to a listserv

- regularly search for your name
- shut down your computer
- store passwords in a file
- try alternate urls to avoid censorship
- understand new features before you try them
- upgrade your email provider
- use a load balancer

B.5 Harmful Advice

The pieces of advice listed below were rated as harmful by **at least one** expert.

All advice is of the form “You should...”:

- base passwords on upcoming events
- buy devices with passwords, preferably passwords that you can change
- change passwords often
- clear your cookies
- create a new email address if your last one is compromised
- create keyboard patterns to help with remembering passwords
- download a filtering software to prevent website access

- draw shapes on your keyboard to generate passwords
- feel comfortable making weak passwords for sites thar don't keep personal info
- install firmware on mobile devices
- isolate iot devices on their own network
- keep sensitive information on removable storage media
- lock your sim card in your smartphone
- not change browser security settings
- not change your passwords unless they become compromised
- not download or execute any files
- not identify yourself to websites
- not open attachments from unknown senders
- not respond to or retaliate against cyberbullies
- not send or forward files you haven't scanned for viruses
- not shut down your computer
- not use a password manager
- not use encryption when sending e-mail to a listserv
- not use extensions or plugins

- obfuscate something meaningful to generate a password
- protect your computer from power surges
- remove improper and/or sensitive information from the web
- store passwords in a file
- store passwords properly
- transfer sensitive files to network shares
- turn off automatic downloads
- use different personas online
- use filters in email
- use less common software
- use private search engines
- use tor
- use tracking applications
- write down passwords on paper

B.6 High Prioirty Advice

The pieces of advice listed below had a median rating of “top 3” or above from the majority of experts.

All advice is of the form “You should...”:

- ask people to remove your personal information and photos
- be careful entering passwords in public computers
- bring proof-of-purchase for computer equipment when travelling
- buy devices with passwords, preferably passwords that you can change
- clear your cookies
- encourage others to use strong passwords
- feel comfortable making weak passwords for sites thar don't keep personal info
- install software in phases
- keep antivirus/antimalware up to date
- monitor network traffic on your router
- not give out your email address without good reason
- not tell anyone your passwords, even it
- not use passwords based on personal information
- protect devices against power surges
- regularly back up your data
- remember your passwords
- scan attachments you open for viruses

- try alternate urls to avoid censorship
- update devices
- use a password manager
- use administrator rights to prevent unauthorized actions
- use anti-malware software
- use different passwords for different accounts/devices
- use unique passwords
- use unique passwords for different accounts

B.7 Computed Priority Ranking of Advice by Experts

Use unique passwords for different accounts	1.81
Update devices	1.88
Use anti-malware software	1.91
Scan attachments you open for viruses	1.99
Use different passwords for different accounts/devices	2.06
Use unique passwords	2.1
Encourage others to use strong passwords	2.17
Not tell anyone your passwords, even IT	2.18
Use end-to-end encryption for communication	2.19
Remember your passwords	2.22
Keep passwords safe if written down	2.35
Not identify yourself to websites	2.36
Keep antivirus/antimalware up to date	2.38
Monitor network traffic on your router	2.41

Not store passwords in a file	2.45
Use strong passwords	2.45
Turn on automatic updates for devices	2.46
Be careful entering passwords in public computers	2.51
Install only trusted software	2.55
Use administrator rights to prevent unauthorized actions	2.55
Not use passwords based on personal information	2.58
Lock your smartphone with passcode or touch ID	2.62
Use 2+ factor authentication	2.62
Not give out your email address without good reason	2.62
Use a combination of letters, numbers, and special characters in passwords	2.64
Test your firewall	2.7
Protect devices against power surges	2.7
Regularly back up your data	2.71
Set rules for your kids about the Internet	2.72
Read install prompts	2.72
Use a VPN	2.73
Use a password manager	2.74
Replace letters with symbols in your passwords	2.78
Use single sign-on SSO	2.78
Beware of revealing personal information unless you know who you're talking to	2.81
Set your antivirus/antimalware to run periodic full scans	2.84
Watch out for phishing	2.9
Lock your computer when you're away from it	2.91
Not do online banking on a public computer	2.92
Change passwords and security questions on compromised accounts	2.92
Not click on ads	2.93
Apply real-world common sense and follow your instincts	2.95
Set antivirus to autoscan email	2.95
Verify who you are talking to	2.96
Not open attachments from unknown senders	2.96
Be wary of false emails from trusted institutions	2.97

Not run or keep unnecessary programs	2.97
Change default passwords on devices/networks/services	2.99
Disable your webcam	3
Only use HTTPS	3.01
Not post vulnerable information (addresses, credit card numbers, etc.)	3.02
Be suspicious if something is too good to be true	3.02
Use passwords	3.1
Be careful with permissions	3.1
Encourage children to talk to you if they feel uncomfortable online	3.13
Not use dictionary words as or in passwords	3.14
Download only trusted programs	3.15
Be suspicious of attachments	3.16
Be suspicious of unusual email if grammar in an email is not good	3.17
Delete phishing or spam emails, even if you might know the sender	3.18
Don’t trust sites with certificate warnings	3.19
Only allow authorized users to access your network	3.19
Disable features you aren’t using (BlueTooth, WiFi, etc.)	3.2
Double check email addresses	3.2
Not store data if you don’t need to	3.24
Not click random or unfamiliar links from unknown senders	3.24
Turn on download notifications	3.26
Only download things you are looking for	3.26
Use antivirus	3.27
Disable Autorun to prevent malicious code from running	3.27
Watch for spelling mistakes in provided URLs	3.28
Use a disposable email service	3.28
Update applications	3.28
Avoid plugging external devices into computers	3.29
Take only devices you need when traveling	3.3
Only reveal financial information to reputable actors	3.31
Delete spam	3.32
Fully reset hacked devices	3.33

Always keep learning about security and privacy	3.33
Make a threat model	3.33
Check if website you're visiting uses HTTPS	3.34
Not friend people you don't know	3.34
Look at the URL bar to verify you're at the intended website	3.35
Not use loyalty cards	3.35
Visit only known websites	3.36
Be suspicious of unusual email	3.37
Verify suspicious email and email contents	3.37
Not share third party personal information i.e. friends and family	3.38
Keep virus definitions up to date	3.38
Disable macros	3.39
Avoid passwords with only numbers	3.39
Avoid illegal or unaffiliated download sites	3.39
Ensure Wifi is secured to at least WPA2 level	3.4
Install a firewall	3.42
Disable automatic download of email attachments	3.43
Understand who to trust online	3.44
Not follow links in spam	3.45
Not trust the From address on an email	3.48
Educate yourself on how to avoid fraud	3.48
Be careful of downloads	3.49
Secure your router	3.49
Physically destroy drives you're done with and wish to erase	3.49
Not use built-in erasing on SSDs	3.49
Be suspicious of links	3.5
Minimize network exposure for control systems	3.5
Not click on flashy things	3.52
Use privacy settings	3.53
Encourage children to follow age limit guidelines for websites	3.54
Clear your cookies	3.54
Be careful who uses your computer	3.54

Look for real-world contact information while online	3.54
Buy devices with passwords, preferably passwords that you can change	3.56
Not enter passwords after clicking links in email	3.56
Encrypt your device data	3.56
Block unwanted users	3.57
Apply the highest level of security that’s practical	3.57
Use a secure machine to hold confidential data	3.58
Use long passwords	3.58
Use a password that’s different from your username	3.58
Monitor credit cards for unauthorized activity	3.6
Not open unnecessary attachments	3.6
Use a password to protect your WiFi	3.6
Set up auto-lock timers for your smartphone	3.61
Encourage others to use Tor	3.62
Turn on automatic updates for applications	3.62
Turn off Bluetooth	3.62
Scan files downloaded from websites for viruses	3.63
Not use repetitive characters in passwords	3.63
Not use look-alike substitutions for your password	3.64
Not enter sensitive information or credentials without HTTPS	3.64
Beware of ”free” products	3.64
Not use debit cards	3.64
Not give out your email address for free samples or products	3.66
Be wary of proxy servers	3.66
Turn off location services	3.67
Change passwords often	3.67
Secure devices and fix vulnerabilities that caused the breach	3.68
Encrypt your email	3.69
Turn off/limit pop-ups	3.69
Be suspicious of unusual email of things even from known people	3.7
Use randomly generated passwords or password generator websites	3.71
Beware of incognito mode	3.71

Backup your password database	3.71
Avoid using open Wi-Fi networks for business, banking, shopping etc.	3.71
Beware of free VPN programs	3.72
Not store mobile passwords directly on the device	3.72
Remain anonymous online	3.72
Disable extension-hiding for known file types	3.72
Use anti-spyware	3.72
Use a virtual machine or even multiple virtual machines	3.73
Remove unsafe devices from the network	3.73
Secure other devices like you would your computer	3.73
Use browsers that protect against phishing	3.74
Encrypt your hard drive	3.74
Not give out your email address for free software downloads	3.77
Not store passwords online	3.77
Verify file signatures	3.77
Check that websites have valid digital certificates	3.78
Watch for unusual posts on your account	3.78
Do sensitive tasks on dedicated and trusted devices	3.79
Encrypt select groups of files	3.79
Use secure payment methods like PayPal, BPay, or credit cards	3.8
Understand what permissions you give new software downloads	3.8
Not blindly trust HTTPS	3.8
Use security extensions	3.8
Not let computers or browsers remember passwords	3.81
Use a proxy server	3.82
Only add people you know in the offline world to contact lists	3.82
Report account breaches or losses to the appropriate people	3.82
Use different computers for work and home use	3.82
Verify URLs you visit	3.84
Monitor where your kids go online	3.84
Exit sites your browser warns are malicious	3.84
Run a virus scan on new devices	3.84

Be suspicious of popups and requests, even from known sources	3.84
Remain calm and talk with kids about bad web experiences	3.86
Not use automatic network log-in	3.87
Use encryption	3.87
Keep your private key safe	3.87
Review your root certificates	3.87
Use ad blocker extensions	3.87
Not send or forward files you haven't scanned for viruses	3.88
Overwrite deleted files	3.88
Encrypt your WiFi	3.9
Unsubscribe from unwanted email lists	3.9
Check the extensions of downloaded files	3.9
Read privacy policies	3.91
Use a paid spam filtering service	3.91
Document cyberbullying incidents	3.91
Disable message and image previews	3.91
Report suspicious things to IT or support	3.91
Not be lulled into a false sense of security from antivirus/firewall	3.91
Turn off automatic downloads	3.91
Use passphrases	3.91
Seek expert help	3.93
Encrypt cloud data	3.95
Make sure no one is watching you enter passwords	3.95
Be wary of using public computers that could be infected	3.95
Research the security of IoT devices before purchase	3.95
Obfuscate something meaningful to generate a password	3.96
Log out of accounts	3.96
Use passwords that are dissimilar to previous ones	3.96
Read terms of service	3.96
Start your PC in Safe Mode when you need to troubleshoot viruses	3.97
Use unusual phonetics in passwords	3.97
Confirm public WiFi information with staff before using	3.99

Not send executable programs with macros	4
Treat security questions like passwords	4.01
Encrypt your other devices	4.01
Develop a mnemonic for complex passwords	4.0199999999999996
Use both upper and lower case in passwords	4.04
Turn off remote access/management features	4.04
Restrict physical access to computers and removable media	4.04
Discard devices with security weaknesses that can't be fixed	4.05
Disable third-party cookies	4.05
Buy devices with security-focused platforms	4.05
Disable 2G support	4.06
Keep your receipts	4.06
Delete originals once a document has been encrypted	4.06
Beware of malware	4.07
Whitelist executable directories to prevent malicious binaries	4.07
Regularly search for your name	4.08
Consider opening a secondary account for shopping etc.	4.09
Use a load balancer	4.09
Install latest OS updates	4.09
Do online banking only on your own computer	4.10
Check camera logs	4.10
Disable "Universal Plug and Play (UPnP)" on your router	4.10
Make sure to overwrite files you want to delete	4.10
Not forward email unnecessarily	4.11
Enable remote data wiping for devices	4.11
Base passwords on upcoming events	4.11
Not respond to or retaliate against cyberbullies	4.12
Use filters in email	4.12
Not forward cyberbullying	4.13
Not sign up for unnecessary accounts	4.13
Request your data from sites or corporations that store it	4.13
Opt out of ad tracking	4.14

Be aware of the Internet	4.15
Not use a password manager	4.15
Not enable file sharing on networks exposed to the Internet	4.15
Only copy files onto machines using physical media	4.16
Use electronic bank statements	4.17
Be aware of what you share	4.17
Pay attention to virus warnings	4.18
Manage and track cookies	4.18
Write down password clues	4.19
Report messages as spam	4.19
Disconnect your computer from the Internet when you're away	4.2
Be suspicious	4.2
Keep the computer in a common room in your house if you have children	4.2
Check your credit report regularly	4.21
Not write down passwords	4.21
Keep track of file extensions	4.22
Perform a factory reset before device disposal	4.22
Not reply to spam	4.23
Be careful using email	4.23
Disable sharing on peer-to-peer apps	4.24
Take note of the countries your VPN providers works in	4.26
Change and rethink security questions	4.27
Create a network demilitarization zone (DMZ)	4.27
Keep your own data locally (not in the cloud or on a remote server)	4.28
Not use your real name online	4.28
Understand data usage and storage	4.28
Change your router name from the default	4.28
Unmount encrypted disks	4.28
Not use extensions or plugins	4.28
Monitor online accounts for unusual activity	4.28
Remove improper and/or sensitive information from the web	4.28
Securely wipe devices before disposal, where possible	4.28

Configure antivirus to scan all files in real time	4.29
Use incognito mode	4.3
Create separate networks for devices	4.31
Not open email from unknown senders	4.31
Consider partitioning your computer into seperate accounts	4.34
Not overwrite SSDs	4.34
Only do business with trusted institutions	4.34
Cancel or change accounts if you are being cyberbullied	4.35
Leave unsafe websites	4.35
Clear your browser history	4.35
Seek professional help for cybersecurity issues	4.36
Understand the Internet	4.36
Limit the number of antivirus applications you install	4.37
Look for the lock icon in the address bar	4.37
Be aware of your online reputation	4.37
Turn down transmission strength	4.38
Use an alarm on your devices	4.39
Disconnect from the Internet	4.39
Discuss identifiable information in private (so you are not overheard)	4.39
Not rely on mobile access as a primary means for email	4.40
Draw shapes on your keyboard to generate passwords	4.40
Set browser to click-to-play for videos and ads	4.40
Consider opening a credit card for online use only	4.40
Keep sensitive information on removable storage media	4.40
Try alternate URLs to avoid censorship	4.40
Be cautious when picking an email address	4.40
Store passwords properly	4.41
Not jailbreak devices	4.43
Pin your SSL certificate	4.44
Cover your camera	4.46
Ask for advice and information about online security and privacy	4.46
Let your children teach you about the Internet too	4.49

Be wary of third-party apps on social networks	4.5
Change your MAC address	4.51
Not run tasks as administrator if unnecessary	4.51
Talk to children about their online habits	4.51
Keep your devices with you when traveling	4.52
Remove sensitive files from your machine	4.52
Pay attention to and follow software warnings	4.52
Use private search engines	4.56
Ask people to remove your personal information and photos	4.57
Not need to use antivirus on Mac	4.57
Disable active content (JavaScript, Flash, etc.)	4.58
Create copies of your websites (mirror sites)	4.59
Limit the amount of personal info being collected about you online	4.61
Create pronounceable passwords	4.61
Use a content delivery network or caching service	4.61
Not post your email address on forums	4.62
Transfer sensitive files to network shares	4.63
Not change browser security settings	4.66
Use an air gap	4.66
Do online banking on a separate computer	4.69
Feel comfortable making weak passwords for sites thar don’t keep personal info	4.69
Manually type links you receive into the URL bar	4.7
Not use encryption when sending e-mail to a listserv	4.71
Not meet up with people you’ve met online	4.71
Use Tor	4.74
Not try to be anonymous if you don?t need to be	4.74
Turn off WiFi	4.75
Not open documents downloaded through Tor while online	4.75
Understand new features before you try them	4.76
Suspend unused accounts	4.76
Upgrade your email provider	4.76
Enable ”Do Not Track” or similar options	4.8

Download a filtering software to prevent website access	4.82
Increase firewall security measures to inspect incoming data	4.83
Avoid common passwords	4.85
Bring proof-of-purchase for computer equipment when travelling	4.86
Use parental controls	4.86
Not download or execute any files	4.8600000000000003
Not include sensitive information in email	4.88
Use a cable lock for your laptop	4.88
Use different browsers for different activities	4.89
Contact police or authority figures in case of a cyberattack or cyberbullying	4.89
Not use Facebook	4.94
Only use open-source software	4.96
Use less common software	4.97
Clear your cache	4.98
Isolate IoT devices on their own network	5.04
Understand where your child accesses Internet out of the house	5.06
Change your username regularly	5.11
Make your email subject lines vague since they are not encrypted	5.11
Use airplane mode in stores with retail tracking	5.11
Use tracking applications	5.11
Install software in phases	5.13
Use different personas online	5.14
Not shut down your computer	5.15
Protect your computer from power surges	5.16
Not change your passwords unless they become compromised	5.18
Not use credit or debit cards online	5.2
Store passwords in a file	5.23
Carry laptops in something other than laptop cases	5.27
Encourage the positive sides of the Internet with children and friends	5.35
Shut down your computer	5.35
Not use banking apps or websites	5.4
Create keyboard patterns to help with remembering passwords	5.46

Lock your SIM card in your smartphone	5.52
Create multiple accounts	5.58
Disable and/or limit caching	5.63
Write down passwords on paper	5.75
File taxes early	5.75
Install firmware on mobile devices	5.81
Use an unbranded smartphone	6.36
Create a new email address if your last one is compromised	7.01

B.8 Computed Priority Ranking of Advice by Users

Buy devices with security-focused platforms	0.748
Not tell anyone your passwords, even IT	0.748
Not open unnecessary attachments	0.84
Be wary of using public computers that could be infected	0.84
Use antivirus	0.871
Not click random or unfamiliar links from unknown senders	0.900
Verify suspicious email and email contents	0.936
Not open email from unknown senders	0.98
Not use passwords based on personal information	1.01
Not store passwords online	1.07
Not friend people you don't know	1.07
Be suspicious if something is too good to be true	1.079
Not try to be anonymous if you don?t need to be	1.08
Use both upper and lower case in passwords	1.123
Be wary of third-party apps on social networks	1.13
Leave unsafe websites	1.15
Make sure no one is watching you enter passwords	1.147
Set your antivirus/antimalware to run periodic full scans	1.16
Install a firewall	1.17
Use passwords	1.204
Secure your router	1.218
Unsubscribe from unwanted email lists	1.24
Lock your computer when you're away from it	1.26
Write down password clues	1.28
Understand where your child accesses Internet out of the house	1.28
Not forward cyberbullying	1.29
Remove improper and/or sensitive information from the web	1.30
Not reply to spam	1.304
Use passwords that are dissimilar to previous ones	1.33
Watch for spelling mistakes in provided URLs	1.331

Encrypt your WiFi	1.33
Log out of accounts	1.337
Not open attachments from unknown senders	1.34
Not do online banking on a public computer	1.355
Not meet up with people you've met online	1.36
Lock your smartphone with passcode or touch ID	1.36
Change passwords often	1.371
Update applications	1.37
Securely wipe devices before disposal, where possible	1.40
Always keep learning about security and privacy	1.41
Start your PC in Safe Mode when you need to troubleshoot viruses	1.407
Not jailbreak devices	1.444
Be suspicious of unusual email	1.45
Be suspicious of unusual email if grammar in an email is not good	1.46
Only allow authorized users to access your network	1.458
Not run or keep unnecessary programs	1.47
Be careful who uses your computer	1.49
Only reveal financial information to reputable actors	1.49
Install firmware on mobile devices	1.49
Visit only known websites	1.502
Encrypt your device data	1.508
Beware of malware	1.52
Understand the Internet	1.524
Not enter sensitive information or credentials without HTTPS	1.53
Not use automatic network log-in	1.536
Not use repetitive characters in passwords	1.544
Use a combination of letters, numbers, and special characters in passwords	1.544
Confirm public WiFi information with staff before using	1.54
Install only trusted software	1.56
Disable features you aren't using (BlueTooth, WiFi, etc.)	1.56
Not use a password manager	1.57
Do sensitive tasks on dedicated and trusted devices	1.573

Scan files downloaded from websites for viruses	1.58
Be suspicious of links	1.579
Watch out for phishing	1.60
Download only trusted programs	1.60
Verify who you are talking to	1.615
Beware of ”free” products	1.62
Use secure payment methods like PayPal, BPay, or credit cards	1.641
Protect your computer from power surges	1.681
Turn off automatic downloads	1.681
Keep virus definitions up to date	1.69
Keep your private key safe	1.70
Store passwords properly	1.70
Use anti-malware software	1.71
Delete spam	1.72
Use different personas online	1.73
Beware of incognito mode	1.742
Be suspicious of attachments	1.744
Manage and track cookies	1.746
Avoid illegal or unaffiliated download sites	1.746
Turn off remote access/management features	1.75
Seek professional help for cybersecurity issues	1.756
Encrypt select groups of files	1.77
Turn off/limit pop-ups	1.78
Turn on automatic updates for applications	1.786
Be suspicious of popups and requests, even from known sources	1.804
Clear your cookies	1.80
Be aware of the Internet	1.81
Be wary of false emails from trusted institutions	1.833
Check that websites have valid digital certificates	1.83
Not post vulnerable information (addresses, credit card numbers, etc.)	1.843
Delete phishing or spam emails, even if you might know the sender	1.85
Use a secure machine to hold confidential data	1.87

Not store mobile passwords directly on the device	1.877
Use long passwords	1.89
Not open documents downloaded through Tor while online	1.893
Beware of revealing personal information unless you know who you're talking to	1.91
Cancel or change accounts if you are being cyberbullied	1.94
Keep antivirus/antimalware up to date	1.94
Not run tasks as administrator if unnecessary	1.944
Use unique passwords	1.97
Use parental controls	1.978
Encrypt your email	1.99
Be aware of what you share	1.99
Not sign up for unnecessary accounts	1.992
Encrypt cloud data	2.00
Whitelist executable directories to prevent malicious binaries	2.00
Not send or forward files you haven't scanned for viruses	2.004
Read terms of service	2.00
Not download or execute any files	2.008
Ensure Wifi is secured to at least WPA2 level	2.01
Use a password to protect your WiFi	2.01
Not use your real name online	2.02
Keep your devices with you when traveling	2.02
Clear your cache	2.03
Only use open-source software	2.04
Backup your password database	2.05
Use a disposable email service	2.06
Document cyberbullying incidents	2.06
Update devices	2.06
Use browsers that protect against phishing	2.07
Look at the URL bar to verify you're at the intended website	2.07
Not post your email address on forums	2.07
Be suspicious	2.08
Use a password that's different from your username	2.09

Report suspicious things to IT or support	2.093
Pin your SSL certificate	2.1
Be cautious when picking an email address	2.10
Change and rethink security questions	2.137
Check the extensions of downloaded files	2.14
Set antivirus to autoscan email	2.165
Disable extension-hiding for known file types	2.17
Perform a factory reset before device disposal	2.17
Understand new features before you try them	2.18
Monitor online accounts for unusual activity	2.181
Minimize network exposure for control systems	2.181
Not share third party personal information i.e. friends and family	2.19
Encourage the positive sides of the Internet with children and friends	2.21
Set rules for your kids about the Internet	2.21
Secure other devices like you would your computer	2.23
Disable and/or limit caching	2.23
Configure antivirus to scan all files in real time	2.23
Not follow links in spam	2.24
Not enter passwords after clicking links in email	2.24
Educate yourself on how to avoid fraud	2.24
Turn on download notifications	2.24
Avoid using open Wi-Fi networks for business, banking, shopping etc.	2.25
Double check email addresses	2.259
Look for the lock icon in the address bar	2.25
Research the security of IoT devices before purchase	2.262
Use anti-spyware	2.27
Test your firewall	2.28
Not change browser security settings	2.28
Disconnect your computer from the Internet when you're away	2.28
Shut down your computer	2.29
Ask for advice and information about online security and privacy	2.29
Limit the amount of personal info being collected about you online	2.31

Report account breaches or losses to the appropriate people	2.31
Set browser to click-to-play for videos and ads	2.31
Take only devices you need when traveling	2.32
Restrict physical access to computers and removable media	2.32
Regularly back up your data	2.32
Make sure to overwrite files you want to delete	2.331
Seek expert help	2.33
Monitor credit cards for unauthorized activity	2.34
Understand what permissions you give new software downloads	2.34
Use an unbranded smartphone	2.35
Monitor network traffic on your router	2.35
Increase firewall security measures to inspect incoming data	2.35
Apply real-world common sense and follow your instincts	2.355
Be aware of your online reputation	2.36
Not be lulled into a false sense of security from antivirus/firewall	2.37
Do online banking only on your own computer	2.39
Develop a mnemonic for complex passwords	2.387
Delete originals once a document has been encrypted	2.387
Protect devices against power surges	2.403
Use strong passwords	2.40
Be careful with permissions	2.41
Not store passwords in a file	2.41
Secure devices and fix vulnerabilities that caused the breach	2.431
Change default passwords on devices/networks/services	2.44
Consider opening a secondary account for shopping etc.	2.44
Physically destroy drives you're done with and wish to erase	2.46
Be careful using email	2.47
Report messages as spam	2.47
Be suspicious of unusual email of things even from known people	2.47
Only use HTTPS	2.476
Not use look-alike substitutions for your password	2.484
File taxes early	2.484

Only do business with trusted institutions	2.48
Discuss identifiable information in private (so you are not overheard)	2.49
Not click on flashy things	2.49
Avoid common passwords	2.508
Create a new email address if your last one is compromised	2.512
Don’t trust sites with certificate warnings	2.52
Limit the number of antivirus applications you install	2.52
Disable automatic download of email attachments	2.52
Check if website you’re visiting uses HTTPS	2.53
Exit sites your browser warns are malicious	2.536
Only download things you are looking for	2.536
Avoid passwords with only numbers	2.55
Turn on automatic updates for devices	2.55
Be careful entering passwords in public computers	2.552
Remember your passwords	2.58
Keep your receipts	2.577
Keep sensitive information on removable storage media	2.585
Remove unsafe devices from the network	2.589
Not use extensions or plugins	2.59
Watch for unusual posts on your account	2.60
Use a proxy server	2.605
Contact police or authority figures in case of a cyberattack or cyberbullying	2.61
Disable your webcam	2.61
Suspend unused accounts	2.62
Obfuscate something meaningful to generate a password	2.63
Create copies of your websites (mirror sites)	2.633
Apply the highest level of security that’s practical	2.637
Use electronic bank statements	2.645
Make your email subject lines vague since they are not encrypted	2.645
Download a filtering software to prevent website access	2.665
Overwrite deleted files	2.68
Remain anonymous online	2.68

Not send executable programs with macros	2.68
Disable sharing on peer-to-peer apps	2.68
Install latest OS updates	2.70
Not blindly trust HTTPS	2.70
Be careful of downloads	2.70
Not give out your email address without good reason	2.70
Check your credit report regularly	2.70
Replace letters with symbols in your passwords	2.71
Use single sign-on SSO	2.71
Not store data if you don't need to	2.71
Not respond to or retaliate against cyberbullies	2.722
Not trust the From address on an email	2.722
Read install prompts	2.72
Fully reset hacked devices	2.74
Pay attention to and follow software warnings	2.73
Disable third-party cookies	2.73
Upgrade your email provider	2.74
Not give out your email address for free software downloads	2.746
Cover your camera	2.754
Turn off location services	2.758
Scan attachments you open for viruses	2.758
Be wary of proxy servers	2.758
Beware of free VPN programs	2.76
Treat security questions like passwords	2.77
Use 2+ factor authentication	2.78
Understand data usage and storage	2.78
Use unusual phonetics in passwords	2.80
Clear your browser history	2.80
Use randomly generated passwords or password generator websites	2.80
Draw shapes on your keyboard to generate passwords	2.81
Change passwords and security questions on compromised accounts	2.84
Use privacy settings	2.86

Not use banking apps or websites	2.86
Only add people you know in the offline world to contact lists	2.87
Read privacy policies	2.87
Only copy files onto machines using physical media	2.87
Unmount encrypted disks	2.88
Buy devices with passwords, preferably passwords that you can change	2.88
Keep your own data locally (not in the cloud or on a remote server)	2.88
Use passphrases	2.88
Use administrator rights to prevent unauthorized actions	2.91
Install software in phases	2.92
Transfer sensitive files to network shares	2.92
Use end-to-end encryption for communication	2.927
Use different computers for work and home use	2.931
Enable ”Do Not Track” or similar options	2.94
Use different passwords for different accounts/devices	2.94
Keep track of file extensions	2.94
Run a virus scan on new devices	2.944
Discard devices with security weaknesses that can’t be fixed	2.95
Look for real-world contact information while online	2.96
Avoid plugging external devices into computers	2.96
Opt out of ad tracking	2.972
Turn down transmission strength	2.99
Make a threat model	3.00
Use encryption	3.00
Use a VPN	3.00
Remain calm and talk with kids about bad web experiences	3.01
Manually type links you receive into the URL bar	3.01
Encrypt your other devices	3.01
Disconnect from the Internet	3.02
Use unique passwords for different accounts	3.02
Talk to children about their online habits	3.04
Remove sensitive files from your machine	3.05

Not shut down your computer	3.05
Block unwanted users	3.07
Use filters in email	3.073
Regularly search for your name	3.08
Not enable file sharing on networks exposed to the Internet	3.09
Use airplane mode in stores with retail tracking	3.09
Disable 2G support	3.09
Verify URLs you visit	3.10
Disable Autorun to prevent malicious code from running	3.10
Isolate IoT devices on their own network	3.10
Pay attention to virus warnings	3.10
Use a password manager	3.11
Request your data from sites or corporations that store it	3.11
Encourage others to use strong passwords	3.12
Not include sensitive information in email	3.12
Use private search engines	3.121
Lock your SIM card in your smartphone	3.15
Use security extensions	3.15
Ask people to remove your personal information and photos	3.15
Not use debit cards	3.16
Consider partitioning your computer into seperate accounts	3.17
Not identify yourself to websites	3.18
Keep passwords safe if written down	3.19
Bring proof-of-purchase for computer equipment when travelling	3.20
Use tracking applications	3.206
Review your root certificates	3.21
Use less common software	3.21
Not click on ads	3.22
Use an alarm on your devices	3.22
Turn off Bluetooth	3.22
Not write down passwords	3.23
Not need to use antivirus on Mac	3.23

Set up auto-lock timers for your smartphone	3.23
Use different browsers for different activities	3.23
Create pronounceable passwords	3.26
Do online banking on a separate computer	3.26
Not forward email unnecessarily	3.27
Encourage children to follow age limit guidelines for websites	3.28
Write down passwords on paper	3.28
Consider opening a credit card for online use only	3.30
Not use Facebook	3.30
Not give out your email address for free samples or products	3.3
Change your username regularly	3.30
Check camera logs	3.306
Encourage children to talk to you if they feel uncomfortable online	3.31
Use a paid spam filtering service	3.31
Enable remote data wiping for devices	3.31
Disable message and image previews	3.32
Take note of the countries your VPN providers works in	3.32
Use a load balancer	3.371
Store passwords in a file	3.375
Feel comfortable making weak passwords for sites that don't keep personal info	3.37
Monitor where your kids go online	3.39
Verify file signatures	3.4
Disable active content (JavaScript, Flash, etc.)	3.43
Encrypt your hard drive	3.432
Not use built-in erasing on SSDs	3.44
Turn off WiFi	3.45
Disable macros	3.46
Understand who to trust online	3.468
Not use dictionary words as or in passwords	3.468
Use incognito mode	3.48
Encourage others to use Tor	3.488
Keep the computer in a common room in your house if you have children	3.49

Create keyboard patterns to help with remembering passwords	3.492
Use a content delivery network or caching service	3.50
Change your router name from the default	3.52
Not use encryption when sending e-mail to a listserv	3.54
Not overwrite SSDs	3.56
Not let computers or browsers remember passwords	3.56
Not change your passwords unless they become compromised	3.589
Create separate networks for devices	3.61
Not use credit or debit cards online	3.62
Disable ”Universal Plug and Play (UPnP)” on your router	3.622
Use a virtual machine or even multiple virtual machines	3.633
Base passwords on upcoming events	3.66
Use ad blocker extensions	3.67
Not rely on mobile access as a primary means for email	3.68
Use an air gap	3.738
Carry laptops in something other than laptop cases	3.798
Create a network demilitarization zone (DMZ)	3.82
Create multiple accounts	3.83
Let your children teach you about the Internet too	3.851
Use Tor	3.87
Use a cable lock for your laptop	4.01
Try alternate URLs to avoid censorship	4.01
Not use loyalty cards	4.05
Change your MAC address	4.173

B.9 Pairwise Comparisons of Priority Rankings by Topic

	Account Security	Antivirus	Browsers	Data Storage	Device Security	Finance	General Security	Incident Response
Antivirus	0.83							
Browsers	0.28	0.62						
Data Storage	0.55	0.54	0.15					
Device Security	0.71	0.65	0.22	0.87				
Finance	0.63	0.85	0.80	0.33	0.53			
General Security	0.37	0.41	0.05	0.97	0.73	0.24		
Incident Response	0.28	0.30	0.08	0.55	0.49	0.19	0.53	
Network Security	0.81	0.99	0.66	0.51	0.68	0.94	0.35	0.30
Passwords	**0.01***	0.11	**0.04***	**0.01***	**0.01***	0.10	<**0.001***	**0.02***
Privacy	0.78	0.69	0.27	0.77	0.98	0.51	0.67	0.40
Software	0.31	0.54	0.78	0.17	0.26	0.70	0.09	0.09

Table B.12: Results of Mann-Whitney pairwise comparisons of median priority rating of advice about each topic. Holm Bonferonni multiple testing correction applied.

B.10 Pairwise Comparisons of Actionability Submetrics By Topic

	Account Security	Antivirus	Browsers	Data Storage	Device Security	Finance	General Security	Incident Response	Network Security	Passwords	Privacy
Antivirus	1.00										
Browsers	**0.03***	0.11									
Data Storage	**<0.001***	**<0.001***	**<0.001***								
Device Security	0.07	0.12	1.00	**<0.001***							
Finance	**0.03**	**0.04***	0.97	**0.02***	1.00						
General Security	**<0.001***	**<0.001***	**<0.001***	**0.03***	**0.02***	1.00					
Incident Response	0.07	0.10	1.00	**<0.001***	1.00	1.00	0.06				
Network Security	**0.01***	**<0.001***	**<0.001***	0.59	**<0.001***	**<0.001***	**<0.001***	**<0.001***			
Passwords	1.00	1.00	**0.01***	**<0.001***	0.05	**0.02***	**<0.001***	0.06	**<0.001***		
Privacy	**<0.001***	**<0.001***	0.08	**0.03***	1.00	1.00	0.74	1.00	**<0.001***	**<0.001***	
Software	**0.04***	**0.04***	**<0.001***	**<0.001***	**<0.001***	**<0.001***	**<0.001***	**<0.001***	**0.03***	**0.04***	**<0.001***

Table B.13: Results of pairwise Mann-Whitney comparisons of confidence ratings of advice about each topic. Holm Bonferonni multiple testing correction applied.

	Account Security	Antivirus	Browsers	Data Storage	Device Security	Finance	General Security	Incident Response	Network Security	Passwords	Privacy
Antivirus	**<0.001***										
Browsers	**<0.001***	**0.01***									
Data Storage	**<0.001***	**<0.001***	**<0.001***								
Device Security	**0.01***	0.08	0.85	**<0.001***							
Finance	0.81	**<0.001***	**0.04***	**<0.001***	**0.04***						
General Security	**<0.001***	**<0.001***	**<0.001***	**<0.001***	**<0.001***	**<0.001***					
Incident Response	**<0.001***	0.51	**0.02***	**<0.001***	**0.04***	**<0.001***	0.29				
Network Security	**0.03***	**<0.001***	**<0.001***	**<0.001***	**0.01***	**0.04***	**<0.001***	**<0.001***			
Passwords	**<0.001***	**0.02***	0.50	**<0.001***	0.49	**0.04***	**<0.001***	**0.03***	**<0.001***		
Privacy	**<0.001***	0.72	0.15	**<0.001***	0.22	**0.02***	**0.01***	0.38	**<0.001***	0.22	
Software	**<0.001***	**<0.001***	**<0.001***	**<0.001***	**<0.001***	**0.02***	**<0.001***	**<0.001***	0.16	**<0.001***	**<0.001***

Table B.14: Results of pairwise Mann-Whitney comparisons of time consumption ratings of advice about each topic. Holm Bonferonni multiple testing correction applied.

	Account Security	Antivirus	Browsers	Data Storage	Device Security	Finance	General Security	Incident Response	Network Security	Passwords	Privacy
Antivirus	0.10										
Browsers	0.23	0.41									
Data Storage	**<0.001***	**<0.001***	**<0.001***								
Device Security	**<0.001***	**<0.001***	**<0.001***	**<0.001***							
Finance	**0.03***	**0.02***	**<0.001***	**<0.001***	**0.01***						
General Security	**<0.001***	**<0.001***	**<0.001***	**<0.001***	0.39	**<0.001***					
Incident Response	**<0.001***	**<0.001***	**<0.001***	**<0.001***	**0.03***	**<0.001***	0.14				
Network Security	0.40	0.08	0.09	**<0.001***	**<0.001***	**0.03***	**<0.001***	**<0.001***			
Passwords	**<0.001***	0.39	**0.01***	**<0.001***	**<0.001***	**0.03***	**<0.001***	**<0.001***	**<0.001***		
Privacy	0.61	0.66	0.83	**<0.001***	**<0.001***	**0.04***	**<0.001***	**<0.001***	0.31	0.19	
Software	**<0.001***	**<0.001***	**<0.001***	**<0.001***	**<0.001***	**<0.001***	**<0.001***	**<0.001***	0.05	**<0.001***	**<0.001***

Table B.15: Results of pairwise Mann-Whitney comparisons of ratings of advice disruptiveness by topic. Holm Bonferonni multiple testing correction applied.

	Account Security	Antivirus	Browsers	Data Storage	Device Security	Finance	General Security	Incident Response	Network Security	Passwords	Privacy
Antivirus	<0.001*										
Browsers	0.88	0.14									
Data Storage	<0.001*	<0.001*	<0.001*								
Device Security	<0.001*	1.00	<0.001*	<0.001*							
Finance	0.05	1.00	1.00	<0.001*	1.00						
General Security	<0.001*	0.26	<0.001*	0.10	0.03*	<0.001*					
Incident Response	<0.001*	1.00	1.00	<0.001*	1.00	1.00	<0.001*				
Network Security	<0.001*	<0.001*	0.02*	0.28	<0.001*	0.04*	<0.001*	0.01*			
Passwords	0.06	<0.001*	0.40	<0.001*	<0.001*	0.67	<0.001*	<0.001*	<0.001*		
Privacy	<0.001*	<0.001*	0.40	<0.001*	0.01*	0.09	<0.001*	0.04*	0.02*	0.04*	
Software	0.30	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*	<0.001*

Table B.16: Results of pairwise Mann-Whitney comparisons of ratings of advice difficulty by topic. Holm Bonferonni multiple testing correction applied.

B.11 Unfollowed Advice

B.11.1 Advice Not Followed by Experts

Below we list the advice that no experts or only one expert reported following at least some of the time.

First, we list the six pieces of advice that no experts reported following at least some of the time.

All advice is of the form “You should...”:

- base passwords on upcoming events
- create a new email address if your last one is compromised
- create keyboard patterns to help with remembering passwords
- file taxes early
- not use banking apps or websites
- write down passwords on paper

Next, we list the 46 pieces of advice that only one expert reported following at least some of the time.

- ask people to remove your personal information and photos
- be aware of the Internet
- beware of malware

- bring proof-of-purchase for computer equipment when travelling
- carry laptops in something other than laptop cases
- clear your cache
- confirm public wifi information with staff before using
- consider opening a credit card for online use only
- contact police or authority figures in case of a cyberattack or cyberbullying
- disable and/or limit caching
- disconnect from the Internet
- disconnect your computer from the Internet when you're away
- do online banking on a separate computer
- encourage the positive sides of the Internet with children and friends
- feel comfortable making weak passwords for sites thar don't keep personal info
- install firmware on mobile devices
- isolate iot devices on their own network
- keep the computer in a common room in your house if you have children
- lock your sim card in your smartphone
- not change browser security settings

- not change your passwords unless they become compromised
- not forward cyberbullying
- not meet up with people you've met online
- not need to use antivirus on mac
- not open email from unknown senders
- not shut down your computer
- not use encryption when sending e-mail to a listserv
- not use facebook
- only use open-source software
- protect your computer from power surges
- remain calm and talk with kids about bad web experiences
- scan files downloaded from websites for viruses
- shut down your computer
- store passwords properly
- suspend unused accounts
- take note of the countries your vpn providers works in
- transfer sensitive files to network shares

- turn off wifi
- understand new features before you try them
- understand the Internet
- upgrade your email provider
- use a load balancer
- use a virtual machine or even multiple virtual machines
- use different personas online
- use tor
- use tracking applications

B.11.2 Advice Not Followed by Respondents

Below we list the advice that no respondents or only one respondent reported following at least some of the time.

First, we list the 31 pieces of advice that no respondents reported following at least some of the time.

All advice is of the form "You should...":

- avoid plugging external devices into computers
- base passwords on upcoming events
- bring proof-of-purchase for computer equipment when travelling

- carry laptops in something other than laptop cases
- change your mac address
- cover your camera
- create a network demilitarization zone (dmz)
- create keyboard patterns to help with remembering passwords
- create multiple accounts
- create separate networks for devices
- disable autorun to prevent malicious code from running
- disconnect from the Internet
- encourage others to use tor
- encrypt your hard drive
- isolate iot devices on their own network
- keep the computer in a common room in your house if you have children
- lock your sim card in your smartphone
- make sure to overwrite files you want to delete
- monitor where your kids go online
- not rely on mobile access as a primary means for email

- not use encryption when sending e-mail to a listserv
- review your root certificates
- set up auto-lock timers for your smartphone
- unmount encrypted disks
- use a cable lock for your laptop
- use a content delivery network or caching service
- use a disposable email service
- use an air gap
- use private search engines
- use single sign-on sso
- verify file signatures

Next, we list the 85 pieces of advice that only one respondent reported following at least some of the time.

- apply the highest level of security that's practical
- ask people to remove your personal information and photos
- be wary of proxy servers
- beware of free vpn programs

- buy devices with security-focused platforms
- change your router name from the default
- change your respondentname regularly
- check camera logs
- check your credit report regularly
- consider opening a credit card for online use only
- create copies of your websites (mirror sites)
- create pronounceable passwords
- disable "universal plug and play (upnp)" on your router
- disable automatic download of email attachments
- disable message and image previews
- disable your webcam
- do online banking on a separate computer
- document cyberbullying incidents
- draw shapes on your keyboard to generate passwords
- enable remote data wiping for devices
- encourage children to follow age limit guidelines for websites

- feel comfortable making weak passwords for sites thar don't keep personal info
- file taxes early
- fully reset hacked devices
- install latest os updates
- keep your own data locally (not in the cloud or on a remote server)
- leave unsafe websites
- let your children teach you about the Internet too
- limit the number of antivirus applications you install
- look for the lock icon in the address bar
- minimize network exposure for control systems
- not change your passwords unless they become compromised
- not do online banking on a public computer
- not give out your email address for free samples or products
- not open documents downloaded through tor while online
- not open unnecessary attachments
- not overwrite ssds
- not send executable programs with macros

- not shut down your computer
- not use banking apps or websites
- not use built-in erasing on ssds
- not use credit or debit cards online
- not use extensions or plugins
- not use facebook
- not use loyalty cards
- not write down passwords
- obfuscate something meaningful to generate a password
- only add people you know in the offline world to contact lists
- pin your ssl certificate
- read privacy policies
- regularly search for your name
- remain calm and talk with kids about bad web experiences
- remember your passwords
- remove sensitive files from your machine
- research the security of iot devices before purchase

- run a virus scan on new devices
- scan attachments you open for viruses
- seek expert help
- store passwords in a file
- suspend unused accounts
- take note of the countries your vpn providers works in
- talk to children about their online habits
- test your firewall
- try alternate urls to avoid censorship
- turn off bluetooth
- turn off wifi
- use a load balancer
- use a password manager
- use a proxy server
- use a virtual machine or even multiple virtual machines
- use ad blocker extensions
- use administrator rights to prevent unauthorized actions

- use airplane mode in stores with retail tracking
- use an alarm on your devices
- use an unbranded smartphone
- use different browsers for different activities
- use different passwords for different accounts/devices
- use encryption
- use end-to-end encryption for communication
- use incognito mode
- use less common software
- use tor
- whitelist executable directories to prevent malicious binaries
- write down passwords on paper

B.12 Short Longevity Advice

Here we list the 38 pieces of advice that had a median rating of "2-5 years" for "how long do you think this advice will remain useful for improving people's security":

- be cautious when picking an email address

- block unwanted users
- change passwords often
- change your mac address
- clear your cookies
- create multiple accounts
- delete spam
- disable "universal plug and play (UPNP)" on your router
- disable active content (javascript, flash, etc.)
- disable message and image previews
- discard devices with security weaknesses that can't be fixed
- disconnect from the Internet
- disconnect your computer from the Internet when you're away
- do sensitive tasks on dedicated and trusted devices
- double check email addresses
- encourage children to follow age limit guidelines for websites
- keep your own data locally (not in the cloud or on a remote server)
- keep your receipts

- make sure to overwrite files you want to delete
- manage and track cookies
- not forward email unnecessarily
- not give out your email address for free software downloads
- not send executable programs with macros
- not store mobile passwords directly on the device
- not store passwords online
- not use loyalty cards
- not use your real name online
- obfuscate something meaningful to generate a password
- only use https
- securely wipe devices before disposal, where possible
- take note of the countries your VPN providers works in
- test your firewall
- use a combination of letters, numbers, and special characters in passwords
- use incognito mode
- use pass-phrases

- use tor
- use unusual phonetics in passwords
- whitelist executable directories to prevent malicious binaries

B.13 Analysis of Advice Priority Rankings by Users

We also asked the respondents who took our actionability evaluation questionnaire to rate the priority of doing the piece of advice they evaluated. They rated 73 of the imperatives (19.5%) as the number one behavior they should do, 143 (38.2%) as being in the top 3 behaviors they should do, 95 (24.5%) as being in the top 5 behaviors, 44 (11.8%) as being in the top ten behaviors, and the rest (18 imperatives, 4.80%) as being advice they should follow but it would not be in the top 10.

B.14 Update Messages used in Chapter 7

Figure B.1 shows all 11 update messages. Crash data is available for 5 versions: Adobe Reader 9.5.1.283 (Figure B.1b), Flash Player 10.3.181.14 (Figure B.1g), Flash Player 11.0.1.152 (Figure B.1h), Firefox 8.0.1.4341 (Figure B.1i) and Opera 11.64.1403.0 (Figure B.1k).

B.15 Chapter 7 Survey Demographics

We have demographics only for the SSI participants as our surveys were conducted in a privacy preserving manner and participant demographics were not collected in the survey directly. SSI produces aggregated reports on sample demographics; AMT does not. Table B.17 presents a comparison of the SSI sample demographics with the U.S. Census [11].

Metric	**SSI**	**Census**	**Metric**	**SSI**	**Census**
Male	49.7%	48.2%	H.S. or below	40.7%	41.3%
Female	50.3%	51.8%	Some college	22.2%	31.0%
			B.S. or above	37.1%	27.7%
Caucasian	67.5%	65.8%	18-29 years	29.6%	20.9%
Hispanic	9.1%	15%	30-49 years	39%	34.7%
African American	12.2%	11.5%	50-64 years	27.8%	26.0%
Other	11.2%	7.6%	65+ years	3.1%	18.4%
<$20k	19.6%	32%			
$20k-$40k	23.6%	19%			
$40k-$75k	28.6%	18%			
$75k-$100k	11.8%	11%			
$100k-$150k	11.8%	12%			
$150k+	4.5%	8%			

Table B.17: Demographics of the 455 respondents in the SSI sample compared to U.S. Census demographics [11].

B.16 Chapter 7 Additional Analysis

B.16.1 RQ4: Regression Models

Table B.18 presents the results for the hierarchical regression modeling conducted using a dataset containing observations or responses to all eleven messages; these models do not include any risk metrics, as this risk data was only available for the five update messages released in 2011 or 2012. Table B.20 presents the results for the risk-related modeling, conducted using only data pertaining to the five update messages with risk metrics available.

B.16.2 Survey Filtering

We mapped the answer choices for the second and third survey questions, which queried why respondents would and would not want to update in response to the given message. This mapping is indicated in Appendix A.9 above. In line with the approach of Fahl et al., who filtered out survey respondents who self-reported not answering their survey honestly, we filter out respondent's who's answers to Q2 and Q3 are clearly illogical: that is we remove (1) any respondents who noted that they would not install the update shown because it required a restart, but the update message they saw explicitly stated that it did not require a restart, (2) any respondents who indicated that they would install the update shown because it did not require a restart, but in fact saw an update message that stated that it did require a restart, (3) any respondent who noted that they would not install the update because it contained features they did not want, but who saw an update message that mentioned only security and no other enhancements, and (4) any respondent who noted that they would install an update because it contained features they would want, but who saw a message that mentioned security and no other enhancements. This filtering results in a dataset consisting of 981 respondents (44% of the original 2,092): 749 (43% of the original) from MTurk and 232 (51% of the original) from SSI.

B.16.3 Survey Internal Consistency

In addition to filtering the dataset, we also checked for internal consistency more broadly by testing for independence (X^2, corrected with Holm-Bonferonni procedure) between responses to Q2 and Q3 and: the actual message features (for security-only, cost, and application) or later responses to Q4-8 (risk and general tendency). Internally consistent responses should not be independent (i.e., should produce a significant X^2 independence test result). Table B.19 shows our results.

Factor	**Full**						**Reader**						**Flash**											
	Baseline			Application			Baseline			Cost (Restart)			Baseline			Length (words)			Security-Only			Security, Length		
	S	SF	WINE	S	SF	WINE	S	SF	WINE	S	SF	WINE	S	SF	WINE	S	SF	WINE	S	SF	WINE	S	SF	WINE
Gen. Tendency	6.09*	4.36*	1.60*	4.39*	4.39*	1.58*	4.82*	4.19*	1.84*	5.61*	4.10*	1.89*	6.47*	4.10*	1.45*	6.48*	4.10*	1.44*	6.48*	4.11*	1.35*	6.47*	4.12*	1.28*
Security-Only																			1.05	0.84	3.33*	0.94	0.65	4.54*
Length											1.02	1.03	0.98*			1.02	1.06	0.93*						
Cost							0.99	1.47	0.86*															
App.: Firefox				0.70*	0.69*	0.66*																		
App.: Opera				0.98	0.29*	1.17																		
App.: Reader				0.71*	0.68*	0.63*																		
Sample: MTurk	1.29	1.17	–	1.30*	1.18	–	0.90	1.39	–	1.23	1.41	–	1.40*	1.16	–	1.41*	1.19	–	1.40*	1.16	–	1.41*	1.17	–
n	2,092	981	517,932	2,092	981	517,932	386	219	306,654	386	219	306,654	1,149	301	343,697	1,149	301	343,697	1,149	301	343,697	1,149	301	343,697

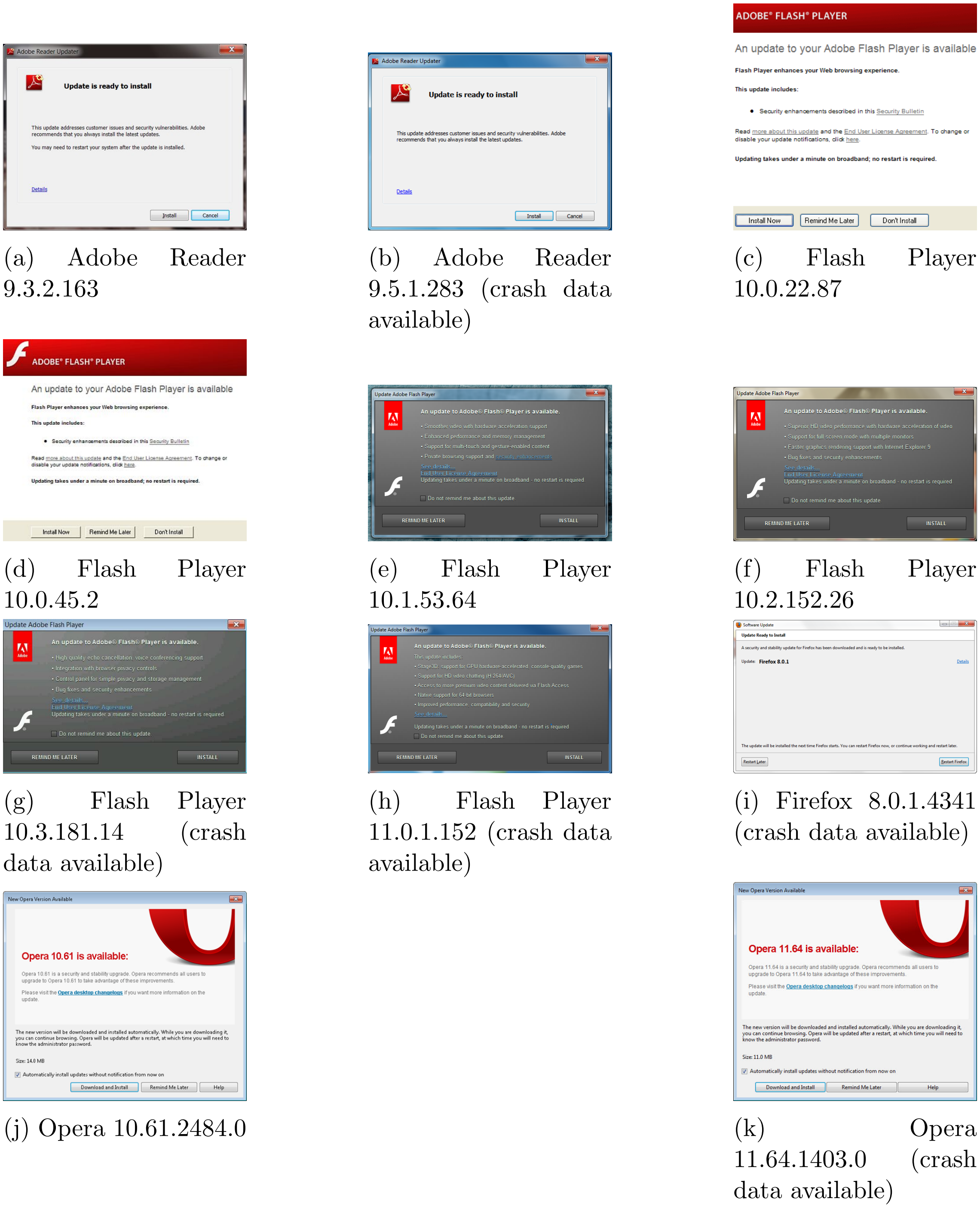
Table B.18: Table of hierarchical regression models on the full datasets, the Reader messages, and the Flash messages. S indicates for models built on the survey data, SF indicates for models built on the filtered survey data, and WINE indicates for models built on the measurement data. p-values significant at $alpha = 0.05$ are marked with *.

		Comparison	X^2	p-value
Detailed Constructs	Cost	Why: Low cost — Message: Restart	6.39	0.024*
		Why Not: High cost — Message: Restart	0.917	0.384
	Security / Features	Why: Features — Security-Only: Restart	4.72	0.067
		Why Not: Features — Security-Only	0.050	0.823
	Application	Why: Application — Application	38.2	<0.001*
		Why Not: Application — Application	11.8	0.019*
General Constructs	General Tendency	Why: Always Update — General Tendency	141.2	<0.001*
		Why Not: Rarely Update — General Tendency	7.77	0.042*
	Risk	Why: Risk — Sys. Crash. Freq.	15.6	0.005*
		Why: Risk — Sys. Crash. More	4.95	0.040*
		Why: Risk — App. Crash. Freq.	15.5	0.005*
		Why: Risk — App. Crash. More	5.56	0.401
		Why Not: Risk — Sys. Crash. Freq.	3.09	0.031*
		Why Not: Risk — Sys. Crash. More	17.3	<0.001*
		Why Not: Risk — App. Crash. Freq.	10.5	0.028*
		Why Not: Risk — App. Crash. More	7.59	0.0166*

Table B.19: X^2 tests comparing respondent's reported reasons for updating with the true message features or their later survey responses.

Factor	**Full**					
		Baseline			Risk	
	Measurement	Survey	Survey: F	Measurement	Survey	Survey: F
Gen. Tendency	1.56*	4.69*	4.36*	1.54 *	4.75*	4.82*
Risk: Sys. Crash Freq.				1.03*	0.82	1.76*
Risk: Sys. Crash More				1.00	0.87	1.09
Risk: App. Crash Freq.				1.00	1.14	1.37
Risk: App. Crash More				0.89*	1.06	0.53*
Application: Firefox				0.66*	0.74*	0.72
Application: Opera				0.29*	1.02	1.22
Application: Reader				0.63*	0.72*	0.64*
Sample: MTurk	–	1.37	1.17	–	1.06	1.10
n	41,551	749	480	41,551	749	480

Table B.20: Table of hierarchical regression models for risk factors in the dataset containing the five messages for which these features are available; p-values significant at $alpha = 0.05$ are marked with *. Survey: F is the filtered survey data.



Update is ready to install

This update addresses customer issues and security vulnerabilities. Adobe recommends that you always install the latest updates.

You may need to restart your system after the update is installed.

(a) Adobe Reader 9.3.2.163

Update is ready to install

This update addresses customer issues and security vulnerabilities. Adobe recommends that you always install the latest updates.

(b) Adobe Reader 9.5.1.283 (crash data available)

An update to your Adobe Flash Player is available

Flash Player enhances your Web browsing experience.

This update includes:

- Security enhancements described in this Security Bulletin

Read more about this update and the End User License Agreement. To change or disable your update notifications, click here.

Updating takes under a minute on broadband; no restart is required.

(c) Flash Player 10.0.22.87

An update to your Adobe Flash Player is available

Flash Player enhances your Web browsing experience.

This update includes:

- Security enhancements described in this Security Bulletin

Read more about this update and the End User License Agreement. To change or disable your update notifications, click here.

Updating takes under a minute on broadband; no restart is required.

(d) Flash Player 10.0.45.2

An update to Adobe® Flash® Player is available.

- Smoother video with hardware acceleration support
- Enhanced performance and memory management
- Support for multi-touch and gesture-enabled content
- Private browsing support and security enhancements

See details...
End User License Agreement
Updating takes under a minute on broadband - no restart is required.

☐ Do not remind me about this update.

(e) Flash Player 10.1.53.64

An update to Adobe® Flash® Player is available.

- Superior HD video performance with hardware acceleration of video
- Support for full screen mode with multiple monitors
- Faster graphics rendering support with Internet Explorer 9
- Bug fixes and security enhancements

See details...
End User License Agreement
Updating takes under a minute on broadband - no restart is required.

☐ Do not remind me about this update.

(f) Flash Player 10.2.152.26

An update to Adobe® Flash® Player is available.

- High quality echo cancellation, voice conferencing support
- Integration with browser privacy controls
- Control panel for simple privacy and storage management
- Bug fixes and security enhancements

See details...
End User License Agreement
Updating takes under a minute on broadband - no restart is required.

☐ Do not remind me about this update.

(g) Flash Player 10.3.181.14 (crash data available)

An update to Adobe® Flash® Player is available.

This update includes:

- Stage3D: support for GPU hardware-accelerated, console-quality games
- Support for HD video chatting (H.264/AVC)
- Access to more premium video content delivered via Flash Access
- Native support for 64-bit browsers
- Improved performance, compatibility and security

See details...

Updating takes under a minute on broadband - no restart is required.
☐ Do not remind me about this update.

(h) Flash Player 11.0.1.152 (crash data available)

Update Ready to Install

A security and stability update for Firefox has been downloaded and is ready to be installed.

Update: **Firefox 8.0.1**

The update will be installed the next time Firefox starts. You can restart Firefox now, or continue working and restart later.

(i) Firefox 8.0.1.4341 (crash data available)

Opera 10.61 is available:

Opera 10.61 is a security and stability upgrade. Opera recommends all users to upgrade to Opera 10.61 to take advantage of these improvements.

Please visit the **Opera desktop changelogs** if you want more information on the update.

The new version will be downloaded and installed automatically. While you are downloading it, you can continue browsing. Opera will be updated after a restart, at which time you will need to know the administrator password.

Size: 14.0 MB

☑ Automatically install updates without notification from now on

(j) Opera 10.61.2484.0

Opera 11.64 is available:

Opera 11.64 is a security and stability upgrade. Opera recommends all users to upgrade to Opera 11.64 to take advantage of these improvements.

Please visit the **Opera desktop changelogs** if you want more information on the update.

The new version will be downloaded and installed automatically. While you are downloading it, you can continue browsing. Opera will be updated after a restart, at which time you will need to know the administrator password.

Size: 11.0 MB

☑ Automatically install updates without notification from now on

(k) Opera 11.64.1403.0 (crash data available)

Figure B.1: Update Messages

Bibliography

[1] CNET. http://www.cnet.com

[2] The Department of Health and Human Services Information Systems Security Awareness Training. http://www.hhs.gov/ocio/securityprivacy/awarenesstraining/issa.pdf

[3] Diffbot. https://www.diffbot.com/

[4] Federal Communications Commission Cyber Security Planning Guide. https://transition.fcc.gov/cyber/cyberplanner.pdf

[5] Mashable. http://mashable.com

[6] Schneier On Security. https://www.schneier.com

[7] Wired. http://www.wired.com

[8] 2013a. American Community Survey 1-year 2013 Census. (2013). https://www.census.gov/acs/www/data/data-tables-and-tools/index.php

[9] 2013. Anonymity Omnibus Dataset. (2013). http://www.pewinternet.org/datasets/july-2013-anonymity-omnibus/

[10] 2013b. Household Income In The Past 12 Months: 2009-2013 American Community Survey 5-Year Estimates. (2013).

[11] 2014. American Community Survey 5-Year Estimates. (2014).

[12] 2015. National Cybersecurity Alliance. (2015). https://staysafeonline.org/

[13] 2015. State and County QuickFacts. (2015). http://quickfacts.census.gov/qfd/states/00000.html

[14] 2016. Chesapeake IRB. (2016). https://www.chesapeakeirb.com/

[15] 2016a. Pew American Trends Panel. (2016). http://www.pewresearch.org/methodology/u-s-survey-research/american-trends-panel/

[16] 2016b. Pew Internet and American Life Project. (2016). http://www.pewinternet.org/

[17] 2016. Reason-Rupe Surveys. (2016). http://reason.com/poll

[18] Hervé Abdi. 2010. Holm's sequential Bonferroni procedure. *Encyclopedia of research design* 1, 8 (2010).

[19] Anne Adams and Martina Angela Sasse. 1999. Users are not the enemy. *Commun. ACM* 42, 12 (1999), 40–46.

[20] H. Akaike. 1974. A new look at the statistical model identification. *IEEE Trans. Automat. Control* (1974). DOI:http://dx.doi.org/10.1109/TAC.1974.1100705

[21] Devdatta Akhawe and Adrienne Porter Felt. 2013. Alice in Warningland: A Large-Scale Field Study of Browser Security Warning Effectiveness.. In *USENIX security symposium*, Vol. 13.

[22] Richard C Anderson. 1972. How to construct achievement tests to assess comprehension. *Review of educational research* 42, 2 (1972), 145–170.

[23] N. A. G. Arachchilage and S. Love. 2013. A Game Design Framework for Avoiding Phishing Attacks. *Comput. Hum. Behav.* (2013). http://dx.doi.org/10.1016/j.chb.2012.12.018

[24] W. A. Arbaugh, W. L. Fithen, and J. McHugh. 2000. Windows of Vulnerability: A Case Study Analysis. *IEEE Computer* 33, 12 (2000), 52–59.

[25] Reg Baker, Stephen Blumberg, and et al. 2010. AAPOR REPORT ON ONLINE PANELS. *The Public Opinion Quarterly* (2010). http://www.jstor.org/stable/40927166

[26] A. Beautement, M. A. Sasse, and M. Wonham. 2008. The compliance budget: Managing security behaviour in organisations. In *workshop on new security paradigms*. DOI:http://dx.doi.org/10.1145/1595676.1595684

[27] Elmer V Bernstam, Dawn M Shelton, Muhammad Walji, and Funda Meric-Bernstam. 2005. Instruments to assess the quality of health information on the World Wide Web: what can our patients actually use? *International journal of medical informatics* 74, 1 (2005), 13–19.

[28] Paul P Biemer and Sharon L Christ. 2008. Weighting survey data. *International handbook of survey methodology* 2008 (2008), 317–341.

[29] JM Blythe and CE Lefevre. 2017. Cyberhygiene Insight Report. *Retrived from https://iotuk. org. uk/wp-content/uploads/2018/01/PETRAS-IoTUK-Cyberhygiene-Insight-Report. pdf* (2017).

[30] John R Bormuth. 1967. Comparable cloze and multiple-choice comprehension test scores. *Journal of Reading* 10, 5 (1967), 291–299.

[31] John R Bormuth. 1968. Cloze test readability: Criterion reference scores. *Journal of educational measurement* 5, 3 (1968), 189–196.

[32] John R Bormuth. 1973. Reading literacy: Its definition and assessment. *Reading research quarterly* (1973), 7–66.

[33] Jasmine Bowers, Bradley Reaves, Imani N Sherman, Patrick Traynor, and Kevin Butler. 2017. Regulators, mount up! analysis of privacy policies for mobile money services. In *Thirteenth Symposium on Usable Privacy and Security ({SOUPS} 2017)*. 97–114.

[34] boyd D., Levy K., and Arwick A. 2013. The Networked Nature of Algorithmic Discrimination. Data and Discrimination: Collected Essays. (2013). `http://newamerica.org/downloads/OTI-Data-an-Discrimination-FINAL-small.pdf`

[35] C. Bravo-Lillo, S. Komanduri, L. F. Cranor, R. W. Reeder, M. Sleeper, J. Downs, and S. Schechter. 2013. Your Attention Please: Designing Security-decision UIs to Make Genuine Risks Harder to Ignore. In *SOUPS*. `DOI:http://dx.doi.org/10.1145/2501604.2501610`

[36] Jonathan C Brown, Gwen A Frishkoff, and Maxine Eskenazi. 2005. Automatic question generation for vocabulary assessment. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. ACL.

[37] Zoran Bursac, C Heath Gauss, David Keith Williams, and David W Hosmer. 2008. Purposeful selection of variables in logistic regression. *Source code for biology and medicine* 3, 1 (2008), 17.

[38] Karoline Busse, Julia Schäfer, and Matthew Smith. 2019. Replication: No One Can Hack My Mind Revisiting a Study on Expert and Non-Expert Security Practices and Advice. In *SOUPS 2019: Symposium on Usable Privacy and Security*. `https://www.usenix.org/conference/soups2019/presentation/busse`

[39] Kelly Erinn Caine. 2009. *Exploring everyday privacy behaviors and misclosures.* Ph.D. Dissertation. Georgia Institute of Technology.

[40] L. O. Campbell, M. Kepple, and C. Herlihy. 2015. Women in technology:An underrepresented population. In *Global Learn 2015*. AACE. `http://www.editlib.org/p/150902`

[41] Fredrik Carlsson. 2010. Design of stated preference surveys: Is there more to learn from behavioral economics? *Environmental and Resource Economics* 46, 2 (2010), 167–177.

[42] K. Charmaz. 2006. *Constructing grounded theory : A. practical guide through qualitative analysis.* Sage Publications, London; Thousand Oaks, Calif. http://www.amazon.com/Constructing-Grounded-Theory-Qualitative-Introducing/dp/0761973532

[43] Chia-Yin Chen, Hsien-Chin Liou, and Jason S Chang. 2006. Fast: an automatic generation system for grammar tests. In *Proceedings of the COLING/ACL on Interactive presentation sessions.* ACL.

[44] C. Ciampa. 2013. A comparison of password feedback mechanisms and their impact on password entropy. *Information Management & Computer Security* (2013). DOI:http://dx.doi.org/10.1108/IMCS-12-2012-0072

[45] Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics* 165, 2 (2014), 97–135.

[46] Kevyn Collins-Thompson and James P Callan. 2004. A language modeling approach to predicting reading difficulty. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004.*

[47] Max Coltheart. 1981. The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A* 33, 4 (1981), 497–505.

[48] Jean M Converse and Stanley Presser. 1986. *Survey questions: Handcrafting the standardized questionnaire.* Number 63. Sage.

[49] Mick P Couper and Peter V Miller. 2008. Web survey methods introduction. *Public Opinion Quarterly* 72, 5 (2008), 831–835.

[50] Elisabeth Coutts and Ben Jann. 2011. Sensitive questions in online surveys: Experimental results for the randomized response technique (RRT) and the unmatched count technique (UCT). *Sociological Methods & Research* 40, 1 (2011), 169–193.

[51] Lorrie Faith Cranor. 2008. A framework for reasoning about the human in the loop. *UPSEC* 8, 2008 (2008).

[52] Lee J Cronbach and Paul E Meehl. 1955. Construct validity in psychological tests. *Psychological bulletin* 52, 4 (1955), 281.

[53] Scott A Crossley, Jerry Greenfield, and Danielle S McNamara. 2008. Assessing text readability using cognitively based indices. *Tesol Quarterly* 42, 3 (2008), 475–493.

[54] Edgar Dale and Ralph W Tyler. 1934. A study of the factors influencing the difficulty of reading materials for adults of limited reading ability. *The Library Quarterly* 4, 3 (1934), 384–412.

[55] Amedeo D'Angiulli*, Linda S Siegel, and Clyde Hertzman. 2004. Schooling, socioeconomic context and literacy development. *Educational Psychology* 24, 6 (2004), 867–883.

[56] S. Das, T. H. Kim, L.A. Dabbish, and J.I. Hong. 2014a. The Effect of Social Influence on Security Sensitivity. In *SOUPS.* `https://www.usenix.org/conference/SOUPS2014/proceedings/presentation/das`

[57] S. Das, A. D.I. Kramer, L. A. Dabbish, and J. I. Hong. 2014b. Increasing Security Sensitivity With Social Proof: A Large-Scale Experimental Confirmation. In *CCS.* `DOI:http://dx.doi.org/10.1145/2660267.2660271`

[58] Richard R Day and Jeong-suk Park. 2005. Developing Reading Comprehension Questions. *Reading in a foreign language* 17, 1 (2005), 60–73.

[59] Orphée De Clercq and Véronique Hoste. 2016. All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch. *Computational Linguistics* 42, 3 (2016), 457–490.

[60] Lawrence T DeCarlo. 1997. On the meaning and use of kurtosis. *Psychological methods* 2, 3 (1997), 292.

[61] Theresa J. DeMaio, Jennifer Rothgeb, and Jennifer Hess. 2003. Improving Survey Quality Through Pretesting. *U.S. Bureau of the Census* (2003). `https://www.census.gov/srd/papers/pdf/sm98-03.pdf`

[62] T. Denning, A. Lerner, A. Shostack, and T. Kohno. 2013. Control-Alt-Hack: The Design and Evaluation of A. Card Game for Computer Security Awareness and Education. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer &#38; Communications Security (CCS '13).* ACM, New York, NY, USA, 915–928. `DOI:http://dx.doi.org/10.1145/2508859.2516753`

[63] Don A Dillman, Robert D Tortora, and Dennis Bowker. 1998. Principles for constructing web surveys. In *Joint Meetings of the American Statistical Association.*

[64] Thomas Dübendorfer and Stefan Frei. 2009. Web Browser Security Update Effectiveness. In *International Workshop on Critical Information Infrastructures Security.*

[65] Nell K Duke and P David Pearson. 2009. Effective practices for developing reading comprehension. *Journal of education* 189, 1-2 (2009), 107–122.

[66] Tudor Dumitras and Darren Shou. 2011. Toward a Standard Benchmark for Computer Security Research: The Worldwide Intelligence Network Environment (WINE). In *Workshop on Building Analysis Datasets and Gathering Experience Returns for Security.*

[67] Zakir Durumeric, James Kasten, David Adrian, J. Alex Halderman, Michael Bailey, and et al. 2014. The Matter of Heartbleed. In *Internet Measurement Conference.*

[68] S. Egelman, L. F. Cranor, and J. Hong. 2008. You'Ve Been Warned: An Empirical Study of the Effectiveness of Web Browser Phishing Warnings. In *CHI.* DOI:http://dx.doi.org/10.1145/1357054.1357219

[69] S. Egelman and E. Peer. 2015. Scaling the Security Wall: Developing a Security Behavior Intentions Scale (SeBIS). In *CHI.* DOI:http://dx.doi.org/10.1145/2702123.2702249

[70] Serge Egelman, Andreas Sotirakopoulos, Ildar Muslukhov, Konstantin Beznosov, and Cormac Herley. 2013. Does my password go up to eleven?: the impact of password meters on password selection. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, 2379–2388.

[71] Warwick B Elley and Cedric Croft. 1989. *Assessing the difficulty of reading materials: The noun frequency method.* ERIC.

[72] Gunther Eysenbach, John Powell, Oliver Kuss, and Eun-Ryoung Sa. 2002. Empirical studies assessing the quality of health information for consumers on the world wide web: a systematic review. *Jama* 287, 20 (2002), 2691–2700.

[73] Michael Fagan, Mohammad Maifi Hasan Khan, and Ross Buck. 2015. A study of users' experiences and beliefs about software update messages. *Computers in Human Behavior* (2015).

[74] Sascha Fahl, Marian Harbach, Yasemin Acar, and Matthew Smith. 2013. On the ecological validity of a password study. In *Proceedings of the Ninth Symposium on Usable Privacy and Security.* ACM, 13.

[75] Adrienne Porter Felt, Alex Ainslie, Robert W Reeder, Sunny Consolvo, Somas Thyagaraja, Alan Bettes, Helen Harris, and Jeff Grimes. 2015. Improving SSL warnings: Comprehension and adherence. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems.* ACM, 2893–2902.

[76] Adrienne Porter Felt, Elizabeth Ha, Serge Egelman, Ariel Haney, Erika Chin, and David Wagner. 2012. Android permissions: User attention, comprehension, and behavior. In *Proceedings of the Eighth Symposium on Usable Privacy and Security.* ACM, 3.

[77] Michael Fendrich and Connie M Vaughn. 1994. Diminished lifetime substance use over time: An inquiry into differential underreporting. *Public Opinion Quarterly* 58, 1 (1994).

[78] Lijun Feng, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the ACL*. ACL, 229–237.

[79] A. Fisher and J. Margolis. 2002. Unlocking the Clubhouse: The Carnegie Mellon Experience. *SIGCSE Bull.* 34 (June 2002), 79–83. `DOI:http://dx.doi.org/10.1145/543812.543836`

[80] Robert J Fisher. 1993. Social desirability bias and the validity of indirect questioning. *Journal of consumer research* 20, 2 (1993), 303–315.

[81] Rudolf Flesch. 1943. Marks of readable style; a study in adult education. *Teachers College Contributions to Education* (1943).

[82] Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology* 32, 3 (1948), 221.

[83] Thomas François and Eleni Miltsakaki. 2012. Do NLP and machine learning improve traditional readability formulas?. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations.* Association for Computational Linguistics, 49–57.

[84] D. G. Freelon. 2010. ReCal: Intercoder reliability calculation as a web service. *International Journal of Internet Science* 5, 1 (2010), 20–33.

[85] Daniela B Friedman and Laurie Hoffman-Goetz. 2006. A systematic review of readability and comprehension instruments used for print and web-based cancer information. *Health Education & Behavior* 33, 3 (2006), 352–373.

[86] Karl Pearson F.R.S. 1900. X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5* (1900). `DOI:http://dx.doi.org/10.1080/14786440009463897`

[87] M. Fujita, M. Yamada, S. Arimura, Y. Ikeya, and M. Nishigaki. 2015. An Attempt to Memorize Strong Passwords while Playing Games. In *NBIS*. `DOI:http://dx.doi.org/10.1109/NBiS.2015.41`

[88] S.M. Furnell, P. Bryant, and A.D. Phippen. 2007. Assessing the security perceptions of personal Internet users. *Computers & Security* (2007). `DOI:http://dx.doi.org/10.1016/j.cose.2007.03.001`

[89] Vaibhav Garg, L. Jean Camp, Katherine Connelly, and Lesa Lorenzen-Huber. 2012. Risk Communication Design: Video vs. Text. In *PETS*. DOI:http://dx.doi.org/10.1007/978-3-642-31680-7

[90] Donna Marie Gates. 2011. How to generate cloze questions from definitions: A syntactic approach. In *AAAI Fall Symposium Series*.

[91] Christos Gkantsidis, Thomas Karagiannis, Pablo Rodriguez, and Milan Vojnovic. 2006. Planet scale software updates. In *ACM SIGCOMM Computer Communication Review*.

[92] Takuya Goto, Tomoko Kojiri, Toyohide Watanabe, Tomoharu Iwata, and Takeshi Yamada. 2010. Automatic generation system of multiple-choice cloze questions and its evaluation. *Knowledge Management & E-Learning: An International Journal (KM&EL)* 2, 3 (2010), 210–224.

[93] Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods* 36, 2 (2004), 193–202.

[94] P Grassi, M Garcia, and J Fenton. 2017. NIST Special Publication 800-63-3 Digital Identity Guidelines. *National Institute of Standards and Technology, Los Altos, CA* (2017).

[95] Jim Gray. 1986. Why do computers stop and what can be done about it?. In *Symposium on reliability in distributed software and database systems*. Los Angeles, CA, USA, 3–12.

[96] G. Guest, A. Bunce, and L. Johnson. 2006. How Many Interviews Are Enough?: An Experiment with Data Saturation and Variability. *Field Methods* 18 (2006), 59–82. DOI:http://dx.doi.org/10.1177/1525822X05279903

[97] Daniel A Hackman, Martha J Farah, and Michael J Meaney. 2010. Socioeconomic status and the brain: mechanistic insights from human and animal research. *Nature reviews neuroscience* 11, 9 (2010), 651.

[98] T. Halevi, J. Lewis, and N. Memon. 2013. A Pilot Study of Cyber Security and Privacy Related Behavior and Personality Traits. In *WWW*. http://dl.acm.org/citation.cfm?id=2487788.2488034

[99] Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of COLING 2012*. 1063–1080.

[100] Marian Harbach, Sascha Fahl, Thomas Muders, and Matthew Smith. 2012. Towards Measuring Warning Readability. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security (CCS '12)*. ACM, New York, NY, USA, 989–991. DOI:http://dx.doi.org/10.1145/2382196.2382301

[101] Marian Harbach, Sascha Fahl, Polina Yakovleva, and Matthew Smith. 2013. Sorry, I don't get it: An analysis of warning message texts. In *FC*.

[102] J. B. Hardee, R. West, and C. B. Mayhorn. 2006. To Download or Not to Download: An Examination of Computer Security Decision Making. *interactions* (2006). DOI:http://dx.doi.org/10.1145/1125864.1125887

[103] E. Hargittai. 2002. Second-Level Digital Divide: Mapping Differences in People's Online Skills. *First Monday* (2002). http://arxiv.org/abs/cs.CY/0109068

[104] E. Hargittai. 2003. *The Digital Divide and What to Do About It.* http://www.eszter.com/research/pubs/hargittai-digitaldivide.pdf

[105] Eszter Hargittai and Yuli Patrick Hsieh. 2012. Succinct Survey Measures of Web-Use Skills. *Soc. Sci. Comput. Rev.* (2012). DOI:http://dx.doi.org/10.1177/0894439310397146

[106] M. C. Harrell and M. A. Bradley. 2009. *Data collection methods. Semi-structured interviews and focus groups.* Technical Report. DTIC Document. http://www.rand.org/content/dam/rand/pubs/technical_reports/2009/RAND_TR718.pdf

[107] Jerome C Harste and others. 1984. *Language stories & literacy lessons.* ERIC.

[108] T. Hastie, R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition.* Springer New York. http://www.springer.com/us/book/9780387848570

[109] Michael Heilman. 2011. *Automatic factual question generation from text.* Ph.D. Dissertation. Carnegie Mellon University.

[110] C. Herley. 2009. So Long, and No Thanks for the Externalities: The Rational Rejection of Security Advice by Users. In *NPSW*. DOI:http://dx.doi.org/10.1145/1719030.1719050

[111] C. Herley. 2014. More is Not the Answer. *IEEE Security and Privacy magazine* (2014). http://research.microsoft.com/apps/pubs/default.aspx?id=208503

[112] C. Herley. 2016a. The Unfalsifiability of Security Claims. (2016). https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/herley

[113] Cormac Herley. 2016b. Unfalsifiability of security claims. *Proceedings of the National Academy of Sciences* (2016).

[114] Cormac Herley and Paul C Van Oorschot. 2018. Science of Security: Combining Theory and Measurement to Reflect the Observable. *IEEE Security & Privacy* 16, 1 (2018), 12–22.

[115] A. L. Holbrook, M. C. Green, and J. 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* (2003). http://poq.oxfordjournals.org/cgi/citmgr?gca=pubopq; 67/1/79

[116] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics* (1979), 65–70.

[117] Ayako Hoshino and Hiroshi Nakagawa. 2007. Assisting cloze test making with a web application. In *Society for Information Technology & Teacher Education International Conference*. AACE.

[118] D. W. Hosmer and S. Lemeshow. 2000. *Applied logistic regression*. http://opac.inria.fr/record=b1128889

[119] A. E. Howe, I. Ray, M. Roberts, M. Urbanska, and Z. Byrne. 2012. The Psychology of Security for the Home Computer User.. In *IEEE S&P*.

[120] Iulia Ion, Rob Reeder, and Sunny Consolvo. 2015. "... no one can hack my mind": Comparing Expert and Non-Expert Security Practices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*. 327–346.

[121] Panagiotis G Ipeirotis. 2010. Demographics of mechanical turk. (2010).

[122] J. Jerome. 2013. Buying and Selling Privacy: Big Data's Different Burdens and Benefits. *Stanford Law Review* (2013). http://www.stanfordlawreview.org/online/privacy-and-big-data/buying-and-selling-privacy

[123] Ruogu Kang, Stephanie Brown, Laura Dabbish, and Sara B Kiesler. 2014. Privacy Attitudes of Mechanical Turk Workers and the US Public.. In *SOUPS*. 37–49.

[124] Rohit J Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond J Mooney, Salim Roukos, and Chris Welty. 2010. Learning to predict readability using diverse linguistic features. In *Proceedings of the 23rd international conference on computational linguistics*. Association for Computational Linguistics, 546–554.

[125] Patrick Gage Kelley, Lorrie Faith Cranor, and Norman Sadeh. 2013. Privacy As Part of the App Decision-making Process. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 3393–3402. DOI:http://dx.doi.org/10.1145/2470654.2466466

[126] Timothy Kelley and Bennett I Bertenthal. 2016. Attention and Past Behavior, not Security Knowledge, Modulate Users? Decisions to Login to Insecure Websites. *Information and Computer Security* (2016). DOI:http://dx.doi.org/10.1108/ICS-01-2016-0002

[127] Aishwarya Ratan Kentaro Toyama. 2009. Kelsa+: Digital Literacy for Low-Income Office Workers. In *International Conference on Information and Communication Technologies and Development*. https://www.microsoft.com/en-us/research/publication/kelsa-digital-literacy-for-low-income-office-workers/

[128] Leslie Kish. 1965. Survey sampling. (1965).

[129] Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine* 15, 2 (2016), 155–163.

[130] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 8 (2009), 30–37.

[131] Frauke Kreuter, Stanley Presser, and Roger Tourangeau. 2008. Social desirability bias in CATI, IVR, and web surveys: The effects of mode and question sensitivity. *Public Opinion Quarterly* 72, 5 (2008).

[132] J. A. Krosnick. 2000. The threat of satisficing in surveys: the shortcuts respondents take in answering questions. Survey Methods Centre Newsletter. (2000).

[133] J. A. Krosnick. 2010. *Handbook of Survey Research*. http://www.sciencedirect.com/science/book/9780125982269

[134] Ivar Krumpal. 2013. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity* 47, 4 (2013), 2025–2047.

[135] P. Kumaraguru, J. Cranshaw, A. Acquisti, L. F. Cranor, J. Hong, M. A. Blair, and T. Pham. 2009. School of Phish: A Real-world Evaluation of Anti-phishing Training. In *SOUPS*. DOI:http://dx.doi.org/10.1145/1572532.1572536

[136] P. Kumaraguru, S. Sheng, A. Acquisti, L. F. Cranor, and J. Hong. 2010. Teaching Johnny Not to Fall for Phish. *ACM Trans. Internet Technol.* (2010). DOI:http://dx.doi.org/10.1145/1754393.1754396

[137] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* (1977), 159–174.

[138] John Lee and Stephanie Seneff. 2007. Automatic generation of cloze items for prepositions. In *Eighth Annual Conference of the International Speech Communication Association*.

[139] Eric Lin, Saul Greenberg, Eileah Trotter, David Ma, and John Aycock. 2011. Does Domain Highlighting Help People Identify Phishing Sites?. In *CHI*. DOI:http://dx.doi.org/10.1145/1978942.1979244

[140] Wen-Pin Lin and Heng Ji. 2010. Automatic Cloze Generation based on Cross-document Information Extraction. In *Asian Conference on Education*.

[141] Michael W Link and Ali H Mokdad. 2005. Effects of survey mode on self-reports of adult alcohol consumption: a comparison of mail, web and telephone approaches. *Journal of Studies on Alcohol* 66, 2 (2005), 239–245.

[142] M. Lombard, J. Snyder-Duch, and C. C. Bracken. 2002. Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability. *Human Communication Research* 28 (2002), 587–604. `http://dx.doi.org/10.1111/j.1468-2958.2002.tb00826.x`

[143] T. Lumley. 2016. R 'survey': Analysis of Complex Survey Samples. (2016). `https://cran.r-project.org/web/packages/survey/survey.pdf`

[144] Anandi Mani, Sendhil Mullainathan, Eldar Shafir, and Jiaying Zhao. 2013. Poverty impedes cognitive function. *science* 341, 6149 (2013), 976–980.

[145] H. B. Mann and D. R. Whitney. 1947a. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Ann. Math. Statist.* (1947). `DOI:http://dx.doi.org/10.1214/aoms/1177730491`

[146] Henry B Mann and Donald R Whitney. 1947b. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* (1947), 50–60.

[147] Alfred C Marcus and Lori A Crane. 1986. Telephone surveys in public health research. *Medical care* (1986), 97–112.

[148] Arunesh Mathur and Marshini Chetty. 2017. Impact of User Characteristics on Attitudes Towards Automatic Mobile Application Updates. In *Symposium on Usable Privacy and Security (SOUPS)*.

[149] Arunesh Mathur, Josefine Engel, Sonam Sobti, Victoria Chang, and Marshini Chetty. 2016. " They Keep Coming Back Like Zombies": Improving Software Updating Interfaces.. In *SOUPS*.

[150] Arunesh Mathur, Nathan Malkin, Marian Harbach, Eyal Peer, and Serge Egelman. 2018. Quantifying User Beliefs about Software Updates. (2018).

[151] D. C. May, N. E. Rader, and S. Goodrum. 2010. A Gendered Assessment of the 'Threat of Victimization': Examining Gender Differences in Fear of Crime, Perceived Risk, Avoidance, and Defensive Behaviors. *Criminal Justice Review* 35, 2 (2010), 159–182. `http://cjr.sagepub.com/content/35/2/159.abstract`

[152] Michelle L Mazurek, Saranga Komanduri, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Patrick Gage Kelley, Richard Shay, and Blase Ur. 2013. Measuring password guessability for an entire university. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*. ACM, 173–186.

[153] Aleecia M. McDonald, Robert W. Reeder, Patrick Gage Kelley, and Lorrie Faith Cranor. 2009. *A Comparative Study of Online Privacy Policies and Formats*. Springer Berlin Heidelberg, Berlin, Heidelberg, 37–55. DOI: http://dx.doi.org/10.1007/978-3-642-03168-7_3

[154] Danielle S McNamara, Arthur C Graesser, Philip M McCarthy, and Zhiqiang Cai. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.

[155] Andreas Möller, Florian Michahelles, Stefan Diewald, Luis Roalter, and Matthias Kranz. 2012. Update behavior in app markets and security implications: A case study in google play. In *Research in the Large, LARGE 3.0: 21/09/2012-21/09/2012*.

[156] David Moore, Colleen Shannon, and Kimberly C. Claffy. 2002. Code-Red: a case study on the spread and victims of an internet worm. In *Internet Measurement Workshop*.

[157] Miraida Morales and Nina Wacholder. 2018. Conceptualizing the Role of Reading and Literacy in Health Information Practices. In *International Conference on Information*. Springer.

[158] Jack Mostow and Hyeju Jang. 2012. Generating diagnostic multiple choice comprehension cloze questions. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*. ACL.

[159] Antonio Nappa, Richard Johnson, Leyla Bilge, Juan Caballero, and Tudor Dumitras. 2015. The Attack of the Clones: A Study of the Impact of Shared Code on Vulnerability Patching. In *IEEE Symposium on Security and Privacy*.

[160] Annamaneni Narendra, Manish Agarwal, and others. 2013. Automatic cloze-questions generation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*.

[161] James Nicholson, Lynne Coventry, and Pam Briggs. 2018. Introducing the Cybersurvival task: assessing and addressing staff beliefs about effective cyber protection. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*. 443–457.

[162] Ana Nuno and Freya AV St John. 2015. How to ask sensitive questions in conservation: A review of specialized questioning techniques. *Biological Conservation* 189 (2015), 5–15.

[163] Ann A O'Connell. 2006. *Logistic regression models for ordinal response variables*. Number 146. Sage.

[164] John W Oller, J Donald Bowen, Ton That Dien, and Victor W Mason. 1972. Cloze Tests in English, Thai, and Vietnamese: Native and Non-Native Performance. *Language Learning* 22, 1 (1972), 1–15.

[165] Kenneth Olmstead and Aaron Smith. 2017. Americans and Cybersecurity. (2017).

[166] JE Overland, PL Hoskins, MJ McGill, and DK Yue. 1993. Low literacy: a problem in diabetes education. *Diabetic Medicine* 10, 9 (1993), 847–850.

[167] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running experiments on amazon mechanical turk. (2010).

[168] Simon Parkin, Elissa M Redmiles, Lynne Coventry, and M Angela Sasse. 2019. Security When it is Welcome: Exploring Device Purchase as an Opportune Moment for Security Behavior Change. In *Proceedings of the Workshop on Usable Security and Privacy (USEC'19)*. Internet Society.

[169] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. 2014. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior research methods* 46, 4 (2014).

[170] Juan Pino, Michael Heilman, and Maxine Eskenazi. 2008. A selection strategy to improve cloze question quality. In *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains. 9th International Conference on Intelligent Tutoring Systems, Montreal, Canada.*

[171] Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the conference on empirical methods in natural language processing*. ACL, 186–195.

[172] Erika Shehan Poole, Marshini Chetty, Tom Morgan, Rebecca E. Grinter, and W. Keith Edwards. 2009. Computer Help at Home: Methods and Motivations for Informal Technical Support *(CHI)*. DOI:http://dx.doi.org/10.1145/1518701.1518816

[173] Stanley Presser, Mick P Couper, Judith T Lessler, Elizabeth Martin, Jean Martin, Jennifer M Rothgeb, and Eleanor Singer. 2004. Methods for testing and evaluating survey questions. *Public opinion quarterly* 68, 1 (2004), 109–130.

[174] Kevin M Quinn, Burt L Monroe, Michael Colaresi, Michael H Crespin, and Dragomir R Radev. 2010. How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science* 54, 1 (2010), 209–228.

[175] E. Rader and R. Wash. 2015. Identifying patterns in informal sources of security information. *J. Cybersecurity* (2015). DOI:http://dx.doi.org/10.1093/cybsec/tyv008

[176] E. Rader, R. Wash, and B. Brooks. 2012. Stories As Informal Lessons About Security. In *SOUPS*. DOI:http://dx.doi.org/10.1145/2335356.2335364

[177] Lee Rainie, Sara Kiesler, Ruogu Kang, Mary Madden, Maeve Duggan, Stephanie Brown, and Laura Dabbish. 2013. Anonymity, privacy, and security online. *Pew Research Center* 5 (2013).

[178] Earl F Rankin and Joseph W Culhane. 1969. Comparable cloze and multiple-choice comprehension test scores. *Journal of Reading* 13, 3 (1969), 193–198.

[179] E.M. Redmiles, S. Kross, and M. L. Mazurek. 2016. How I Learned to be Secure: a Census-Representative Survey of Security Advice Sources and Behavior. In *CCS*. `http://dl.acm.org/citation.cfm?id=2978307`

[180] E.M. Redmiles, L. Maszkiewicz, E. Hwang, D. Kuchhal, E. Liu, M. Morales, D. Peskov, S. Rao, R. Stevens, K. Gligoric, S. Kross, M.L. Mazurek, and H. Daume III. 2019. Comparing and Developing Tools to Measure the Readability of Domain-Specific Texts. In *EMNLP 2019: Conference on Empirical Methods in Natural Language Processing*.

[181] Elissa M Redmiles. 2019. " Should I Worry?" A Cross-Cultural Examination of Account Security Incident Response. In *IEEE Security and Privacy*.

[182] Elissa M Redmiles, Neha Chachra, and Brian Waismeyer. 2018. Examining the Demand for Spam: Who Clicks?. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 212.

[183] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. 2017. Where is the Digital Divide?: A Survey of Security, Privacy, and Socioeconomics. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 931–936.

[184] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. 2019. How Well Do My Results Generalize? Comparing Security and Privacy Survey Results from MTurk, Web, and Telephone Samples. In *IEEE Security & Privacy*.

[185] Elissa M Redmiles, Sean Kross, Alisha Pradhan, and Michelle L Mazurek. 2017. *How Well Do My Results Generalize? Comparing Security and Privacy Survey Results from MTurk and Web Panels to the US*. Technical Report. `https://drum.lib.umd.edu/handle/1903/19164`

[186] Elissa M Redmiles, Amelia R Malone, and Michelle L Mazurek. 2016. I Think They're Trying to Tell Me Something: Advice Sources and Selection for Digital Security. In *Security and Privacy (SP), 2016 IEEE Symposium on*. IEEE, 272–288.

[187] Elissa M Redmiles, Michelle L Mazurek, and John P Dickerson. 2018a. Dancing pigs or externalities?: Measuring the rationality of security decisions. In *Proceedings of the 2018 ACM Conference on Economics and Computation*. ACM, 215–232.

[188] Elissa M Redmiles, Ziyun Zhu, Sean Kross, Dhruv Kuchhal, Tudor Dumitras, and Michelle L Mazurek. 2018b. Asking for a Friend: Evaluating Response Biases in Security User Studies. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 1238–1255.

[189] Robert W Reeder, Iulia Ion, and Sunny Consolvo. 2017. 152 simple steps to stay safe online: security advice for non-tech-savvy users. *IEEE Security & Privacy* (2017).

[190] Luz Rello, Martin Pielot, and Mari-Carmen Marcos. 2016. Make it big!: The effect of font size and line spacing on online readability. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM.

[191] Eric Rescorla. 2003. Security holes... who cares. In *USENIX Security Symposium*.

[192] R. E. Rice. 2006. Influences, usage, and outcomes of Internet health information searching: Multivariate results from the Pew surveys. *International J. Medical Informatics* (2006). DOI:http://dx.doi.org/10.1016/j.ijmedinf.2005.07.032

[193] Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 193–203.

[194] Stefan A. Robila and James W. Ragucci. 2006. Don'T Be a Phish: Steps in User Education. In *SIGCSE*. DOI:http://dx.doi.org/10.1145/1140124.1140187

[195] Ronald W Rogers and Steven Prentice-Dunn. 1997. Protection motivation theory. (1997).

[196] Joel Ross, Lilly Irani, M Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI'10 extended abstracts on Human factors in computing systems*. ACM, 2863–2872.

[197] Sukamol S. and S. Jakobsson. 2008. Using Cartoons to Teach Internet Security. *Cryptologia* 32, 2 (2008), 137–154. DOI:http://dx.doi.org/10.1080/01611190701743724

[198] Armin Sarabi, Ziyun Zhu, Chaowei Xiao, Mingyan Liu, and Tudor Dumitras. 2017. Patch Me If You Can: A Study on the Effects of Individual User Behavior on the End-Host Vulnerability State. In *International Conference on Passive and Active Network Measurement*. Springer, 113–125.

[199] Loukia Sarroub and P David Pearson. 1998. Two steps forward, three steps back: The stormy history of reading comprehension assessment. *The Clearing House* 72, 2 (1998), 97–105.

[200] Jeff Sauro and Joseph S Dumas. 2009. Comparison of three one-question, post-task usability questionnaires. In *Proceedings of the SIGCHI conference on human factors in computing systems.* ACM, 1599–1608.

[201] N.C. Schaeffer and S. Presser. 2003. The Science of Asking Questions. *Annual Review of Sociology* (2003). DOI:http://dx.doi.org/10.1146/annurev.soc.29.110702.110112

[202] Stuart Schechter and Joseph Bonneau. 2015. Learning Assigned Secrets for Unlocking Mobile Devices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015).* 277–295.

[203] S. E. Schechter, R. Dhamija, A. Ozment, and I. Fischer. 2007. The Emperor's New Security Indicators. *IEEE S&P* (2007). DOI:http://dx.doi.org/10.1109/SP.2007.35

[204] A.J. Scott and J.N.K. Rao. 1984. On Chi-squared Tests For Multiway Contingency Tables with Proportions Estimated From Survey Data. *Annals of Statistics* (1984). https://www.jstor.org/stable/2241033

[205] Muhammad Shahzad, Muhammad Zubair Shafiq, and Alex X. Liu. 2012. A large scale exploratory analysis of software vulnerability life cycles. In *International Conference on Software Engineering.*

[206] Cyrus Shaoul. 2010. The westbury lab wikipedia corpus. *Edmonton, AB: University of Alberta* (2010).

[207] S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. Downs. 2010. Who Falls for Phish?: A Demographic Analysis of Phishing Susceptibility and Effectiveness of Interventions. In *CHI.* http://doi.acm.org/10.1145/1753326.1753383

[208] S. Sheng, B. Magnien, P. Kumaraguru, A. Acquisti, L. F. Cranor, J. Hong, and E. Nunge. 2007. Anti-Phishing Phil: The Design and Evaluation of a Game That Teaches People Not to Fall for Phish. In *SOUPS.* DOI:http://dx.doi.org/10.1145/1280680.1280692

[209] Yue Shi, Martha Larson, and Alan Hanjalic. 2014. Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. *ACM Computing Surveys (CSUR)* 47, 1 (2014), 3.

[210] Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin* 86, 2 (1979), 420.

[211] Robert Siciliano. 2012. 17 Percent of PCs Are Exposed. (2012). https://blogs.mcafee.com/consumer/family-safety/17-of-pcs-are-exposed/

[212] RI Singh, M Sumeeth, and J Miller. 2012. Evaluating the readability of privacy policies in mobile environments. *Developments in Technologies for Human-Centric Mobile Computing and Applications* 56 (2012).

[213] Frank Smith. 2004. *Understanding reading: A psycholinguistic analysis of reading and learning to read.* Routledge.

[214] Scott Smith. 4 Ways to Ensure Valid Responses for your Online Survey. *Qualtrics* (????). https://www.qualtrics.com/blog/online-survey-valid-responses/

[215] L.D. Stanley. 2003. Beyond Access: Psychosocial Barriers to Computer Literacy Special Issue: ICTs and Community Networking. *The Information Society* (2003). DOI:http://dx.doi.org/10.1080/715720560

[216] A. Strauss and J. Corbin. 1998. *Basics of qualitative research: Procedures and techniques for developing grounded theory.*

[217] Mega Subramaniam, Natalie Greene Taylor, Beth St. Jean, Rebecca Follman, Christie Kodama, and Dana Casciotti. 2015. As simple as that?: Tween credibility assessment in a complex online world. *Journal of Documentation* 71, 3 (2015), 550–571.

[218] Joshua Sunshine, Serge Egelman, Hazim Almuhimedi, Neha Atri, and Lorrie Faith Cranor. 2009. Crying Wolf: An Empirical Study of SSL Warning Effectiveness.. In *USENIX security symposium.*

[219] Wilson L Taylor. 1953. Cloze procedure: A new tool for measuring readability. *Journalism Bulletin* 30, 4 (1953), 415–433.

[220] Wilson L Taylor. 1956. Recent developments in the use of "Cloze Procedure". *Journalism Quarterly* 33, 1 (1956), 42–99.

[221] Chaffai Tekfi. 1987. Readability formulas: An overview. *Journal of documentation* 43, 3 (1987), 261–273.

[222] Tuck Meng Tham. 1987. *Linguistic variables as predictors of Chinese text readability.* Ph.D. Dissertation.

[223] Yuan Tian, Bin Liu, Weisi Dai, Blase Ur, Patrick Tague, and Lorrie Faith Cranor. 2015. Supporting privacy-conscious app update decisions with user reviews. In *Proceedings of the 5th Annual ACM CCS Workshop on Security and Privacy in Smartphones and Mobile Devices.* ACM.

[224] R. Tourangeau and T. Yan. 2007. Sensitive Questions in Surveys. *Psychological Bulletin* (2007).

[225] B. Ur, P. G. Kelley, S. Komanduri, J. Lee, M. Maass, M. L Mazurek, T. Passaro, R. Shay, T. Vidas, L. Bauer, N. Christin, and L. F. Cranor. 2012. How does your password measure up? the effect of strength meters on password creation. In *USENIX Sec.* https://www.usenix.org/system/files/conference/usenixsecurity12/sec12-final209.pdf

[226] Sowmya Vajjala and Detmar Meurers. 2013. On the applicability of readability models to web texts. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*. 59–68.

[227] Jan van Dijk and Kenneth Hacker. 2003. The Digital Divide as a Complex and Dynamic Phenomenon. *The Information Society* (2003). DOI:http://dx.doi.org/10.1080/01972240309487

[228] Kami Vaniea and Yasmeen Rashidi. 2016. Tales of software updates: The process of updating software. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM.

[229] Kami E Vaniea, Emilee Rader, and Rick Wash. 2014. Betrayed by updates: how negative experiences affect future security. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 2671–2674.

[230] Jessica Vitak, Yuting Liao, Mega Subramaniam, and Priya Kumar. 2018. 'I Knew It Was Too Good to Be True: The Challenges Economically Disadvantaged Internet Users Face in Assessing Trustworthiness, Avoiding Scams, and Developing Self-Efficacy Online. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 176.

[231] R. Wash. 2010. Folk models of home computer security. In *SOUPS*. DOI:http://dx.doi.org/10.1145/1837110.1837125

[232] R. Wash and E. Rader. 2015. Too Much Knowledge? Security Beliefs and Protective Behaviors Among United States Internet Users. In *SOUPS*. https://www.usenix.org/conference/SOUPS2015/proceedings/presentation/wash

[233] Gordon B Willis. 2004. *Cognitive interviewing: A tool for improving questionnaire design*. Sage Publications.

[234] George Y Wong and William M Mason. 1985. The hierarchical logistic regression model for multilevel analysis. *J. Amer. Statist. Assoc.* 80, 391 (1985), 513–524.

[235] M. Wu, R. C. Miller, and S. L. Garfinkel. 2006. Do Security Toolbars Actually Prevent Phishing Attacks?. In *CHI*. DOI:http://dx.doi.org/10.1145/1124772.1124863

[236] Scott Yilek, Eric Rescorla, Hovav Shacham, Brandon Enright, and Stefan Savage. 2009. When private keys are public: Results from the 2008 Debian OpenSSL vulnerability. In *Internet Measurement Conference*.

[237] Ming Yuan and Yi Lin. 2006. Model selection and estimation in regression with grouped variables. *J. Royal Statistical Society* (2006). `DOI:http://dx.doi.org/10.1111/j.1467-9868.2005.00532.x`

[238] L. Zhang-Kennedy, Sonia Chiasson, and Robert Biddle. 2014. *Persuasive Technology: 9th International Conference, PERSUASIVE 2014, Padua, Italy, May 21-23, 2014. Proceedings.* Springer International Publishing, Cham, Chapter Stop Clicking on "Update Later": Persuading Users They Need Up-to-Date Antivirus Protection, 302–322. `DOI:http://dx.doi.org/10.1007/978-3-319-07127-5_27`

[239] L. Zhang-Kennedy, S. Chiasson, and R. Biddle. 2016. The Role of Instructional Design in Persuasion: A. Comics Approach for Improving Cybersecurity. *Int. J. Hum. Comput. Interaction* 32, 3 (2016), 215–257.