A Comprehensive Quality Evaluation of Security and Privacy Advice on the Web

Elissa M. Redmiles, Noel Warford, Amritha Jayanti, and Aravind Koneru, *University of Maryland;* Sean Kross, *University of California, San Diego;* Miraida Morales, *Rutgers University;* Rock Stevens and Michelle L. Mazurek, *University of Maryland*

https://www.usenix.org/conference/usenixsecurity20/presentation/redmiles

This paper is included in the Proceedings of the 29th USENIX Security Symposium.

August 12–14, 2020

978-1-939133-17-5

Open access to the Proceedings of the 29th USENIX Security Symposium is sponsored by USENIX.

A Comprehensive Quality Evaluation of Security and Privacy Advice on the Web

Elissa M. Redmiles[1], *Noel Warford*[1], *Amritha Jayanti*[1], *Aravind Koneru*[1], *Sean Kross*[2], *Miraida Morales*[3], *Rock Stevens*[1], *Michelle L. Mazurek*[1]

[1]*University of Maryland*
[2]*University of California San Diego*
[3]*Rutgers University*

Abstract

End users learn defensive security behaviors from a variety of channels, including a plethora of security advice given in online articles. A great deal of effort is devoted to getting users to follow this advice. Surprisingly then, little is known about the *quality* of this advice: Is it comprehensible? Is it actionable? Is it effective? To answer these questions, we first conduct a large-scale, user-driven measurement study to identify 374 unique recommended behaviors contained within 1,264 documents of online security and privacy advice. Second, we develop and validate measurement approaches for evaluating the *quality* – comprehensibility, perceived actionability, and perceived efficacy – of security advice. Third, we deploy these measurement approaches to evaluate the 374 unique pieces of security advice in a user-study with 1,586 users and 41 professional security experts. Our results suggest a crisis of advice prioritization. The majority of advice is perceived by the most users to be at least somewhat actionable, and somewhat comprehensible. Yet, both users and experts struggle to *prioritize* this advice. For example, experts perceive 89% of the hundreds of studied behaviors as being effective, and identify 118 of them as being among the "top 5" things users should do, leaving end-users on their own to prioritize and take action to protect themselves.

1 Introduction

It is often considered ideal to remove end users from the security loop, reducing both their burden and the chance of potentially harmful errors [12]. However, removing the user entirely has proven difficult, if not impossible. Users are still responsible for protecting themselves in a variety of situations, from choosing and protecting passwords, to recognizing phishing emails, to applying software updates, and many more.

Researchers and practitioners have spent significant time and effort encouraging users to adopt protective behaviors. Examples include redesigning warnings to make them harder to ignore [7,8,15,61,62], testing scores of alternative authentication methods intended to reduce user burden [5], "nudging" users toward better behavior [1, 17], and even using unicorns to promote secure authentication in encrypted messaging [64]. Despite all this encouragement, user adoption of protective behaviors remains inconsistent at best [43, 54, 67].

If we wish to improve overall outcomes, it is insufficient to consider protective behaviors independently from each other; we must instead consider the cumulative ecosystem of security-behavior messaging and its effect on users. For example, there are limits to how much time and effort users can spend on protective behaviors [3], and some protective behaviors may require more effort than they are worth [23, 47]. Further, recommended behaviors are sometimes conflicting [10, 26, 50], change over time (e.g., from changing passwords frequently to limiting password changes except in cases of breach [9, 20], and (as with any topic on which people provide advice to others) there is likely to be significant misinformation available.

It is critical, therefore, to understand where users get their security information, and what they are learning. Previously, researchers have identified several key sources of security information and advice: friends and family, fictional media, device prompts, and of course, the web [18,41,43,46]. However, the content of this advice has remained largely unexamined.

We make three primary contributions:

1. We create the first comprehensive taxonomy of end-user-focused security and privacy advice. To do so, we scraped 1,264 documents of security advice from the web, identified based on user-generated search queries from 50 users and via recommendations from vetted expert. We then manually annotated 2,780 specific pieces of advice contained in these 1,264 documents, ultimately identifying 374 unique advice imperatives, 204 of which were documented for the first time in this work [4, 10, 11, 26, 35, 50].

2. We develop measurement approaches for and validate a novel set of advice quality metrics: perceived actionabil-

ity, perceived efficacy, and comprehensibility. We show that these metrics correlate with the ultimate goal of security advice: end-user adoption of secure behaviors.

3. We conduct a study with 1,586 users and 41 professional security experts to evaluate the quality of the current body of security advice: we evaluate all 374 advice imperatives along these quality axes, examining the relative quality of different topics (e.g., passwords vs. privacy) and advice-givers (e.g., the government vs. popular media), identifying areas needing improvement.

Our results suggest the key challenge is not in the *quality* of security advice, but in the *volume* and *prioritization* of that advice. While users find the majority of the 374 advice imperatives they evaluate fairly actionable and somewhat comprehensible, they struggle to identify which advice is most important, listing 146 pieces of advice as being among the top 3 things they should attempt to do. Yet, we know that they do not adopt anywhere near this many protective behaviors [43, 56, 67, 68], nor would doing so be practical [3].

We find little evidence that experts are any better off than end-users on the subject of security advice: experts identify 118 pieces of security advice as being among the top 5 things they would recommend to a user, consider 89% of the 374 pieces of advice to be useful, and struggle with internal consistency and alignment with the latest guidelines (for example, claiming that failing to change passwords is harmful, despite the latest NIST advice to the contrary). Thus, users – whose priority ratings of advice have little to no correlation with expert priority ratings – are left to fend for themselves, navigating through a sea of reasonably well-crafted but poorly organized advice. These findings suggest that the path forward for security advice is one of data-driven measurement, minimality and practicality: experts should rigorously measure the impact of suggested behaviors on users' risk and ruthlessly identify only the minimal set of highest impact, most practical advice to recommend.

2 Related Work

In this section, we review related work on security education and advice, as well as measurement of text quality.

Security education and advice. Users receive security advice from a variety of different sources, including from websites, TV, and peers, depending on their level of expertise, access to resources, and demographics [18, 36, 43, 45, 46]. People also learn from negative experiences — their own and others' — through stories about security incidents [41]. The negative experiences that inform these stories are effective but carry undesirable emotional and practical costs. Some researchers have thus explored comic strips and interactive approaches as effective means of teaching security lessons [13, 29, 34, 53, 57, 71]; others have used visual media to teach security [2, 19].

Rader and Wash [40] found that the types of security information users encounter depends strongly on the source, with websites seeking to impart information from organizations, news articles focusing on large breaches or attacks, and interpersonal stories addressing who is hacking whom and why. While there are many sources of security *information*, prior work has shown that websites are one of the most common sources of *advice* specifically [43]. We therefore aim to characterize advice that is available on the Internet. Rather than use topic modeling, as in prior work [40], we manually coded each document we collected in order to deeply understand the online security advice ecosystem.

In addition to studying where and how people get security advice, researchers have studied what is in that advice. Ion et al. [26, 50] found that experts and non-experts consider different practices to be most important; Busse et al. replicated this work in 2019 and found this was still true [10]. Reeder et al. [50] additionally report on advice imperatives provided by security experts. We leverage this work as a starting point for our taxonomy, while also examining what users might find by directly seeking security advice.

Prioritizing advice is important, because people and organizations have a limited "compliance budget" with which to implement security practices [3]. It has been shown that users make time-benefit tradeoffs when choosing a security behavior [47], and may find it irrational to follow all, or even most, security advice [23]. Further, advice can be difficult to retract once disseminated, creating a continuously increasing burden for users and organizations [24, 25].

Text evaluation. There are many ways to define and measure text quality. Louis and Nenkova [32], for example, investigate the quality of science journalism articles using both general measures, like grammar or spelling correctness, and domain-specific measures, like the presence of narrative. Tan et al. define quality using linguistic features — like Jaccard similarity, number of words, and number of first person pronouns — of successfully persuasive arguments on Reddit [63].

Perhaps the most common measure of text quality is *comprehensibility*: how easy or difficult it is for people to comprehend a document. Prior work has considered the comprehensibility of three types of security- and privacy-relevant text: privacy policies [33, 60], warning messages [21], and data breaches [72]. These investigations have shown that security and privacy content is often difficult to read, and that problems of readability may also be compounded by other factors such as the display constraints of mobile devices [60]. In this work, we consider a broader class of security-relevant documents — security advice from the web — and we apply multiple measures of quality along three axes: comprehensibility, actionability, and accuracy. There are a number of different mechanisms for measuring the comprehensibility of

adult texts. Redmiles et al. [49] evaluate the validity of these different mechanisms. We leverage their proposed decision strategy and tools for our measurements (see Section 4.4 for more detail).

3 Identifying Security Advice

We used two approaches to collect text-based security advice aimed at end users: (1) We collected search queries for security advice from 50 crowdworkers and scraped the top 20 articles surfaced by Google for each query, and (2) we collected a list of authoritative security-advice sources from computer security experts and librarians and scraped articles accordingly.

User search query generation. We recruited 50 participants from Amazon Mechanical Turk (AMT) 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 (See Appendix A), 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, 2.5 min completion time) or $0.50 (second survey, 4 min completion time). Our protocol was approved by the University of Maryland IRB.

From these surveys, we collected 140 security-advice search queries. After manual cleaning to remove 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 used the Diffbot API[1] to scrape and parse the top twenty Google search results for these queries[2]. Our collection was conducted in September 2017.

Expert advice recommendations. To identify the types of articles users might be referred to if they asked an authority figure for advice, 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.

Initial corpus & cleaning. The resulting corpus contained 1,896 documents. Examples include Apple and Facebook help pages, news articles from Guardian and the New York Times, advice or sales material from McAfee, Avast, or Norton, U.S. CERT pages, FBI articles, and articles from Bruce Schneier's blog. To ensure that all of the documents in our corpus actually pertained to online security and privacy, we recruited CrowdFlower crowdworkers[3] 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.' When two of the three initial workers answered 'Yes,' we recruited an additional two workers to review the document, retaining documents for which four of the five workers answered 'Yes.' After this cleaning, 1,264 of the initial 1,896 documents were retained in our corpus.

Extracting & evaluating advice imperatives. Next, we decomposed these documents into specific advice imperatives (e.g., "Use a password manager"). Two members of the research team manually annotated each of the 1,264 documents in our corpus to extract the advice imperatives contained within them.

We constructed an initial taxonomy of advice imperatives based on prior work that had identified user security behaviors [4, 11, 26, 35]. We manually reviewed each of these articles, made a list of all described behaviors, and reached out to the article authors to ask for any additional behaviors not reported in the papers. The authors of [26] shared their codebook with us. After merging duplicates, our initial list contained 196 individual advice imperatives. 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, adding to the taxonomy as needed.We reached a Krippendorff's alpha agreement of 0.69 (96.36% agreement) across the 12 high-level code categories, which is classified as substantial agreement [30]. Given this substantial agreement, and the large time burden of double annotating all 1,264 documents, the researchers proceeded to independently code the remaining documents. To evaluate the consistency of our in-

[1] https://www.diffbot.com/

[2] Diffbot uses a variety of approaches to maximize stability of search results and minimize personalization impact.

[3] We use CrowdFlower instead of AMT because the CrowdFlower platform is designed to allow for the validation of work quality for Natural Language Processing text cleaning processes like this one, and the workers are more used to such tasks.

dependent annotations, we compute the intraclass correlation (ICC), a commonly used statistical metric [59] 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 [27].

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 targeting end users: 204 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 document corpus, set of advice imperatives, together with linked evaluation metrics, can be found here: https://securityadvice.cs.umd.edu.

As part of this 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 (news reports, help pages for specific software, etc.) containing no actionable advice. To maintain our focus on end-user advice, we also discarded imperatives targeting, e.g., system administrators or corporate IT departments. This resulted in a final corpus of 614 documents containing security advice.

It is important to note that we use manual annotation to analyze this data because (a) we cannot use supervised automated classification, as there exists at present no labeled training data from which to build a classifier (this work establishes such labeled data) and (b) unsupervised modeling of advice "topics" and automated tagging of non-standardized open text with those topics, with a very large number of possible classes as in our case, remains an open, unsolved problem [31].

Twelve topics of security advice from 476 unique web domains. Our annotation process identified 374 security advice imperatives relevant to end-users. These pieces of advice occurred 2780 times overall, with an average of 4.53 imperatives per document. We categorized these pieces of advice into 12 high-level topics, which are summarized in Table 1. Figure 1 (left) shows the distribution of topics across the documents in our corpus. We identified 476 unique web domains in our corpus; we manually grouped these domains into broader categories, while retaining certain specific, high-frequency domain owners of interest, such as Google and the Electronic Frontier Foundation (EFF). Hereafter, we use "domain" to refer to these groupings. Figure 1 (right) shows the distribution of domains in our corpus.

4 Evaluating Security Advice

After identifying and categorizing the broad set of security advice being offered to users, we next evaluated its quality. Specifically, we measure the perceived actionability and perceived efficacy of the imperatives, as well as the comprehensibility of the documents. Below we describe our measurement approach, including the novel metrics we developed, the user study (1,586 users and 41 professional security experts) we conducted to instantiate these metrics, and our assessment of the metrics' validity.

4.1 Measurement Approach

Perceived actionability. We assess perceived actionability by asking users from the general population to report how hard they think it would be to put a given imperative into practice. In particular, our actionability questionnaire incorporates 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.

each evaluated on a 4-point Likert scale from "Not at All" to "Very." The full questionnaire, which included an example to help respondents distinguish among the different sub-metrics, is included in Appendix C. Each imperative was evaluated by three respondents, and each respondent evaluated five randomly drawn imperatives.

These four sub-metrics align with theoretical foundations relevant to security behavior. The confidence sub-metric is drawn from Protection Motivation Theory [52], which identifies perceived ability to protect oneself as a key component of protective behavior implementation, and from the Human in the Loop model [12], 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 [3, 23, 47, 48]. Finally, the difficulty sub-metric is used to align with the capabilities component of the Human in the Loop model [12].

Perceived efficacy. We also use human-generated data to measure the perceived efficacy of the advice imperatives. We asked professional security experts (see qualification criteria below) to answer an evaluation questionnaire for each piece of security advice. Each advice imperative was again evaluated by three respondents. The efficacy questionnaire evaluated, for each advice imperative, *Perceived efficacy:* whether the expert believed that a typical end user following this advice would experience an improvement in, no effect on, or harm to their

Topic	Examples
Account Security	Identify compromise on your social media account, Avoid spam in your email account, don't sign up for "unnecessary" accounts (this does *not* include mechanisms for authentication, e.g., passwords, 2FA)
Browsers	Clear browser history, only download things you are looking for, verify website signatures and certificates
Data Storage	Keep sensitive information on removable storage media, use backups and SSDs, encrypt data
Device Security	Cover your webcam, keep your devices with you, lock your smartphone
Finance	Do online banking only on certain devices, use secure payment methods, type banking links manually
General Security	Seek out expert help, avoid overconfidence online, use parental controls for children
Incident Response	Cancel or change accounts, report suspicious incidents to IT/support, document the incident
Network Security	Use a password to protect your wifi, change your router name from the default, turn off Bluetooth, how to set up firewalls
Passwords	Use strong passwords (including specific imperatives regarding how to construct such a password), use unique passwords, how to store passwords, use 2FA
Privacy	Use Tor, read privacy policies, and act anonymously online
Software	Update applications, only install trusted software, remove unnecessary programs

Table 1: The 12 categories of security advice we identified.

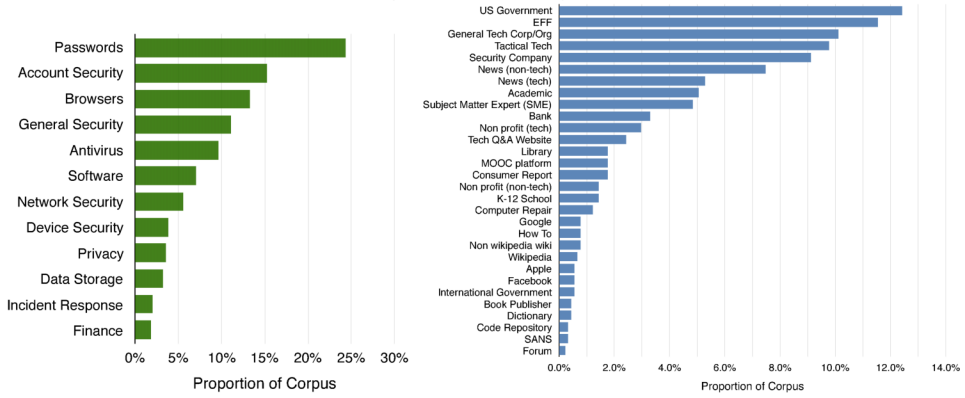

Figure 1: Distribution of topics (left) and domain categories (right) across the corpus.

security. For advice the expert thought would reduce security risk, the expert also provided *Perceived risk reduction:* how much the expert estimated risk would be reduced if the advice was followed (numerically, on a scale from 0% to more than 50%)[4]; and *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)[5]. Finally, for advice the expert thought would increase security risk, we measured *perceived risk increase:* how much the expert estimated risk would be increased if the advice was followed. Each survey contained 10 randomly selected imperatives. The full questionnaire is included in Appendix D.

Comprehensibility. Human-written comprehension questions are often considered the most effective method for measuring comprehensibility, but are difficult to scale [16]. As an alternative, prior work on security-relevant texts have used computed measures such as the Flesch Reading Ease Score (FRES) [16], which can be calculated directly from a document without human input. However, measures like FRES are based on features like syllables and word length, which may not be appropriate for technical words (e.g., 'bot' may be more difficult to comprehend than 'bat,' but this distinction would not be observed in a computed measure). As an alternative, natural language processing and linguistics researchers often construct automatically-generated comprehension tests, answered by human readers. The most commonly used procedure for constructing such tests is the Cloze procedure [65, 66]. Following guidance from the literature regarding domain-specific texts such as ours [49], we use two comprehensibility metrics: comprehension tests generated using the *Smart Cloze* variant of the Cloze procedure and perceived ease assessments.

Smart Cloze replaces every 5th word of a document with a blank. The test-taker fills in the blank from provided multiple-choice options. Unlike other Cloze variants [6, 22, 38, 42, 65], Smart Cloze has been validated on domain-specific documents, including security documents [49]. Smart Cloze scores (hereafter, Cloze) are computed as the percentage of blanks filled in correctly.

To complement the Smart Cloze scores, we use perceived ease. To measure perceived ease, we use a single item [51, 55]: "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). This metric is more effective than Smart Cloze for measuring "first-glance" perception of difficulty, which can affect whether or not the reader will continue to read the entire document [39, 49]. Each document was evaluated by three respondents for each metric, and each respondent evaluated

[4] While it would be ideal to measure risk reduction quantitatively – see more discussion in Section 10 – we measured experts perception of risk reduction as a first cut toward quantitatively evaluating efficacy.

[5] We used this assessment mechanism to get relative, rather than explicit, prioritization and that we used this phrasing to stay in line with prior work [10, 26] asking experts to prioritize advice so that we could compare our results.

four randomly drawn documents.

4.2 Human Subjects Recruitment

For measurements conducted with the general population (measurements of actionability and comprehensibility), we recruited users from the survey research firm Cint's[6] survey panel, which allows for the use of quota sampling to ensure our respondents' demographics were representative of the U.S. population within 5% on age, gender, race, education and income. We recruited a total of 1,586 users in June 2019 to evaluate the actionability and comprehensibility of our security advice. Participants were compensated in accordance with their agreement with Cint.

The efficacy measurements were conducted with professional security experts. We recruited experts during May and June 2019. We did so by tweeting from the lead author's Twitter account, asking well-known security Twitter accounts to retweet, and leveraging our personal networks. We also posted in multiple professional LinkedIn groups and contacted authors of security blogs. All recruited individuals completed a screening questionnaire to assess 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, and 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 practitioners; only three were academics. Our experts have diverse workplace contexts: engineer through director-level information security professionals for large corporations and government agencies, red team/pen testers, independent security consultants, and privacy-focused professionals at large and well-known non-profit/advocacy organizations. 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 batches as desired and were able to skip previously-evaluated pieces of advice. On average, experts evaluated 38 pieces of advice each.

4.3 Measurement Validity

We evaluate the validity of our measurements in two ways: (1) we check the reliability of ratings provided by our user and expert evaluators, again using the ICC metric (see Section 3) and (2) we examine whether these quality measures are discriminant, whether they correlate with behavior adoption (the ultimate goal of security advice), and, where possible, whether we reproduce results of prior work on security advice. We report on (1) here and point the reader to Section 9 for the results of (2).

Overall, all of our evaluators achieved at least "good" reliability in evaluating our three metrics of advice quality [27]. For actionability, reliability was "very good": ICC = 0.896, 0.854, 0.868, and 0.868 for confidence, time consumption, disruption, and difficulty, respectively. For efficacy, the experts achieved "very good" reliability, with an ICC of 0.876, and for comprehensibility, our Cloze raters had "excellent" reliability (ICC=0.989), while our ease raters achieved "good" reliability (ICC = 0.757).

4.4 Limitations

As with all measurement and user studies, our work has certain inherent limitations. First, it is possible that our security advice corpus is not fully representative of the ecosystem of security advice. We used multiple techniques — soliciting advice recommendations from experts, two methods for collecting search queries from users — to ensure broad coverage of advice in order to mitigate this potential limitation. Second, it is possible that our manual annotation process was inaccurate. We conducted a double annotation of over 10% of our documents, achieving "sufficient" inter-annotator agreement before proceeding to annotate independently, to mitigate this risk; further, both coders conducted a full review of each other's taxonomies once annotation was finished, and reached a final, cohesive taxonomy that was applied to all documents. Third, we cannot capture all possible types of relevant expertise. To minimize data collection, we screened our experts for expertise but explicitly did not collect demographic data; examining how experts' sociodemographic backgrounds may affect how experts prioritize advice may be an exciting direction for future work.

Fourth, due to the volume of advice, experts and users evaluated advice in the abstract and did not evaluate all advice. We find, through a X^2 proportion test, that there is not a statistically significant difference between the priority ratings of the 26 experts who rated less than 30 pieces of advice and the 15 who rated more advice; however, lack of a full sense of the dataset may still have affected prioritization.

Fifth, our instantiations of our metrics may not provide a full picture of the comprehensibility, perceived efficacy, and perceived actionability of our documents. To mitigate this limitation, we relied upon established, validated tools from library science and NLP [49,65] and constructed questionnaires that we robustly pre-tested using techniques such as cognitive interviewing, following survey methodology best practice for mitigating self-report biases such as social desirability [44].

[6] https://www.cint.com/reach-survey-respondents/

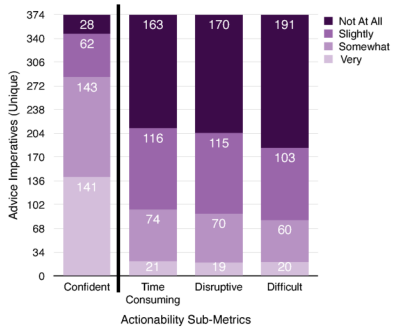

Figure 2: User ratings of actionability.

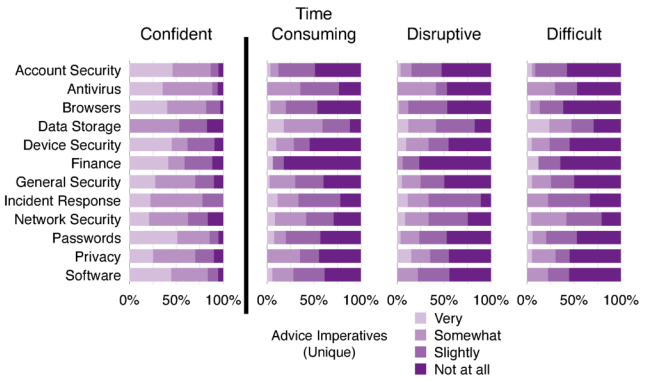

Figure 3: Advice actionability by topic across 374 unique advice imperatives.

Future work may wish to consider evaluating these metrics even more comprehensively, for example evaluating whether users feel that the advice is effective. As with any new metric, repeated validation in multiple contexts is necessary to firmly establish validity. While self report biases may indeed remain, recent work on survey methodology specifically for security topics suggests that self-report data provides a reasonable, if not entirely precise, picture of real security behavior [48].

5 Perceived Actionability

In this section, we summarize users' perceptions of the actionability of the 374 pieces of advice in our corpus.[7]

5.1 Overall Actionability

The majority of advice is perceived as fairly actionable. Overall, the advice imperatives in our corpus had a median rating of "somewhat" confident the rater could implement the advice, as well as median ratings of "slightly" time consuming, "slightly" disruptive, and "not at all" difficult. The distribution of actionability ratings across advice is shown in Figure 2. People were "very" (37.7%) or "somewhat" (38.2%) confident about implementing around three quarters of the advice imperatives, with only 7.5% of advice receiving a median rating of "not at all" confident. Further, around half of the imperatives were considered "not at all" time consuming (43.6%), "not at all" disruptive (45.5%), and "not at all" difficult (51.1%).

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 imperatives were rated "very" time consuming, disruptive, and difficult, respectively; 28 pieces of advice received a median rating of "not at all" confident. This totals 49 imperatives (13.1%) that were unactionable on at least one metric, and 25 (6.7%) that were unactionable on two or more metrics. This unactionable advice is summarized in Table 2.

[7] In the interest of space, in the remainder of the paper we refer to perceived actionability, as judged by our survey respondents, as "actionability."

5.2 Actionability by Frequency

In addition to considering individual imperatives, we consider actionability across the frequency of advice (2780 pieces total) collected in our corpus. This allows us to consider the extent to which the most commonly presented advice is actionable.

Most advice in the corpus is fairly actionable. We found that 81.3%, 81.0%, and 84.0% of the 374 imperatives were considered at most "slightly" time consuming, disruptive, and difficult to implement, respectively. Participants were "somewhat" or "very" confident about implementing 86.2% of the advice. Further, the 49 pieces of advice we consider unactionable made up only 3.6% of the advice in the corpus by frequency, despite making up 13.1% of the unique advice imperatives. The less actionable advice, however, was spread throughout the corpus. 20.5% of documents contained at least one piece of advice we consider unactionable.

5.3 Actionability by Topic

Each of the four sub-metrics of actionability differ significantly by topic overall, 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 3. Pairwise comparison tables, including specific topic differences that were significant after correction — which are detailed below — can be found in Appendix F.

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

The most-actionable advice is about account security, finance, and passwords. People were at least "somewhat" confident they could implement more than 80% of the advice about account security, and people found more than 80% of the advice about account security at most "slightly" time consuming and "slightly" difficult to implement. Advice about fi-

Advice	Not Confident	Very Time Consuming	Very Disruptive	Very Difficult	Efficacy	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 2: List of the most unactionable advice based on user ratings. The first four columns indicate advice with median rating of "not at all" confident and"very" time consuming, disruptive, and/or difficult. The fifth column indicates expert-perceived efficacy and the sixth column provides expert-estimated median risk reduction for efficacious advice (negative for harmful advice).

nance was also considered quite actionable: 94.1% of finance advice was perceived as at most "slightly" time-consuming or disruptive to implement, and more than 80% of this advice was perceived as at most "slightly" difficult to implement. Advice about passwords scored well on two of the four actionability submetrics: for more than 80% of passwords advice people were at least "somewhat" confident they could implement it and perceived it as at most "slightly" difficult to implement.

The least-actionable advice is about data storage and network security. The topic with the highest proportion of poor (lowest two ratings on Likert scale) actionability ratings, across all four metrics, was data storage. More than half the data storage imperatives received confidence responses of "slightly" or "not at all," there was no advice about data storage for which people were "very" confident. Similarly, 58.8%, 41.2%, and 47.1% of the imperatives about data storage were rated at least "somewhat" time consuming, disruptive, and difficult to implement, respectively. Advice about network security performed nearly as badly on three of the four actionability submetrics; participants were confident they could implement barely half the advice about network security, and they perceived at least 40% of network security advice as "very" time consuming or difficult to implement.

Privacy advice polarizing in perceived actionability. It is additionally interesting to note that the actionability ratings for privacy advice were quite split. Near-equal proportions of privacy advice were rated as at least "somewhat" time

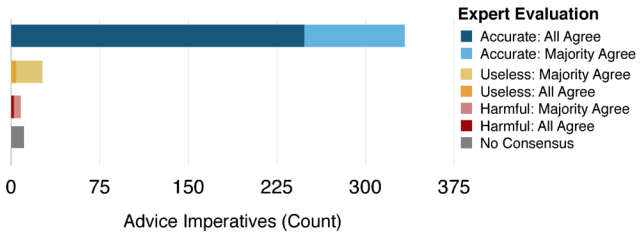

Figure 4: Expert ratings of advice efficacy.

consuming (35%) and "not at all" time consuming (45%). Similarly, while data storage had the highest proportion of advice considered "very" or "somewhat" disruptive, privacy had a higher proportion of advice considered "very" disruptive to implement than any other topic (15%) and no advice that was considered "somewhat" disruptive to implement; on the other hand, 45% of the advice about privacy was considered "not at all" disruptive. Advice about browsers, too, exhibited interesting patterns, although it was not as starkly polarized as privacy advice. A majority of advice about browsers was rated "not at all" difficult to implement, but a (partially overlapping) majority of this advice was also rated as at least "slightly" disruptive.

5.4 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: $p = 0.873$). There was also no significant difference when considering just unactionable advice by domain ($p = 0.684$).

6 Perceived Efficacy

In this section we summarize experts' perceptions of the efficacy[8] of the 374 advice imperatives we identified in our corpus. We caution the reader that this measurement captures experts' perceptions: often the best information available, but not an objective *measurement* of the true efficacy of these behaviors against security threats.

6.1 Overall Efficacy

Almost all advice perceived as effective, almost none as harmful. Overall, 248 imperatives (66.3%) were rated as effective by all three experts who evaluated them, 85 imperatives (22.7%) were identified as effective by two of three experts, and 31 imperatives (8.3%) were identified as effective by one expert. Experts reported that following these imperatives would lead to a median 37.5% reduction in users' security risk.

[8]For space, in the remainder of the paper we refer to perceived efficacy, as judged by our expert evaluators, as "efficacy."

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," and "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. All 26 pieces of advice identified as useless by the majority (2 of 3) of experts are listed at `securityadvice.cs.umd.edu`.

Two imperatives (0.5%) were considered 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: basing passwords on upcoming events, creating a new email address when compromised, using tracking applications (to monitor your online activity), using different personas online, storing 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 **??**). Finally, eight pieces of advice had no consensus (each expert gave a different evaluation). Figure 4 summarizes these results.

6.2 Efficacy by Frequency

Of the 2,780 instances of advice imperatives throughout our corpus, only 3% are perceived by experts as harmful or useless. By frequency, 95.8% of the advice provided was considered effective by a majority of experts (77.7% was considered effective by all experts). 2.5% of advice was considered useless by a majority of experts, and 0.5% was considered harmful. Thus, even though 8.8% of imperatives were perceived as useless or harmful by a majority of experts, these imperatives make up only 3.0% of the advice in our corpus.

Every document contained at least one advice imperative rated as effective by a majority of experts. Further, 82.7% of documents contained exclusively advice evaluated as effective; on average, these documents contained 4.7 pieces of advice. In contrast, documents that contained at least one piece of useless, harmful, or no-consensus advice contained an average of 11.2 pieces of advice, of which an average of 9.8 pieces (83.3%) were perceived as effective.

About 10% of documents in the corpus contained at least one of the 26 imperatives our experts identified as useless. On average, these documents contained 1.4 pieces of useless advice. Also, 6.1% of documents contained one of the eight pieces of advice about which there was no consensus, these documents contained an average of 1.2 pieces of no consensus advice. Finally, 2.8% of of documents contained harmful

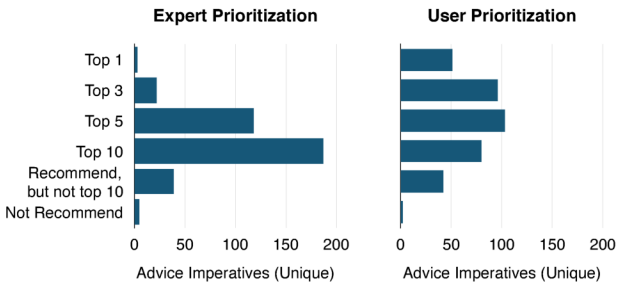

Figure 5: Advice priority rating: experts (left), users (right).

advice; on average, 1.1 such imperatives.

6.3 Efficacy by Topic and by Domain

Perhaps unsurprisingly given that the vast majority of advice is considered effective by a majority of experts, efficacy does not vary significantly by topic ($p = 0.245$, Kruskal-Wallis test) or by domain ($p = 0.958$, KW test). There is also not a significant difference in the median risk reduction of the effective advice across different domains ($p = 0.210$, KW test), nor about different topics ($p = 0.312$, KW test).

6.4 Advice Priority

Twenty-five imperatives considered high priority. Experts considered twenty-five pieces of advice (6.7%) as being among the "top 3 behaviors I would recommend" (see `securityadvice.cs.umd.edu` for the full list). Almost a third of all imperatives (31.6%, 118) were rated as being in the "top 5" behaviors experts would recommend. Of the remaining 231 pieces of advice, 187 (half of all advice) were rated as being among the "top 10" behaviors experts would recommend (Figure 5).

In addition to consulting experts' priority ratings, 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 [28, 58] used to determine a full ranking of items when individual users have provided ratings on only a small portion.[9]

This approach identified the following pieces of advice as the top 10: using unique passwords, updating devices, using anti-malware software, scanning email attachments for viruses, encouraging others to use strong passwords, not telling anyone your passwords, using end-to-end encryption, remembering your passwords, and keeping passwords that are written down safe.

Passwords, antivirus, finance, network, and software-security advice given highest priority. Priority differs significantly by topic ($p = 0.044$, Kruskal-Wallis test) but not by domain ($p = 0.089$); Table 3 in Appendix F shows the pairwise (Mann-Whitney, Holm-Bonferonni corrected) comparisons among the topics. Advice about passwords is given especially high priority, along with advice about antivirus, finance, network security, and software security. More detail can be found at `securityadvice.cs.umd.edu`.

Users struggle to discern priority; 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, collected from the perceived actionability survey. Users identified 51 pieces of advice (13.5% of imperatives) as the "number one" behavior they should follow, and an additional 96 (25.6%) pieces of advice as among the "top 3" behaviors they should follow. This high proportion of advice perceived as being in the "top 3" suggests that users have an even more difficult time differentiating the importance of advice than do experts, as illustrated in Figure 5.

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 ($tau = 0.175$).[10] This result aligns with the findings of Ion et al., who considered a smaller sample (20 pieces) of advice [26].

7 Adoption

We also asked both our expert and general-population advice-evaluators to report whether or not they themselves followed each imperative "at least some of the time".[11] We also used cushioning language to reassure respondents that their answer would not affect their survey compensation or qualification (see Appendices D and C for full questionnaires).

Despite these efforts, self-reports of behavior likely still contain biases from true behavior. Recent work has shown self-reports in the security context to correlate with true behavior, providing insights correct in direction and approximate effect size but off in precise magnitude [48, 69]. As such, the results below should be taken as an indication of how users strive to behave, and a rough approximation of how they actually behave, but not as a precise estimate of daily behavior.

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 reported followed by none of the evaluating experts; by frequency, this represents less than 1% of the advice in the corpus. At a document level, 90.5% of documents have at least one piece of advice that evaluating experts reported

[9] We performed matrix factorization with 1,000 iterations, $alpha = 0.0002$, and $beta = 0.02$, following best practices [37].

[10] 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.

[11] Wording selected after several rounds of cognitive testing [70] with both experts and non-experts.

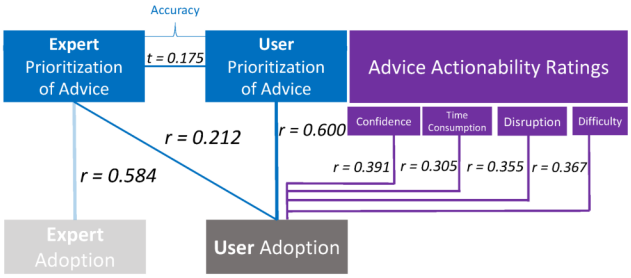

Figure 6: Correlation between security advice adoption, actionability, and priority rankings.

following at least some of the time, and only 2.1% of documents contain any piece of advice that no experts reported following.

General population. All three evaluators claimed to adopt 34.4% of imperatives (129) at least some of the time; another 34.7% were reported as adopted by two of three raters. Only 31 imperatives (8.27% of advice) were reported followed by no respondents.

By frequency, only 3.2% of the advice in the corpus was reported as unfollowed by all evaluators, and 15.4.% was reported followed by only one evaluator. The remaining 81.4% was reported followed by the majority of evaluators (48.0% by all three). We find that 13.5% of documents contain at least one piece of advice not followed by any respondent, while 72.4% of documents contain at least one piece of advice that all three evaluators reported following.

Harmful and useless advice. Of the 40 pieces of advice that were rated harmful or useless by the majority of experts, or which had 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. Among general users, 14 of 40 such imperatives were followed by one respondent, eight were followed by two respondents, and 11 were followed by all three respondents.

Actionability and adoption. There is a significant (Pearson) correlation between actionability and general users' reported adoption of 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 reported as adopted by general users. Actionability correlates significantly, but with much smaller effect size, with adoption by experts only for user-rated confidence ($r = 0.110, p = 0.034$) and difficulty ($r = 0.124$, $p = 0.017$).

Priority and adoption. Finally, expert and user priority rankings also correlate with their respective adoption of the advice (users: $r = 0.600$, $p < 0.001$; experts: $r = 0.584$, $p < 0.001$). While experts' and users' reported behavior correlates with their own priority rankings, users' adoption of behaviors correlates only weakly with experts' prioritization of advice ($p < 0.001$, $r = 0.212$). That is, user behavior does not follow experts' preferred pattern. This is perhaps unsurprising given that users' and experts' priority rankings barely correlate ($p < 0.001, tau = 0.175$), Figure 6 summarizes the correlations among adoption, priority, and actionability.

8 Comprehensibility

Our final results address the comprehensibility of the documents in which the security advice imperatives evaluated in the prior sections were contained.

8.1 Overall Comprehensibility

We measure the comprehensibility of our documents via Smart Cloze scores and ease perception measurements (see Section 4.4 for more detail). Cloze scores below 50% are interpreted as low comprehension, 50-60% indicate partial but incomplete comprehension, and scores above 60% indicate sufficient comprehension [14].

Wide variance in comprehensibility; on average, partially comprehensible to the general public. Overall, our documents had an average Cloze score of 47.9% (median 51.4%, s.d. 17.9%), equivalent to low-to-partial comprehension. About a quarter of documents (27%) were rated as partially comprehensible (Cloze > 50%); another 28% were rated as fully comprehensible (Cloze > 60%). Figure 7 summarizes the overall Cloze results. People's perception of the ease of reading the documents, as compared to their performance on the Cloze tests, were a bit more positive. The mean document in the corpus was perceived as "somewhat easy" to read. However, we observe very high variance in this metric: the standard deviation of 1.05 is a little more than one step on the Likert scale.

8.2 Comprehensibility by Topic

The most comprehensible documents contain advice about account security, browsers, data storage, device security, and finance. There is a significant difference in the Cloze scores for security documents containing advice about different topics ($p < 0.001$ for all topics, ANOVA; all pairwise tests remain significant after Holm-Bonferonni correction[12]), with the mean Cloze scores by topic varying from lowest to highest by 13.1 percentage points. Figure 8 summarizes the differences in Cloze scores by topic. Documents containing advice about account security, browsers, data storage, device

[12] We omit the p-value tables for the pairwise comparisons here and in the following subsection, as all p-values were below 0.05, even after correction.

security, and finance achieved at least partial comprehension on average (Cloze scores, mean across the topic, above 50%). Finance-related documents had particularly low variance in scores, with a standard deviation of 6.22%.

The remaining topics 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, with very wide score spreads. Passwords was the most popular topic in the corpus and also had the highest standard deviation in Cloze scores; we therefore hypothesize that the low scores may be at least partially about quantity. On the other hand, network security is a particularly technical topic, so the low scores may relate to additional complexity or jargon.

There was no significant difference in reading ease perceptions among different topics ($p = 0.999$, Kruskal-Wallis[13]).

8.3 Comprehensibility by Domain

The most comprehensible sources are general news channels, subject-matter experts (SMEs), non-profits, and security and computer-repair companies. To understand whether some advice-givers provided more readable advice than others, we examined Cloze scores grouped by domain. Figure 9 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). Of the 30 domain groups we considered, seven scored above 50% (mean across documents): SMEs, general news outlets, how-to websites, non-tech-focused and tech-focused nonprofit organizations, security companies, and computer-repair companies. Within particular categories, we see that some organizations perform better than others (Appendix B); we discuss the more notable cases of variance below. As with topics, there was not a significant difference in ease perception by domain ($p = 0.999$, Kruskal-Wallis).

Government organizations. Among U.S. government organizations, ic3.gov, whitehouse.gov, ftc.gov, and dhs.gov had average scores mapping to partial comprehension or better; the remaining domains perform worse. 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 partial comprehension 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. That said, content collected from school websites was not particularly readable, with mean Cloze scores indicating low comprehension, suggesting that schools may be better off obtaining content from child-focused nonprofit organizations.

Technical non-profits. Documents from non-profit organizations with technical focus had wider variance. Documents from the Tor Project, GNU, and techsoup.org had mean Cloze scores of at least partial comprehension. However, documents from nine other technical non-profits, including Mozilla, Chromium, and Ubuntu as well as organizations focused specifically on helping non-experts (e.g., libraryfreedomproject.org) had mean Cloze scores well below this threshold. Documents from the EFF and Tactical Tech-sponsored organizations also had mean Cloze scores mapping to low comprehension. This is important, as documents from these two organizations make up 21% of our corpus.

Corporations. Security-focused companies and those offering computer-repair services both scored very high on comprehensibility. We hypothesize that for these companies, which focus on lay users as customers, providing readable materials may be tantamount to a business requirement. On the other hand, non-security-focused companies — including some frequently under fire for privacy and security issues — scored poorly: mean Cloze scores for Google, Facebook, and Apple were 45.1%, 37.9%, and 41.7%, respectively.

Low-comprehension platforms. Finally, seven of the 30 advice-givers we examined provided particularly difficult to read advice (mean Cloze scores under 40%): SANS (sans.org), security forums (e.g., malwaretips.com, wilderssecurity.com), MOOC platforms (e.g., lynda.com, khanacademy.org), consumer rating sites (e.g., consumerreports.org, av-comparatives.org), Facebook, Technical Q&A websites (e.g., stackoverflow.com, stackexchange.com), and academic publications.

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 the average person, low readability 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.

9 Discussion

This work makes three primary contributions.

We create a taxonomy of 374 pieces of security advice. This work provides a comprehensive point-in-time taxonomy of 374 end-user security behaviors, including 204 pieces of security advice that were not previously catalogued in the literature. The full set of behaviors can be explored here: https://securityadvice.cs.umd.edu. This taxonomy provides (i) insight into the scope and quantity of advice received by users, (ii) a tool for researchers to consult when considering what security and privacy behaviors to study or analyze, and (iii) a mechanism for the broader security community to move forward with improving security advice by

[13] Ease scores were not normally distributed, so we use a non-parametric test; Cloze scores, on the other hand, were near-normal in a QQ plot and are thus evaluated with an ANOVA.

identifying advice in need of repair or retirement.

We develop and evaluate axes of security advice quality. Our approach to evaluating security advice is in itself a contribution: the axes of quality that we identify (comprehensibility, actionability, and efficacy) and the measurement approaches we designed to assess them can be applied to new advice that is created to ensure that as we move forward in advice-giving, we create higher-quality, more effective advice. Before we can recommend further use of these evaluation strategies, however, we must be convinced of their validity. Specifically, do the quality measurements correlate with behavior adoption (the ultimate goal of security advice), are the measurements discriminant, and are the measurements consistent with prior work (where applicable)? In an initial validation using the results of our work, we find that our metrics indeed 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, for both general users and experts.

We also find that our quality metrics are indeed 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 scoring consistently low on actionability. Similarly, documents containing advice about software security and antivirus were among the more difficult to read, but were not high in implementation difficulty, indicating that readability of the document containing the advice is different from the actionability of the advice itself.

Further, we examine whether we can replicate the results of prior studies in which security experts were asked to prioritize 20 pieces of security advice [10, 26, 50]. We find that our prioritization results replicate these quite closely. 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.

Of course, this preliminary validation connects these axes of advice quality to reported, rather than actual, behavior. Replication is necessary to fully validate any new metrics, and to examine how they perform in broader application (e.g., having both users and experts rate the efficacy of the advice).

We rigorously evaluate the comprehensibility, perceived efficacy, and perceived actionability of our corpus. By applying our metrics to the taxonomy we developed, we provide a thorough and novel characterization of the quality of the security-advice ecosystem. While prior work focused on expert and user prioritization of a small set of security advice (at most, 20 topics) [10, 26, 50], we evaluate a much larger set of advice and conduct a more comprehensive evaluation that considers not only prioritization, but also comprehensibility, perceived actionability, perceived efficacy, and how these factors interact. Further, our metrics allow us (differently from prior work) to characterize both generalized advice imperatives and specific wording within particular documents.

Overall, we find that security advice is perceived as fairly actionable — only 49 advice imperatives were rated by users as 'very' unactionable on one of our four metrics – as well as effective. The majority of security advice (89%) was perceived as effective by professional security experts.

Yet, we know that users do not adopt even a fraction of this advice consistently, despite their best intentions [43, 56, 67, 68]. This may be due in part to mis-comprehension of the instructions: the hundreds of documents we evaluate exhibit only low to partial comprehensibility for the general public. A larger factor, however, appears to be a crisis of advice prioritization. The 41 professional security experts consulted in this study not only evaluated 89% of the advice in our corpus as accurate, but reported that 118 pieces of advice were in the top 5 items they would recommend to users. By asking people to implement an infeasible number of behaviors, with little guidance on which is the most important, we slowly chip away at compliance budgets [3], leaving users haphazardly selecting among hundreds of "actionable," "effective," "high-priority" behaviors.

10 Next Steps

Our results suggest two key directions of focus for moving toward a healthier ecosystem of security advice.

Measurement and a new focus on minimality. We as security experts and advice givers have failed to narrow down a multitude of relatively actionable, but half-heartedly followed, security behaviors to a key, critical set that are most important for keeping users safe. The U.S. government alone offers 205 unique pieces of advice to end users, while non-technical news media, such as CNN and Forbes, offers over 100 unique pieces of advice to users. This overload of advice affects a large portion of the user population: prior work [43, 45] suggests the government is a primary source of advice for more than 10% of users, while 67.5% of users report getting at least some of their security advice through the news media.

Our struggle as experts to distinguish between more and less helpful advice may be due to unfalsifiability: being unable to identify whether a piece of advice is actually useful, or prove when it is not. Without measurement of impact on actual security, or proven harm, we presume that everything is slightly useful against potential harms. Fixing this problem will require rigorous measurement (e.g., comparing the effect of different practices on frequency of compromise) to evaluate which behaviors are the most effective, for which users, in which threat scenarios. It will also require a strong commitment among security advice givers to minimality and

practicality: empirically identifying the smallest and most easily actionable set of behaviors to provide the maximum user protection.

If we do not make changes to our advice-giving approach, this situation is destined to get worse. As new attacks continue to emerge, we are likely to continue to issue new, reactive advice without deprecating old advice (that might still be at least somewhat useful) or reevaluating overall priorities [25]. Further, we need to explore how to better disseminate updates to best practices. For example, many experts in our study were still emphasizing password changes and avoiding storing passwords, despite this advice having been updated and disproven in the most recent NIST standards [20]. Delays in propagating new priorities among experts will surely translate into even more severe lags in end-user behavior.

Based on our analysis, the U.S. government is currently the giver of the most advice. Unifying the voices across the government into a single central authority for both end-users and external experts to turn to for validated best practices – similar to the role police departments serve for community education on car break-ins, or the role of the surgeon general for health advice – may help to cut down on inconsistent or delayed updates to advice. A similar effort could be made to reduce redundancy across trusted non-profits and advocacy groups by encouraging such groups to all support a centralized advice repository rather than each providing their own.

Fixing existing advice. While the primary outcome of this work is that we need less advice and more empirical measurement, we do note that a few topics of advice performed consistently worse than others across our evaluations and thus are good candidates for revision and improvement. 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").

Privacy advice was more of a mixed bag. While a quarter of the advice about privacy was rated as unactionable, a significant proportion of the remaining 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.

Additionally, across all topics, many advice articles combined a diverse set of advice types that could be appropriate to different users; future work may wish to examine whether this is effective or whether articles focused on a single context are most appropriate. Relatedly, future work may wish to pursue mechanisms for personalizing advice to users or helping users filter to advice that is most relevant to them, as searches for security advice are likely to surface context-broad advice that may or may not have direct relevance.

Acknowledgements

We are grateful to the reviewers and especially to our shepherd Mary Ellen Zurko for their feedback and guidance. This material is based upon work supported by a UMIACS contract under the partnership between the University of Maryland and DoD. Elissa M. Redmiles additionally wishes to acknowledge support from the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 1322106 and a Facebook Fellowship.

References

[1] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, et al. Nudges for privacy and security: Understanding and assisting users? choices online. *ACM Computing Surveys (CSUR)*, 50(3):44, 2017.

[2] Elham Al Qahtani, Mohamed Shehab, and Abrar Aljohani. The effectiveness of fear appeals in increasing smartphone locking behavior among saudi arabians. In *SOUPS 2018: Symposium on Usable Privacy and Security*, 2018.

[3] A. Beautement, M. A. Sasse, and M. Wonham. The compliance budget: Managing security behaviour in organisations. In *NSPW 2009: New Security Paradigms Workshop*, 2008.

[4] JM Blythe and CE Lefevre. Cyberhygiene insight report. 2017.

[5] Joseph Bonneau, Cormac Herley, Paul C Van Oorschot, and Frank Stajano. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In *2012 IEEE Symposium on Security and Privacy*, pages 553–567. IEEE, 2012.

[6] John R Bormuth. Comparable cloze and multiple-choice comprehension test scores. *Journal of Reading*, 1967.

[7] C. Bravo-Lillo, S. Komanduri, L. F. Cranor, R. W. Reeder, M. Sleeper, J. Downs, and S. Schechter. Your attention please: Designing security-decision UIs to make genuine risks harder to ignore. In *SOUPS*, 2013.

[8] Cristian Bravo-Lillo, Lorrie Faith Cranor, Julie Downs, Saranga Komanduri, and Manya Sleeper. Improving computer security dialogs. In *IFIP Conference on Human-Computer Interaction*. Springer, 2011.

[9] William Burr, Donna Dodson, and W Polk. Electronic authentication guideline. Technical report, National Institute of Standards and Technology, 2004.

[10] Karoline Busse, Julia Schäfer, and Matthew Smith. 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*, 2019.

[11] Kelly Erinn Caine. *Exploring everyday privacy behaviors and misclosures*. PhD thesis, 2009.

[12] Lorrie Faith Cranor. A framework for reasoning about the human in the loop. *UPSEC*, (2008), 2008.

[13] T. Denning, A. Lerner, A. Shostack, and T. Kohno. Control-alt-hack: The design and evaluation of a. card game for computer security awareness and education. In *SIGSAC 2013: Conference on Computer & Communications Security*. ACM, 2013.

[14] Warwick B Elley and Cedric Croft. *Assessing the difficulty of reading materials: The noun frequency method*. 1989.

[15] Adrienne Porter Felt, Robert W Reeder, Hazim Almuhimedi, and Sunny Consolvo. Experimenting at scale with google chrome's ssl warning. In *CHI 2014: Conference on Human Factors in Computing Systems*. ACM, 2014.

[16] Rudolph Flesch. A new readability yardstick. *Journal of applied psychology*, 1948.

[17] Alisa Frik, Serge Egelman, Marian Harbach, Nathan Malkin, and Eyal Peer. Better late (r) than never: Increasing cyber-security compliance by reducing present bias. In *Symposium on Usable Privacy and Security*, 2018.

[18] Kelsey R. Fulton, Rebecca Gelles, Alexandra McKay, Yasmin Abdi, Richard Roberts, and Michelle L. Mazurek. The effect of entertainment media on mental models of computer security. In *SOUPS 2019: Symposium on Useable Privacy and Security*, 2019.

[19] Vaibhav Garg, L. Jean Camp, Katherine Connelly, and Lesa Lorenzen-Huber. Risk communication design: Video vs. text. In *PETS 2012: Performance Evaluation of Tracking and Surveillance*, 2012.

[20] P Grassi, M Garcia, and J Fenton. Nist special publication 800-63-3 digital identity guidelines. *National Institute of Standards and Technology, Los Altos, CA*, 2017.

[21] Marian Harbach, Sascha Fahl, Thomas Muders, and Matthew Smith. Towards measuring warning readability. In *ACM CCS 2019: Conference on Computer and Communications Security*, 2012.

[22] Michael Heilman. *Automatic factual question generation from text*. PhD thesis, Carnegie Mellon University, 2011.

[23] Cormac Herley. So long, and no thanks for the externalities: the rational rejection of security advice by users. In *NSPW 2009: New Security Paradigms Workshop*. ACM, 2009.

[24] Cormac Herley. More is not the answer. *IEEE Security and Privacy magazine*, 2014.

[25] Cormac Herley. Unfalsifiability of security claims. National Academy of Sciences, 2016.

[26] Iulia Ion, Rob Reeder, and Sunny Consolvo. "... no one can hack my mind": Comparing expert and non-expert security practices. In *SOUPS 2015: Symposium On Usable Privacy and Security*, 2015.

[27] Terry K Koo and Mae Y Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 2016.

[28] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 2009.

[29] P. Kumaraguru, J. Cranshaw, A. Acquisti, L. F. Cranor, J. Hong, M. A. Blair, and T. Pham. School of phish: A real-world evaluation of anti-phishing training. In *SOUPS 2009: Symposium on Usable Privacy and Security*, 2009.

[30] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *Biometrics*, 1977.

[31] Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 115–124, 2017.

[32] Annie Louis and Ani Nenkova. What makes writing great? first experiments on article quality prediction in the science journalism domain. *Transactions of the ACL*, 2013.

[33] Aleecia M. McDonald, Robert W. Reeder, Patrick Gage Kelley, and Lorrie Faith Cranor. *A Comparative Study of Online Privacy Policies and Formats*. 2009.

[34] Christine Mekhail, Leah Zhang-Kennedy, and Sonia Chiasson. Visualizations to teach about mobile online privacy. In *Persuasive Technology Conference (poster)*, 2014.

[35] James Nicholson, Lynne Coventry, and Pam Briggs. Introducing the cybersurvival task: assessing and addressing staff beliefs about effective cyber protection. In *SOUPS 2018: Fourteenth Symposium on Usable Privacy and Security*, 2018.

[36] James Nicholson, Lynne Coventry, and Pamela Briggs. If it's important it will be a headline: Cybersecurity information seeking in older adults. In *CHI 2019: Conference on Human Factors in Computing Systems*. ACM, 2019.

[37] Tony Ojeda, Sean Patrick Murphy, Benjamin Bengfort, and Abhijit Dasgupta. *Practical data science cookbook*. 2014.

[38] John W Oller, J Donald Bowen, Ton That Dien, and Victor W Mason. Cloze tests in english, thai, and vietnamese: Native and non-native performance. 1972.

[39] Emily Pitler and Ani Nenkova. Revisiting readability: A unified framework for predicting text quality. In *EMNLP 2008: Conference on Empirical Methods in Natural Language Processing*. ACL, 2008.

[40] E. Rader and R. Wash. Identifying patterns in informal sources of security information. *J. Cybersecurity*, 2015.

[41] E. Rader, R. Wash, and B. Brooks. Stories as informal lessons about security. In *SOUPS*, 2012.

[42] Earl F Rankin and Joseph W Culhane. Comparable cloze and multiple-choice comprehension test scores. *Journal of Reading*, 1969.

[43] E. M. Redmiles, S. Kross, and M. L. Mazurek. How i learned to be secure: a census-representative survey of security advice sources and behavior. In *CCS*, 2016.

[44] Elissa M Redmiles, Yasemin Acar, Sascha Fahl, and Michelle L Mazurek. A summary of survey methodology best practices for security and privacy researchers. Technical report, 2017.

[45] Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. Where is the digital divide?: A survey of security, privacy, and socioeconomics. In *CHI 2017: Conference on Human Factors in Computing Systems*. ACM, 2017.

[46] Elissa M Redmiles, Amelia R Malone, and Michelle L Mazurek. I think they're trying to tell me something: Advice sources and selection for digital security. In *IEEE 2016: Symposium on Security and Privacy (SP)*. IEEE, 2016.

[47] Elissa M Redmiles, Michelle L Mazurek, and John P Dickerson. Dancing pigs or externalities?: Measuring the rationality of security decisions. In *EC 2018: Conference on Economics and Computation*. ACM, 2018.

[48] Elissa M Redmiles, Ziyun Zhu, Sean Kross, Dhruv Kuchhal, Tudor Dumitras, and Michelle L Mazurek. Asking for a friend: Evaluating response biases in security user studies. In *SIGSAC 2018: Conference on Computer and Communications Security*. ACM, 2018.

[49] 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. Comparing and developing tools to measure the readability of domain-specific texts. In *EMNLP 2019: Conference on Empirical Methods in Natural Language Processing*, 2019.

[50] Robert W Reeder, Iulia Ion, and Sunny Consolvo. 152 simple steps to stay safe online: security advice for non-tech-savvy users. *IEEE Security & Privacy*, 2017.

[51] Luz Rello, Martin Pielot, and Mari-Carmen Marcos. Make it big!: The effect of font size and line spacing on online readability. In *CHI 2016: Conference on Human Factors in Computing Systems*. ACM, 2016.

[52] Ronald W Rogers and Steven Prentice-Dunn. Protection motivation theory. 1997.

[53] Sukamol S. and S. Jakobsson. Using cartoons to teach internet security. *Cryptologia*, 2008.

[54] Armin Sarabi, Ziyun Zhu, Chaowei Xiao, Mingyan Liu, and Tudor Dumitras. 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*, pages 113–125. Springer, 2017.

[55] Jeff Sauro and Joseph S Dumas. Comparison of three one-question, post-task usability questionnaires. In *SIGCHI 2009: Conference on Human Factors in Computing Systems*. ACM, 2009.

[56] Yukiko Sawaya, Mahmood Sharif, Nicolas Christin, Ayumu Kubota, Akihiro Nakarai, and Akira Yamada. Self-confidence trumps knowledge: A cross-cultural study of security behavior. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2202–2214, 2017.

[57] S. Sheng, B. Magnien, P. Kumaraguru, A. Acquisti, L. F. Cranor, J. Hong, and E. Nunge. Anti-phishing phil: The design and evaluation of a game that teaches people not to fall for phish. In *SOUPS 2007: Symposium on Usable Privacy and Security*, 2007.

[58] Yue Shi, Martha Larson, and Alan Hanjalic. Collaborative filtering beyond the user-item matrix: A survey of the state of the art and future challenges. *ACM Computing Surveys (CSUR)*, 2014.

[59] Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

[60] RI Singh, M Sumeeth, and J Miller. Evaluating the readability of privacy policies in mobile environments. *Developments in Technologies for Human-Centric Mobile Computing and Applications*, 2012.

[61] Andreas Sotirakopoulos, Kirstie Hawkey, and Konstantin Beznosov. On the challenges in usable security lab studies: lessons learned from replicating a study on ssl warnings. In *SOUPS 2011: Symposium on Usable Privacy and Security*. ACM, 2011.

[62] Joshua Sunshine, Serge Egelman, Hazim Almuhimedi, Neha Atri, and Lorrie Faith Cranor. Crying wolf: An empirical study of ssl warning effectiveness. In *USENIX 2009: Security Symposium*, 2009.

[63] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *WWW 2016: International Conference on World Wide Web*, 2016.

[64] Joshua Tan, Lujo Bauer, Joseph Bonneau, Lorrie Faith Cranor, Jeremy Thomas, and Blase Ur. Can unicorns help users compare crypto key fingerprints? In *CHI 2017: Conference on Human Factors in Computing Systems*. ACM, 2017.

[65] Wilson L Taylor. Cloze procedure: A new tool for measuring readability. *Journalism Bulletin*, 1953.

[66] Wilson L Taylor. Recent developments in the use of "cloze procedure". *Journalism Quarterly*, 1956.

[67] R. Wash and E. Rader. Too much knowledge? security beliefs and protective behaviors among united states internet users. In *SOUPS*, 2015.

[68] Rick Wash, Emilee Rader, Ruthie Berman, and Zac Wellmer. Understanding password choices: How frequently entered passwords are re-used across websites. In *Twelfth Symposium on Usable Privacy and Security ({SOUPS} 2016)*, pages 175–188, 2016.

[69] Rick Wash, Emilee Rader, and Chris Fennell. Can people self-report security accurately?: Agreement between self-report and behavioral measures. In *CHI 2017: Conference on Human Factors in Computing Systems*. ACM, 2017.

[70] Gordon B Willis. *Cognitive interviewing: A tool for improving questionnaire design*. Sage Publications, 2004.

[71] L. Zhang-Kennedy, S. Chiasson, and R. Biddle. The role of instructional design in persuasion: A. comics approach for improving cybersecurity. *Int. J. Hum. Comput. Interaction*, 2016.

[72] Yixin Zou, Shawn Danino, Kaiwen Sun, and Florian Schaub. You "might" be affected: An empirical analysis of readability and usability issues in data breach notifications. In *CHI 2019: Conference on Human Factors in Computing Systems*. ACM, 2019.

Appendix

A Search Query Generation Prompt Articles

- https://www.zdnet.com/article/previously-unseen-malware-\behind-cyberattack-against-uks-biggest-hospital-group/
- https://mobile.wnd.com/2017/03/operating-system-movie-computer-virus-stored-on-dna/
- https://www.pbs.org/newshour/show/ransomware-attack-takes-down-la-hospital-for-hours
- https://www.mysanantonio.com/business/local/article/Computer-hackers-steal-San-Antonio-Symphony-10931790.php
- https://www.marketwatch.com/story/your-childs-teddy-bear-may-now-be-hacked-2017-03-01
- https://www.wired.com/2017/03/Internet-bots-fight-theyre-human/

B Advice Comprehensibility

Figure 7 and 9 summarize the comprehensibility of the corpus.

Figure 10 summarizes the mean Cloze scores across specific advice providers who are members of the U.S. Government, non-tech non-profits, and technical non-profits.

C User Perceived Actionability Questionnaire

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."

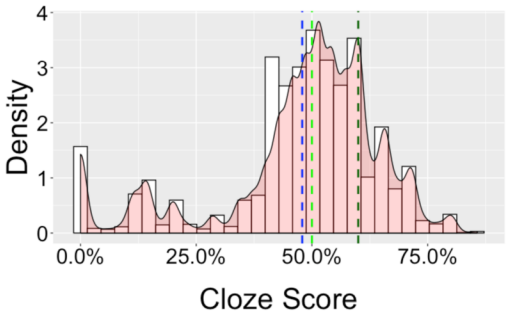

Figure 7: Security advice corpus Cloze scores. Higher scores indicate more comprehensible documents, light green shading and dotted line indicates mean Cloze score >50% which signifies partial comprehensibility, brighter green shading and line indicates mean score >60% which signifies full comprehensibility. Dashed blue line indicates corpus mean.

Figure 8: Mean Cloze scores by topic.

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. [Answer choices: Yes (at least some of the time), No (never), 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? [Answer choices: Very difficult, Somewhat difficult, Slightly difficult, Not at all difficult]

3. 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**.

4. How **time consuming** do you think it would be for you to follow this advice on your personal/non-work computer or mobile device? [Answer choices: Very time consuming, Somewhat time consuming, Slightly time consuming, Not at all time consuming]

5. How **disruptive** do you think it would be for you to follow this advice on your personal/non-work computer or mobile device? [Answer choices: Very disruptive, Somewhat disruptive, Slightly disruptive, Not at all disruptive]

6. How **confident** do you feel that you could implement this advice? [Answer choices: Very confident, Somewhat confident, Slightly confident, Not at all confident]

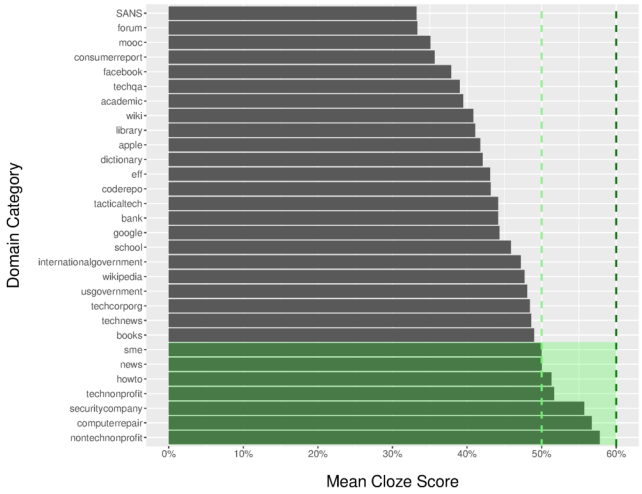

Figure 9: Mean Cloze scores by domain.

D Expert Perceived Efficacy Questionnaire

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. [Answer choices: Yes (at least some of the time), No (never), Not applicable]

2. Please select the option that best matches your opinion. [Answer choices: Following this advice would... IMPROVE someone's digital security or privacy at least a little bit (e.g., this advice is beneficial), would HARM someone's digital security or privacy at least a little bit (e.g., this advice is harmful), have ABSOLUTELY NO EFFECT on someone's digital security or privacy (e.g., this advice is useless)]

3. *[If they answered "IMPROVE" to Q2]*:

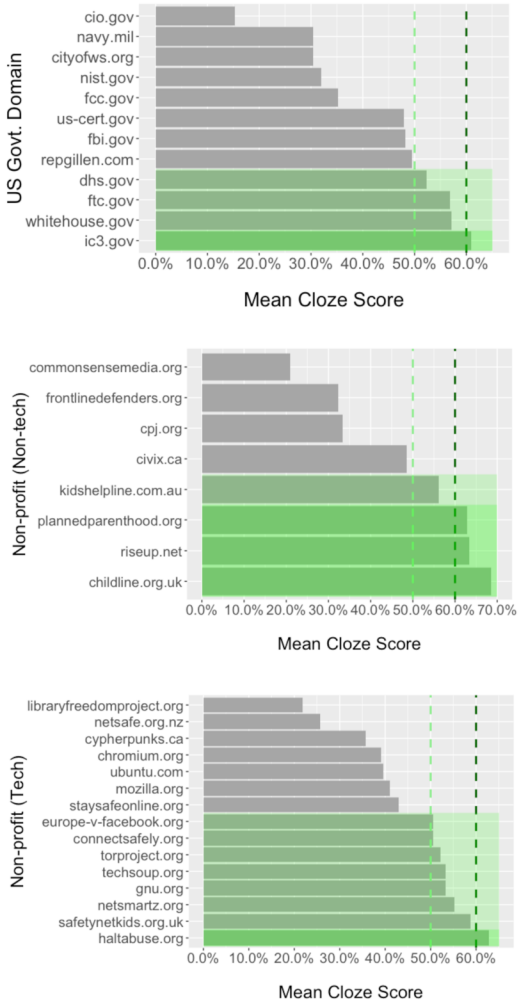

Figure 10: Mean Cloze scores for U.S. Government domains (left), non-tech non-profits (center), and technical non-profits (right).

- 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)? [Answer choices: 0% decrease in risk, 5% decrease in risk, 10% decrease in risk, 15% decrease in risk, 20% decrease in risk, 25% decrease in risk, 30% decrease in risk, 40% decrease in risk, 50%+ decrease in risk]
- For how long do you think this advice will remain useful for improving people's security? [Answer choices: For the next...year (0-1 years), few years (2-5 years), five to ten years (5-10 years), few decades (10+ years), Other: [text entry], I don't know]

4. *[If they answered "HARM" 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)? [Answer choices: 0% increase in risk, 5% increase in risk, 10% increase in risk, 15% increase in risk, 20% increase in risk, 25% increase in risk, 30% increase in risk, 40% increase in risk, 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? [Answer choices: the number 1 behavior I would recommend, in the top 3 behaviors I would recommend, in the top 5 behaviors I would recommend, in the top 10 behaviors I would recommend, I would recommend it, but it's not in the top 10 behaviors I would recommend, I would not recommend this advice]

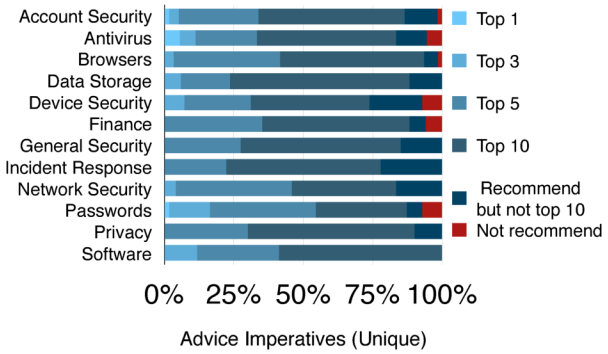

Figure 11: Expert priority rating of advice by topic.

E Priority by Topic

Figure 11 summarizes experts priority rating of security advice by topic.

F Pairwise Comparisons of Actionability and Priority By Topic

	Account Security	Antivirus	Browsers	Data Storage	Device Security	Finance	General Security	Incident Response	Network Security	Passwords	Privacy
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***	0.07		
Privacy	0.78	0.69	0.27	0.77	0.98	0.51	0.67	0.40	0.67	**0.02***	
Software	0.31	0.54	0.78	0.17	0.26	0.70	0.09	0.09	0.53	0.30	0.27

Table 3: Results of Mann-Whitney pairwise comparisons of median priority rating 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	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 4: 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 5: 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 6: 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 7: Results of pairwise Mann-Whitney comparisons of ratings of advice difficulty by topic. Holm Bonferonni multiple testing correction applied.