A Systematic Literature Review of Empirical Methods and Risk Representation in Usable Privacy and Security Research

VERENA DISTLER, University of Luxembourg
MATTHIAS FASSL, CISPA Helmholtz Center for Information Security
HANA HABIB, Carnegie Mellon University
KATHARINA KROMBHOLZ, CISPA Helmholtz Center for Information Security
GABRIELE LENZINI, University of Luxembourg
CARINE LALLEMAND, Eindhoven University of Technology & University of Luxembourg
LORRIE FAITH CRANOR, Carnegie Mellon University
VINCENT KOENIG, University of Luxembourg

Usable privacy and security researchers have developed a variety of approaches to represent risk to research participants. To understand how these approaches are used and when each might be most appropriate, we conducted a systematic literature review of methods used in security and privacy studies with human participants. From a sample of 633 papers published at five top conferences between 2014 and 2018 that included keywords related to both security/privacy and usability, we systematically selected and analyzed 284 full-length papers that included human subjects studies. Our analysis focused on study methods; risk representation; the use of prototypes, scenarios, and educational intervention; the use of deception to simulate risk; and types of participants. We discuss benefits and shortcomings of the methods, and identify key methodological, ethical, and research challenges when representing and assessing security and privacy risk. We also provide guidelines for the reporting of user studies in security and privacy.

CCS Concepts: • **Security and privacy → Human and societal aspects of security and privacy**; • **Human-centered computing → Empirical studies in HCI**;

Additional Key Words and Phrases: Usable privacy and security, Human-Computer Interaction (HCI), user experience (UX) research

ACM Reference format:
Verena Distler, Matthias Fassl, Hana Habib, Katharina Krombholz, Gabriele Lenzini, Carine Lallemand, Lorrie Faith Cranor, and Vincent Koenig. 2021. A Systematic Literature Review of Empirical Methods and Risk Representation in Usable Privacy and Security Research. *ACM Trans. Comput.-Hum. Interact.* 28, 6, Article 43 (December 2021), 50 pages.
https://doi.org/10.1145/3469845

This work is supported by the Fonds National de la Recherche (PRIDE15/10621687) and the Carnegie Corporation of New York.

Authors' addresses: V. Distler, G. Lenzini, and V. Koenig, University of Luxembourg; M Fassl and K. Krombholz, University of Luxembourg Maison des Sciences Humaines 11, Porte des Sciences L-4366 Esch-sur-Alzette Luxembourg; H. Habib and L. F. Cranor, Carnegie Mellon University, 4720 Forbes Avenue, Pittsburgh, PA 15213, United States; emails: htq@cs.cmu.edu, lorrie@cs.cmu.edu; C. Lallemand, Eindhoven University of Technology & University of Luxembourg.

1073-0516/2021/12-ART43
https://doi.org/10.1145/3469845

1 INTRODUCTION

As the use of digital technology evolves, so does the number and the type of risks to which users and their data are exposed. Studying how people perceive those risks and what interventions will help people better understand and respond to them is thus an essential factor in achieving better security. To understand the extent to which study participants will take security- or privacy-protective steps often requires that participants are exposed to a scenario that provides realistic cues such that they will behave in the study in a similar way that they would behave in real life. This often requires exposing participants to real or simulated risk. Participants may be able to navigate menus, click buttons, and follow instructions to use a security tool such as an encrypted chat client or a web browser security feature, but testing the effectiveness of such a tool often requires exposing participants to a real or simulated attack to determine how they will respond and whether the tool helps prevent them from being deceived by the attacker. Usable privacy and security (UPS) researchers regularly encounter challenges when they design studies—whether in the lab, online, or in-situ—that focus on people's perceptions of and response to security and privacy risks.

Since users usually have a primary objective that is not related to security or privacy, instructing participants to pay particular attention to privacy and security would lead to bias and "security priming" [Sotirakopoulos et al. 2011]. Designing a realistic experience for a laboratory or online study of security and privacy risk is difficult. For additional realism, deception can be used to simulate attacks and to make participants believe they are at risk [Cranor and Buchler 2014]. When using deception, researchers must find a balance between preserving the realism of an attack and ethically exposing study participants to it. This is often achieved by debriefing participants about the deception promptly so they do not spend much time worrying about having been the victim of an attack or take unnecessary steps attempting to recover from a perceived attack.

In-situ studies are also challenging because attacks occur rarely and thus require collecting user activity logs over extended periods of time, potentially raising logistical and privacy concerns. In addition, observing real attacks without interfering and mitigating potential harm to participants can be ethically questionable. On the other hand, if researchers inject simulated attacks into a participant's real-world activities, participants may become accustomed to being attacked for research purposes and may ignore real attacks.

These are just some of the challenges that researchers in usable privacy and security face when designing user studies; we describe these in more detail in Section 2.2. In light of these challenges, we investigated which methods researchers use for privacy and security studies that include human participants, with a particular focus on the approaches researchers use to represent risk. We conducted a systematic literature review of 284 full-length research papers published at five top conferences between 2014 and 2018, with the goal of accumulating the knowledge from a large body of studies and providing an analysis of the characteristics of empirical user studies in usable privacy and security. Systematic literature reviews efficiently integrate existing knowledge [Mulrow 1994], and can aid in understanding the breadth of research on a topic. They can also be used to develop theories or conceptual background for subsequent research, and identify topics that require more investigation [Paré et al. 2015]. A better understanding of how risk is represented in privacy and security user studies will allow the community to critically assess and discuss the validity and ethics of different approaches, as well as help researchers as they design future studies.

Our literature review suggests that risk representation in the analyzed papers was mostly based on naturally occurring or simulated risk, varying with study methods and research objectives of the paper. Papers with an experimental objective mostly used simulated risk, and descriptive studies mostly relied on naturally occurring risk. Few studies relied only on mentioned risk or did

not represent risk at all. Common tools used to represent risks to research participants included security/privacy-related tasks, prototypes, and scenarios. Deception, educational interventions, and incentives for secure behavior were only rarely used. Based on our systematic review, we discuss the implications of our findings and suggest guidelines for designing and reporting UPS user studies.

Our study makes the following contributions:

(1) A systematic review of the methods employed in UPS papers from 2014 to 2018 for inducing a perception of privacy and security risks.
(2) A structure for systematically analyzing methods in UPS studies, with a focus on risk representation.
(3) Identification of six approaches used in UPS studies (individually or in combination) for inducing a perception of risk: assigned tasks, prototypes, scenarios, deception, educational interventions, and incentives for secure behavior.
(4) Guidelines for designing and reporting UPS studies.

In this article, we first position our research within the broader context of related work in the field of usable privacy and security, considering previous literature reviews, methodological challenges faced by researchers in the field and how deception is used in UPS studies (Section 2). We then present the methods we used for this systematic literature review (Section 3). Next, we describe the approaches found in our literature review for representing risk to research participants (Section 4). We then discuss the choice of methods, participant recruitment and ethics, and consider limitations of our study (Section 5) before concluding (Section 6).

2 BACKGROUND AND RELATED WORK

We review related work in three areas relevant to our study: UPS literature reviews, methodological challenges and risk representation in UPS research, and the use of deception when studying perceived risk.

2.1 Literature Reviews in Usable Privacy and Security

Although several literature reviews have been conducted in areas related to usable privacy and security, none address the need for a systematic review of methodological issues specific to risk representation in UPS. Iachello and Hong [2007] summarize research on the topic of privacy in Human Computer Interaction, with a focus on current approaches, results, and trends. They identify future grand challenges in HCI and privacy, such as developing better ways of helping end-users manage their privacy, creating stronger analysis techniques and survey tools, or developing a theory of technology acceptance, specifically related to privacy. Garfinkel and Lipford [2014] review past UPS research and identify important research directions within UPS. They also describe important challenges of research in UPS, ranging from authentication, adversary modelling, system administration, consumer privacy, social computing, ecological validity, and teaching. Acquisti and colleagues [2017] review literature pertaining to privacy and security decision making with a focus on research assisting individuals' privacy and security choices with soft paternalistic interventions that nudge users toward more beneficial choices. The authors discuss potential benefits and shortcomings, as well as identify ethical, design, and research challenges.

In the subfield of authentication research, Biddle et al. [2012] provide an overview of published research in the area of graphical passwords, including usability and security aspects as well as system evaluation. Bonneau et al. [2012] evaluate two decades of proposals to replace text passwords for general-purpose user authentication. They provide a framework enabling researchers to

evaluate the methods and to benchmark future web authentication proposals. Finally, Velásquez et al. [2018] present a systematic literature review of authentication schemes.

While a number of HCI papers have reviewed methods for user experience evaluation generally ([Alves et al. 2014], [Obrist et al. 2009], and [Pettersson et al. 2018]), we know of no review of the methods used in UPS studies in particular. We fill this gap by reviewing methods used in UPS studies from both the HCI literature and from specialized privacy and security publication venues, focusing on approaches to risk representation.

2.2 Methodological Challenges and Risk Representation in Usable Privacy and Security

The methodological challenges in UPS are different from those in other focus areas of user-centered design. In particular, collecting data in an ecologically valid way remains particularly complex in UPS. While lab studies allow researchers to create a controlled environment and isolate the effect of certain variables, participants may face different threats and motivations than in the field. Using fictitious personal data rather than a participant's real data may reduce privacy risks but will impact the ecological validity of the study. In addition, simulated attacks in a lab environment will be experienced at a significantly higher rate than in a real-world setting, further jeopardizing ecological validity [Garfinkel and Lipford 2014].

Another difficulty is that, for many relevant scenarios, security or privacy is not the primary goal of the individual, and thus any mention of security or privacy may prime participants and cause them to behave differently than they would normally [Egelman et al. 2007]. In addition, it is difficult to simulate a situation in which users would both fulfil their primary tasks and respond to potential risks [Schechter 2013].

A lab setting (as well as participant briefings, instructions, and research framing) can also frequently lead participants to state that they care more about security than they would in a real-world setting. Indeed, studies have found inconsistencies between what people say and the actions they actually take related to privacy and security ([Egelman et al. 2007] and [Sotirakopoulos et al. 2011]). Social desirability of privacy and security behaviors may play a role in this context [Egelman and Peer 2015]. Lastly, while the use of self-reported data about security and privacy behaviors can provide rich insights, these data can sometimes lack reliability for many reasons, including participants misremembering their past behaviors, or feeling uncomfortable making accurate disclosures.

UPS researchers use a variety of approaches to create a realistic experience of risk in their studies. In some studies, researchers introduce hypothetical scenarios and use role playing to simulate a real-life situation. Schechter and colleagues [2007] describe the use of role playing to create a perception of risk, but find that role playing has a significant negative effect on the security vigilance of study participants. Another approach is the use of deception, which can be beneficial to understand how people react to attacks in a realistic setting, for instance by simulating the presence of an adversary [Cranor and Buchler 2014] or by launching attacks on research participants [Egelman et al. 2007]. While deceptive studies raise ethical concerns, these studies are often justified because it would be difficult or impossible to conduct some types of studies without deception, and harm to participants can be minimized through timely debriefing. For example, Cranor and Buchler [2014] argue that in the context of computer security warnings, it is important to use simulated attack scenarios to observe how participants respond to warnings when they have been led to believe they are actually at risk. Another approach is long-term in-situ studies. Forget et al. built the Security Behavior Observatory (SBO) to recruit and observe a panel of consenting home computer users, allowing for the study of security-related behavior in a real-world setting [Forget et al. 2014].

2.3 The Use of Deception When Studying Perceived Risk

Deception can be defined as deliberately misleading participants or not informing them about the purpose of the investigation, usually to avoid the possibility that responses might be given to meet perceived expectations of the researchers [Deception Research—APA Dictionary of Psychology n.d.]. In UPS studies, deception is often used to mislead participants so they believe the study is unrelated to security or that simulated risks are actually real. Generally, participants are debriefed promptly at the conclusion of the study to prevent psychological harm, for example, from worrying about actual harm from simulated risks they have been misled to believe are real. Debriefing participants prevents mistrust in the researcher resulting from the use of deception [American Psychological Association 2017].

Researchers have long emphasized the far-reaching ethical issues of studies using deception in psychological research, including a decrease of trust in researchers, lack of informed consent, and the insufficient effect of the debriefing [Baumrind 1985]. On the other end of the spectrum, Christensen [1988] argues that research participants often do not have negative feelings after participating in a deception experiment and that the acceptability of deception depends on the behaviors being investigated, the setting of the investigation (public vs. private place), and the outcome of the experiment. He concludes that deception should be avoided in studies investigating personal information or in studies that potentially harm the subject. Athanassoulis and Wilson [2009] argue that the fact that a research study uses deception does not necessarily make it morally problematic. Rather, they suggest that ethics committees should focus on the reasonableness of withholding information from a participant, which is context-dependent.

In the field of Human-Computer Interaction, Adar and colleagues [2013] argue that deception is understudied and present a view of deception that takes into account motive (why deception happens), means (how deception is designed), and opportunity (when it works).

In the usable privacy and security community, the importance of rigorous reporting of deception has been underlined, including reporting how participants were debriefed, how they reacted, and how data was protected [Schechter 2013]. Deception is usually used to avoid security priming, which could impact participants' responses and reactions. Schechter et al. explored the effect of priming, where one experimental group was instructed to pay attention to security during the study which included banking tasks, while the other group was not "primed" in this way. The authors did not find a statistically significant difference between the groups [Schechter et al. 2007]. Fahl et al. [2013] asked students to role-play that they had enrolled in a new university, and thus needed to create passwords. Similar to Schechter and colleagues, they did not find an effect of priming. Naiakshina et al. [2018] conducted an experiment with student developers, where half of the developers were primed to consider security when implementing the user registration functionality of a social network, while the other half were not primed. Contrary to the previously described studies, priming clearly had an effect on the number of participants who attempted to implement a secure solution. One should note, however, that while some of these studies use the notions of priming and deception almost interchangeably, we consider the lack of priming to be *partial disclosure* rather than *deception*, as described in detail in the results section.

In the quest to avoid priming participants, some deception studies do not ask participants for informed consent since informing them about the study could prime them. While most UPS studies undergo Institutional Review Board (IRB) or ethics board review, IRB approval does not guarantee that non-consenting participants do not feel violated, as described by Garfinkel and Lipford [2014]. They emphasize that the question of consent in a field setting, where it is often avoided in an attempt to avoid priming participants, is a serious concern. The authors describe a study where the requirement for informed consent was waived through IRB approval, however, the experiment

still resulted in significant negative attention because participants resented being involved without their consent [Jagatic et al. 2007].

3 RESEARCH APPROACH

3.1 Research Objectives

The goal of this article is to summarize and extract knowledge from a large corpus of UPS work between 2014 and 2018. The objective is to analyze the characteristics of empirical studies in UPS to better understand how risk is represented in user studies and how researchers navigate the tension between realistic exposure to risk and ethical, legal, and practical considerations. We conduct a systematic literature review of a sample of recent papers published at top peer-reviewed UPS venues. We review the methods used in these studies and how they allow the authors to represent risk.

Our analysis focuses on three research questions:

RQ1: Which methods do researchers in the UPS community use?
RQ2: How do researchers in UPS represent risk?
RQ3: How do researchers in the UPS community use deception in their user study protocols?

The first two research questions cover the entire range of methods used by researchers in UPS and all types of risk representations (e.g., naturally occurring, simulated, mentioned, no intentionally designed risk perception), including an analysis of participant recruitment. Our third research question focuses on deceptive studies, examining the details of how deception is used. Based on our analysis we provide guidelines for researchers when designing and reporting UPS user studies.

3.2 Review Process

Our systematic literature review approach includes the following phases: (1) identification, (2) filtering, (3) review, and (4) analysis. In the identification phase we constructed the initial set of papers using keyword searches, during the filtering phase we checked whether the papers fulfilled our eligibility criteria, in the review phase we read all papers in detail and categorized them, and in the analysis phase we explored trends we observed during our review and developed guidelines for future UPS study design.

3.2.1 Phase 1: Identification of Potentially Relevant Papers.

3.2.1.1 Source Selection. We selected the five most relevant peer-reviewed conference publication venues for UPS papers. We did not consider journal papers, as most UPS papers are published at conferences. We selected top tier privacy and security conferences that also invite UPS papers, namely ACM Conference on Computer and Communications Security (ACM CCS), IEEE Symposium on Security and Privacy (IEEE S&P), and USENIX Security Symposium (USENIX Security). In addition, we included the Symposium on Usable Privacy and Security (SOUPS), a conference that focuses on UPS papers specifically. We also included the ACM Conference on Human Factors in Computing Systems (CHI), the top HCI conference where UPS papers are regularly published. Our selection of conferences includes three of the "big four" security conferences. We did not select the other big-four conference, the Network and Distributed Systems Security Symposium (NDSS), or the Privacy Enhancing Technologies Symposium (PETS) because they (a) have tended to publish fewer UPS papers than the other conferences, and (b) their publishers do not provide a searchable database of their papers. While UPS papers have also appeared in other top conferences in particular application areas such as The Web Conference and UbiComp, we limited our selection to conferences primarily focused on either security/privacy or HCI. As our focus is on current

Table 1. Exclusion of Papers per Round of Exclusion and per Publication Venue

	ACM CCS	IEEE S&P	USENIX Security	SOUPS	CHI	Total
Phase 1: Identification of potentially relevant papers	237	44	117	118	117	633
Phase 2: Papers filtered based on title and abstract to remove those without user data and those that are not full conference papers	194	20	87	0	4	305
Phase 3: Papers filtered after detailed review	15	4	5	3	17	44
Included Papers	28	20	25	115	96	284

practices and methods, we only considered papers from the last 5 years. Since the publication year 2019 was still ongoing at the time of our data collection, we limited the search results to the period from 2014 to 2018.

3.2.1.2 Search Procedure. We used the keyword search provided by the ACM Digital Library and the IEEE Computer Society Digital Library to construct our initial set of potentially relevant UPS papers in July 2019. As we are interested in UPS papers with a clear focus on user perceptions or behavior, we used a search query designed to select papers mentioning privacy or security in addition to at least one user-related term (user, usability, usable, user experience, UX) in title or abstract: *(privacy OR security) AND (user OR usability OR usable OR ux OR "user experience").* We conducted a pilot in which we added specific terms related to security (e.g., encryption, passwords, authentication) to the search query. We decided against adding these terms as they retrieved only a small number of additional papers, most of which were not relevant to our research questions. The search query resulted in 633 potentially relevant papers (shown in Table 1).

3.2.2 Phase 2: Filtering Initial Set of Papers. The first author reviewed the titles and abstracts of the 633 papers identified in Phase 1 and removed papers that met one or more the following three exclusion criteria:

- Study does not involve any user data (data obtained directly as part of the study, data previously collected in other studies, or data obtained through naturally generated datasets).
- Paper is not a full conference paper (e.g., workshop paper, extended abstract). We decided to exclude such papers since it helped us avoid duplicates if a paper was first published as an extended abstract and later as a full paper. Short papers also tend to include less details on the methodology, and thus provide less insights into risk representations.
- Paper presents theoretical models or simulations without including a user study.

The first author coded all papers as to whether or not they met each of the exclusion criteria. In addition, the remaining authors double-coded 77 papers (12%). Cohen's kappa with the first author ranged between 0.80 (substantial agreement) and 1 (perfect agreement). Remaining conflicts were resolved in discussion. This resulted in 305 papers being removed and 328 advancing to Phase 3.

3.2.3 Phase 3: Detailed Review of the Papers. The first three authors split up the 328 included papers between them. The first author read and coded 213 papers, the second and third author read and coded 72 and 41 papers respectively.

Based on the full paper, the authors excluded a total of 44 additional papers for the following reasons:

- User data was used purely to demonstrate the technical feasibility or effectiveness of a protocol ($n = 11$).
- No user data was used ($n = 17$).
- While the paper may mention privacy or security perceptions or behaviors, the authors did not design their study with the intention of studying privacy and security perceptions or behaviors ($n = 15$).
- The publication was not a full paper ($n = 1$).

The authors reviewed the remaining 284 papers (Table 1) in detail, filling out a spreadsheet row for each paper with information on the dimensions of our analysis structure. 22% ($n = 63$) of papers were analyzed by two coders and any disagreements were discussed and resolved. All three coders participated in bi-weekly calls where they discussed unclear papers.

Our analysis includes the following dimensions which correspond to our research questions. Dimensions A–B describe the dataset, dimensions C–F respond to RQ1 (Which methods do researchers in the UPS community use?), dimensions G–M respond to RQ2 (How do researchers in UPS represent risk?) and dimension N responds to RQ3 (How do researchers in the UPS community use deception in their user study protocols?).

A. Publication Venue (Section 4.1.1)
B. Topic: privacy-enhancing technologies, encryption, authentication, access control, privacy transparency and choice mechanisms, security indicators and warnings, social engineering, security perceptions, attitudes and behaviors, privacy perceptions, attitudes and behaviors, privacy and security for special populations, security for admins and developers, multiple topics (Section 4.1.2)
C. Objective of the study: descriptive, relational, experimental, combination (Section 4.1.3)
   - Replication: yes, no, partial
D. Study method: survey, interview, experiment, focus group, workshop, analysis of existing datasets, log analysis, diary study, co-creation methods, vignette study, observation study, list experiment, vignette experiment, other (Section 4.1.4)
E. Participants: representative sample, non-representative convenience sample, students, computer science students, developers, university employees, employees, security experts, other experts, MTurkers, Prolific, other crowdsourcing, Google Consumer Survey (GCS), Security Behavior Observatory (SBO), users of specific technology, disabled users, children or teenagers, women in particular, LGBTQ+, recruitment not mentioned, other (Section 4.1.5)
   - Number of participants
F. IRB or ethics board approval: ethics board approval, approved exempt, exempt from needing approval, not mentioned, corporate internal review, other (Section 4.1.6)
G. Risk representation: naturally occurring, simulated, mentioned, no induced risk representation (Section 4.2.1)
H. Risk response assessment: observational data, self-reported, assigned security or privacy task, assigned unrelated task, combination (Section 4.2.4)

I. Participants complete an assigned task: security- or privacy-related task, unrelated task, both, no task (Section 4.3.1)
J. Participants interact with prototype: yes, no (Section 4.3.2)
K. Participants asked to respond to one or more hypothetical scenarios: yes, no (Section 4.3.3)
L. Educational intervention: yes, no (Section 4.3.4)
M. Participants received an incentive for secure behavior: yes, no (Section 4.3.5)
N. Deception used: yes, no (Section 4.3.6)
    - Type of deception: deception about the objective of the study, deception about the presence of risk, lack of consent
    - Debriefing (for deception studies): yes, no

3.2.4 Phase 4: Analysis. In this phase, we explored the trends we observed during our review. We focused our analysis on how risk was represented and measured, and how researchers combined approaches for risk representation (assigned tasks, prototypes, scenarios, deception, educational interventions, incentives for secure behavior). We describe the results of this analysis in Section 4.

3.3 Limitations

This article is based on the analysis of a large corpus of papers. Although we report quantitative results on the frequency of papers with various attributes, we caution that our categorization of papers was somewhat of a subjective process, largely due to the fact that some authors did not provide complete information about their methods and that authors use terms like "deception" and "exempt" in inconsistent ways. Some papers fell into gray areas with details that could be interpreted in multiple ways. Such cases were resolved by discussion between the co-authors.

We took a number of steps to promote consistency between our coders in their interpretation of these papers. To arrive at our dataset we double coded 12% of the papers according to our phase 2 exclusion criteria, an approach commonly used in systematic literature reviews to ensure the reliability of the inclusion/exclusion process. During the detailed analysis phase, we used bi-weekly calls of all coders to discuss ambiguous papers and ensure consistency. In addition, we also discussed difficult cases with the co-authors who were not otherwise participating in the coding process. Twenty-two percent of the papers were coded by two coders and conflicts were resolved through discussions, further clarifying any discrepancies in the coders' understanding of the categories. After all papers were coded, the first author also reviewed all code assignments to check for plausibility and consistency. Despite these efforts to maintain data accuracy, the frequencies and percentages in this article are meant to describe trends in the data, rather than to be interpreted as exact indicators due to a certain level of subjective interpretation in the coding.

One might also question why we analyzed papers in the 5-year period between 2014 and 2018, rather than including a longer time period. Since the publication year 2019 was still ongoing at the time of our data collection, we did not include papers from 2019. As our objective was to analyze recent research trends and methods in the UPS field rather than taking a long-term, historic perspective, we limited the search results to the period from 2014 to 2018.

As described previously, we included papers from five top-tier peer-reviewed conferences that welcome UPS papers. While there are other venues that publish UPS papers (e.g., Network and Distributed System Security Symposium, The Web Conference, UbiComp), we limited our selection to conferences primarily focused on either security/privacy or HCI and that provided a searchable database. We also did not consider journals as most UPS papers are published at conferences and some that are published in journals are extended versions of conference papers. A search of the ACM "Transactions" journals found that we omitted relatively few papers by omitting journals. Nevertheless, a review of UPS papers in a wider array of journals might be insightful.

There were some types of data that we did not code for, but that should be considered for future research. For instance, future studies could investigate differences between online or in-person studies, and single-session versus longitudinal studies. A detailed analysis of where participants are located would also give compelling insights into certain geographic areas that are understudied. We did not analyze whether studies reference certain theories or frameworks (e.g., grounded theory, mental models, self-determination theory), which would be an interesting focus point for future research. Looking back at our results, we can also see that drawing tasks seemed to be used in some of the studies in our sample, and we have observed these tasks in more recent studies as well. We did not specifically focus on drawing tasks, but analyzing how drawing tasks are used in UPS studies seems to be a relevant analysis to conduct. In our sample, we could see that a wide variety of compensation styles was employed, ranging from voluntary participation, to course credit, to raffles, to direct financial compensation, which we did not systematically compare and analyze. Future studies could analyze research participant compensation in UPS in more detail, and perhaps contribute to a "standard" of participant compensation. In addition, we recorded the number of participants in each study but did not record the number of experimental conditions. An analysis of participants per experimental condition and associated statistical power could provide added insights. Finally, our sample includes a number of studies that recruited experts as participants. We did not focus in detail on how experts contributed during their study participation.

4 RESULTS

In this section, we first provide an overview of our dataset and the methods used (responding to RQ1), and then focus on how researchers represent risk and assess participants' responses to risk in their studies. We describe the "tools" used by researchers to represent risks to research participants (prototypes, scenarios, educational interventions) and which study methods (e.g., experiments, surveys) coincide with which risk representation modes (RQ2). Finally, we analyze the use of deception (RQ3).

4.1 Dataset Description

In this subsection, we provide an overview of our dataset. We include descriptive statistics about the distribution of papers across venues and publication years. Further, we summarize high-level information about the papers, including the topics studied, research objectives, and study methods. In short, our dataset included 284 UPS papers, with a large percentage published at SOUPS or CHI. The most frequent topics were authentication and privacy or security attitudes. Most of the papers had an experimental or descriptive objective and replications were rare. Experiments, interviews, and surveys were common study methods, and crowdsourcing and non-representative convenience samples were frequently used to recruit participants.

4.1.1 Publication Venue. As shown in Table 2, the *most frequent conference venues* for UPS over the past 5 years, as defined by our search query, were SOUPS ($n = 115$ included papers) and CHI ($n = 96$). Not surprisingly, our dataset includes almost all of the papers published at SOUPS during this time period, the only publication venue that specifically focuses on usable privacy and security topics. The papers were fairly well distributed across the 5 years of the study, as shown in the Appendix, Table 16.

4.1.2 Topics. Papers were coded into mutually exclusive, broad categories to obtain a high-level overview of frequently studied topics. The most frequently addressed topics in our analysis were *authentication* (25% of papers, $n = 72$), followed by papers on *privacy perceptions, attitudes, and behaviors* (19%, $n = 55$) and *security perceptions, attitudes, and behaviors* (16%, $n = 46$). *Access control* (7%, $n = 20$) *and security for admins and developers* (6%, $n = 18$) were other frequent topics,

Table 2. Number of Papers Published at Conference and Papers Included in Our Sample

	Papers published at conference (2014–2018)	Included papers	Percent included
SOUPS	119	115	97
CHI	2675	96	4
ACM CCS	664	28	4
USENIX security	391	25	6
IEEE security and privacy	277	20	7

Table 3. Topics Addressed in Papers (N = 284)

	Frequency	Percent
Authentication	72	25
Privacy perceptions, attitudes, and behaviors	55	19
Security perceptions, attitudes, and behaviors	46	16
Access control	20	7
Security for admins and developers	18	6
Encryption	15	5
Privacy transparency and choice mechanisms	14	5
Security indicators and warnings	12	4
Multiple	11	4
Privacy-enhancing technologies	11	4
Social engineering	10	4

as well as encryption (5%, $n = 15$) *and privacy transparency and choice mechanisms* (5%, $n = 14$). For the remaining topics, refer to Table 3. In Section 4.2.3. we explain how risk was represented within each topic.

4.1.3 Objectives. We categorized papers according to their most important objective, which was either experimental, descriptive, or relational. Experimental research partitions participants into equivalent groups and measures the influence of different experimental manipulations applied to each group [Stangor and Walinga 2018]. Descriptive research provides a snapshot or summary of participants and their opinions or behavior with respect to a particular context or setting. Relational research is designed to discover relationships among variables [Stangor and Walinga 2018], for example the impact of certain demographics or past experiences on behavior. Overall, most included papers had an experimental (41%, $n = 115$) or descriptive (31%, $n = 89$) objective. 5% ($n = 13$) of papers had a relational objective, and the remaining papers combined multiple objectives (see Appendix Table 17). Note that we did not classify experimental research as combined with descriptive if descriptive results were used only to characterize the experimental population and were not an important objective of the study.

Replication studies appear to be rare in UPS. Our dataset includes four replications, and one partial replication. Replications were conducted for multiple reasons. For example, Bravo-Lillo et al. replicated the experimental methodology documented in an earlier study, but added new conditions [Bravo-Lillo et al. 2014]. Another study replicated an earlier experiment in order to assess its robustness [Canfield et al. 2017].

Table 4. Combinations of Study Methods in Our Sample
(Full Table in Appendix, Table 18) (N = 284)

	Frequency	Percent
Experiment	99	35
Interview	36	13
Survey	34	12
Survey and datalogs	12	4
Analyze dataset	11	4
Survey and interview	11	4
Experience sampling method	8	3
Survey and experiment	8	3
Methods and combinations with six or less occurrences	65	23

4.1.4 Study Methods. The papers from our sample predominantly used *experiments, surveys,* and *interviews*, or combinations of these study methods (Table 4). We classify *experiments* as procedures where experimental conditions were manipulated, and the effect of this manipulation was measured. When study authors referred to an "online experiment" in their study, but without apparent experimental conditions, we instead classified the respective studies as *surveys*. *Surveys*, in our analysis, are different from *interviews* in that they usually took place in a written questionnaire form (on paper or online) without a conversation-style interaction between the researcher and the research participant. We coded experiments involving an oral debriefing phase as *experiments* only, not *experiments and interviews,* as we did not consider this debriefing phase an interview study in its own right, and debriefing phases were not always analyzed as rigorously as interview studies typically would be.

Less common study methods included analyses of existing datasets and log analysis. In addition, we found occasional use of focus groups, co-creation methods, list experiments, observation studies, workshops, vignette studies, and diary studies. In Section 4.2.2. we explain how risk was represented in studies using each method, and provide examples.

4.1.5 Participants. As shown in Table 5, the analyzed papers relied heavily on easily accessible populations, in particular crowdsourcing (n = 106), non-representative convenience samples (n = 79), and students (n = 66). Most crowdsourcing studies used Amazon Mechanical Turk (n = 86). Non-representative convenience samples here refer to recruitment of easily accessible, undefined population groups (e.g., through flyers in the neighborhood of the university, or general snowball sampling). In contrast, when researchers specifically recruited students, we used the separate category students. Users of specific technology (referring for instance to users of VR glasses, Android users, or users of specific social networks) were studied as well (n = 37). Employees (n = 33) also played a frequent role.

The number of participants varied considerably between studies, as shown in Appendix, Table 19. Interview studies tended to have the fewest participants among the frequently used study methods, with a median of 21 and a maximum of 200 participants. Surveys and log analysis studies tended to have many more participants. The median number of participants for surveys and log analysis was 307 and 110, respectively. However, some of these studies had over 10,000 participants.

We were interested in understanding to what extent underrepresented populations were studied in user-centered privacy and security studies and how researchers represented risk to them. For the purpose of this article, understudied populations include geographically rarely included

Table 5. Number of Papers that Include a Certain type of Participants (N = 284)

Type of participants	Frequency
Crowdsourcing (including MTurk, Prolific, Google Consumer survey, other crowdsourcing)	106
Non-representative convenience sample	79
Students (including Computer Science)	66
Users of specific technology	37
Employees (including university employees)	33
Experts (security experts and other experts)	22
Other	13
Special user groups (including people with impairments, children or teenagers, women in particular, LGBTQ)	12
Developers	10
Representative sample	7
Security Behavior Observatory (SBO)	4
Recruitment not mentioned	3

Note. One paper can include multiple types of participants.

populations (based on our sample from top-tier UPS venues), disabled persons, members of the LGBTQ+ community, certain age groups (older adults, children and teenagers), and any other special population that has not been widely studied. In total, 20 (7%) papers focused on these understudied populations, 8 of which include geographically understudied, 4 include people with disabilities, 4 papers include children, 2 include members of the LGTBQ+ community, 2 papers include survivors of intimate partner abuse.

4.1.6 IRB or Ethics Board Approval. Recently, some publication venues have started requiring that authors mention ethics board reviews for all papers with human-subjects studies. However, this was not commonly required during the time period in which the papers we reviewed were published. About two-thirds of the papers we analyzed discussed IRB or ethics board approval. 56% (n = 159) of papers stated they had obtained approval or received exempt approval, 35% (n = 99) did not mention whether they had approval, 4% (n = 10) of papers were from an institution without approval procedure. The remaining papers either described a corporate internal review process, or claimed to be exempt from needing approval (see Appendix, Table 20). A number of papers that describe research conducted in the United States stated that they were "approved exempt" or "received exempt approval." As this is actually a category of IRB approval in the United States that requires review by the IRB, we include these in the papers that received approval and distinguish them from those that are exempt from needing approval. From talking to some of the study authors who did not mention IRB or ethics board approval in their studies, we learned that they did actually receive approval but did not mention it in their papers. It is likely that the percentage of papers that received IRB or ethics board approval is actually higher than what we report based on the statements in the papers.

4.2 How Risk Is Represented and Measured

Many UPS studies focus on understanding how participants perceive security or privacy risk or how they use a tool or otherwise respond to a situation involving privacy or security risk. We were interested in how researchers represent risk to participants and how they approach the assessment of risk response. In addition, we investigated how risk was represented to understudied populations.

We categorized the way researchers represented risk to their participants. The categories included *simulated risk* (e.g., through the use of scenarios participants should imagine themselves in), *naturally occurring risk* (e.g., through observation or self-reported measures of naturally occurring behavior), *mentioned risk* (e.g., a questionnaire where participants were presented with hypothetical situations), or *no representation of risk*. In some cases, researchers using simulated risk in their studies did not inform participants about a scenario but instead used deception to make a simulated risk appear to be naturally occurring; we classify these as simulated risk.

The majority of papers used either naturally occurring or simulated risk. Certain study objectives coincided with certain types of risk representation. For example, experimental studies used mostly simulated risk, and descriptive studies used naturally occurring risk.

In addition, risk representation also varied by topics. For instance, studies on privacy transparency and choice mechanisms and studies on authentication mostly used simulated risk, while studies on access control, privacy-enhancing technologies, and security perceptions, attitudes, and behaviors used mostly naturally occurring risk. Response to risk was measured mostly through self-reported measures, either on their own or in combination with observed measures. A smaller proportion of papers relied on purely observational measures.

4.2.1 Risk Representation. We see that the vast majority of papers represent risk to participants in some way: 37% of papers used *naturally occurring risk*, 35% used *simulated risk*, 16% combined *multiple* approaches, 7% did not attempt to represent risk in any way to their participants, and only 6% *mentioned* risk to participants (Table 6).

It might seem surprising that there are studies that do not attempt to create a perception of privacy and security risk. But indeed, there were studies that focused solely on the instrumental aspects of usability of a privacy or security tool. Fuller et al. [2017] tested the usability of cryptographically protected search systems with participants who were not made aware of the privacy features. The authors evaluated participants' perception of the performance of the search system, rather than their perception of potential security and privacy risks. Others opted for evaluating users' perception of security practices. Oltrogge et al. [2015] conducted a survey with 45 developers for qualitative feedback on the implementation of TLS certificate pinning with the goal of creating a usable tool for implementing secure certificate validation. Chatterjee et al. [2016] had MTurk workers type leaked passwords under time pressure, yet without informing them about the security- and privacy-related rationale of the task. Lastly, some studies had participants talk about experiences without mentioning security or privacy, thus not creating any perception of risk. In all these studies, there was no attempt, indeed no need, to involve users in any security rationale and perception of risks.

4.2.2 Risk Representation by Study Objective and Method. Some risk representation approaches were frequently associated with particular study objectives, as shown in Figure 1. *Experimental* studies mostly use *simulated risk* (64%), *descriptive* studies frequently rely on *naturally occurring* (67%), while *relational* studies rely on *naturally occurring risk* (23%) and risk *combinations* (46%).

We observed that the approach to risk representation also varied considerably based on the type of study methods used, as detailed further.

4.2.2.1 Experiments. Most papers using only *experiments* used *simulated risk* (73%), as shown in Table 7, as this allowed researchers to introduce risk in a controlled way in all experimental treatments. However, 12% of experiments had *no representation of risk*. Indeed, some *experimental* studies had participants complete an *assigned security or privacy task*, yet with *no induced privacy and security risk* perception. In these cases, participants did not know that the task they were completing was privacy or security-relevant, and the authors did not intentionally create a perception

Table 6. Risk Representation and Examples (N=284)

	Frequency	Percent	Examples
Naturally occurring	105	37	A password reset email is sent to LinkedIn users, and its effectiveness is measured through an online survey of LinkedIn users [Huh et al. 2017]. Threat modeling is introduced in an enterprise setting, and its effectiveness is evaluated [Stevens et al. 2018].
Simulated	98	35	Participants in an online experiment are asked to imagine they are creating a password for an account they "care a lot about, such as their primary email account." [Ur et al. 2017] Developers are asked to roleplay and imagine they are responsible for creating the code for user registration and authentication of a social networking platform [Naiakshina et al. 2017a]. Participants are asked to test a banking prototype for one week and are led to believe that the objective was to test the usability of the application (deception). After some days, the authors simulate a phishing attack to test the effect of personalized security indicators [Marforio et al. 2016].
Multiple	45	16	Participants in an online survey self-report behaviors in updating workplace passwords (naturally occurring), and their attitudes toward four password-management behaviors (mentioned) [Habib et al. 2018].
None	19	7	Researchers analyze multiple gesture recognizers and evaluate them based on various security criteria, and use pre-existing datasets to verify how well their prototype of a new authentication system works [Liu et al. 2017]. Participants were asked to type sentences on phones provided to them by the researchers without knowing what the purpose was. The researchers used the data to understand the effect of participant movement on keystroke dynamics [Crawford and Ahmadzadeh 2017].
Mentioned	17	6	In an online survey, participants are first provided with a description of the "legalese" language, and are then asked to encode clauses of a privacy policy in legalese terms [Sen et al. 2014].

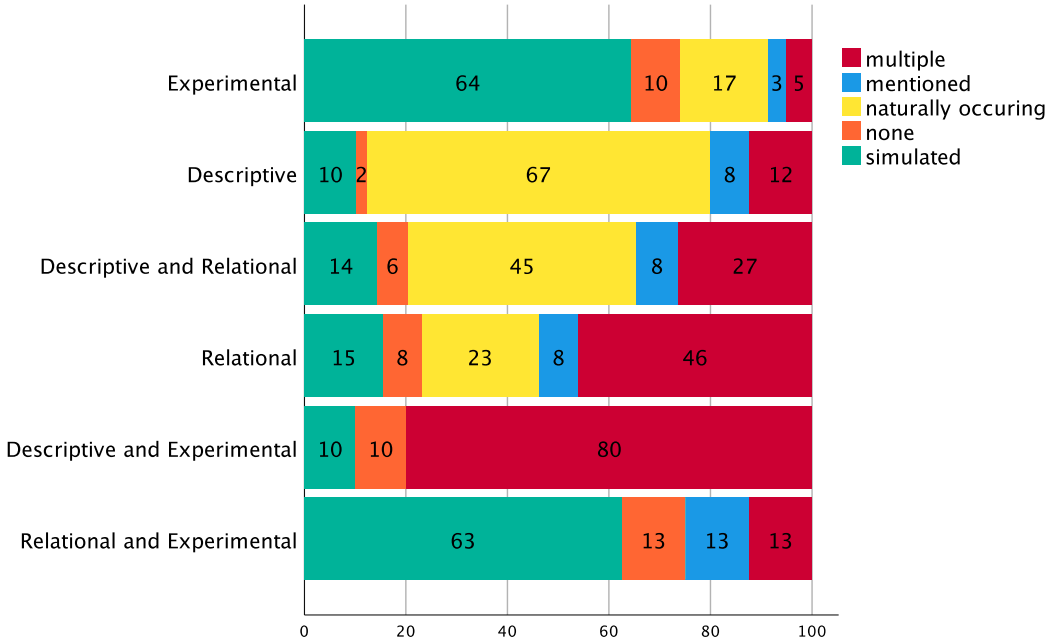

Fig. 1. Crosstab objective and risk representation (percentage of all studies with certain objective).

Table 7. Risk Representation for Experiments
(Papers that Only Use Experiments, $n = 99$)

	Frequency	Percent
Simulated	72	73
None	12	12
Naturally occurring	9	9
Multiple	3	3
Mentioned	3	3

of privacy and security risk. For instance, Shirvanian and Saxena [2014] asked their participants to read out checksums, but did not inform participants as to why this was necessary. Similarly, in another study, participants were asked to transcribe audio-captchas, but they were not aware of what they were doing [Meutzner et al. 2015].

One seemingly contradictory combination in the experimental category concerns a study with an *assigned security or privacy task*, yet *naturally occurring risk*. In this case, researchers collect and analyze participants' real passwords using semantic transformation, as well as their reasoning behind their habits [Hanamsagar et al. 2018]. Given that participants used their own passwords to connect to their real accounts, the risk was naturally occurring even though the login task was assigned to them by the researchers.

4.2.2.2 Surveys. Papers that used a survey-based approach, most frequently used *naturally occurring* risk or *multiple* risk representations, but were less likely to *mention* or *simulate* risk, as shown in Table 8.

Table 8. Risk Representation for Surveys
(Papers that Only Use Surveys, $n = 34$)

	Frequency	Percent
Naturally occurring	11	32
Multiple	9	27
Mentioned	5	15
Simulated	5	15
None	4	12

When risk is naturally occurring, participants are asked about real-world behaviors in actual situations. For example, Felt et al. [2016] surveyed Chrome users about existing security indicators. Similarly, Redmiles and colleagues [2016] investigate how users' security beliefs, knowledge, and demographics correlate with their sources of security advice, and how all these factors influence security behaviors.

Studies with mentioned risk do not ask participants about their own experiences and also do not involve a scenario to simulate risk. For example, Eiband et al. [2017] presented participants with a sketch depicting a person watching another person's screen to introduce the concept of shoulder surfing.

The difference between mentioned and naturally occurring risk can best be explained through a study that combines both mentioned and naturally occurring risk. Shay et al. [2014] for instance conducted a survey regarding account hijacking experiences. They first asked participants whether somebody had broken into one of their personal accounts. Participants who had experienced a compromise were asked about their experience (naturally occurring risk), those who had not yet experienced a compromise were asked to think about their primary personal email or social networking account throughout the survey. These participants were then asked about whom they were concerned might break into their accounts, how they thought accounts were compromised and other hypothetical questions. This second group was thus exposed to mentioned risk.

When risk was simulated in a survey study, a prototype or scenario was used. For instance, in a study by Karunakaran et al. [2018], participants were asked to imagine that they were victims of a data breach. This scenario simulated the risk.

In some cases, the survey was not situated in a privacy or security-relevant context for participants so we classify it as having no mention of risk. Oltrogge et al. [2015] for instance surveyed a sample of developers about their knowledge of certificate pinning, obstacles to pinning implementation, and how to help developers implement certificate pinning. Given that the questions concerned knowledge, obstacles, and wishes in general, there was no induced risk perception.

4.2.2.3 Interviews. Interviews most frequently used naturally occurring risk, as shown in Table 9, to investigate people's real-life privacy and security experiences.

For instance, Rashidi and colleagues [2018] interviewed undergraduates to understand their real-life privacy workarounds in the context of pervasive photography. Similarly, Ahmed and colleagues [2015] interviewed people with visual impairments about their real-life privacy concerns. In another study with naturally occurring risk, kids played with connected toys in a lab setting with their parents present. The parents were interviewed about their mental model of the toys, with questions about parental controls, privacy, and monitoring of what the child says to the toy. The children were interviewed about their mental model of the toy and privacy perceptions, asking them if they thought the toy could remember what they told it, if they would tell the toy a secret, and whether their parents could find out what they told the toy [McReynolds et al.

Table 9. Risk Representation for Interviews
(Studies that Only Use Interviews, $n = 36$)

	Frequency	Percent
Naturally occurring	26	72
Mentioned	4	11
Multiple	4	11
Simulated	2	6

2017]. Naturally occurring risk was also used by two studies exploring security and privacy in an organizational context. Conway et al. [2017] interviewed bank employees about organizational privacy and security practices, and Haney et al. [2018] interviewed employees in a company for cryptographic products.

Interview studies with simulated risk typically use scenarios to *simulate* risk. For example, Vaniea et al. [2014] use a set of hypothetical scenarios to elicit stories about software update experiences. In this study the interviewer asked participants to imagine how they would respond to scenarios such as being prompted to restart an internet browser mid-task or seeing that a large number of urgent Windows updates were available. Sometimes interview studies combined *naturally occurring* risk with *simulated* risk. In one study, participants had to create passwords for three hypothetical websites while thinking aloud (simulated risk), and were then interviewed about their strategies, as well as general habits related to password creation (naturally occurring risk) [Ur et al. 2015].

4.2.2.4 Log Analysis. 24 papers include the use of *log analysis,* and 4 papers use log analysis alone. Datalogs usually use *naturally occurring risk*. One study, for instance, created and deployed a privacy-preserving advertising platform. The authors report on the number of opt-in users and describe their behavior by analyzing usage logs [Reznichenko and Francis 2014].

4.2.2.5 Analysis of Existing Datasets. In total, 22 papers include the *analysis of existing datasets,* and 11 papers use the analysis of existing datasets alone. The analyses of datasets usually use *naturally occurring risk*. One example is a study where researchers study the reaction to news articles by analyzing public comments [Fiesler and Hallinan, 2018].

4.2.2.6 Rarely Used Methods. Our dataset includes eight *experience sampling* studies. Seven of the experience sampling studies used *naturally occurring risk*, and one used *simulated* risk. As an example for naturally occurring risk, Reeder et al. [2018] conducted an experience sampling study investigating people's reaction to web browser security warnings where they surveyed users in-situ (after being exposed to a warning) to understand their reasons for adhering or not to real warnings. Yang et al. [2016] conducted an experience sampling study using *simulated* risk. Participants were alerted multiple times a day to complete password creation or recall tasks. The passwords were for accounts that were used purely for the study, thus simulating the risk to participants.

Our dataset includes six *focus group* studies, five of which combine focus groups with other methods. The majority of these papers use *naturally occurring risk*. For instance, Sambasivan et al. [2018] conducted focus groups with 199 women from India, Pakistan, and Bangladesh focused on understanding how women perceive, manage, and control their personal privacy on shared phones. The authors identified five performative practices that participants employed to maintain individuality and privacy.

Our dataset also includes three *diary* studies *all in combination with other methods*, which use *naturally occurring or combinations of risk*. For example, Mare et al. [2016] gave participants smartwatches to log any authentication events as they went about their daily lives as part of a digital diary study.

We examined only three studies that used *workshops*, all in combination with other methods. The studies mostly *combined risk representation*. For example, Pearson et al. [2017] conducted workshops in which they presented design probes to explore the notion of "chameleon devices," mobile devices that blend into their background with the objective of making them more secure and private.

We examined two each of *vignette studies, list experiments*, and *co-creation studies*. The vignette studies were experimental studies that used *simulated risk*. For example, Votipka et al. [2018] conducted a *vignette* study to investigate user comfort level with resource accesses that happen in the background, without any visual indication of resource use. They find that both when and why a resource is accessed influences user comfort. The list experiments *mentioned risk* to participants. For example, Usmani et al. [2017] conducted a list experiment to investigate the prevalence of social insider attacks, where attackers know their victims, and gain access to their account through directly using their device. The list experiment method allowed the authors to explore the sensitive topic of social insider attacks, finding that an estimated 24% of participants had perpetrated social insider attacks.

The *co-creation* studies used *simulated risk* and *naturally occurring risk*. For example, Egelman et al. [2015] created and evaluated a set of privacy indicators for ubiquitous sensing platforms. Using a crowdsourcing approach, they collected 238 sketches from participants based on 14 ubiquitous sensing concepts to understand how end users conceptualize the concepts. Using the themes identified in participants' sketches, the researchers then created icons for each concept and evaluated their comprehension rate in comparison to icons created by a designer. The icon sets performed similarly well at conveying the privacy concepts, with certain crowdsourced icons even outperforming designer-made icons.

4.2.3 Risk Representation by Topic. Depending on the topic being studied, risk representation varied, as shown in Figure 2. Studies on *privacy transparency and choice mechanisms* and *authentication* mostly used simulated risk (57%). Studies on *access control, privacy-enhancing technologies,* and *security perceptions, attitudes and behaviors* applied mostly *naturally occurring risk* (50%, 64%, and 52%, respectively).

Not surprisingly, studies on the topics that were mostly studied experimentally (Figure 2), such as *authentication, encryption, privacy transparency and choice mechanisms,* and *social engineering*, were more likely to use *simulated risk*, which is induced through the use of an experimental setup. On the other hand, studies on topics that frequently had descriptive objectives often applied *naturally occurring* or *mentioned* risk since descriptive methods usually offer less opportunity for risk simulation and are better suited to evaluate real-life risks or mentioned risks using methods such as interviews or surveys.

Additional analysis by topic, beyond risk representation, can be found in the appendix.

4.2.4 How Response to Risk Is Measured. We categorize papers based on their approach to collecting data about how participants perceive and respond to risks. We analyze whether papers use self-reported measures, observed measures, or combine both for data collection (Table 10).

4.2.5 Understudied Populations and Risk Representation and Measurement. 20 (7%) papers focused on understudied populations (Section 4.1.5.). In terms of risk representation, 13 of these 20 papers used naturally occurring risk, 4 used multiple risk representations, 2 mentioned risk to

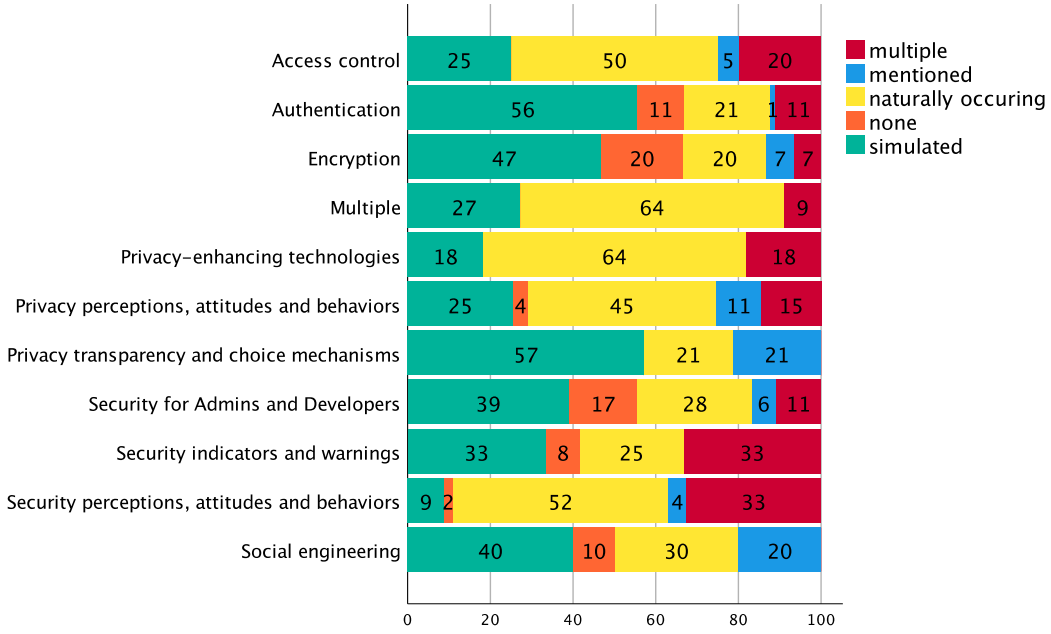

Fig. 2. Crosstab topic and risk representation (percentage of all papers in topic).

Table 10. Approach to Collecting Data Regarding Risk Response (N = 284)

	Frequency	Percent	Examples
Both observation and self-report	131	46	Authors combine semi-structured interviews and observe participants' use of a login procedure [Holz and Bentley 2016]. The researchers observed the participants and also asked questions about the authentication process. Participants were asked to create passwords in a lab setting, and were then interviewed about their process [Ur et al. 2015]. Participants were observed and self-reported their experience.
Self-report	119	42	Researchers use a combination of interviews and a survey to identify privacy panic situations [Angulo and Ortlieb 2015]. Researchers conduct interviews with visually impaired participants to understand their privacy concerns and techniques to protect their privacy [Ahmed et al. 2015].
Observation	34	12	Researchers evaluate HTTPS adoption from a user perspective by collecting aggregate user metrics from major web browsers [Felt et al. 2017]. Researchers analyze Twitter data to understand longitudinal exposure and withdrawal of socially shared data [Mondal et al. 2016].

research participants and 1 used no representation of risk. In line with this observation, 17 of these 20 studies had a *descriptive* or *descriptive and relational* objective and mostly used methods like *interviews* or *surveys*. Only 2 studies had an *experimental* objective ([Lastdrager et al. 2017] and [Qahtani et al. 2018]), suggesting that it is rare to assess the suitability of security and privacy tools for these populations. Lastdrager and colleagues [2017] study the effectiveness of anti-phishing training for children, and asked pupils to distinguish phishing emails from non-phishing emails after receiving training to recognize phishing. Qahtani et al. [2018] studied the effectiveness of fear appeals for the smartphone locking behavior of Saudi-Arabians, while also highlighting some of the methodological challenges of conducting research in Saudi-Arabia. Note that, while these populations are rarely included in the papers of our analysis, our sample of papers did not include specialized conferences that focus specifically on these populations. Nevertheless, there seems to be a research gap of these population groups at the venues we studied.

In comparison with these results, papers involving crowdworkers mostly relied on simulated risk, and, less frequently on naturally occurring or combinations of risk.

4.3 Tools for Risk Representation and Measurement

In this section, we describe and analyze the "tools" UPS researchers use to represent risk to participants, focusing on assigned tasks, prototypes, scenarios, deception, educational interventions or incentives for secure behavior. We describe the characteristics and trends that were associated with these tools.

The majority of papers either used no assigned task or a security/privacy related task. Approximately a third of papers involved a prototype and almost a third of papers involved a scenario. The use of deception was relatively rare in our sample and was associated with experimental studies and simulated risk. Only a small number (approximately 2%) of papers used educational interventions or incentives for secure behavior.

4.3.1 Assigned Tasks. We categorize papers based on the tasks (if any) that are assigned to research participants. We analyze whether papers assign tasks to participants that are relevant to *security or privacy*, whether the tasks are *unrelated* to security or privacy, or whether *both* security and privacy related and unrelated tasks are used, as shown in Table 11. Not surprisingly, most papers that included an assigned task, assigned a task related to security or privacy. Sometimes participants were assigned tasks unrelated to security or privacy, often so that researchers could observe routine or incidental security tasks that participants had to perform as part of completing the assigned task without focusing participants' attention on security or privacy.

4.3.2 Studies Including Prototypes. In total, 90 (32%) studies included a *prototype*, that is a new solution such as a textual message, an icon, or an interface that the authors present to participants, sometimes in a low-fidelity or non-interactive form. The risk representation of studies involving prototypes was usually simulated (54%) or naturally occurring (20%). 16% of the studies combined multiple ways of risk representation. For example, Vaziripour et al. [2018] made changes to the authentication ceremony in the secure messaging app Signal and evaluated the effect of these changes in a between-subjects experiment. Harbach et al. [2014] explored the effect of novel personalized security decision dialogues on the Android app installation process. Overall, studies involving prototypes follow the overall trends in the data with mainly convenience samples, many experimental and descriptive studies, and a high percentage of authentication papers. It is interesting to note that while prototypes appear in about a third of UPS studies we analyzed, the majority of the studies in our sample did not include prototypes, suggesting a focus on understanding user perceptions, attitudes, and behaviors as they relate to general concepts or to existing systems rather than proposed new solutions.

Table 11. Assigned Tasks that Participants Were Asked to Complete (N = 284)

	Frequency	Percent	Examples
No assigned task	137	48	Participants' datalogs are collected over multiple months to understand realistic security and privacy behaviors, and some of these participants are invited to participate in interviews [Forget et al. 2016]. In an experience sampling study, participants answer to in-situ surveys throughout the day in response to certain trigger events related to permission settings [Bonné et al. 2017].
Security/privacy related task	102	36	Novice participants are instructed to attempt to send encrypted email [Ruoti et al. 2016]. Participants are asked to use a new authentication method [Das et al. 2017].
Unrelated task	27	10	Participants are asked to perform unrelated tasks on an email platform. They are asked to find a certain email, schedule a calendar appointment and look up a contact. These tasks served as a reason for participants to login to the email platform, which was the interaction researchers were interested in [Holz and Bentley 2016]. Children are invited to the lab and asked to play with connected toys [McReynolds et al. 2017].
Both security/ privacy related and unrelated task	18	6	Participants are asked to complete unrelated tasks that represent common smartphone activities (text entry activity, email reading activity). During the unrelated tasks, the participants were triggered mid-task to re-authenticate (S/P task) [Agarwal et al. 2016]. In a developer study, participants were asked to complete coding tasks. One did not have direct security implications (URL shortener), and is thus an unrelated task. The other tasks had direct security implications (credential storage, string encryption) and are thus security-relevant tasks [Acar et al. 2017].

Studies involving prototypes were usually experimental (70%) or descriptive (16%). Risk response assessment for prototype studies usually included both self-reported and observed measures (68%), while 18% used self-reported measures alone and 14% used observed measures alone. Most studies that include prototypes study topics related to authentication (37%) or privacy perceptions, attitudes and behaviors (12%). In total, 58% of papers including prototypes asked participants to complete security or privacy related tasks, 16% assigned no specific tasks, and 13% asked participants to complete unrelated tasks, 13% asked participants to complete a combination of security and privacy-related and unrelated tasks.

4.3.3 Studies Including Scenarios. In total, 75 (26%) studies included *scenarios* in which the researchers asked participants to imagine themselves being in a certain situation. Studies involving scenarios mostly used simulated risk representation (64%) or combined multiple ways of representing risk (25%). A smaller proportion of the studies also used naturally occurring risk (9%). Several studies asked participants to imagine that their email account had been compromised and that they were asked to change the password (Komanduri et al. [2014], Melicher et al. [2016], Segreti et al. [2017], Shay et al. [2015], and Ur et al. [2017]). Another study asked participants to play the role of managers responsible for access review in an organization [Jaferian et al. 2014]. Hang et al. [2015] recruited participants in pairs who had a close relationship for a study on a secure fallback authentication scheme. The authors asked participants to engage in a roleplay, instructing one of them to play an adversary, whereas the other participant played a legitimate user.

Overall, scenarios were used by researchers as an easy way to simulate risk in a wide variety of research settings. For instance, scenarios were used in a lab setting by asking participants to roleplay and attempt to send each other a fictitious credit card number via secure messaging [Vaziripour et al. 2017]. Scenarios could also be used in a survey or interview setting to introduce hypothetical scenarios that participants should situate themselves in, as used for instance in one study that presented interview participants with hypothetical scenarios related to software updates, in combination with probing questions [Wash et al. 2014].

4.3.4 Studies Including an Educational Intervention. Seven papers included an *educational intervention*. Wash and Cooper [2018] attempt to educate their participants on how to detect phishing attempts, and Lastdrager et al. [2017] evaluate the effectiveness of anti-phishing training for children. Stevens et al. [2018] describe the effects of introducing staff of a digital defense organization to threat modelling, and Warshaw et al. [2016] use teaching sessions in an effort to improve adults' inference literacy (i.e., the beliefs and misconceptions people have about how companies collect and make inferences from their data). Two papers use informational videos to educate participants about smartphone locking ([Albayram et al. 2017] and [Qahtani et al. 2018]). In one paper, participants first took part in a user test including three secure email systems, and in the post-study interview, the researchers described the actual security model of the system to the participants. After hearing these descriptions, participants were asked whether their opinions regarding any of the systems had changed [Ruoti et al. 2018].

Almost all of the papers including an educational intervention had an *experimental* objective (86%). In terms of risk representation, there was no clear tendency: two papers *mentioned* risk to research participants, two used *naturally occurring* risk and two *simulated* risk to their participants.

Although educational interventions were fairly rare, we observed (but did not code) a number of papers that included nudges or small interventions as part of a prototype tool. For example, several papers ([Shay et al. 2015] and [Ur et al. 2017]) included interventions that provided feedback to users on password strength as part of a password-creation interface. These sorts of integrated interventions may be easier to deploy and more likely to be seen by users than a training program or educational video.

4.3.5 Studies Including Incentives for Secure Behavior. The field of behavioral economics frequently uses financial incentives to model real world incentives in an experimental setting. However, these seem to be relatively rare in UPS studies. Only five (2%) papers included financial incentives for secure behavior, usually with the intent of motivating participants to try to perform well on a security or privacy-related task that was part of the study. Indeed, all five papers included an *assigned security or privacy-relevant task*. Three of the papers used *simulated* risk, one used *combinations* of risk representation, and one did not attempt to simulate risk. All of the studies had an

experimental objective, and four of them also included a *scenario* participants should situate themselves in. Two of the papers included *prototypes*. Four of the papers were related to *authentication* and one to *encryption*. None of these papers involved *deception*.

These studies all encouraged participants to try to perform well on their assigned security or privacy task by making a part of their compensation contingent upon successful completion. In the absence of this compensation structure, participants might not be motivated to try to perform the assigned task well as there would be no consequences for poor performance. A user who performs a real-world security task poorly risks a security-related consequence (e.g., an account compromise) or an inconvenience (e.g., having to reset a forgotten password). A financial incentive provides a substitute risk to participants: the risk of losing the contingent compensation. For instance, Vaziripour et al. [2018] asked participants to complete the authentication ceremony of a secure messaging app, which they had attempted to improve for better usability. Participants received a base pay of $7, and they could receive a bonus of $3 if they managed to perform the task safely. Tan et al. [2017] conducted an online between-subjects experiment in which participants were asked to imagine they were an accountant who was requesting social security numbers from employees using a secure messaging system. Participants were asked to do a fingerprint verification for each request and researchers were interested in whether participants noticed mismatched fingerprints. As an incentive for fast and accurate performance, participants were told that the fastest 15% of participants who performed the task correctly would receive a $1 bonus in addition to a base compensation of $3. Similarly, Huh et al. [2015] simulated the PIN setup page of a made-up bank, informing participants that they would use the PIN for card purchases. Each participant was assigned a specific technique for memorizing the PIN, and they received an incentive of $0.25 if they came back later, and an additional $0.25 if they were able to remember the PIN.

Some security economics papers go a step further and use financial incentives as part of a model of participants' valuation of privacy or security protections. For example, in a UPS paper that was published at a conference on economics and computation (thus, not in our sample), Redmiles et al. [2018] gave participants a small deposit into an online account and offered them the opportunity to add two-factor authentication (2FA) to their account. They were told the probability that their account would be hacked and they would lose the balance, both with and without 2FA enabled. The researchers varied these probabilities across the experimental treatments and were able to observe whether participants made rational decisions about whether it was worth their time to enable 2FA.

4.3.6 Studies Involving Deception. Sixteen papers included *deception*, comprising between 4% and 10% of UPS papers from each of the five venues we studied. In five additional papers, authors referred to their own protocols as deceptive; however, we instead categorized them as *partial disclosure* to avoid priming participants to think about privacy and security specifically. For instance, one of the studies split software developers into two groups, which they refer to as "non-priming" and "priming." The non-priming group was told that the study was about API usability, whereas the priming group was told that the study was about secure password storage. While the researchers called this deception, we are not considering studies that simply avoid priming participants as deception studies [Naiakshina et al. 2017b].

We first provide an overview of the characteristics of deception studies, before going into the details of how exactly participants were deceived and for which objectives. Ten papers describe a *debriefing procedure* that exposes the deception to participants at the end of the study, the remaining six papers do not describe any participant debriefing.

Table 12. Assigned Tasks in Deception Studies (n = 16)

	Frequency	Percent	Percentage of papers in that category that use deception
No assigned task	6	38	4
Unrelated task	6	38	22
Both security/privacy related and unrelated task	4	25	22
Security/privacy-related task	0	0	0

Table 13. Risk Representation of Deception Studies (n = 16)

	Frequency	Percent	Percentage of papers with each risk representation that use deception
Simulated	11	69	11
Naturally occurring	4	25	4
Multiple	1	6	2
None	0	0	0
Mentioned	0	0	0

Table 14. Types of Deception Studies and Examples (n = 16)

	Frequency	Percent	Examples
Deception about the objective of the study	8	50	Anderson et al. [2015], Marforio et al. [2016]
Deception about the presence of risk	4	25	Rashtian et al. [2014], Samat and Acquisti [2017]
Lack of consent (deception about study participation)	4	25	Han et al. [2016], Hu and Wang [2018], and Wash and Cooper [2018]

Papers involving deception typically mentioned IRB approval: 11 studies had *IRB or ethics board approval*, three studies were from an *institution without approval procedure*, one study went through an *ethics review in industry*, and one study *did not mention IRB-related information*.

The use of deception was highest for papers on *social engineering* (30% of papers in this category used deception) and *privacy transparency and choice mechanisms* (21% of papers in this category used deception). More information on the topics of deception studies can be found in Appendix, Table 21. In total, 75% (n = 12) of deception papers had an *experimental* objective, and measurements for papers including deception often combined *observed and self-reported measures* (50%). As shown in Table 12, most deception studies were framed to avoid focusing participants on security or privacy tasks and most did *not use assigned tasks* (38%) or assigned only *unrelated tasks* (38%). As shown in Table 13, the majority (69%) of papers involving deception used *simulated risk*. In terms of study methods, the majority of papers involving deception were based on an *experiment* (69%, see Appendix, Table 22).

We categorized *deception* papers according to the *type of deception* they used in their study, as shown in Table 14. Sixteen were papers coded as deceptive, eight of these included *deception about*

the objective of the study, four papers deceived participants about the *presence of risk* and in four papers, participants were *not aware they were participating in a study* (lack of consent).

A frequent approach was to deceive participants about the *objective of the study*. For instance, Marforio et al. [2016] instructed participants to use their online banking prototype for one week and simulated a phishing attack on the prototype on the fourth day. Similarly, in another study participants were led to believe that their objective was to evaluate browser extensions, but the authors performed a man-in-the-middle attack to spoof Google search results to ensure that only the experimental extensions were installed [Anderson et al. 2015]. Participants were asked to find 20 weather extensions within the spoofed search results and evaluate their usability and aesthetics. Three of the manipulated search results at random present an unreasonable permission warning. The control group received conventional warnings that did not change their appearance, the treatment group received polymorphic warnings. The objective was to understand whether polymorphic warnings performed better at encouraging secure behavior.

Some papers deceived participants about the *presence of risk*. For example, Samat and Acquisti [2017] told participants that their information would be shared with a specified audience; however, the data was not shared with anyone outside the primary researchers of the study. In another study, the researchers sent Facebook friend requests from a "fake" account to participants before an interview study. They then confronted participants with inconsistencies in their self-reported interview answers and their observed reactions to the friend request [Rashtian et al. 2014]. This study deceived participants because they did not know that the friend request was part of the study. In addition, participants had not consented to a friend request being sent on Facebook as part of the study.

All four papers that did not obtain *consent* from their participants (shown in Table 15) were situated in the context of social engineering and three of them focused on attacks via email such as phishing or email spoofing ([Han et al. 2016], [Hu and Wang 2018], and [Wash and Cooper 2018]). Two of these papers included real attackers among their non-consenting participants. The study of attackers in human-subjects experiments raises ethical issues that may warrant further exploration.

Han et al. [2016] leveraged a web honeypot to attract real attackers into installing phishing kits in a compromised web application. They then presented a sandbox designed to neutralize a phishing kit while keeping it functional. The approach was designed to preserve the victim's privacy, without interfering with the attack process in order to make sure that attackers can compromise the honeypot, install phishing kits, and conduct functional tests without being alerted about the sandbox configuration. The researchers collected and analyzed data from two-categories of unwitting study participants: attackers and victims. The study was conducted at a company and received approval from the company's legal department but not an ethics board. The authors do not describe a debriefing procedure.

Wash and Cooper [2018] sent four simulated phishing emails to university employees over a 30-day period. The employees did not know they were participating in a study. The first phishing email led to an education page where participants were educated about phishing. The authors tested the effectiveness of text variants that explained phishing to potential phishing victims. The study was IRB-approved and the authors discuss ethics. They explain that they did not obtain informed consent to avoid biasing the participants' response to a phishing email. In addition, they did not debrief participants to prevent participants from thinking that all future phishing attempts are part of a research study. In contrast, a 2009 phishing study that also involved sending simulated phishing emails to a university community took a different approach, first recruiting participants for a study advertised as helping protect the university from identity theft, and later debriefing participants via email [Kumaraguru et al. 2009].

Table 15. Papers that Did Not Obtain Informed Consent Before the Study Began

	IRB or ethics board approval	Ethics discussed	Participant debriefing	Mention of "deception"
PhishEye: Live monitoring of sandboxed phishing kits [Han et al. 2016]	No (Institution without approval procedure)	Yes	No	No
Who provides phishing training? Facts, stories, and people like me [Wash and Cooper 2018]	Yes	Yes	No	No
End-to-end measurements of email spoofing attacks [Hu and Wang 2018]	Yes	Yes	Yes	Yes
Using chatbots against voice spam: Analyzing Lenny's effectiveness [Sahin et al. 2017]	Not mentioned	No	No (used existing dataset)	No

Hu and Wang [2018] describe a study in which participants took part in an online survey on their email usage. Participants were led to believe that this was the entire survey and they were done participating. However, 10 days later the participants were sent a spoofed email impersonating MTurk technical support. After the study, they were sent a debriefing email, which explained the true purpose of the experiment and obtained informed consent retroactively. The study received IRB approval. We classified this as lack of consent as the participants had not consented at the time of their participation.

Sahin et al. [2017] tried to understand why a phonebot ("Lenny") was so successful in dealing with spam calls. They used a publicly available dataset of calls where spammers were deceived into thinking they were talking to a human, when in reality, they were talking to the phonebot. The study was conducted by a company and was not reviewed by an ethics board. Spammers were not debriefed either in this study or when the calls were originally recorded.

5 DISCUSSION

Our discussion focuses on four observations from our study. First, we discuss the choice of methods in our sample and how they correlated with certain types of risk representation. We also point to some methods that were rarely used in the papers we reviewed that may have advantages for

UPS studies. Second, we discuss participant recruitment, including populations that appear to be understudied, and how risk was represented to them. Third, we discuss ethical issues faced in UPS studies, especially those involving deception or involving attackers as human subjects. Finally, we suggest guidelines for the reporting of empirical UPS studies, and propose a structure for their systematic categorization, with a focus on risk representation.

5.1 Choice of Methods and Risk Representation

One of our research objectives was to explore how researchers navigate the tension between realistic exposure to risk and ethical, legal, and practical considerations. Overall, the choice of method usually coincided with certain types of risk representation. When picking a method, researchers will thus often face tradeoffs with regards to the risk representation they can possibly use in their study design. This review can make such tradeoffs more explicit so that researchers can choose accordingly. For instance, experimental studies and simulated risk often coincided, whereas descriptive studies often relied on naturally occurring or mentioned risk. Experimental setups lend themselves to simulating risky situations, for instance through the use of scenarios and prototypes that allow participants to situate themselves in a risky situation. On the other hand, descriptive studies frequently employ methods such as interviews or surveys, which offer less opportunity for risk simulation, but are highly suitable to study real-life risks or mention risky situations.

When measuring the response to risk, researchers frequently used self-reported measures alone or in combination with observed measures. One might think that a combination of self-reported and observed measures would always be the best choice, but the studies in our sample that used self-reported measures clearly focused on subjective perceptions, and did not have the objective of evaluating behavior. In these cases, self-reported measures were most suitable and least intrusive, for instance when understanding privacy panic situations [Angulo and Ortlieb 2015] or evaluating people's privacy concerns and strategies they use to mitigate these concerns [Ahmed et al. 2015].

Naturally occurring risk was frequently used in self-report studies, for instance in a survey on sources of security advice and behaviors [Redmiles et al. 2016]. Using self-report measures in studies involving naturally occurring risk can be a good option, as it minimizes logistical issues and allows participants to control what information they share with researchers. However, participants do not always self-report information accurately for a variety of reasons (e.g., social desirability bias, inaccurate memory). Direct observation of risk response usually offers the most accurate way to observe participants' responses to naturally occurring risk, but depending on the data being collected, it may pose logistical challenges. A study on private-by-design advertising [Reznichenko and Francis 2014] for instance built a functional prototype of a privacy-preserving ad system, and ran into the challenge of incentivizing potential users to install the prototype on a large scale. They deployed their prototype by bundling it with a popular Firefox add-on that allows viewing documents (e.g., doc, ppt) in the browser without downloading them. Users updating this browser extension were asked whether they wanted to join the experiment, allowing the researchers to collect a large dataset using naturally occurring risk. Felt et al. [2017] used telemetry data from Google Chrome and Mozilla Firefox, which provides user metrics from a subset of users who opted in (for Firefox) or did not opt out (for Google Chrome) to understand the state of HTTPS adoption. As the users were using their browsers to carry out their real-life activities, the risk in this study was naturally occuring. Dunphy et al. [2015] used the Twitter Search API to collect "#password" tweets or the keyword "password," in combination with pronouns and possessive pronouns, to ensure that the data was connected to personal experiences. They collected 500,000 publicly available tweets, which they analyzed qualitatively. As the dataset was public and twitter users freely shared their thoughts on passwords, risk was naturally occurring.

Using *simulated* risk is often a good option when using participants' real accounts could be too invasive, for instance when the researchers would be able to see participants' real passwords, email inboxes, or bank account balances. Simulated risk was often induced through the use of scenarios, for instance by Ur et al. [2017], who asked participants to imagine they are creating a password for an account they "care a lot about, such as their primary email account." Another example where simulating risk is necessary is when the phenomenon of interest doesn't often occur naturally or involves a prototype that has not yet been deployed. An example is a developer-centered study by Naiakshina et al. [2017b], who asked a group of student developers who received a carefully designed set of instructions to imagine they were responsible for creating the user registration and authentication of a social networking platform. The authors told half of the participants that the study was about the usability of Java frameworks, while priming the other half by telling them that the study was about secure password storage. By situating all of the participants in the same context, and only varying the task instructions, the researchers were able to isolate the effect of priming participants to think about security, demonstrating the advantage of simulated risk representation.

Mentioned risk was used rarely in our dataset. One example is a study evaluating the effectiveness of anti-phishing training with children. The authors first provided cybersecurity training for the children on a variety of security topics (e.g., phishing, hacking, cyberbullying). They then evaluated the ability of the children to detect phishing attempts. The authors did not create a scenario for the children and asked them to imagine a situation where they might be led to distinguish the legitimacy, but instead introduced the task as a "cybersecurity test," asking them to decide whether or not "action should be taken" [Lastdrager et al. 2017]. If possible, in terms of risk representation, it seems preferable to attempt to simulate risk to research participants, which may explain that mentioned risk was comparatively rare. Simulating risks can help participants situate themselves in a hypothetical situation (e.g., through the use of scenarios, as described above), allowing them to comment on real-life motivations or obstacles that may play a role if they were exposed to the scenario in everyday life. In addition, simulating risks can feel more engaging for research participants, thus potentially leading to more in-depth insights.

Finally, a small number of studies used *no representation of risk*. These studies mostly focused on evaluating the usability of a prototype such as gesture recognizers [Liu et al. 2017] or keystroke dynamics [Crawford and Ahmadzadeh, 2017]. While these prototypes are components of authentication systems, these studies focused only on evaluating usability of the prototypes on their own, without providing the context to participants and without any mention of risk. Nonetheless, it might still be relevant to simulate risk as it could impact participants' motivations to complete tasks correctly.

In our sample, researchers creatively combined a variety of tools aimed at helping participants perceive risk, ranging from scenarios and deception to incentives for secure behavior. Educational interventions were tested, and prototypes were frequently used to create relatively realistic interactions for participants.

One takeaway from our analysis is that, while prototypes appear in about a third of the studies we analyzed, the majority of studies did not include prototypes. This might suggest a focus on understanding user perceptions, attitudes and behaviors in terms of general concepts or existing systems, rather than proposing and testing new solutions. Research that does not involve prototypes is often used to explore and define the problem space, as for example by Matthews et al. [2017], who studied privacy and security practices of survivors of intimate partner abuse. Exploring and defining a privacy- and security-related problem space holds much value, without necessarily proposing a new solution in the same paper. Exploratory UPS papers may eventually

be followed-up with proposed solutions, either by the same authors or by others inspired by the exploratory paper.

Prototypes can also be a valuable tool even in more exploratory phases of research. Most studies involving a prototype in our sample had an experimental objective, but prototypes can be useful in combination with a variety of methods going beyond experiments. A prototype could for instance also be used to enhance the discussion in focus groups or interviews, or a deliberately imperfect prototype could serve as a basis that participants build upon in co-creation methods. Low-fidelity prototypes can be helpful to solicit more fundamental feedback on a scenario than a functional interface. Prototypes can also help participants situate themselves in hypothetical security or privacy-critical situations and make them seem more concrete, thus allowing researchers to explore participant reactions to the prototype as an artefact. Overall, the value of a prototype is also enhanced by the process that led up to its creation; user-centered approaches and extended pilot testing can improve the quality of the prototype that is ultimately exposed to research participants. The description of how prototypes and other tools were used in Section 4.3. can provide inspiration for researchers planning UPS user studies.

Most of the papers we surveyed adopt traditional study methods: interview, experiment, and surveys. Methods such as focus groups, diary studies, vignette studies, list experiments, co-creation methods, and workshops were used only rarely. UPS studies, in this regard, do not diverge much from trends in HCI, where the same set of methods are most prevalent ([Caine 2016] and [Pettersson et al. 2018]). Research on how to adapt a larger variety of HCI and design methods to the UPS field would help broaden the methodological spectrum currently used.

Some of the methods that do not occur frequently in our sample may nonetheless be useful to the UPS community and could hold potential for novel approaches to represent and measure risk. *Diary methods*, for instance, could help provide longitudinal insights into how participants perceive security or privacy risks over a longer time period. The method could be used for naturally occurring risks, but researchers might also equip participants with a new technology for the duration of the study and explore their long-term perceptions of security and privacy risks. Co-creation/participatory design and group methods can also hold advantages for use in UPS studies, we will consider these in the next two subsections.

5.1.1 Co-creation and Participatory Design Methods. Methods including co-creation could help end users make an active contribution to the creation of effective privacy and security mechanisms and for instance help design more user-centered descriptions of privacy and security concepts. Such methods can hold value for UPS, in particular when the objective is to elicit and unveil user needs throughout the activity. Note that the creation of a final solution is usually not the objective of participatory or co-creative design methods. Quite frequently, participants are asked to create prototypes "in order for participants to gain knowledge for critical reflection, and provide users with concrete experience of the future design in order for them to specify demands for it" [Hansen et al. 2019]. In terms of risk representation, co-design and participatory design activities can help users reflect and build upon the security and privacy risks that naturally occur in their lives, and contribute ideas leading to potential solutions. Going beyond naturally occurring risk, co-design and participatory design can also simulate or mention new risky situations to participants, helping researchers understand participant thought processes when exposed to risks.

Two papers in our sample used a form of co-creation. Egelman et al. [2015] asked crowdworkers to design icons to communicate what type of data devices with recording capabilities were currently recording. Adams et al. [2018] conducted a co-design study with Virtual Reality (VR) developers who were asked to contribute to a VR code of ethics on a shared online document.

5.1.2 Group Methods. Few studies used group methods such as workshops and focus groups. However, workshops and focus groups hold the potential of gathering qualitative in-depth insights into privacy and security attitudes that might help the community obtain even richer results. In comparison to interviews, which are already frequently used, such group activities allow researchers to confront and contrast different privacy and security attitudes and behaviors. By confronting various attitudes and behaviors, participants also naturally explain contradictions in their behavior and attitudes. This study method can help reveal how participants perceive naturally occurring risks and how they weigh advantages and disadvantages. Group methods are not limited to naturally occurring risks, however, they can also mention or simulate novel or futuristic risk situations. One could also imagine participants acting out scenarios with security or privacy risks in the group. Group methods can also provide insights on topics where users' attitudes seemingly contradict their behavior. Recent studies have used group methods in this way to understand privacy trade-offs better ([Distler et al. 2020] and [Rainie and Duggan 2015]). These examples used multiple scenarios of potential privacy tradeoffs that focus group participants should imagine themselves confronted with, for instance the possibility of using a smart thermostat that shares their data with undefined parties online. Focus group participants first noted advantages and shortcomings individually, and then discussed and confronted their opinions in the group setting.

Examples in our sample included [Sambasivan et al. 2018] who conducted focus groups with women in South Asian countries to explore performative practices used to maintain individuality and privacy in contexts where devices were frequently borrowed and monitored by their social relations. Another paper used focus groups to understand how abusers in intimate partner violence exploit technology in order to gain a better understanding of threat models in this context and find mitigation strategies for such attacks [Freed et al. 2018].

5.2 Participant Recruitment and Risk Representation

One remarkable observation within our sample showed that researchers often rely on easily accessible populations (e.g., MTurkers, convenience samples, students). While this is understandable from a researcher's point of view, including a more diverse set of research participants holds value, since minority groups often face specific risks in their daily life. Here, we discuss some approaches to including more understudied groups and provide some observations on the use of crowdworkers as UPS study participants.

5.2.1 Understudied Groups. Within our sample of papers in top-tier security and privacy conferences that did not include special interest venues for these populations, non-Western populations, disabled persons, members of the LGBTQ+ community, and certain age groups (older adults, children, and teenagers) were rarely studied. When these groups were included, they mostly participated in descriptive (rather than experimental) studies, such as interviews or surveys. Accordingly, risk representation was mostly based on naturally occurring risk. This means that most security or privacy tools are likely not tested by members of these groups, who might have special needs and thus might not be able to take advantage of their privacy and security-enhancing properties. They could also perceive or react to risks differently, further underlining the importance of including these groups. Other researchers may build on these results by striving to include understudied groups at all steps of research and design, including exploration, generation of ideas, and iterating on the development of prototypes and final tools. At SOUPS 2020, Fanelle et al. [2020] presented one such experimental study in which people with visual impairments tested the usability of audio captchas.

In addition to traditional recruitment approaches, such as contacting communities of understudied groups directly, using crowdsourcing solutions might hold potential. As the filtering options on crowdsourcing platforms such as Prolific Academic continue to become more fine-grained, researchers might take advantage of these filtering options to recruit understudied groups. Researchers should also make sure to advertise research in ways that are accessible to people with various types of impairments whenever possible. Including a wider variety of participants might also make it necessary to adapt the research methods to the abilities and strengths of the research participants, such as described for instance in research on participatory design with autistic children [Spiel et al. 2017].

5.2.2 Use of Crowdworkers. In our sample, 106 papers used crowdsourcing platforms to recruit participants. Many of these papers in our sample discuss shortcomings of using crowdworkers in the limitations section, such as Tan et al. [2017] who point out that MTurkers are not representative of the general U.S. population. Habib et al. [2018] also recognize the limitations of using convenience samples such as MTurk, but point to research demonstrating that MTurk is a valid source of high-quality human subjects data [Kittur et al. 2008]. Papers involving crowdworkers mostly relied on simulated risk, or, to a lesser extent, naturally occurring or combinations of risk. Indeed, the use of crowdworkers does not exclude realistic risk representations, and researchers combine tools such as prototypes, scenarios, or educational interventions. Based on the sample of papers we analyzed, it appears that using crowdworking platforms has become an accepted practice for UPS studies, especially when researchers need to create a controlled experimental setup that requires sufficient statistical power, thus calling for large numbers of participants.

We did not systematically analyze compensation for crowdworkers, but we observed anecdotally that compensation seemed to vary substantially between studies, with one study, for example, compensating participants with $0.70 for a 10-minute survey on MTurk [Shrestha et al. 2016], and another study compensating participants $4 for a 10-to-15-minute MTurk study [Lyastani et al. 2018]. While Prolific enforces payments of at least $6.50 per hour, MTurk does not currently enforce a minimum compensation. In addition to ethical concerns related to compensation for crowdworkers, low pay might also affect data quality and participants' willingness to disclose private information [Sannon and Cosley 2018], thus potentially influencing the validity of UPS studies. A discussion of how to define "fair" compensation of crowdworkers who participate in UPS studies thus seems important.

5.3 Ethics

Here we focus on two ethical issues of particular importance for UPS studies: deception and use of attackers as human subjects in studies.

5.3.1 Deception. Based on our analysis, it seems that the tension between deception and ethics in UPS remains. In particular, the community is not consistent with the definition of deception. Five papers with broad or partial disclosure (not considered deception in our analysis) stated that they used deception, whereas not all papers that we coded as using deception stated that they did. We refer to studies with partial disclosure when the study objective was stated in a relatively broad manner to avoid priming participants. In one study for instance, the authors recruited participants "for a study on personal finance and credit bureaus" and purposefully omitted that they were specifically interested in Equifax or identity theft to avoid priming participants and limit self-selection bias [Zou et al. 2018]. In another study, the researchers recruited Android users without mentioning that the study would focus on permissions [Harbach et al. 2014]. Both these studies mention that they use forms of deception, but we consider these instances of partial disclosure and not deception, as they did not mislead participants or withhold information important for them

to understand their participation in the study. It is important to note that we discuss these papers with partial disclosure here because they referred to their own approaches as forms of deception. Unless a paper mentioned deception, we did not specifically look for partial disclosure; thus, we expect there may thus be many more papers with partial disclosure in our sample.

The community would profit from a clearer definition of what constitutes deception, and more discussion on what types of deceptions are ethically acceptable in UPS studies. Such a discussion should also include harm-mitigation strategies put in place by researchers, including the use of trained experimenters and requirements for strict debriefing protocols. In addition, clear reporting guidelines for deception studies should be established, as proposed in the next section (also refer to the list of guidelines in the Appendix).

Despite the utility of using deception to make study participants subjected to simulated risk believe they are actually at risk, we find relatively few studies that use deception. Some papers in our sample explain why they decided not to use deception in their studies, mostly pointing to the lack of necessity and avoidance of potential ethical concerns. Dechand et al. [2016] explain that they opted not to obfuscate the goal of their study since they wanted to find the best possible comparison of key-fingerprints in a security context, and the question of how to motivate users to do so was out of scope for their paper. Therefore, the authors argue that it was not necessary to use deception, and thus opted for what they call the "honest" approach. Similarly, Haque et al. [2014] argue that they did not use deception since it was not necessary for their study in which they created a scale to measure participants' comfort when constructing a strong password. The authors argue that given that since there was a relative lack of consequences (e.g., no embarrassment, no reason to respond dishonestly), they considered that it was unnecessary to hide the true intent of the study. Volkamer et al. [2018] also explain that they opted to avoid deception in their study, during which they observed people using ATMs in public. The authors explain that they intentionally avoided conducting a researcher-as-participant study, which uses deception and makes it more difficult to preserve anonymity of subjects. Instead, the authors decided to conduct a pure observation study without any type of deception, thus avoiding ethical concerns. Petracca et al. [2017] also do not use or mention deception in their paper, but describe an interesting approach that we consider partial disclosure. The authors performed a lab study in which they evaluated the effectiveness of their authorization framework in supporting users in avoiding attacks by malicious apps. Before starting their experiment, the authors informed participants that attacks targeting sensitive audio or video data were possible during the interaction with the apps involved in the experimental tasks, but did not inform participants of the attack source. By revealing the possibility of attacks, yet without mentioning the attack source, the authors thus managed to simulate attacks without the use of deception.

5.3.2 Attackers as Human Subjects. When analyzing the use of deception in our sample, we found that in four papers, all related to social engineering, the authors did not obtain consent from their participants. Two of these papers included attackers as non-consenting participants ([Han et al. 2016] and [Sahin et al. 2017]). The inclusion of real (not simulated) attackers in UPS studies raises ethical questions. Some researchers might not think of attackers as research participants for which they require IRB approval or need to obtain informed consent, especially when they "only" analyze existing datasets (e.g., [Sahin et al. 2017]) or when observing the attackers' naturally occurring behavior (e.g., [Han et al. 2016]). Compared to general HCI research, the issue of including attackers as research participants seems somewhat unique to UPS and non-UPS security studies. However, researchers in criminology have discussed the ethical implications of including criminals, prisoners, or persons exhibiting potentially criminal behavior in their research. For instance, [Ray et al. 2010] discuss the legal, ethical, and methodological considerations when studying child

pornography offenders online. They underline that the Belmont Report, which provides ethical principles to which all researchers are bound, applies for all human subjects research, even when participants exhibit criminal behaviors. The report includes the principles of respect for persons, beneficence, and justice. The principle of respect for persons requires researchers to ensure that participants are autonomously and voluntarily consenting to take part in research. Beneficence requires researchers to minimize the risk of harm resulting from participation in research studies. The authors also discuss the tension between the legal perspective, which does not recognize participant privilege, and the ethical perspective, which requires the researcher to reduce the risk participants might incur from taking part in studies. Roberts and Indermaur [2003] discuss how signed consent forms, while usually required by human research ethics committees, can pose a threat to research participants in criminological research, especially offenders. Written documentation that proves participation in a research study can threaten the offender's future well-being and create a barrier to participation. The authors thus suggest developing alternative approaches for obtaining informed consent. Furthermore, in UPS studies, it may be impossible to contact attackers to obtain permission to observe their illegal behavior for research purposes without scaring them away.

IRB review for studies with attackers as participants may focus on whether collecting data on attackers is done in a way that minimizes the potential that these unwitting participants are harmed, for example, by avoiding the collection of data that would allow the attackers to be identified, as identifying attackers is the job of law enforcement, but should be avoided by human-subjects researchers. In addition, identifying attackers will make them less likely to willingly participate in future research. Similar issues and remediations arise when unwitting participants are employees of a company being probed by UPS researchers. For example, a recent paper describing a study in which research assistants called customer service representatives (CSRs) at several mobile phone carriers to attempt to carry out "SIM swap attacks" includes an ethics section in which the authors explain that they did not record the phone calls with the CSRs and their notes on these calls do not include time of the call, any information that might help identify the CSR, or the phone number discussed on the call. In this way they minimized the chance that the CSRs (who in some cases made mistakes that allowed for successful attacks) might be identified by their employers [Lee et al. 2020].

The UPS community would profit from a constructive discussion on the ethics of research in which attackers (or other non-consenting people) are used as research participants. In contrast with research including criminal offenders, "attackers" in UPS do not always engage in criminal behavior—e.g., an attacker may be someone who snoops on their friend's phone or a child who circumvents access controls put in place by their parents [Schechter 2013b]. Thus, a nuanced consideration of the ethical aspects is necessary. The discussion should also address the use of existing datasets that log attacker behavior without obtaining their consent.

5.4 Reporting User Study Methods

To analyze how researchers represent risk to their participants, it is essential to have a clear understanding of how the authors recruited participants, what the participants were told or led to believe, and how tasks or questions were framed. In some of the papers we reviewed, these details were not clear, and we suggest improving reporting standards for better replicability and understandability of research. Conferences and journals should request (or require) more detailed reporting and encourage (and provide space for) the inclusion of research material (recruitment material, questionnaires, prototypes) in appendixes or as supplemental materials.

We suggest that the following questions should be answered for user studies in UPS (in addition to a typical description of the methods) to provide a clear understanding of risk representation and thus allow for an informed interpretation of the results. We also provide these questions in the form of a checklist in the appendix for researchers and reviewers to use.

- How were participants recruited?
- Were measures taken to include under-studied groups? If yes, what measures were taken?
- Was informed consent obtained? If yes, how?
- Did participants have an accurate understanding of when the data collection started and ended?
- Did participants receive a broad disclosure to avoid security or privacy priming? If so, what was it?
- In the participants' mind, whose data was at risk (if any)?
- Were participants led to believe something that was not the case (use of deception)?
- How did the research protocol mitigate potential harm to participants?
- What other ethical issues were discussed within the author team or the IRB and how were they treated?
- Did participants receive fair compensation? Report time needed for study participation and compensation. What constitutes fair compensation may also depend on factors such as the minimum wage in the area from which participants are recruited and the nature of the tasks they are asked to complete, as well as demographics and how challenging it is to recruit the target population (e.g., a student sample vs. senior doctors with a specific specialization). We suggest providing these details where relevant.
- Is the study protocol (including the instructions given to participants) available in the appendix?

In addition, we include a structure for categorizing UPS studies with respect to their methods and their treatment of risk. Publication venues that welcome research from the field of UPS (e.g., CHI, SOUPS, IEEE S&P, ACM CCS, USENIX Security) could use these guidelines to encourage better reporting of user studies. After reading a paper, reviewers should be able to easily categorize a paper according to these guidelines. This would improve the quality of user studies and encourage replicability and ethical approaches in user studies. In addition, it is useful for students to consider these guidelines as they read papers and start writing research papers of their own.

6 CONCLUSION

Studying how users of digital systems perceive privacy and security risks is a challenging endeavor for researchers and practitioners, who need to balance realistic risk representation and ethical concerns. Studying such questions in a context where research participants perceive no privacy and security risks would impact the validity of the study's results.

On the other hand, exposing people to realistic cyberattacks (e.g., having their identities or credit card numbers stolen) and letting them feel the cost of recovering a sense of normality after a crime, may be unethical unless done carefully to minimize the risk of harm. The issue is also intrinsically related to the use of deception in security research, a practice that seems inevitable in certain contexts to preserve the study validity and to avoid priming participants to look for or expect an attack. We conducted a systematic literature review investigating how recent research in UPS addresses this issue, analyzing 284 papers with regards to their study methods; how the study represents and assesses response to risk; and the use of prototypes, scenarios, educational interventions, and deception.

Important findings include that, across our sample, risk representation was mostly based on naturally occurring or simulated risk. Risk representation varied with the study methods and objectives of a paper. Papers with an experimental objective mostly used simulated risk, and descriptive studies mostly used naturally occurring risk. Response to risk was measured mostly through a combination of observed and self-reported measures, or self-reported measures on their own.

Researchers used a variety of "tools" to represent risk to participants. Security/privacy-related tasks were used in more than a third of the papers, and approximately a third of the papers involved a prototype. Scenarios were also frequently used to represent risk. Deception was only rarely used to create a perception of risk, and only a small number of papers used educational interventions or incentives for secure behavior. In terms of participant recruitment, researchers frequently chose crowdworkers and non-representative convenience samples.

By reviewing the wide array of methods adopted by researchers interested in how users perceive privacy and security, we give an overview of the tradeoffs researchers frequently face and present the community's response to them. Through a discussion of the advantages and shortcomings of the approaches used, our review helps the community be more cognizant of the plethora of different approaches for user studies in UPS, and of how papers discussed the validity of their approaches for risk representation and associated ethical choices. The systematic approach we followed allowed us to suggest guidelines for researchers who aim to report on user studies in privacy and security, and in particular, risk representation and assessment. In addition, we identify key methodological, ethical and research challenges.

Of course, there is no such thing as a "perfect" method. Rather, there is a large set of tradeoffs to consider when choosing a research method. Our review of the methods in UPS studies offers transparency and improves the community's awareness of the adopted practices. We provide a checklist with methodological information we suggest should be included in all empirical UPS studies. On a larger level, we are convinced that fostering an ongoing discussion regarding methods and their potential to represent risk to participants will help the UPS community continuously improve toward a common understanding of valid, ethical and replicable science, and ultimately a richer understanding of how people behave in the presence of privacy and security risk.

APPENDIX

GUIDELINES FOR REPORTING OF USER STUDIES IN USABLE PRIVACY AND SECURITY—PAGE 1

CATEGORIZATION OF RISK REPRESENTATION

We suggest the following shared vocabulary for describing the risk representation in UPS studies. UPS Venues can also ask reviewers to fill out this categorization and use it for descriptive statistics about the published studies.

Objective of the study (check as many as apply)

__**Descriptive**: provides a snapshot of the current state of affairs
__**Relational**: designed to discover relationships among variables
__**Experimental**: participants are placed into multiple groups who experience different manipulations of a given experience so that the influence of each manipulation can be measured[1]

Risk response assessment method (check as many as apply)

__**Observational data**
__**Self-reported data**
__**Assigned security or privacy task**: e.g., password creation, send encrypted message
__**Assigned unrelated task**: e.g., drawing task, buy something on an online store, other non-security or privacy-related tasks

Risk representation (check as many as apply)

__**Naturally occurring risk:** e.g., through observation or self-reported measures of naturally occurring behavior
__**Simulated risk**: e.g., through the use of scenarios participants should imagine themselves in
__**Mentioned risk**: e.g., a questionnaire where participants were presented with hypothetical situations
__**No induced risk representation**

Incentives for secure behavior

Were research participants incentivized to adopt a certain secure behavior, e.g., they would receive financial compensation if they managed to send an encrypted message?
__ yes __ no

Prototype

Does the study involve exposing participants to a prototype of any fidelity (interactive or non-interactive)? A prototype is defined as a new solution such as a textual message, an icon, or an interface.
__ yes __ no

Scenario

Do researchers ask participants to imagine themselves being in a certain situation?
__ yes __ no

Educational Intervention

Did the researchers attempt to educate research participants on privacy and security related topics?
__ yes __ no

[1]Definitions based on Stangor & Walinga [2018].

GUIDELINES FOR REPORTING OF USER STUDIES IN USABLE PRIVACY AND SECURITY—PAGE 2

CHECKLIST FOR ESSENTIAL METHODOLOGICAL DETAILS AND ETHICS

The following information should be clearly stated in usable privacy and security studies.

This checklist can be used by study authors and reviewers in usable privacy and security.

Recruitment

How exactly were participants recruited? E.g., flyers on university campus inviting students only, recruitment panel including parents in specific geographic area, undefined convenience sampling through flyers in an entire city, representative sample, purposive sample

Explain the recruitment strategy: ……

Were measures taken to include under-studied groups (e.g., LGBTQ, older adults, kids, disabled persons…)?

__ yes __ no

Explain: ……

Informed consent

Was informed consent obtained?

__ yes __ no

If yes, how? ……

Did participants have an accurate understanding of when the data collection started and ended?

__ yes __ no

Explain: ……

Did participants receive a broad disclosure to avoid security or privacy priming?

Explain: ……

Methodological details

In the participants' mind, whose data was at risk (if any)?

Explain: ……

Were participants led to believe something that was not the case (use of deception)?

Explain: ……

How did the research protocol mitigate potential harm to participants?

Explain: ……

Which other ethics issues discussed within the author team or the IRB and how were they treated?

Explain: ……

Did participants receive fair compensation? Report time needed for study participation and compensation.

Time needed for participation: ……. Amount of compensation: ……

Replicability

Is the study protocol (including the instructions given to participants) available in the appendix?

__ yes __ no

List of Included papers

We provide the list of included papers as supplemental material.

Dataset

We provide the full dataset as supplemental material.

Additional Results

Table 16. Number of Included Papers per Year ($N = 284$)

	Frequency	Percent
2014	43	15
2015	56	20
2016	54	19
2017	70	25
2018	61	22

Table 17. Objectives of the Papers (N = 284)

	Frequency	Percent
Experimental	115	41
Descriptive	89	31
Descriptive and relational	49	17
Relational	13	5
Descriptive and experimental	10	4
Relational and experimental	8	3

Table 18. Combinations of Study Methods in Our Sample (N = 284)

	Frequency	Percent
Experiment	99	35
Interview	36	13
Survey	34	12
Survey and datalogs	12	4
Analyze dataset	11	4
Survey and interview	11	4
Experience sampling method	8	3

(Continued)

Table 18. Continued

	Frequency	Percent
Survey and experiment	8	3
Other	6	2
Datalogs	4	1
Survey and analyze dataset	4	1
Interview and experiment	3	1
Interview and focus group	3	1
Interview and other	3	1
Survey and other	3	1
Observation and interview	2	1
Experiment and datalogs	2	1
Experiment and other	2	1
Survey and interview and datalogs	2	1
Survey and list experiment	2	1
Vignette experiment	2	1
Workshop and other	1	0.4
Focus group	1	0.4
Survey and datalogs	1	0.4
Survey and experiment and analyze dataset	1	0.4
Interview and diary	1	0.4
Interview and experiment and analyze dataset	1	0.4
Interview and observation	1	0.4
Survey and Observation	1	0.4
Experiment and analyze dataset and datalogs	1	0.4
Experiment and focus group	1	0.4
Observation	1	0.4
Survey and co-creation	1	0.4
Survey and interview and analyze dataset	1	0.4
Survey and interview and other	1	0.4
Survey and workshop	1	0.4
Datalogs and other	1	0.4
Experiment and analyze dataset	1	0.4
Experiment and diary	1	0.4
Interview and co-creation	1	0.4
Interview and analyze dataset	1	0.4
Interview and datalogs	1	0.4
Interview and workshop and observation	1	0.4
Survey and interview and diary and observation	1	0.4
Survey and interview and experiment	1	0.4
Survey and interview and experiment and other	1	0.4
Survey and interview and focus group and analyze dataset	1	0.4
Survey and interview and observation	1	0.4

Table 19. Average Number of Participants per Paper per Study Method

	Experiment (n = 119)	Survey (n = 85)	Interview (n = 72)	Log analysis (n = 20)
Mean	653	846	29	808
Median	80	307	21	110
Std. dev.	1514	1538	32	2882
Min	6	7	4	19
Max	9114	10763	200	13000

Table 20. IRB or Ethics Board Approval of Papers (N = 284)

	Frequency	Percent
Ethics board approval	159	56
Not mentioned	99	35
Institution without approval procedure	10	4
Corporate/industry internal review process	10	4
Exempt from needing approval	4	1
Other	1	0.4
Multiple	1	0.4

Table 21. Topics of Deception Studies (n = 16)

	Frequency	Percent	Percentage of papers in that category that use deception
Authentication	3	19	4
Privacy perceptions, attitudes and behaviors	3	19	5
Privacy transparency and choice mechanisms	3	19	21
Social engineering	3	19	30
Security indicators and warnings	1	6	8
Security perceptions, attitudes and behaviors	1	6	2
Access control	1	6	5
Multiple	1	6	9

Table 22. Study Methods Used in Deception Studies (n = 16)

	Frequency	Percent	Percentage of papers using each method that used deception
Experiment	11	69	11
Analyze Dataset	1	6	9
Interview	1	6	3
Other	1	6	17
Survey and Experiment	1	6	13
Survey and Interview	1	6	9

Additional Analysis by Topics

As shown in Figure 3, most research topics had a dominant study objective. Access control, authentication, encryption, privacy transparency and choice mechanisms, social engineering, and security indicators and warnings are all investigated using experimental studies about two-thirds of the time. Papers that study privacy and security perceptions, attitudes and behaviors frequently use a descriptive objective (45% and 54%, respectively). Studies of privacy-enhancing technologies and studies of multiple topics also tended to use a descriptive objective. Relational and combination objectives were not used frequently for any topic.

One might assume that in topic areas where experimental studies are predominant (e.g., authentication, social engineering, security indicators and warnings), researchers mostly seek to test and validate solutions, for instance new authentication schemes. While almost half (46%) of the authentication studies and 58% of papers on security indicators and warnings include a prototype, other topics with a large proportion of experimental studies (e.g., social engineering) do not include many prototypes. In these cases, the researchers might actually use experimental approaches to test and find difficulties users have with existing products. Research on topics with many descriptive studies (e.g., privacy and security perceptions, attitudes and behaviors)

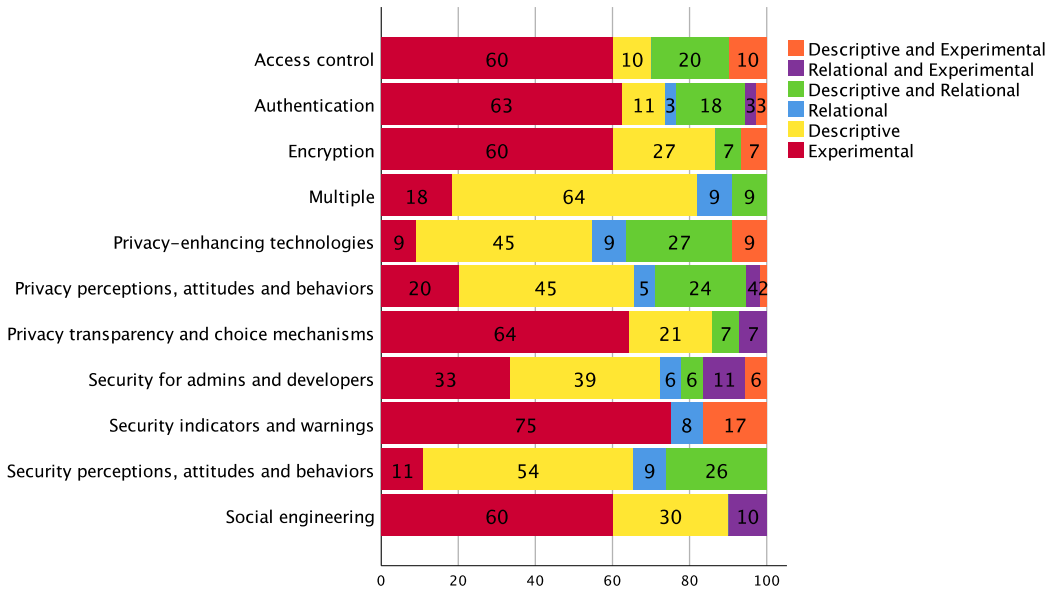

Fig. 3. Crosstab research topic and objective (percentage).

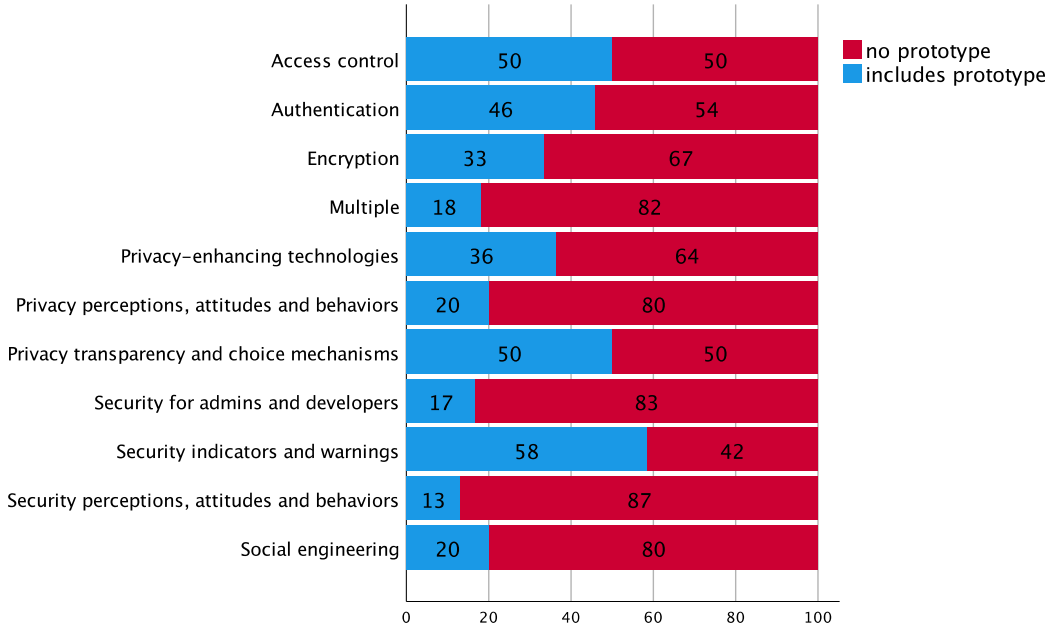

Fig. 4. Proportion of studies including a prototype per topic.

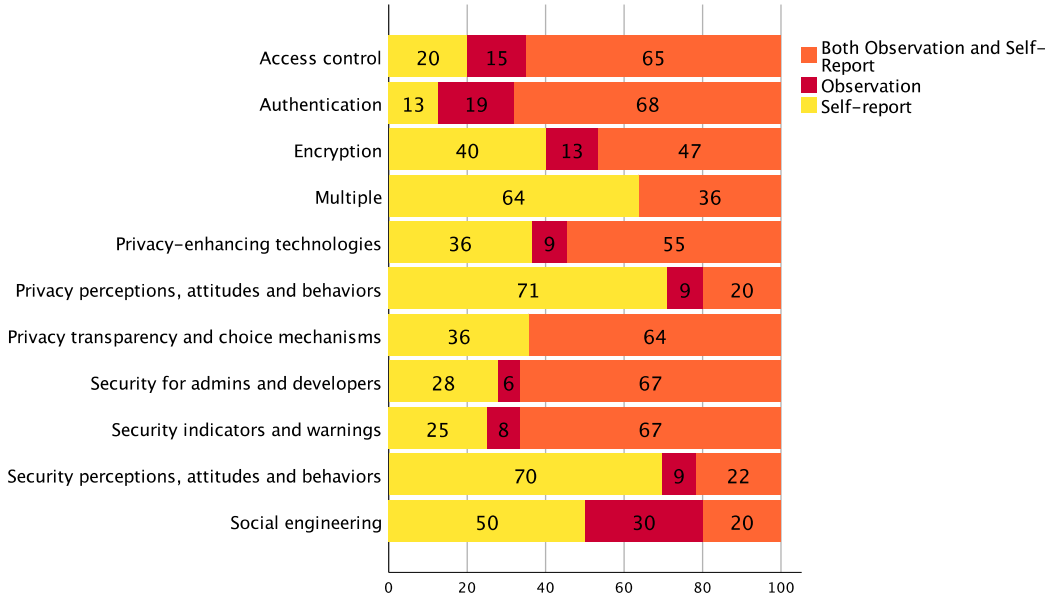

Fig. 5. Crosstab risk response assessment and topic (percent of all studies on topic).

have a tendency to seek to describe the current state of affairs, without necessarily evaluating solutions in an experimental setting. Figure 4, which correlates the number of prototypes per topic, corroborates this hypothesis, showing that privacy and security perceptions, attitudes and behaviors indeed include very few prototypes (20% and 13%, respectively).

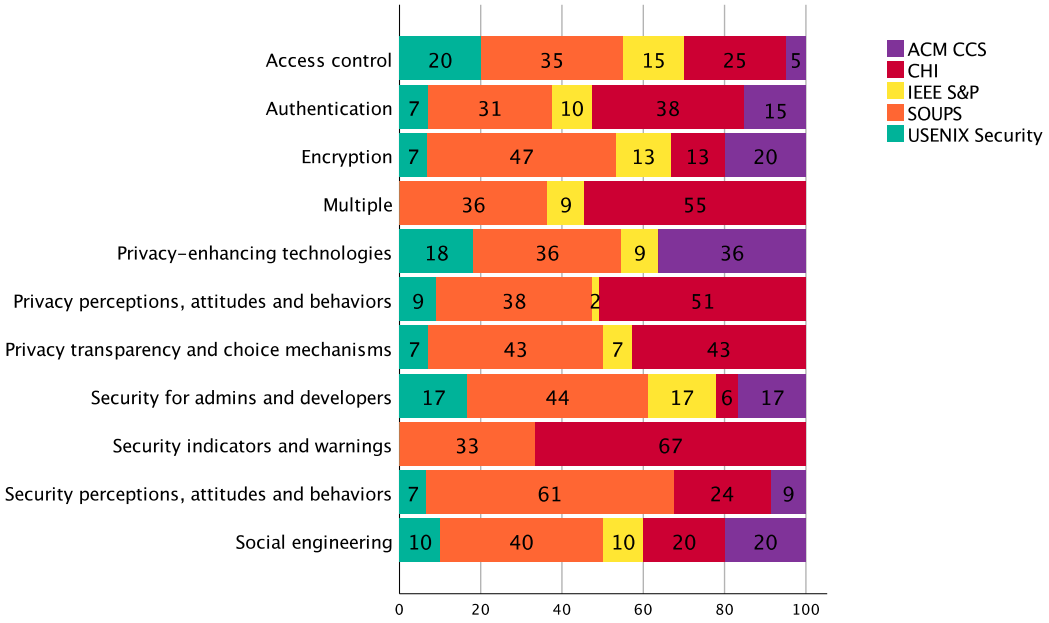

Fig. 6. Crosstab topic and venue.

Figure 5 shows that observation occurs rarely on its own and that most topic areas rely on a combination of self-reported and observed measures. Self-report studies dominate two topic areas: privacy perceptions, attitudes and behaviors (71%) and security perceptions, attitudes, and behaviors (70%). Authentication and social engineering are the topics with the highest percentage of papers that use observed measures only (19%, 30%).

As shown in Figure 6, most topics appeared at all five of the venues we reviewed. However, a small number of topics were more likely to appear or not appear at certain venues. For example, papers on *privacy perceptions, attitudes and behaviors* were more likely to appear at CHI and rarely appeared at IEEE S&P or CCS. Papers on *security perceptions, attitudes and behaviors* never appeared at IEEE S&P. Papers on the topics *security for admins and developers*, and *privacy-enhancing technologies* were rarely published at CHI. Papers on *security indicators and warnings* appeared only at CHI and SOUPS. Papers on *privacy transparency and choice mechanisms* and papers on *access control* rarely appeared at ACM CCS. SOUPS was the only venue that had published papers on all topic areas. Overall, our data show that certain topics were more likely to be published at certain publications venues over others. Our data does not show whether authors self-selected and did not submit papers at certain venues because the topics did not seem suitable, or whether reviewers at certain venues were likely to judge papers with certain research topics more favorably.

REFERENCES

Y. Acar, M. Backes, S. Fahl, S. Garfinkel, D. Kim, M. L. Mazurek, and C. Stransky. 2017. Comparing the usability of cryptographic apis. In *Proceedings of the 38th IEEE Symposium on Security and Privacy*.

A. Acquisti, M. Sleeper, Y. Wang, S. Wilson, I. Adjerid, R. Balebako, L. Brandimarte, L. F. Cranor, S. Komanduri, P. G. Leon, N. Sadeh, and F. Schaub. 2017. Nudges for privacy and security: Understanding and assisting users' choices online. *ACM Computing Surveys* 50, 3 (2017), 1–41. https://doi.org/10.1145/3054926

D. Adams, A. Bah, C. Barwulor, N. Musaby, K. Pitkin, and E. M. Redmiles. 2018. Ethics emerging: The story of privacy and security perceptions in virtual reality. In *Proceedings of the 14th Symposium on Usable Privacy and Security (SOUPS'18)*. 427–442. https://www.usenix.org/conference/soups2018/presentation/adams.

E. Adar, D. S. Tan, and J. Teevan. 2013. Benevolent deception in human computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'13)*. 1863. DOI: https://doi.org/10.1145/2470654.2466246

L. Agarwal, H. Khan, and U. Hengartner. 2016. Ask me again but don't annoy me: Evaluating re-authentication strategies for smartphones. In *Proceedings of the 12th Symposium on Usable Privacy and Security (SOUPS'16)*. 221–236. https://www.usenix.org/conference/soups2016/technical-sessions/presentation/agarwal.

T. Ahmed, R. Hoyle, K. Connelly, D. Crandall, and A. Kapadia. 2015. Privacy concerns and behaviors of people with visual impairments. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 3523–3532.

Y. Albayram, M. M. H. Khan, T. Jensen, and N. Nguyen. 2017. "...better to use a lock screen than to worry about saving a few seconds of time": Effect of fear appeal in the context of smartphone locking behavior. In *Proceedings of the 13th Symposium on Usable Privacy and Security (SOUPS'17)*. 49–63. https://www.usenix.org/conference/soups2017/technical-sessions/presentation/albayram.

R. Alves, P. Valente, and N. J. Nunes. 2014. The state of user experience evaluation practice. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction Fun, Fast, Foundational (NordiCHI'14)*. 93–102. DOI: https://doi.org/10.1145/2639189.2641208

American Psychological Association. 2017. Ethical principles of psychologists and code of conduct (2002, Amended Effective June 1, 2010, and January 1, 2017). http://www.apa.org/ethics/code/index.html.

B. B. Anderson, C. B. Kirwan, J. L. Jenkins, D. Eargle, S. Howard, and A. Vance. 2015. How Polymorphic Warnings Reduce Habituation in the Brain: Insights from an fMRI Study. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2883–2892. DOI: https://doi.org/10.1145/2702123.2702322

J. Angulo and M. Ortlieb. 2015. "WTH..!?!" Experiences, reactions, and expectations related to online privacy panic situations. In *Proceedings of the 11th Symposium on Usable Privacy and Security (SOUPS'15)*. 19–38. https://www.usenix.org/conference/soups2015/proceedings/presentation/angulo.

N. Athanassoulis and J. Wilson. 2009. When is deception in research ethical? *Clinical Ethics* 4, 1 (2009), 44–49. https://doi.org/10.1258/ce.2008.008047

D. Baumrind. 1985. Research using intentional deception. *American Psychologist* 40, 2 (Feb. 1985), 165–174.

R. Biddle, S. Chiasson, and P. C. Van Oorschot. 2012. Graphical passwords: Learning from the first twelve years. *ACM Computing Surveys* 44, 4 (2012), 1–41. DOI: https://doi.org/10.1145/2333112.2333114

B. Bonné, S. T. Peddinti, I. Bilogrevic, and N. Taft. 2017. Exploring decision making with Android's runtime permission dialogs using in-context surveys. In *Proceedings of the 13th Symposium on Usable Privacy and Security (SOUPS'17)*. 195–210. https://www.usenix.org/conference/soups2017/technical-sessions/presentation/bonne.

J. Bonneau, C. Herley, P. C. van Oorschot, and F. Stajano. 2012. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy*. 553–567. DOI: https://doi.org/10.1109/SP.2012.44

C. Bravo-Lillo, L. Cranor, S. Komanduri, S. Schechter, and M. Sleeper. 2014. Harder to ignore? Revisiting pop-up fatigue and approaches to prevent it. In *Proceedings of the 10th Symposium on Usable Privacy and Security (SOUPS'14)*. 105–111. https://www.usenix.org/conference/soups2014/proceedings/presentation/bravo-lillo.

K. Caine. 2016. Local standards for sample size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI'16)*. 981–992. DOI: https://doi.org/10.1145/2858036.2858498

C. Canfield, A. Davis, B. Fischhoff, A. Forget, S. Pearman, and J. Thomas. 2017. Replication: Challenges in using data logs to validate phishing detection ability metrics. In *Proceedings of the 13th Symposium on Usable Privacy and Security (SOUPS'17)*. 271–284. https://www.usenix.org/conference/soups2017/technical-sessions/presentation/canfield.

R. Chatterjee, A. Athayle, D. Akhawe, A. Juels, and T. Ristenpart. 2016. PASSWORD tYPOS and how to correct them securely. In *Proceedings of the 2016 IEEE Symposium on Security and Privacy (SP'16)*. 799–818.

L. F. Cranor and N. Buchler. 2014. Better together: Usability and security go hand in hand. *IEEE Security Privacy* 12, 6 (2014), 89–93. DOI: https://doi.org/10.1109/MSP.2014.109

H. Crawford and E. Ahmadzadeh. 2017. Authentication on the go: Assessing the effect of movement on mobile device keystroke dynamics. In *Proceedings of the 13th Symposium on Usable Privacy and Security (SOUPS'17)*. 163–173. https://www.usenix.org/conference/soups2017/technical-sessions/presentation/crawford.

S. Das, G. Laput, C. Harrison, and J. I. Hong. 2017. Thumprint: Socially-inclusive local group authentication through shared secret knocks. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3764–3774.

Deception Research—APA Dictionary of Psychology. (n.d.). Retrieved January 20, 2020 from https://dictionary.apa.org/deception-research.

S. Dechand, D. Schürmann, K. Busse, Y. Acar, S. Fahl, and M. Smith. 2016. An empirical study of textual key-fingerprint representations. In *Proceedings of the 25th USENIX Security Symposium (USENIX Security, 16)*. 193–208.

V. Distler, C. Lallemand, and V. Koenig. 2020. How acceptable is this? How user experience factors can broaden our understanding of the acceptance of privacy trade-offs. *Computers in Human Behavior* 106, 106227. DOI: https://doi.org/10.1016/j.chb.2019.106227

S. Egelman, R. Kannavara, and R. Chow. 2015. Is This thing on? Crowdsourcing privacy indicators for ubiquitous sensing platforms. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* 1669–1678. DOI: https://doi.org/10.1145/2702123.2702251

S. Egelman, J. King, R. C. Miller, N. Ragouzis, and E. Shehan. 2007. Security user studies: Methodologies and best practices. In *Proceedings of the CHI'07 Extended Abstracts on Human Factors in Computing Systems (CHI'07)*. 2833. DOI: https://doi.org/10.1145/1240866.1241089

S. Egelman and E. Peer. 2015. Scaling the security wall: Developing a security behavior intentions scale (SeBIS). In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2873–2882. DOI: https://doi.org/10.1145/2702123.2702249

M. Eiband, M. Khamis, E. von Zezschwitz, H. Hussmann, and F. Alt. 2017. Understanding shoulder surfing in the wild: stories from users and observers. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 4254–4265. DOI: https://doi.org/10.1145/3025453.3025636

S. Fahl, M. Harbach, Y. Acar, and M. Smith. 2013. On the ecological validity of a password study. In *Proceedings of the Ninth Symposium on Usable Privacy and Security (SOUPS'13)*. 1. DOI: https://doi.org/10.1145/2501604.2501617

V. Fanelle, S. Karimi, A. Shah, B. Subramanian, and S. Das. 2020. Blind and human: Exploring more usable audio CAPTCHA designs. In *Proceedings of the 16th Symposium on Usable Privacy and Security (SOUPS'20)*. https://www.usenix.org/conference/soups2020/presentation/fanelle.

A. P. Felt, R. Barnes, A. King, C. Palmer, C. Bentzel, and P. Tabriz. 2017. Measuring HTTPS adoption on the web. In *Proceedings of the 26th USENIX Security Symposium (USENIX Security'17)*. 1323–1338. https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/felt.

A. P. Felt, R. W. Reeder, A. Ainslie, H. Harris, M. Walker, C. Thompson, M. E. Acer, E. Morant, and S. Consolvo. 2016. Rethinking connection security indicators. In *Proceedings of the 12th Symposium on Usable Privacy and Security (SOUPS'16)*. 1–14. https://www.usenix.org/conference/soups2016/technical-sessions/presentation/porter-felt.

C. Fiesler and B. Hallinan. 2018. "We are the product": Public reactions to online data sharing and privacy controversies in the media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 53:1–53:13. DOI: https://doi.org/10.1145/3173574.3173627

A. Forget, S. Komanduri, A. Acquisti, N. Christin, L. F. Cranor, and R. Telang. 2014. *Security Behavior Observatory: Infrastructure for Long-Term Monitoring of Client Machines*. Technical Report CMU-CyLab-14-009. CyLab, Carnegie Mellon University, p. 11.

A. Forget, S. Pearman, J. Thomas, A. Acquisti, N. Christin, L. F. Cranor, S. Egelman, M. Harbach, and R. Telang. 2016. Do or do not, there is no try: User engagement may not improve security outcomes. In *Proceedings of the 12th Symposium on Usable Privacy and Security (SOUPS'16)*. 97–111. https://www.usenix.org/conference/soups2016/technical-sessions/presentation/forget.

D. Freed, J. Palmer, D. Minchala, K. Levy, T. Ristenpart, and N Dell. 2018. "A stalker's paradise'': How intimate partner abusers exploit technology. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 667:1–667:13. DOI: https://doi.org/10.1145/3173574.3174241

B. Fuller, M. Varia, A. Yerukhimovich, E. Shen, A. Hamlin, V. Gadepally, R. Shay, J. D. Mitchell, and R. K. Cunningham. 2017. Sok: Cryptographically protected database search. In *Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP)*. 172–191.

S. Garfinkel and H. R. Lipford. 2014. *Usable Security: History, Themes, and Challenges*. Morgan & Claypool Publishers.

H. Habib, J. Colnago, V. Gopalakrishnan, S. Pearman, J. Thomas, A. Acquisti, N. Christin, and L. F. Cranor. 2018. Away from prying eyes: Analyzing usage and understanding of private browsing. In *Proceedings of the 14th Symposium on Usable Privacy and Security (SOUPS'18)*. 159–175. https://www.usenix.org/conference/soups2018/presentation/habib-prying.

X. Han, N. Kheir, and D. Balzarotti. 2016. PhishEye: Live monitoring of sandboxed phishing kits. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. 1402–1413. DOI: https://doi.org/10.1145/2976749.2978330

A. Hanamsagar, S. S. Woo, C. Kanich, and J. Mirkovic. 2018. Leveraging semantic transformation to investigate password habits and their causes. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 570:1–570:12. DOI: https://doi.org/10.1145/3173574.3174144

A. Hang, A. D. Luca, M. Smith, M. Richter, and H. Hussmann. 2015. Where have you been? Using location-based security questions for fallback authentication. In *Proceedings of the 11th Symposium on Usable Privacy and Security (SOUPS'15)*. 169–183. https://www.usenix.org/conference/soups2015/proceedings/presentation/hang.

N. B. Hansen, C. Dindler, K. Halskov, O. S. Iversen, C. Bossen, D. A. Basballe, and B. Schouten. 2019. How participatory design works: Mechanisms and effects. In *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*. 30–41. DOI: https://doi.org/10.1145/3369457.3369460

S. M. T. Haque, S. Scielzo, and M. Wright. 2014. Applying psychometrics to measure user comfort when constructing a strong password. In *Proceedings of the 10th Symposium on Usable Privacy and Security (SOUPS'14)*. 231–242. https://www.usenix.org/conference/soups2014/proceedings/presentation/haque.

M. Harbach, M. Hettig, S. Weber, and M. Smith. 2014. Using personal examples to improve risk communication for security & privacy decisions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2647–2656.

C. Holz and F. R. Bentley. 2016. On-Demand biometrics: Fast cross-device authentication. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 3761–3766. DOI: https://doi.org/10.1145/2858036.2858139

H. Hu and G. Wang. 2018. End-to-end measurements of email spoofing attacks. In *Proceedings of the 27th USENIX Security Symposium (USENIX Security'18)*. 1095–1112.

J. H. Huh, H. Kim, R. B. Bobba, M. N. Bashir, and K. Beznosov. 2015. On the memorability of system-generated PINs: Can chunking help? In *Proceedings of the 11th Symposium on Usable Privacy and Security (SOUPS'15)* 197–209. https://www.usenix.org/conference/soups2015/proceedings/presentation/huh.

J. H. Huh, H. Kim, S. S. V. P. Rayala, R. B. Bobba, and K. Beznosov. 2017. I'm too busy to reset my LinkedIn password: On the effectiveness of password reset emails. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 387–391. DOI: https://doi.org/10.1145/3025453.3025788

G. Iachello and J. Hong. 2007. End-user privacy in human-computer interaction. *Foundations and Trends® in Human-Computer Interaction* 1, 1 (2007), 1–137. DOI: https://doi.org/10.1561/1100000004

P. Jaferian, H. Rashtian, and K. Beznosov. 2014. To authorize or not authorize: Helping users review access policies in organizations. In *Proceedings of the 10th Symposium on Usable Privacy and Security (SOUPS'14)*. 301–320. https://www.usenix.org/conference/soups2014/proceedings/presentation/jaferian.

T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer. 2007. Social phishing. *Communications of the ACM* 50, 10 (2007), 94–100.

S. Karunakaran, K. Thomas, E. Bursztein, and O. Comanescu. 2018. Data breaches: User comprehension, expectations, and concerns with handling exposed data. In *Proceedings of the 14th Symposium on Usable Privacy and Security (SOUPS'18)*. 217–234. https://www.usenix.org/conference/soups2018/presentation/karunakaran.

A. Kittur, E. H. Chi, and B. Suh. 2008. Crowdsourcing user studies with mechanical turk. In *Proceeding of the 26th Annual CHI Conference on Human Factors in Computing Systems (CHI'08)*. 453. DOI: https://doi.org/10.1145/1357054.1357127

S. Komanduri, R. Shay, L. F. Cranor, C. Herley, and S. Schechter. 2014. Telepathwords: Preventing weak passwords by reading users' minds. In *Proceedings of the 23rd USENIX Security Symposium (USENIX Security'14)*. 591–606.

P. Kumaraguru, J. Cranshaw, A. Acquisti, L. Cranor, J. Hong, M. A. Blair, and T. Pham. 2009. School of phish: A real-world evaluation of anti-phishing training. In *Proceedings of the 5th Symposium on Usable Privacy and Security (SOUPS'09)*. DOI: https://doi.org/10.1145/1572532.1572536

E. Lastdrager, I. C. Gallardo, P. Hartel, and M. Junger. 2017. How effective is anti-phishing training for children? In *Proceedings of the 13th Symposium on Usable Privacy and Security (SOUPS'17)*. 229–239. https://www.usenix.org/conference/soups2017/technical-sessions/presentation/lastdrager.

K. Lee, B. Kaiser, J. Mayer, and A. Narayanan. 2020. An empirical study of wireless carrier authentication for SIM swaps. In *Proceedings of the 16th Symposium on Usable Privacy and Security (SOUPS'20)*. https://www.usenix.org/conference/soups2020/presentation/lee.

C. Liu, G. D. Clark, and J. Lindqvist. 2017. Where usability and security go hand-in-hand: Robust gesture-based authentication for mobile systems. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 374–386. DOI: https://doi.org/10.1145/3025453.3025879

S. G. Lyastani, M. Schilling, S. Fahl, M. Backes, and S. Bugiel. 2018. Better managed than memorized? Studying the Impact of Managers on Password Strength and Reuse. In *Proceedings of the 27th USENIX Security Symposium (USENIX Security'18)*. 203–220.

S. Mare, M. Baker, and J. Gummeson. 2016. A study of authentication in daily life. In *Proceedings of the 12th Symposium on Usable Privacy and Security (SOUPS'16)*. 189–206. https://www.usenix.org/conference/soups2016/technical-sessions/presentation/mare.

C. Marforio, R. Jayaram Masti, C. Soriente, K. Kostiainen, and S. Čapkun. 2016. Evaluation of personalized security indicators as an anti-phishing mechanism for smartphone applications. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 540–551. DOI: https://doi.org/10.1145/2858036.2858085

E. McReynolds, S. Hubbard, T. Lau, A. Saraf, M. Cakmak, and F. Roesner. 2017. Toys that listen: A study of parents, children, and internet-connected toys. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 5197–5207. DOI: https://doi.org/10.1145/3025453.3025735

W. Melicher, D. Kurilova, S. M. Segreti, P. Kalvani, R. Shay, B. Ur, L. Bauer, N. Christin, L. F. Cranor, and M. L. Mazurek. 2016. Usability and security of text passwords on mobile devices. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 527–539.

H. Meutzner, S. Gupta, and D. Kolossa. 2015. Constructing secure audio captchas by exploiting differences between humans and machines. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2335–2338.

M. Mondal, J. Messias, S. Ghosh, K. P. Gummadi, and A. Kate. 2016. Forgetting in social media: Understanding and controlling longitudinal exposure of socially shared data. In *Proceedings of the 12th Symposium on Usable Privacy and Security (SOUPS'16)*. 287–299. https://www.usenix.org/conference/soups2016/technical-sessions/presentation/mondal.

C. D. Mulrow. 1994. Systematic reviews: Rationale for systematic reviews. *BMJ* 309, 6954 (1994), 597–599. DOI: https://doi.org/10.1136/bmj.309.6954.597

A. Naiakshina, A. Danilova, C. Tiefenau, M. Herzog, S. Dechand, and M. Smith. 2017a. Why do developers get password storage wrong? A qualitative usability study. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 311–328. DOI: https://doi.org/10.1145/3133956.3134082

A. Naiakshina, A. Danilova, C. Tiefenau, M. Herzog, S. Dechand, and M. Smith. 2017b. Why do developers get password storage wrong? A qualitative usability study. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 311–328. DOI: https://doi.org/10.1145/3133956.3134082

A. Naiakshina, A. Danilova, C. Tiefenau, and M. Smith. 2018. Deception task design in developer password studies: Exploring a student sample. In *Proceedings of the 14th Symposium on Usable Privacy and Security (SOUPS'18)*. 297–313. https://www.usenix.org/conference/soups2018/presentation/naiakshina.

M. Obrist, V. Roto, and K. Väänänen-Vainio-Mattila. 2009. User experience evaluation: Do you know which method to use? In *Proceedings of the 27th International Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA'09)*. 2763. DOI: https://doi.org/10.1145/1520340.1520401

M. Oltrogge, Y. Acar, S. Dechand, M. Smith, and S. Fahl. 2015. To pin or not to pin—Helping app developers bullet proof their TLS connections. In *Proceedings of the 24th USENIX Security Symposium (USENIX Security'15)*. 239–254.

G. Paré, M.-C. Trudel, M. Jaana, and S. Kitsiou. 2015. Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management* 52, 2 (2015), 183–199. DOI: https://doi.org/10.1016/j.im.2014.08.008

J. Pearson, S. Robinson, M. Jones, A. Joshi, S. Ahire, D. Sahoo, and S. Subramanian. 2017. Chameleon devices: Investigating more secure and discreet mobile interactions via active camouflaging. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 5184–5196.

G. Petracca, A.-A. Reineh, Y. Sun, J. Grossklags, and T. Jaeger. 2017. Aware: Preventing abuse of privacy-sensitive sensors via operation bindings. In *Proceedings of the 26th USENIX Security Symposium (USENIX Security'17)*. 379–396.

I. Pettersson, F. Lachner, A.-K. Frison, A. Riener, and A. Butz. 2018. A bermuda triangle? *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18)*. 1–16. DOI: https://doi.org/10.1145/3173574.3174035

E. A. Qahtani, M. Shehab, and A. Aljohani. 2018. The effectiveness of fear appeals in increasing smartphone locking behavior among Saudi Arabians. In *Proceedings of the 14th Symposium on Usable Privacy and Security (SOUPS'18)*. 31–46. https://www.usenix.org/conference/soups2018/presentation/qahtani.

L. Rainie and M. Duggan. 2015. *Privacy and Information Sharing*. Pew Research Center.

H. Rashtian, Y. Boshmaf, P. Jaferian, and K. Beznosov. 2014. To befriend or not? A model of friend request acceptance on Facebook. In *Proceedings of the 10th Symposium on Usable Privacy and Security (SOUPS'14)*. 285–300. https://www.usenix.org/conference/soups2014/proceedings/presentation/rashtian.

J. V. Ray, E. R. Kimonis, and C. Donoghue. 2010. Legal, ethical, and methodological considerations in the Internet-based study of child pornography offenders. *Behavioral Sciences & the Law* 28, 1 (2010), 84–105. DOI: https://doi.org/10.1002/bsl.906

E. M. Redmiles, S. Kross, and M. L. Mazurek. 2016. How I learned to be secure: A census-representative survey of security advice sources and behavior. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. 666–677. DOI: https://doi.org/10.1145/2976749.2978307

E. M. Redmiles, M. L. Mazurek, and J. P. Dickerson. 2018. Dancing pigs or externalities? Measuring the rationality of security decisions. In *Proceedings of the 2018 ACM Conference on Economics and Computation*. 215–232. DOI: https://doi.org/10.1145/3219166.3219185

R. W. Reeder, A. P. Felt, S. Consolvo, N. Malkin, C. Thompson, and S. Egelman. 2018. An experience sampling study of user reactions to browser warnings in the field. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 512:1–512:13. DOI: https://doi.org/10.1145/3173574.3174086

A. Reznichenko and P. Francis. 2014. Private-by-design advertising meets the real world. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. 116–128. DOI: https://doi.org/10.1145/2660267.2660305

L. Roberts and D. Indermaur. 2003. Signed consent forms in criminological research: Protection for researchers and ethics committees but a threat to research participants? *Psychiatry, Psychology and Law* 10, 2 (2003), 289–299. DOI: https://doi.org/10.1375/pplt.2003.10.2.289

S. Ruoti, J. Andersen, S. Heidbrink, M. O'Neill, E. Vaziripour, J. Wu, D. Zappala, and K. Seamons. 2016. "We're on the same page": A usability study of secure email using pairs of novice users. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 4298–4308. DOI: https://doi.org/10.1145/2858036.2858400

S. Ruoti, J. Andersen, T. Monson, D. Zappala, and K. Seamons. 2018. A comparative usability study of key management in secure email. In *Proceedings of the 14th Symposium on Usable Privacy and Security (SOUPS'18)*. 375–394. https://www.usenix.org/conference/soups2018/presentation/ruoti.

M. Sahin, M. Relieu, and A. Francillon. 2017. Using chatbots against voice spam: Analyzing lenny's effectiveness. In *Proceedings of the 13th Symposium on Usable Privacy and Security (SOUPS'17)*. 319–337. https://www.usenix.org/conference/soups2017/technical-sessions/presentation/sahin.

S. Samat and A. Acquisti. 2017. Format vs. Content: The impact of risk and presentation on disclosure decisions. In *Proceedings of the 13th Symposium on Usable Privacy and Security (SOUPS'17)*. 377–384. https://www.usenix.org/conference/soups2017/technical-sessions/presentation/samat-disclosure.

N. Sambasivan, G. Checkley, A. Batool, N. Ahmed, D. Nemer, L. S. Gaytán-Lugo, T. Matthews, S. Consolvo, and E. Churchill. 2018. "'Privacy is not for me, it's for those rich women'": Performative privacy practices on mobile phones by women in South Asia. In *Proceedings of the 14th Symposium on Usable Privacy and Security (SOUPS'18)*. 127–142. https://www.usenix.org/conference/soups2018/presentation/sambasivan.

S. Sannon and D. Cosley. 2018. "It was a shady HIT": Navigating work-related privacy concerns on MTurk. In *Proceedings of the Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* 1–6. DOI: https://doi.org/10.1145/3170427.3188511

S. Schechter. 2013a. *Common Pitfalls in Writing about Security and Privacy Human Subjects Experiments, and How to Avoid Them. MSR-TR-2013-5. Microsoft Technical Report.* https://www.microsoft.com/en-us/research/publication/common-pitfalls-in-writing-about-security-and-privacy-human-subjects-experiments-and-how-to-avoid-them/.

S. Schechter. 2013b. The user is the enemy, and (S)he keeps reaching for that bright shiny power button! In *Proceedings of the Workshop on Home Usable Privacy and Security (HUPS'13)*. https://www.microsoft.com/en-us/research/publication/the-user-is-the-enemy-and-she-keeps-reaching-for-that-bright-shiny-power-button/.

S. E. Schechter, R. Dhamija, A. Ozment, and I. Fischer. 2007. The emperor's new security indicators. In *Proceedings of the IEEE Symposium on Security and Privacy (SP'07)*. 51–65. DOI: https://doi.org/10.1109/SP.2007.35

S. M. Segreti, W. Melicher, S. Komanduri, D. Melicher, R. Shay, B. Ur, L. Bauer, N. Christin, L. F. Cranor, and M. L. Mazurek. 2017. Diversify to survive: Making passwords stronger with adaptive policies. In *Proceedings of the 13th Symposium on Usable Privacy and Security (SOUPS'17)*. 1–12. https://www.usenix.org/conference/soups2017/technical-sessions/presentation/segreti.

S. Sen, S. Guha, A. Datta, S. K. Rajamani, J. Tsai, and J. M. Wing. 2014. Bootstrapping privacy compliance in big data systems. In *Proceedings of the 2014 IEEE Symposium on Security and Privacy*. 327–342.

R. Shay, I. Ion, R. W. Reeder, and S. Consolvo. 2014. "My religious aunt asked why i was trying to sell her viagra": Experiences with account hijacking. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems*. 2657–2666. DOI: https://doi.org/10.1145/2556288.2557330

R. Shay, B. Ur, L. Bauer, N. Christin, L. F. Cranor, A. Forget, S. Komanduri, M. L. Mazurek, W. Melicher, and S. M. Segreti. 2015. A spoonful of sugar? The impact of guidance and feedback on password-creation behavior. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI'15)*. 2903–2912. DOI: https://doi.org/10.1145/2702123.2702586

M. Shirvanian and N. Saxena. 2014. Wiretapping via mimicry: Short voice imitation man-in-the-middle attacks on crypto phones. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. 868–879.

B. Shrestha, M. Shirvanian, P. Shrestha, and N. Saxena. 2016. The sounds of the phones: Dangers of zero-effort second factor login based on ambient audio. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. 908–919. DOI: https://doi.org/10.1145/2976749.2978328

A. Sotirakopoulos, K. Hawkey, and K. Beznosov. 2011. On the challenges in usable security lab studies: Lessons learned from replicating a study on SSL warnings. In *Proceedings of the 7th Symposium on Usable Privacy and Security* 3.

K. Spiel, L. Malinverni, J. Good, and C. Frauenberger. 2017. Participatory evaluation with autistic children. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 5755–5766. DOI: https://doi.org/10.1145/3025453.3025851

C. Stangor and J. Walinga. 2018. *Introduction to Psychology* (1st Canadian ed.). BCcampus Open Publishing. https://opentextbc.ca/introductiontopsychology/.

R. Stevens, D. Votipka, E. M. Redmiles, C. Ahern, P. Sweeney, and M. L. Mazurek. 2018. The battle for New York: A case study of applied digital threat modeling at the enterprise level. In *Proceedings of the 27th USENIX Security Symposium (USENIX Security'18)*. 621–637.

J. Tan, L. Bauer, J. Bonneau, L. F. Cranor, J. Thomas, and B. Ur. 2017. Can unicorns help users compare crypto key fingerprints? In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3787–3798.

Blase Ur, Felicia Alfieri, Maung Aung, Lujo Bauer, Nicolas Christin, Jessica Colnago, Lorrie Faith Cranor, Henry Dixon, Pardis Emami Naeini, Hana Habib, Noah Johnson, and William Melicher. 2017. Design and Evaluation of a Data-Driven

Password Meter. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI'17)*. Association for Computing Machinery, New York, NY, USA, 3775–3786. DOI: https://doi-org.proxy.bnl.lu/10.1145/3025453.3026050

B. Ur, F. Noma, J. Bees, S. M. Segreti, R. Shay, L. Bauer, N. Christin, and L. F. Cranor. 2015. "I Added '!' at the end to make it secure": Observing password creation in the lab. In *Proceedings of the 11th Symposium on Usable Privacy and Security (SOUPS'15)*. 123–140. https://www.usenix.org/conference/soups2015/proceedings/presentation/ur.

W. A. Usmani, D. Marques, I. Beschastnikh, K. Beznosov, T. Guerreiro, and L. Carriço. 2017. Characterizing social insider attacks on Facebook. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3810–3820. DOI: https://doi.org/10.1145/3025453.3025901

K. E. Vaniea, E. Rader, and R. Wash. 2014. Betrayed by updates: How negative experiences affect future security. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2671–2674.

E. Vaziripour, J. Wu, M. O'Neill, D. Metro, J. Cockrell, T. Moffett, J. Whitehead, N. Bonner, K. Seamons, and D appala. 2018. Action needed! Helping users find and complete the authentication ceremony in signal. In *Proceedings of the 14th Symposium on Usable Privacy and Security*. 17.

E. Vaziripour, J. Wu, M. O'Neill, J. Whitehead, S. Heidbrink, K. Seamons, and D. Zappala. 2017. Is that you, Alice? A usability study of the authentication ceremony of secure messaging applications. In *Proceedings of the 13th Symposium on Usable Privacy and Security (SOUPS'17)*. 29–47. https://www.usenix.org/conference/soups2017/technical-sessions/presentation/vaziripour.

I. Velásquez, A. Caro, and A. Rodríguez. 2018. Authentication schemes and methods: A systematic literature review. *Information and Software Technology*, Vol. 94. 30–37. DOI: https://doi.org/10.1016/j.infsof.2017.09.012

M. Volkamer, A. Gutmann, K. Renaud, P. Gerber, and P. Mayer. 2018. Replication study: A cross-country field observation study of real world PIN usage at ATMs and in various electronic payment scenarios. In *Proceedings of the 14th Symposium on Usable Privacy and Security (SOUPS'18)*. 1–11. https://www.usenix.org/conference/soups2018/presentation/volkamer.

D. Votipka, S. M. Rabin, K. Micinski, T. Gilray, M. L. Mazurek, and J. S. Foster. 2018. User comfort with android background resource accesses in different contexts. In *Proceedings of the 14th Symposium on Usable Privacy and Security (SOUPS'18)*. 235–250. https://www.usenix.org/conference/soups2018/presentation/votipka.

D. Warshaw, N. Taft, and A. Woodruff. 2016. Intuitions, analytics, and killing ants: Inference literacy of high school-educated adults in the US. In *Proceedings of the 12th Symposium on Usable Privacy and Security (SOUPS'16)*. 271–285. https://www.usenix.org/conference/soups2016/technical-sessions/presentation/warshaw.

R. Wash and M. M. Cooper. 2018. Who provides phishing training? Facts, stories, and people like me. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18)*. 1–12. DOI: https://doi.org/10.1145/3173574.3174066

R. Wash, E. Rader, K. Vaniea, and M. Rizor. 2014. Out of the Loop: How automated software updates cause unintended security consequences. In *Proceedings of the 10th Symposium on Usable Privacy and Security (SOUPS'14)*. 89–104. https://www.usenix.org/conference/soups2014/proceedings/presentation/wash.

Y. Yang, G. D. Clark, J. Lindqvist, and A. Oulasvirta. 2016. Free-form gesture authentication in the wild. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 3722–3735.

Y. Zou, A. H. Mhaidli, A. McCall, and F. Schaub. 2018. "I've got nothing to lose": Consumers' risk perceptions and protective actions after the Equifax data breach. In *Proceedings of the 14th Symposium on Usable Privacy and Security (SOUPS'18)*. 197–216. https://www.usenix.org/conference/soups2018/presentation/zou.

Received July 2020; revised May 2021; accepted June 2021