Empirical Research Methods in Usable Privacy and Security

Verena Distler, Matthias Fassl, Hana Habib, Katharina Krombholz, Gabriele Lenzini, Carine Lallemand, Vincent Koenig, and Lorrie Faith Cranor

1 Introduction

Researchers in the usable privacy and security (UPS) field study privacy- and security-relevant perceptions and behaviors and aim to design systems that simultaneously address requirements for usability/user experience, security, and privacy. Human-computer interaction (HCI) and social science research methods are well-suited to study many of the types of questions that are relevant in UPS, which often involve concepts such as subjective experience, attitudes, understanding, behavior and behavior change. However, there are many challenges specific to UPS that are not usually described in more generic methods textbooks. We highlight techniques

V. Distler (✉)
University of Luxembourg, HCI Research Group, Esch-sur-Alzette, Luxembourg

University of Bundeswehr, Munich, Germany
e-mail: verena.distler@unibw.de

M. Fassl · K. Krombholz
CISPA Helmholtz Center for Information Security, Saarbrücken, Germany
e-mail: matthias.fassl@cispa.de; krombholz@cispa.de

H. Habib · L. F. Cranor
Carnegie Mellon University, Pittsburgh, PA, USA
e-mail: htq@cs.cmu.edu; lorrie@cs.cmu.edu

G. Lenzini · V. Koenig
University of Luxembourg, HCI Research Group, Esch-sur-Alzette, Luxembourg
e-mail: gabriele.lenzini@uni.lu; vincent.koenig@uni.lu

C. Lallemand
University of Luxembourg, HCI Research Group, Esch-sur-Alzette, Luxembourg

Department of Industrial Design, Eindhoven University of Technology, Eindhoven, Netherlands
e-mail: carine.lallemand@uni.lu

N. Gerber et al. (eds.), *Human Factors in Privacy Research*,
https://doi.org/10.1007/978-3-031-28643-8_3

for risk representation, options for participant recruitment, ethics-related topics in study design, and biases that may play a role in UPS studies with human participants.

Structure of this Chapter We first highlight specific challenges in applying research methods with human participants to the field of UPS (section "Common Challenges in UPS Research"). We then describe the most frequently used research methods, providing a general overview for each method, potential challenges when applying the method to UPS, examples of prior UPS work, and references for further reading on the methods (Sect. 2). Next, we describe methodological "techniques" that can be used in conjunction with these methods (Sect. 3). We then describe options for recruiting participants (Sect. 4). We also discuss important ethical considerations (Sect. 5), and psychological biases (Sect. 6).

Common Challenges in UPS Research UPS studies often set out to study motivations, perceptions, or reactions related to security or privacy threats. The methodological challenges in UPS differ from other areas of human-computer interaction, as including privacy and security threats in an ecologically valid way significantly increases complexity of study designs. While lab studies, for example, can help control the environment to exclude the influence of external variables, the threat participants perceive may be vastly different than what they experience in their everyday lives. When study participants are given fictitious personal information and credentials for the purposes of the study so as to keep their personal data confidential, participants are likely to feel and behave differently than they would if their own data was at risk. In addition, when researchers simulate attacks in a lab, these attacks typically happen at a higher rate than outside of a study context, which further jeopardizes ecological validity [34]. Another challenge is that when people interact with technology, they usually aim to complete tasks that are not related to security or privacy. Thus, mentioning security or privacy may prime participants to think about these topics and cause them to behave differently than they normally would [24]. Privacy and security-related behaviors can also be perceived as socially desirable by participants, and they are likely to adapt answers and behaviors accordingly [25] (more on the social desirability bias in Sect. 6). While self-reported data is crucial to explore participants' experiences and perceptions, it is also not sufficient to quantify and fully understand past behaviors, which are difficult to remember and report accurately. UPS researchers have followed a variety of approaches to tackle these challenges, and to create a realistic experience of risk in their studies, including the use of role-playing to simulate realistic situations [90], the use of deception to simulate the presence of an adversary [16], or launching simulated attacks on research participants [24]. These approaches can bring highly valuable insights, but they need to be balanced with ethical and feasibility considerations.

2 Research Methods in UPS Studies

This section provides readers with an overview of a selection of research methods. These methods are described in more detail in many methods books. Broadly, we can think of empirical methods on a continuum of qualitative to more quantitative methods. Qualitative methods generally have the objective of exploring topics in depth and generating hypotheses inductively, while quantitative methods often build upon such qualitative research to generate research results that generalize to larger samples, or test hypotheses. This distinction is relatively fluid. For example, while interviews are typically used to gain deep understanding of a topic and generate hypotheses, they could also be used in the evaluation of a technology. Similarly, questionnaires (a quantitative method) can be used for descriptive, relational, and experimental objectives.

Regardless of method, studies generally build on previous work. Reviewing existing literature is an essential process. We begin by describing a specific type of literature review: systematic literature reviews. We then describe empirical methods, starting from qualitative methods that are used more frequently to explore and understand topics in depth, and then moving on to quantitative methods that are used more frequently to evaluate hypotheses. In each subsection, we provide a short description of the method, give examples of use of this method in UPS, and discuss how risk is represented. We then point out references that provide detailed descriptions of how to apply these methods. We include methods that are used frequently in UPS, as well as methods that are used less frequently, but seem promising.

2.1 Systematic Literature Reviews

Systematic literature reviews refer to an approach that reviews the literature by adopting explicit procedures [12]. It provides a log of the researcher's decisions, procedures, and conclusions [98] and can help avoid biases on the part of the researcher conducting the literature review [12]. Broadly, systematic literature reviews follow the process of (1) defining the purpose and scope of the review, (2) seeking out relevant studies according to keywords and inclusion criteria (e.g., articles containing the keywords within certain scientific databases), (3) narrowing down the selection of studies from step 2, and (4) analyzing each study based on the study objectives [12].

The authors of this chapter have previously conducted a systematic literature review of recent UPS conference papers [20]. We will describe the method we used and some trends in the UPS research literature we identified through this process. The objective of this systematic literature review was to understand the methods used in the UPS community, how researchers represent risk, and how deception is used in study protocols. We included papers from five top-tier UPS

venues if they mentioned security or privacy in addition to one user-related term in the title or abstract. After filtering out papers, we conducted a detailed review of the papers, excluding additional papers that did not meet our criteria. We analyzed the remaining papers according to an analysis structure to respond to our research questions.

The analysis structure researchers use to make sense of a large corpus of work can be a contribution in itself. For example, to investigate how risk was represented to research participants, we used a categorization of risk representation that is useful when considering user research methods in UPS. *Naturally occurring risk* refers to studies that rely on risk perceptions that already exist in a participant, for instance, when observing or self-reporting on naturally occurring behavior. *Simulated risk* is used in studies in which participants are asked to situate themselves in a scenario that would trigger a perception of risk (e.g., role-playing a security-critical interaction), as well as in studies that use deception to make participants believe that a simulated risk is real. *Mentioned risk* is used in studies that describe risk to research participants (e.g., "some people's partners install tracking apps on their phone"). Mentioned risk can be compared to presenting a hypothetical situation to participants to "anchor" [32] their responses in this context, but without asking them to place themselves in the situation. Some studies also *do not represent risk* at all, for instance, when simply investigating the usability of a new tool. In this chapter, we use this categorization when describing study methods.

In our literature review we found that the included papers predominantly used experiments, surveys, and interviews. The majority of papers involved naturally occurring or simulated risk. The literature review also provided an analysis structure that could be used by venues to encourage better reporting of user studies, and a checklist for researchers and reviewers to use for detailed reporting of methods. We discussed some methods that were used rarely in the sample (e.g., co-creation and participatory design, group methods such as focus groups), some participant groups that were rarely included in the analyzed sample (non-Western populations, people with disabilities, members of the LGBTQ+ community, older adults, and minors). We invite the reader to read the original paper for details [20].

2.2 Interviews

Interviewing is the most widely used method in qualitative research and also one of the top-three methods used in empirical studies in UPS. In our previous systematic literature review, 13% of papers used interviews on their own [20]. An interview typically takes the form of a conversation between an interviewer, who asks questions, and an interviewee, who answers these questions. An individual interview usually lasts between 45 and 90 min. Interviews are a powerful, yet labor-intensive data collection technique [53]. While it seems easy to have a "conversation" with a participant, research interviews require skills and the awareness of best practices to

follow, especially related to the formulation of unbiased questions (see Sect. 6) and qualitative data analysis.

The most common interviewing format in HCI and UPS is the semi-structured interview, which offers a compromise between the level of flexibility to reflect the interviewee's perspective and the comparability between interviewees. In a semi-structured interview, the interviewer has a list of topics to address, formulated as a set of questions compiled in an interview guide. There is some flexibility in the sequence and in the way the interviewee can respond, but all questions will be asked in a similar way to all interviewees. In-between the questions from the interview guide, the interviewer will use follow-up questions aimed at clarifying or exploring some points further.

Interviews most frequently use naturally occurring risks to investigate people's real-life privacy and security experiences [20]. For example, Rashidi et al. [76] interviewed undergraduates to understand their real-life privacy workarounds in the context of pervasive photography. Similarly, Ahmed et al. [1] enquired about the real-life privacy concerns of people with visual impairments. Interviews can also be used to explore naturally occurring risks in organizational settings [15, 41]. Interviewers sometimes make use of material or situations to support the interview [53], especially when involving specific users. McReynolds et al. [59] invited kids to play with connected toys in a lab setting, with their parents present. Both parents and children were interviewed about their mental model of the toys and privacy perceptions. Less frequently, interview studies use scenarios to simulate risk. For example, Vaniea et al. [99] used a set of hypothetical scenarios to elicit stories about software update experiences.

Many texts offer guidelines about preparing, conducting, and analyzing interviews [5, 12, 50, 53].

2.3 Focus Groups

Focus groups are a form of interviewing where several participants are invited to share and discuss their ideas and opinions around a predefined subject in a "group interview" [12]. The sessions are run by a moderator, who guides the participants in a rather flexible and non-intrusive way. Group interaction is essential: it helps participants to explore, clarify, and build their points of view, revealing insights to researchers in the process. Focus groups might involve a combination of individual and group tasks as well as discussion. The main challenges lie in the adequate setup of the sessions and the transcription and analysis of the data [53].

Focus groups have been used only rarely in UPS (less than 1% of papers analyzed in our previous literature review used focus groups on their own [20]), generally in studies employing naturally occurring risk. In most cases, focus groups aim at gathering qualitative in-depth insights from lay participants. For instance, Freed et al. [31] conducted 11 focus groups with 39 survivors of intimate partner violence to understand how abusers exploit technology for harmful purposes. While

discussing sensitive topics in a group setting can sometimes trigger discomfort, group interviews were useful in this context. Focus groups questions usually revolve around high-level topics, in this case how technology was used in these abusive relationships, the strategies used by participants to defend themselves, and ideas to cope with technology-enabled abuse. The findings unveiled abuser strategies and provided a nuanced, yet detailed view of these attacks, helpful for designing mitigation strategies. Sambasivan et al. [86] conducted focus groups to gain an understanding of how South Asian 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 in contexts where devices were frequently borrowed and monitored by social relations.

Focus groups can also be conducted with security and privacy experts. Following interviews and a drawing task with end-users, Murillo et al. [62] ran two focus groups with seven data deletion experts. The aim of the focus groups was to set a baseline to compare the interview findings against. Prior to the session, the experts were asked to draw how they think online data deletion works. The drawings were discussed during the focus group and acted as a support to decide the most important aspects that a lay person should know to have a good understanding about deleting online data. Finally, many disciplines have shown a growing interest in conducting focus groups online, and we expect this practice to develop in the UPS field as well. Studying privacy trade-offs, Distler et al. [22] used private social media groups to conduct asynchronous online focus groups in combination with in-person focus groups. Confronted with several scenarios, participants first noted advantages and shortcomings individually, and then discussed and confronted their opinions in the online focus group setting.

Focus groups in UPS are an underused yet insightful method, holding the potential of gathering qualitative in-depth insights into privacy and security attitudes that might help the community obtain rich qualitative results. Methodological research in social sciences has produced worthwhile recommendations to guide the setup of focus groups. Questions about number and type of participants to involve, how to moderate focus groups, and how to analyze focus group data are addressed in textbooks [5, 12, 36, 53].

2.4 Co-Creation Methods

Co-creation refers to a wide range of practices that usually bring together both designers and people who were not trained in design [54, 87]. The involvement of users goes from a temporary involvement during a workshop to an in-depth participation across all stages of the design process. Co-creation approaches treat users as co-creators and partners who are experts in their lived experience [87, 88]. In contrast, user-centered design broadly sees users as more passive subjects whose role is to inform the designers about requirements or to give feedback on their ideas. Co-creation involves users during knowledge development, idea generation, and

concept development, with designers facilitating the process and providing tools for expression. Participatory design is a related approach that attempts to understand a problem from the participants' perspective and sees methods as means for users to gain influence in the design process [40]. Participatory design allows participants to contribute to building solutions [53]. Originally, participatory design was a form of co-creation with political goals, but today, the term participatory design is often used as a synonym for co-design [9].

As there are no fixed rules for conducting co-creation studies, it is crucial to document the co-creation process in detail and justify choices. In UPS research, user understanding and acceptance of security mechanisms are reoccurring issues. Co-creation methods may contribute to improving them. Design research often focuses on providing design outcomes, but researchers can also use research-through-design approaches [109] throughout the design process itself, e.g., by engaging with wicked problems to reframe problem statements. When recruiting participants who are not experts in UPS for co-creation studies, researchers should consider additionally obtaining security and UX experts' input to create user-centered, appropriate, and functional designs.

Currently, co-creation methods are still used rarely in UPS. Only two papers in the corpus of our UPS literature review [20] used them: Egelman et al. [23] crowd-sourced 274 sketches of potential privacy icons, and Adams et al. [19] asked VR developers to contribute to a VR code of ethics on a shared online document. Outside the scope of our literature review, Weber et al. [105] used a participatory approach to create SSL warning messages. Fassl et al. [26] used collaborative design workshops to ideate potential authentication ceremonies for secure messengers. In Distler et al.'s [21] study, security and HCI experts co-created textual and visual representations of security mechanisms.

2.5 Surveys

Surveys (also called questionnaires) are one of the most common research methods in social sciences, as well as HCI and UPS [20]. 12% of papers in our previous systematic literature review used surveys on their own [20]. Surveys consist of a well-defined and carefully written set of questions to which individuals are invited to respond [53]. They are mainly used to measure people's attitudes, to gauge the existing knowledge around a topic, to profile users, or to gather feedback about user experiences [53]. Surveys are mostly administered in a written form (on paper or online) but can also be conducted face-to-face or by telephone [77]. Surveys are an effective and flexible technique that can address several research objectives and reach a larger audience than most other empirical methods. Surveys can be used as a standalone data collection method or in combination with other techniques. Noteworthy, the validity of the data produced by surveys highly depends on the rigor of the survey design: often misperceived as easy to create, poorly designed surveys

can provide meaningless or inaccurate information [5]. Methodological textbooks provide extensive guidance about survey design [77, 97].

In our systematic literature review [20], papers in UPS 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. In the case of naturally occurring risk, participants are asked about real-world behaviors in actual situations. Redmiles et al. [78], for example, investigated how users' profile, security beliefs, and knowledge influenced security behaviors. Surveys have been used in UPS to explore user's understanding, behaviors, and attitudes towards a myriad of topics, from password expiration policies and authentication [14, 39], data security practices [66], data breaches [47], private browsing [37], security indicators [28], targeted ads [72], privacy perception [85, 96], or encryption [82] to only mention a few. Researchers often compare several user profiles (using demographics questions) or populations (through sampling), for instance, expert and non-expert security practices or mental models [18, 46]. Surveys can, for example, be combined with data logs (see Sect. 2.6) or interviews (Sect. 2.2), both methods being complementary to the type of data produced by survey research. When risk is simulated in a survey study, a prototype or scenario is usually used [20]. For instance, Karunakaran et al. [47] used a scenario to invite participants to imagine that they were victims of a data breach.

2.6 Analyzing Measurement Data (Data Logs)

Measurement data can be analyzed to provide important insights into the use of a particular system, as well as the response of a system to user behaviors. In UPS research, this may involve the measurement of a specific privacy/security-related user behavior or evaluation of a privacy/security enhancing technology. Research studies using measurement data are typically conducted as field studies to analyze events of interest as they occur in the real world, where data is collected through an application developed specifically for data collection—such as a browser extension—or logs integrated into a developed system. Measurement data captured over a period of time can then be analyzed to report on frequency of occurrence of certain behaviors, longitudinal trends in the data, or performance metrics.

As with all research methods, analysis of measurement data offers advantages and disadvantages. Such studies offer the opportunity to capture users' behavior in their ordinary use of a system. This mitigates data quality issues associated with other study methods, such as participant behavior being influenced by a lab setting or poor recall of past behaviors in self-report methods. Furthermore, measurement studies analyze behaviors in environments where users encounter privacy and security risk as they naturally occur, rather than in settings where such risk must be mentioned or simulated. Logs and other data measurement tools can be instrumented to collect data from a large sample over a period of time. While this can allow for important quantitative and longitudinal insights, measurement data

may need to be accompanied with other research methods (e.g., a survey) to explain the context behind the observed data. Additionally, developing and maintaining a data collection infrastructure for measurement studies may require more resources and technical expertise relative to other methods. While there may be opportunities to repurpose existing logs and measurements to answer new research questions, any data sharing and analysis should be done following guidelines for ethical research.

Logs and measurement data have been previously used to explore several usable privacy and security topics. In our literature review, 24 papers included the use of log analysis, and 4 papers used datalogs on their own. These papers usually used naturally occurring risk. Data logs from deployed systems have been analyzed to explore users' adoption and usage of multi-factor authentication [14, 80], user responses to quarantining actions by Internet Service Providers [13], and reidentifiability risk of browsing data [7, 70]. Field studies using measurement data have also been used to evaluate prototype systems, such as those helping users manage Android permissions [69, 91]. Others have developed applications and browser extensions to measure privacy and security behaviors occurring in users' daily online activities, including characteristics about passwords used in real-world accounts [55, 71, 104], usage of private browsing mode [38], and effectiveness of phishing attacks [68]. In addition to user behavior, previous work has used measurement data to evaluate system behavior and characteristics related to different privacy and security topics, such as adoption of HTTPS [27], phishing detection [45, 67], and unsolicited phone calls [73]. Altogether, this past work—which is only a small sample of recent measurement studies in UPS—demonstrates that measurement and log data can be used to explore a wide range of research questions.

2.7 Extracting Online Datasets

In some social sciences, it is common for researchers to conduct research using datasets that are collected by government agencies or other official entities [53]. This is not (yet) a common practice in HCI or UPS, and most researchers collect their own data. Some researchers use creative approaches to extract online datasets. For example, Fiesler and Hallinan [29] collected and analyzed public comments to news articles discussing online data sharing and privacy controversies in the media. Researchers can also use other online data sources to extract and analyze datasets, for instance, from forums [74] or Twitter. In our previous systematic literature review, we found that 4% of papers analyzed datasets. Ethical concerns should be carefully considered when using "public" data for research purposes, as it is often infeasible to obtain informed consent from the people whose data are being studied and these types of studies are often not subject to ethics board review [30]. Indeed, there seems to be a lack of consensus on the role of ethical review boards in this type of research [101]. It seems promising to evaluate research approaches that use online data through the lens of contextual integrity [65, 108].

2.8 Experience Sampling Method

The experience sampling method (ESM) was developed in psychology during the 1960s and 1970s [44, 51]. It focuses on the individual participant's ongoing behavior and everyday experiences in their natural environment rather than laboratory settings. While the method was developed in psychology, it is applicable to a wide variety of questions. ESM studies usually take several weeks to collect sufficiently large samples. Participants receive a signal and then have a limited time to complete a short questionnaire about their current situation and experience. The method to receive these signals evolved over time: Earlier approaches used timers and beepers, while more recent approaches employ mobile phone apps and desktop software. Researchers regularly use time-based, situation-based, or random sampling strategies—depending on the study purpose. The primary benefit of experience sampling is the increased ecological validity compared to lab studies. Other methods also investigate naturally occurring situations, e.g., participant observation and diary studies, a qualitative approach where participants record entries about their daily lives. In contrast, ESM requires less effort from researchers and participants. However, the resulting data usually contains fewer details, and researchers often have limited control over the sampled situations and must rely on participants' situation descriptions. In practice, some qualitative diary studies and mixed-methods ESM studies may look very similar.

Experience sampling is an infrequently used method in UPS. Our UPS literature review [20] uncovered eight studies that used experience sampling (3% of the corpus). The following examples illustrate how experience sampling can provide valuable insights into naturally occurring situations: Bonné et al. [11] explored user decisions in runtime permission dialogs on Android. They used experience sampling to survey participants after each permission dialog. Reeder et al. [79] investigated users' experience with browser security warnings, using a browser extension to survey users immediately after the warnings appeared. Gallagher et al. [33] studied users' problems and frustrations with Tor Browser. They asked participants to explain their reasons every time users ended a browsing session or switched browsers. Most of these studies in UPS use a context-specific trigger for sampling experiences, but there are examples of random sampling as well: Yang et al. [107] evaluated a smartphone authentication mechanism under natural conditions, asking participants to complete a short questionnaire after each authentication task. All of the approaches repeatedly sample user experiences over long study periods, asking users for immediate and necessarily short feedback in situ.

2.9 Experiments

Experiments are an important method to identify the causes of a situation or a set of events [53], in which participants are randomly assigned to experimental

conditions. The researchers then compare the effects of these conditions on one or multiple dependent variables. Experiments are important tools to test hypotheses and research questions, and they are a highly flexible method that can be used to address many types of research questions. Studies in HCI often use experiments to compare types of devices or versions of an interface [53], using both measured variables (e.g., time to complete) and subjective variables such as satisfaction, user experience, perceived security, and acceptance of a technology. While the term experiment is sometimes used more loosely in HCI to describe any type of user test, the random assignment of participants to experimental conditions is typically a defining characteristic of an experiment [12].

Experiments are commonly used in UPS research. Our previous systematic literature review found that experimental approaches were used in more than a third of the analyzed papers [20]. An example of an experiment is a study that asked computer science students to role-play and imagine that they were responsible for creating the code for user registration of a social networking platform. Students were randomly assigned one of four experimental conditions that varied on the instructions they were given (e.g., level of security priming). The researchers compared the effects of the conditions on dependent variables such as functionality or use of secure coding practices [63]. Experimental research can also investigate perceptions of security and privacy-relevant topics. A recent study [21] used a vignette experiment to investigate various approaches to displaying encryption to non-experts. Participants were randomly assigned to conditions that varied in the visual and textual representation of encryption. The authors then compared the effects of these representations on participants' user experience, perceived security and understanding of encryption.

Our 2021 systematic literature review found that papers focusing on experiments used simulated risk representation to a large extent (35% of analyzed papers used experiments on their own [20]). As exemplified by the studies mentioned previously, risk is often simulated using scenarios that participants should situate themselves in.

Insightful experiments are challenging to design, and research on experimental design is constantly progressing. Textbooks provide relevant instructions for conducting experiments [12, 35, 53].

3 Techniques that Can Be Used in Combination with Methods

Researchers often creatively combine the methods in Sect. 2 with a variety of techniques or "tools" that can help study security- and privacy-relevant interactions and represent risk to research participants in UPS.

Assigned Tasks UPS researchers sometimes ask participants to complete tasks that can be related or unrelated to security or privacy. Examples of such security/privacy-related tasks could be attempting to send an encrypted email [81] or asking

participants to try using a new authentication method [17]. Sometimes researchers also ask participants to complete tasks that are not directly related to security or privacy so that they can observe routine tasks or tasks that incidentally have a security component, without drawing participants' attention to the security or privacy aspect.

Prototypes Many studies make use of prototypes developed by the study authors. These are often new low- or high-fidelity interface designs, icons, notices, or instructions [20]. Many of these studies use simulated risk.

Scenarios Many studies include scenarios in which researchers ask participants to imagine themselves being in a certain situation [20]. Such scenarios can be helpful to simulate risk by describing situations that would be risky to research participants. For example, some passwords research papers involve a scenario in which researchers ask participants to imagine that their email account had been hacked, and to change their password in response [49, 60]. Researchers also sometimes ask participants to role-play. For example, in one study participants were recruited together with a friend. These pairs of participants were placed in separate rooms and asked to communicate with each other using a secure email system. They were asked to role-play a scenario about completing taxes. They were instructed to share sensitive information (e.g., social security number) and treat it as if it was their own sensitive information [100].

Educational Interventions Some studies use interventions designed to educate participants about security/privacy topics. For example, Wash and Cooper educated their participants on how to detect phishing attempts [103], and Lastdrager et al. [52] evaluated how effective anti-phishing training can be with children. Warshaw et al. [102] tried to improve adults' understanding about how companies collect and make inferences about their data.

Financial Incentives While the field of behavioral economics frequently uses financial incentives in experiment designs to model real-world incentives, these approaches are still relatively rare in UPS studies [20]. In UPS, researchers often provide some compensation to research participants, but they do rarely provide additional Financial incentives can be used, for example, to provide encouragement for participants to complete privacy and security tasks well by making a part of their compensation contingent upon successful completion of a task (bonus in addition to base compensation). The consequences a user might face in real life for badly performing security and privacy tasks are different and go beyond financial disadvantage (e.g., account compromise, having to reset a password after forgetting it). However, a financial incentive can provide some additional motivation or perception of risk to research participants in a research setting. For example, Tan et al. [94] conducted an online experiment in which participants were asked to play the role of an accountant who requested confidential information from employees using a secure messaging system. The participants were asked to complete fingerprint verification for each request, and the researchers investigated whether the participants were able to recognize mismatched fingerprints. To simulate on-the-job

pressures to get the task done quickly, the fastest 15% of participants were promised a $1 bonus in addition to their base compensation ($3).

Deception When using technology, users often have primary objectives that are unrelated to security or privacy. Letting participants know that a study's objective concerns privacy and security would prime them to think about these topics, when they might not have done so in a non-research context ("security priming" [93]). For additional realism, some studies use deception. There is a grey area between avoiding priming participants and the use of deception [20]. Deception is typically defined as deliberately misleading participants or not informing them about the purpose of the investigation, usually to avoid participants giving answers that they perceive are expected by the researchers [3]. In UPS studies, deception is often used to make participants believe a risk to their security or privacy is real. Studies involving deception need to carefully consider ethics. We describe these considerations in Sect. 5.2.

4 Participant Recruitment

There are a variety of approaches to participant recruitment, and different approaches can be used in combination. In addition to study objectives, logistical concerns play a role when deciding on a recruitment strategy. It is important to report on participant recruitment and compensation transparently in research papers, as this helps readers situate study results and aids replicability. We have previously provided a checklist that can help report all relevant information [20]. We describe common sampling approaches below.

Convenience Samples Convenience samples refer to populations that the researchers can easily access, and no systematic selection criteria are applied. Examples of convenience sampling include recruiting students at the researchers' own institutions, posting flyers in an institution's neighborhood, or recruiting colleagues or friends. Currently, a large proportion of studies in UPS are based on convenience samples [20]. Note that the question of whether a convenience sample will suffice for a study depends on the research question the researchers set out to address. For example, if the objective is to understand whether a novel gesture authentication mechanism is, in principle, usable while walking, a sample of students might be sufficient. However, if the objective is to determine the proportion of adults in the United States who hold certain views on privacy, convenience sampling is not appropriate. Snowball sampling is often closely related to convenience sampling. In this approach, research participants introduce the researcher to other potential participants, who fulfil certain criteria [8].

Representative Samples If the study objective is to generalize results to a larger population, researchers need to recruit a sample of participants representative of that population. Recruiting a randomly selected, representative sample is ideal.

Representative sampling is currently rare in UPS [20] but is now easier to (partially) achieve through new crowdsourcing solutions. Some crowdsourcing solutions, for instance, offer representative sampling options based on indicators such as gender, ethnicity, and age [75]. It is important to remember that these samples may not generalize to less tech-savvy parts of the population.

Crowd Workers Platforms such as Prolific, Mechanical Turk, or Crowdflower provide access to crowd workers who volunteer to take part in paid research studies. They allow researchers to filter potential participants regarding specific characteristics or recruit large, unfiltered samples easily. They also allow researchers to recruit for a variety of methods, from remote interviews to survey experiments. When recruiting on these types of platforms, it is important to keep in mind that participants are likely more tech-savvy than the average person in the population [75].

Specific Groups of People Some studies also include specific groups of people, such as employees in an organization, experts in a relevant field (e.g., [21]), or focus on specific groups such as women, children, or older adults. Some groups, for example, minors, are considered vulnerable populations, and researchers need special ethics board approvals for their participation (also refer to Sect. 5.3).

5 Basics of Ethical Research Design with Human Participants

5.1 Ethical Core Principles

There are numerous frameworks for ethical research, such as APA's ethical principles of psychologists [2] or the Belmont Report [95], which share common principles. The ethics statements of major security conferences reference the 2012 Menlo Report [48], which we use here to illustrate four ethical core principles: (1) Respect for Persons, (2) Beneficence, (3) Justice, and (4) Respect for Law and Public Interest. While the full report provides guidelines to operationalize these core principles, this section summarizes issues that frequently arise in UPS research.

Respect for Persons Human research subjects are autonomous persons; researchers should treat them as such. They have the right to decide if and in what research they want to participate. To ensure this, researchers inform them of the type of research and ask for explicit consent to participate. Written consent forms are a common way to document this procedure. Similarly, human subjects always have the right to withdraw from a research study—researchers must remind them of this and provide an easy way to do so, e.g., contact information or link to an opt-out website. However, some research designs make a detailed informed consent procedure difficult. First, it may be an unreasonable effort for researchers to collect informed consent from all participants (e.g., analyzing existing large

datasets or when participants are hard to identify or contact). Second, research designs that deceive participants cannot collect meaningful informed consent. Researchers should debrief participants to prevent mistrust in researchers [2]. Schechter et al. [89] recommend rigorous reporting on debriefing, participants' reactions and data protection measures in these cases. In addition, we recommend allowing participants to withdraw from the study after debriefing them.

Beneficence Research should have the best interest of the participants and society in mind. Identifying benefits and potential harms is the initial step to deliberating a research project's beneficence. While participants' welfare is the primary focus, researchers should also consider other stakeholders' welfare. In the UPS field, this often concerns participants' family members, malicious actors, security product vendors, or the research community itself. These stakeholders may be involved in or affected by harm in different ways: research may reveal too much personal information or harm their social standing with their relatives or workplace. Research into users' security behavior may also reveal issues with security products and inform future attacks. Deceptive research designs may also impact researchers' reputation, making future research projects on the same topic difficult. Researchers should avoid any potential harm resulting from their research. If some harm is unavoidable, researchers must do their best to mitigate it, e.g., by anonymizing personal information, rephrasing quotes, or debriefing participants after deceptive research.

Justice Research is a burden for participants—they invest time to inform researchers about their privacy and security experiences. There is a danger of exploitative researcher-participant relationships when vulnerable participants bear the burden of research to benefit a different group of people. Hence, research participants should benefit from the research in one way or the other. Fair monetary compensation for the participants' time is a common way to ensure that. In UPS research, this is especially important when using crowdsourcing platforms (e.g., Prolific) for recruitment. Crowdworkers participate in research to make income and might not benefit in any other way. Paying workers at least the minimum wage at their location is a regular (currently not met [42]) demand of workers and researchers alike [92]. Noteworthy, participants may also reap the benefits of security research in other ways, such as personalized security education or improved security and privacy of the tools they use. We recommend reflecting and reporting on these issues when conducting studies.

Respect for Law and Public Interest Researchers need to identify and adhere to laws that apply to their research. Legal issues might be hard to navigate when security research involves reverse engineering or vulnerability disclosure. However, in UPS research, data protection laws and terms of service for crowdsourcing platforms, survey platforms, and transcriptions services are the most common issues. Collecting participants' informed consent about data collection, storage, sharing, and deletion is a best practice. In addition, data protection laws (e.g., GDPR in the European Union) give data subjects certain rights, e.g., information about data

practices and data deletion upon request. At most research institutions, support for legal issues is available. To provide transparency and accountability, researchers should responsibly share research methods, ethical evaluations, collected data, and results with researchers or policy-makers.

5.2 Ethical Considerations for Deceptive Research

In rare cases, deceiving participants about the research purpose may be necessary to address underlying research questions, e.g., knowing a study purpose might influence user behavior under investigation. Deception in research designs is a contentious issue and should remain a highly regulated exception. Guidelines recommend debriefing participants as soon as possible after the deception to avoid any psychological harm (e.g., feelings of shame, guilt, and worrying about negative impacts from assumed but not real risks). Debriefing avoids participants losing trust in researchers [2]. Research protocols involving deception usually require approval from an institution's ethics board.

Studies may deceive participants about the study objective or the presence of risk. For instance, Samat and Acquisti [84] told participants that they would share their information with a specific group of people, when in reality they did not. Marforio et al. [57] asked participants to test a banking prototype for one week and simulated a phishing attack on the prototype. Our 2021 literature review [20] shows that 6% of the sampled UPS papers used deception. Researchers mostly used deception in experimental studies on social engineering and privacy inform-and-consent mechanisms. Two-thirds of these papers mention a debriefing process, and most have IRB-approval.

5.3 Ethical Review Boards

Many universities and other organizations have an Ethical Review Board (ERB) or an Institutional Review Board (IRB) to support researchers with ethical deliberations and legal compliance. While the application process (incl. duration and required level of detail) differs between countries and universities, most applications require at least a description of the research process and the identified ethical concerns. Applying for an ERB/IRB decision should be one of the first steps of starting a research project. Continued communication is necessary when changing the research design.

The call for papers of top-tier security conferences (IEEE S&P, USENIX Security, ACM CCS, and NDSS) includes ethics statements. They require that all submissions that experiment on human subjects, analyze data derived from human subjects, or put humans at risk should discuss their research ethics and disclose whether they have an ERB/IRB-approval or exemption. These ethics

statements describe minimum requirements for human-subjects' studies; ethical community standards and review processes are continuously under scrutiny and updated accordingly.

In our 2021 review of UPS literature [20], we found that 56% of the papers obtained approval or received exempt approval, 35% did not mention the topic, and 4% were from institutions without approval procedure. The remaining papers either described a corporate internal review process or claimed to be exempt from needing approval.

When a researcher's institution does not have its own approval process, we advise stating so in the paper and transparently discussing potential ethical issues and how they were justified and mitigated. Especially in sensitive cases (e.g., studying the privacy and security concerns of political activists or survivors of intimate partner violence (IPV), or designing privacy/security solutions for an elderly population) an informal advisory board with research peers and subject-matter experts may be useful for providing feedback to ethical questions as they arise.

6 Biases in Research with Human Participants

The validity and reliability of what we measure in user studies can be compromised by systematic errors or biases. We highlight some of the most common biases that can impact UPS research.

The Method The choice of an appropriate method or the construction of an ad-hoc method bears numerous opportunities for biasing measures. For instance, issues can arise from the question order, priming effects, inappropriate response types or scale formats.

Biases Related to Study Procedures Biases are grounded in insufficient procedures, too. For instance, a lack of care in the design of a study procedure can lead to problems of *validity* (e.g., when instructions are unclear, excessively complicated, or tasks too numerous) and *reliability* (when a missing or insufficiently defined protocol leads one or several researchers to administer a method variably across participants). A possible remedy is to train and prepare researchers for data collections in a standardized way. Researchers should also pre-test all parts of their procedure.

Biases Related to the Experimental Environment Ecological validity in data collection refers to the extent to which the experimental environment is similar to the real world. It is much easier to obtain high levels of ecological validity in a field study than a lab study, but researchers can take steps to introduce enough realism to have some confidence that results may generalize to real-world conditions.

Biases Related to Sampling It is a frequent limitation that researchers unknowingly overestimate how representative of the population their sample is, thus introducing bias. This can lead to underestimating problems in UPS as these

participants are on average higher educated than the general population. Additional aspects of sample bias can result from specific geographic or time frames. It is important to be aware of the selection and the self-selection bias. The selection bias describes how any sample that was selected by rules that are not fully random sampling do not accurately describe the true population [43], and participants who voluntarily self-select to take part in research studies may represent a subset of the population that differs in terms of privacy and security perceptions and behaviors from the general population. When these biases cannot be avoided, it is important to be transparent about the sampling procedure and to discuss findings in relation to the sample.

Biases Related to the Participant Biases can influence how research participants behave and respond to questions. Typical response biases include, for example, the tendency to acquiescence, i.e., participants tend to give a positive answer more frequently than a negative one [61, 106]. Participants can also alter their behavior, e.g., towards higher performance, when they are aware of being observed. The Hawthorne effect refers to the effect of participants' awareness of study participation on their behavior [58]. Participants also tend to provide answers or show behavior they feel is positively connoted by the researchers or social norms; this is the social desirability bias. For example, in a non-anonymous setting, participants may prefer not to say that they use simplistic passwords rather than complex ones [10, 61, 106]. The halo effect describes the propensity of participants to transfer their impression (positive or negative) to a wider set of properties that they have not experienced before. Participants who feel sympathy for a brand may extend their trust to a specific service [4, 6, 106]. The recall or retrospection bias refers to memory limitations leading to vague or wrong responses. Distortion of recollections may also occur under the influence of emotional tainting [83, 97]. A specific problem in UPS research is that of security priming, which can lead to increased risk awareness in participants and can trigger unnatural responses or behavior.

Biases Related to the Researcher When conducting research with human participants, researchers can underestimate the influence of their own subjectivity when applying procedures, formulating questions, taking notes, or coding answers. Researchers should be aware of the confirmation bias, which refers to the propensity of humans to seek evidence in support of their hypotheses or beliefs, compromising neutrality in identifying all relevant data points [56]. Especially in quantitative studies, pre-registration of the research questions and methodology can be helpful for researchers to document their initial assumptions, and transparently describe how they might have adapted their analysis procedure during the study, and why. The cultural bias describes a propensity to underestimate the diversity in language, behavior, values, or conventions among the participants in a study. It can lead to inappropriate wording or color codes, or a lack of support for right-to-left languages [64]. When interacting with a participant, a frequent problem is the use of leading questions, which increase the risk of socially desirable responses [61, 77].

7 Conclusion

Studying behaviors and subjective experiences in the context of UPS is complex. Researchers need to balance realistic risk representation with practical, ethical, and legal concerns. This chapter describes social science and HCI research methods that are relevant for UPS. We describe a variety of methods, discuss how they can be used in UPS, and direct the reader to relevant resources. We discuss additional methodological tools to enhance the methods, participant recruitment, ethical challenges, and biases that could play a role in UPS research. While there is no such thing as a one-size-fits-all method, this chapter aims to provide readers with the tools to weigh the advantages and shortcomings of the described methods. We hope that UPS, as an inherently interdisciplinary field, can use the rich insights in methods research to conduct valid, ethical, and replicable science.

References

1. Ahmed, T., Hoyle, R., Connelly, K., Crandall, D., & Kapadia, A. (2015). Privacy concerns and behaviors of people with visual impairments. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 3523–3532). ACM.
2. 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
3. American Psychological Association. (2022). Deception research—APA dictionary of psychology. https://dictionary.apa.org/deception-research
4. Balzer, W. K., & Sulsky, L. M. (1992). Halo and performance appraisal research: A critical examination. *Journal of Applied Psychology, 77*(6), 975.
5. Baxter, K., Courage, C., & Caine, K. (2015). *Understanding your users* (2nd ed.). Morgan Kaufmann.
6. Bellé, N., Cantarelli, P., & Belardinelli, P. (2017). Cognitive biases in performance appraisal: Experimental evidence on anchoring and halo effects with public sector managers and employees. *Review of Public Personnel Administration, 37*(3), 275–294.
7. Bird, S., Segall, I., & Lopatka, M. (2020). Replication: Why we still can't browse in peace: On the uniqueness and reidentifiability of web browsing histories. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)* (pp. 489–503).
8. Blandford, A., Furniss, D., & Makri, S. (2016). Qualitative HCI research: Going behind the scenes. *Synthesis Lectures on Human-Centered Informatics, 9*(1), 1–115.
9. Bødker, S., Dindler, C., Iversen, O. S., & Smith, R. C. (2022). *Participatory design*. Number 1946-7680 in synthesis lectures on human-centered informatics. Springer .
10. Bogner, K., & Landrock, U. (2016). *Response biases in standardised surveys (version 2.0)*. GESIS—Leibniz-Institut für Sozialwissenschaften.
11. Bonné, B., Peddinti, S. T., Bilogrevic, I., & Taft, N. (2017). Exploring decision making with Android's runtime permission dialogs using in-context surveys. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)* (pp. 195–210). USENIX Association.
12. Bryman, A. (2015). *Social research methods*. Oxford University Press.
13. Cetin, O., Ganán, C., Altena, L., Tajalizadehkhoob, S. & van Eeten, M. (2018). Let me out! Evaluating the effectiveness of quarantining compromised users in walled gardens. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)* (pp. 251–263).

14. Colnago, J., Devlin, S., Oates, M., Swoopes, C., Bauer, L., Cranor, L., & Christin, N. (2018). "it's not actually that horrible" exploring adoption of two-factor authentication at a university. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–11).
15. Conway, D., Taib, R., Harris, M., Yu, K., Berkovsky, S., & Chen, F. (2017). A qualitative investigation of bank employee experiences of information security and phishing. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)* (pp. 115–129). USENIX Association.
16. Cranor, L. F., & Buchler, N. (2014). Better together: Usability and security go hand in hand. *IEEE Security Privacy, 12*(6), 89–93.
17. Das, S., Laput, G., Harrison, C., & Hong, J. I. (2017). Thumprint: Socially-inclusive local group authentication through shared secret knocks. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 3764–3774).
18. De Luca, A., Das, S., Ortlieb, M., Ion, I., & Laurie, B. (2016). Expert and non-expert attitudes towards (secure) instant messaging. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)* (pp. 147–157).
19. Devon A., Bah, A., & Barwulor, C. (2018). Ethics emerging: The Story of privacy and security perceptions in virtual reality. In *SOUPS '18: Proceedings of the Fourteenth USENIX Conference on Usable Privacy and Security* (p. 17).
20. Distler, V., Fassl, M., Habib, H., Krombholz, K., Lenzini, G., Lallemand, C., Cranor, L. F., & Koenig, V. (2021). A Systematic literature review of empirical methods and risk representation in usable privacy and security research. *ACM Transactions on Computer-Human Interaction, 28*(6), 1–50.
21. Distler, V., Gutfleisch, T., Lallemand, C., Lenzini, G., & Koenig, V. (2022). Complex, but in a good way? How to represent encryption to non-experts through text and visuals—Evidence from expert co-creation and a vignette experiment. *Computers in Human Behavior Reports, 5*, 100161.
22. Distler, V., Lallemand, C., & Koenig, V. (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.
23. Egelman, S., Kannavara, R., & Chow, R. (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* (pp. 1669–1678). ACM.
24. Egelman, S., King, J., Miller, R. C., Ragouzis, N., & Shehan, E. (2007). Security user studies: Methodologies and best practices. In *CHI '07 Extended Abstracts on Human Factors in Computing Systems - CHI '07* (p. 2833). ACM Press.
25. Egelman, S., & Peer, E. (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* (pp. 2873–2882). ACM.
26. Fassl, M., Gröber, L. T., & Krombholz, K. (2021). Exploring user-centered security design for usable authentication ceremonies. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI (pp. 1–15). ACM.
27. Felt, A. P., Barnes, R., King, A., Palmer, C., Bentzel, C., & Tabriz, P. (2017). Measuring https adoption on the web. In *26th USENIX Security Symposium (USENIX Security 17)* (pp. 1323–1338).
28. Felt, A. P., Reeder, R. W., Ainslie, A., Harris, H., Walker, M., Thompson, C., Acer, M. E., Morant, E., & Consolvo, S. (2016). Rethinking connection security indicators. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)* (pp. 1–14). USENIX Association.
29. Fiesler, C., & Hallinan, B. (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* (pp. 53:1–53:13). ACM.
30. Fiesler, C., & Proferes, N. (2018). "Participant" perceptions of Twitter research ethics. *Social Media + Society, 4*(1), 205630511876336.

31. Freed, D., Palmer, J., Minchala, D., Levy, K., Ristenpart, T., & Dell, N. (2018). "a stalker's paradise": How intimate partner abusers exploit technology. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18 (pp. 1–13). Association for Computing Machinery.
32. Furnham, A., & Boo, H. C. (2011). A literature review of the anchoring effect. *The Journal of Socio-Economics, 40*(1), 35–42.
33. Gallagher, K., Patil, S., Dolan-Gavitt, B., McCoy, D., & Memon, N. (2018). Peeling the onion's user experience layer: Examining naturalistic use of the Tor browser. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security* (pp. 1290–1305). ACM.
34. Garfinkel, S., & Lipford, H. R. (2014). *Usable security : History, themes, and challenges*. Morgan & Claypool Publishers.
35. Gerber, A. S., & Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation*. W. W. Norton & Company.
36. Goodman, E., Kuniavsky, M., & Moed, A. (2012). *Observing the user experience* (2nd ed.). Morgan Kaufmann.
37. Habib, H., Colnago, J., Gopalakrishnan, V., Pearman, S., Thomas, J., Acquisti, A., Christin, N., & Cranor, L. F. (2018). Away from prying eyes: Analyzing usage and understanding of private browsing. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)* (pp. 159–175). USENIX Association.
38. Habib, H., Colnago, J., Gopalakrishnan, V., Pearman, S., Thomas, J., Acquisti, A., Christin, N., & Cranor, L. F. (2018). Away from prying eyes: Analyzing usage and understanding of private browsing. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)* (pp. 159–175).
39. Habib, H., Naeini, P. E., Devlin, S., Oates, M., Swoopes, C., Bauer, L., Christin, N., & Cranor, L. F. (2018). User behaviors and attitudes under password expiration policies. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)* (pp. 13–30). USENIX Association.
40. Halskov, K., & Hansen, N. B. (2015). The diversity of participatory design research practice at PDC 2002–2012. *International Journal of Human-Computer Studies, 74*, 81–92.
41. Haney, J. M., Theofanos, M., Acar, Y., & Prettyman, S. S. (2018). "we make it a big deal in the company": Security mindsets in organizations that develop cryptographic products. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)* (pp. 357–373). USENIX Association.
42. Hara, K., Adams, A., Milland, K., Savage, S., Callison-Burch, C., & Bigham, J. P. (2018). A data-driven analysis of workers' earnings on Amazon Mechanical Turk. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (pp. 1–14). Montreal, QC: ACM.
43. Heckman, J. J. (1990). Selection bias and self-selection. In J. Eatwell, M. Milgate, & P. Newman (Eds.), *Econometrics* (pp. 201–224). Palgrave Macmillan.
44. Hormuth, S. E. (1986). The sampling of experiences in situ. *Journal of Personality, 54*(1), 262–293.
45. Hu, H., & Wang, G. (2018). End-to-end measurements of email spoofing attacks. In *27th USENIX Security Symposium (USENIX Security 18)* (pp. 1095–1112).
46. Ion, I., Reeder, R., & Consolvo, S. (2015). ". . .no one can hack my mind": Comparing expert and non-expert security practices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)* (pp. 327–346). USENIX Association.
47. Karunakaran, S., Thomas, K., Bursztein, E., & Comanescu, O. (2018). Data breaches: User comprehension, expectations, and concerns with handling exposed data. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)* (pp. 217–234). USENIX Association.
48. Kenneally, E., & Dittrich, D. (2012). The Menlo Report: Ethical principles guiding information and communication technology research. *SSRN Electronic Journal*, 14. https://www.scinapse.io/papers/2292723020

49. Komanduri, S., Shay, R., Cranor, L. F., Herley, C., & Schechter, S. (2014). Telepathwords: Preventing weak passwords by reading users' minds. In *23rd USENIX Security Symposium (USENIX Security 14)* (pp. 591–606).
50. Kvale, S. (1996). *Interviews: An introduction to qualitative research interviewing*. Sage Publications.
51. Larson, R., & Csikszentmihalyi, M. (2014). *The experience sampling method* (pp. 21–34). Springer Netherlands.
52. Lastdrager, E., Gallardo, I. C., Hartel, P., & Junger, M. (2017). How effective is anti-phishing training for children? In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)* (pp. 229–239). USENIX Association.
53. Lazar, J., Feng, J. H., & Hochheiser, H. (2017). *Research methods in human computer interaction* (2nd ed.). Burlington: Morgan Kaufmann.
54. Lee, J.-J., Jaatinen, M., Salmi, A., Mattelmäki, T., Smeds, R., & Holopainen, M. (2018). Design choices framework for co-creation projects. *International Journal of Design, 12*(2), 17.
55. Lyastani, S. G., Schilling, M., Fahl, S., Backes, M., & Bugiel, S. (2018). Better managed than memorized? Studying the impact of managers on password strength and reuse. In *27th USENIX Security Symposium (USENIX Security 18)* (pp. 203–220).
56. MacCoun, R. J. (1998). Biases in the interpretation and use of research results. *Annual Review of Psychology, 49*(1), 259–287.
57. Marforio, C., Masti, R. J., Soriente, C., Kostiainen, K., & Čapkun, S. (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* (pp 540–551). ACM.
58. McCambridge, J., Witton, J., & Elbourne, D. R. (2014). Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects. *Journal of Clinical Epidemiology, 67*(3), 267–277.
59. McReynolds, E., Hubbard, S., Lau, T., Saraf, A., Cakmak, M., & Roesner, F. (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*, CHI '17 (pp. 5197–5207). ACM.
60. Melicher, W., Kurilova, D., Segreti, S. M., Kalvani, P., Shay, R., Ur, B., Bauer, L., Christin, N., Cranor, L. F., & Mazurek, M. L. (2016). Usability and security of text passwords on mobile devices. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 527–539).
61. Müller, H., Sedley, A., & Ferrall-Nunge, E. (2014). Survey research in HCI. *Ways of knowing in HCI* (pp. 229–266). Springer.
62. Murillo, A., Kramm, A., Schnorf, S., & De Luca, A. (2018). "if i press delete, it's gone": User understanding of online data deletion and expiration. In *Proceedings of the Fourteenth USENIX Conference on Usable Privacy and Security*, SOUPS '18 (pp. 329–339). USENIX Association.
63. Naiakshina, A., Danilova, A., Tiefenau, C., Herzog, M., Dechand, S., & Smith, M. (2017). Why do developers get password storage wrong?: A qualitative usability study. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (pp. 311–328). ACM.
64. Newcomer, K. E., Hatry, H. P., & Wholey, J. S. (2015). Culturally responsive evaluation. In *Handbook of practical program evaluation* (p. 281). https://www-wiley-com.proxy.bnl.lu/en-us/Handbook+of+Practical+Program+Evaluation%2C+4th+Edition-p-9781118893609
65. Nissenbaum, H. (2004). Privacy as contextual integrity. *Washington Law Review, 79*(1), 119–157.
66. Nthala, N., & Flechais, I. (2018). Informal support networks: An investigation into home data security practices. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)* (pp. 63–82). USENIX Association.

67. Oest, A., Safaei, Y., Zhang, P., Wardman, B., Tyers, K., Shoshitaishvili, Y., & Doupé, A. (2020). PhishTime: Continuous longitudinal measurement of the effectiveness of anti-phishing blacklists. In *29th USENIX Security Symposium (USENIX Security 20)* (pp. 379–396).
68. Oest, A., Zhang, P., Wardman, B., Nunes, E., Burgis, J., Zand, A. Thomas, K., Doupé, A., & Ahn, G.-J. (2020). Sunrise to sunset: Analyzing the end-to-end life cycle and effectiveness of phishing attacks at scale. In *29th USENIX Security Symposium (USENIX Security 20)*.
69. Olejnik, K., Dacosta, I., Machado, J. S., Huguenin, K., Khan, M. E., & Hubaux, J.-P. (2017). SmarPer: Context-aware and automatic runtime-permissions for mobile devices. In *2017 IEEE Symposium on Security and Privacy (SP)* (pp. 1058–1076). IEEE.
70. Olejnik, L., Castelluccia, C., & Janc, A. (2012). Why Johnny can't browse in peace: On the uniqueness of web browsing history patterns. In *5th Workshop on Hot Topics in Privacy Enhancing Technologies (HotPETs 2012)*.
71. Pearman, S., Thomas, J., Naeini, P. E., Habib, H., Bauer, L., Christin, N., Cranor, L. F., Egelman, S., & Forget, A. (2017). Let's go in for a closer look: Observing passwords in their natural habitat. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (pp. 295–310).
72. Plane, A. C., Redmiles, E. M., Mazurek, M. L., & Tschantz, M. C. (2017). Exploring user perceptions of discrimination in online targeted advertising. In *26th USENIX Security Symposium (USENIX Security 17)* (pp. 935–951).
73. Prasad, S., Bouma-Sims, E., Mylappan, A. K., & Reaves, B. (2020). Who's calling? Characterizing robocalls through audio and metadata analysis. In *29th USENIX Security Symposium (USENIX Security 20)* (pp. 397–414).
74. Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media + Society, 7*(2), 205630512110190.
75. Prolific (2022). Representative samples. https://researcher-help.prolific.co/hc/en-gb/articles/360019236753-Representative-samples
76. Rashidi, Y., Ahmed, T., Patel, F., Fath, E., Kapadia, A., Nippert-Eng, C., & Su, N. M. (2018). "you don't want to be the next meme": College students' workarounds to manage privacy in the era of pervasive photography. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)* (pp. 143–157). USENIX Association.
77. Rea, L. M., & Parker, R. A. (2014). *Designing and conducting survey research: A comprehensive guide*. Wiley.
78. Redmiles, E. M., Kross, S., & Mazurek, M. L. (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*, CCS '16 (pp. 666–677). ACM.
79. Reeder, R. W., Felt, A. P., Consolvo, S., Malkin, N., Thompson, C., & Egelman, S. (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* (pp. 1–13). ACM.
80. Reynolds, J., Samarin, N., Barnes, J., Judd, T., Mason, J., Bailey, M., & Egelman, S. (2020). Empirical measurement of systemic 2FA usability. In *29th USENIX Security Symposium (USENIX Security 20)* (pp. 127–143).
81. Ruoti, S., Andersen, J., Heidbrink, S., O'Neill, M., Vaziripour, E., Wu, J., Zappala, D., & Seamons, K. (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* (pp. 4298–4308). ACM Press.
82. Ruoti, S., O'Neill, M., Zappala, D., & Seamons, K. (2016). User attitudes toward the inspection of encrypted traffic. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)* (pp. 131–146). USENIX Association.
83. Russell, D. M., & Chi, E. H. (2014). Looking back: Retrospective study methods for HCI. In *Ways of knowing in HCI* (pp. 373–393). Springer.

84. Samat, S., & Acquisti, A. (2017). Format vs. content: The impact of risk and presentation on disclosure decisions. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)* (pp. 377–384). USENIX Association.
85. Samat, S., Acquisti, A., & Babcock, L. (2017). Raise the curtains: The effect of awareness about targeting on consumer attitudes and purchase intentions. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)* (pp. 299–319). USENIX Association.
86. Sambasivan, N., Checkley, G., Batool, A., Ahmed, N., Nemer, N., Gaytán-Lugo, L. S., Matthews, T., Consolvo, S., & Churchill, E. (2018). "privacy is not for me, it's for those rich women": Performative privacy practices on mobile phones by women in South Asia. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)* (pp. 127–142). USENIX Association.
87. Sanders, E. B.-N., & Stappers, P. J. (2008). Co-creation and the new landscapes of design. *CoDesign, 4*(1), 5–18.
88. Sanders, L. (2008). An evolving map of design practice and design research. *Human Factors, 15*, 7.
89. Schechter, S. (2013). Common pitfalls in writing about security and privacy human subjects experiments, and how to avoid them. https://www.microsoft.com/en-us/research/publication/common-pitfalls-in-writing-about-security-and-privacy-human-subjects-experiments-and-how-to-avoid-them/ Tech report nr MSR-TR-2013-5.
90. Schechter, S. E., Dhamija, R., Ozment, A., & Fischer, I. (2007). The emperor's new security indicators. In *2007 IEEE Symposium on Security and Privacy (SP'07)* (pp. 51–65). IEEE.
91. Sikder, A. K., Aksu, H., & Uluagac, A. S. (2017). 6thSense: A context-aware sensor-based attack detector for smart devices. In *26th USENIX Security Symposium (USENIX Security 17)* (pp. 397–414).
92. Silberman, M. S., Tomlinson, B., LaPlante, R., Ross, J., Irani, L., & Zaldivar, A. (2018). Responsible research with crowds: Pay crowdworkers at least minimum wage. *Communications of the ACM, 61*(3), 39–41.
93. Sotirakopoulos, A., Hawkey, K., & Beznosov, K. (2011). On the challenges in usable security lab studies: Lessons learned from replicating a study on SSL warnings. In *Proceedings of the Seventh Symposium on Usable Privacy and Security* (p. 3). ACM.
94. Tan, J., Bauer, L., Bonneau, J., Cranor, L. F., Thomas, J., & Ur, B. (2017). Can unicorns help users compare crypto key fingerprints? In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 3787–3798).
95. The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. The Belmont Report. Technical report, 1979.
96. Torabi, S., & Beznosov, K. (2016). Sharing health information on Facebook: Practices, preferences, and risk perceptions of North American users. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)* (pp. 301–320). USENIX Association.
97. Tourangeau, R., Rips, L. J., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press
98. Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management, 14*(3), 207–222.
99. Vaniea, K. E., Rader, E., & Wash, R. (2014). Betrayed by updates: How negative experiences affect future security. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2671–2674). ACM.
100. Vaziripour, E., Wu, J., O'Neill, M., Whitehead, J., Heidbrink, S., Seamons, K., & Zappala, D. (2017). Is that you, Alice? A usability study of the authentication ceremony of secure messaging applications. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)* (pp. 29–47). USENIX Association.
101. Vitak, J., Proferes, N., Shilton, K., & Ashktorab, Z. (2017). Ethics regulation in social computing research: Examining the role of institutional review boards. *Journal of Empirical Research on Human Research Ethics, 12*(5), 372–382.

102. Warshaw, J., Taft, N., & Woodruff, A. (2016). Intuitions, analytics, and killing ants: Inference literacy of high school-educated adults in the US. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)* (pp. 271–285). USENIX Association.
103. Wash, R., & Cooper, M. M. (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* (pp. 1–12). ACM Press.
104. Wash, R., Rader, E., Berman, R., & Wellmer, Z. (2016). Understanding password choices: How frequently entered passwords are re-used across websites. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)* (pp. 175–188).
105. Weber, S., Harbach, M., & Smith, M. (2015). Participatory design for security-related user interfaces. In *Proceedings 2015 Workshop on Usable Security*. Internet Society.
106. Wetzel, E., Böhnke, J. R., & Brown, A. (2016). Response biases. In F. T. L. Leong, D. Bartram, F. M. Cheung, K. F. Geisinger, & D. Iliescu (Eds.), *The ITC international handbook of testing and assessment* (pp. 349–363). Oxford University Press.
107. Yang, Y., Clark, G. D., Lindqvist, J., & Oulasvirta, A. (2016). Free-form gesture authentication in the wild. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 3722–3735). ACM.
108. Zimmer, M. (2018). Addressing conceptual gaps in big data research ethics: An application of contextual integrity. *Social Media + Society, 4*(2), 205630511876830.
109. Zimmerman, J., Forlizzi, J., & Evenson, S. (2007). Research through design as a method for interaction design research in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 493–502). ACM.