Towards robust experimental design for user studies in security and privacy

Kat Krol, Jonathan M. Spring, Simon Parkin and M. Angela Sasse
University College London

Abstract

Background: Human beings are an integral part of computer security, whether we actively participate or simply build the systems. Despite this importance, understanding users and their interaction with security is a blind spot for most security practitioners and designers.
Aim: Define principles for conducting experiments into usable security and privacy, to improve study robustness and usefulness.
Data: The authors' experiences conducting several research projects complemented with a literature survey.
Method: We extract principles based on relevance to the advancement of the state of the art. We then justify our choices by providing published experiments as cases of where the principles are and are not followed in practice to demonstrate the impact. Each principle is a discipline-specific instantiation of desirable experiment-design elements as previously established in the domain of philosophy of science.
Results: Five high-priority principles – (i) give participants a primary task; (ii) incorporate realistic risk; (iii) avoid priming the participants; (iv) perform double-blind experiments whenever possible and (v) think carefully about how meaning is assigned to the terms *threat model*, *security*, *privacy*, and *usability*.
Conclusion: The principles do not replace researcher acumen or experience, however they can provide a valuable service for facilitating evaluation, guiding younger researchers and students, and marking a baseline common language for discussing further improvements.

1 Introduction and aims

Security mechanisms are incorporated into IT systems to protect them or the information they contain. Protection can extend to the regular activities of users and the systems they interact with, relying on them to behave in a secure manner. In the past, humans have been referred to as the "weakest link" in security, where insecure actions – be they malicious or unintended – can jeopardise systems and expose them to threats [55]. Research from 1999 onwards [1, 64] has shown that the systems themselves can introduce security weaknesses, by being *unusable* and a *bad fit* to the tasks performed by users of those systems [53], in turn making secure behaviour difficult. Usability is then an important factor in the design and deployment of security mechanisms, where treating usability as an after-thought to be added to an existing system can instead impact security [65].

Particularly for complex IT systems, users tend to fail – rather than knowingly refuse – to comply with security expectations [33]. Potential reasons why users do not comply include: security-compliant behaviour demands too much of them, the need to comply is not obvious to the individual, or the definition of compliant behaviour is simply unworkable. In all of these cases, individuals may rationalise that behaving securely is not worth their time or effort if there are no perceived personal benefits [24]. The result of this rationalisation is often that users develop coping strategies to reduce the demands of security, work around security systems, or become disenfranchised with security if it continues to act as a distraction and a barrier [2]. For instance, users may rationalise when considering whether to maintain a written note of a password as a recall aid, even though humans are not adapted to such memory tasks.

The users' efforts to make use of the system, by rationalisation or otherwise, often conflicts with the security architects' efforts to secure the system [61]. User efforts to cope with the demands of security are only exacerbated by security architects who insist that users can be trained to perform security tasks (e.g., [6]), even though usable security research demonstrates that there are cases where this is untrue [9]. Individuals and groups develop their own alternatives to security if the security architecture does not accommodate the users' security needs. Often groups use their own approximation of what a secure system should do while attempting to respect the need to behave securely [34]. Mandates and restricted systems further undermine users when security managers and designers do not understand the user [25, 52]. Thus the user aspect of the system cannot be avoided, ignored, or designed out by the security architect.

The body of evidence in security usability is growing both in general knowledge on how users make security and privacy choices and also for the use and challenges of specific technologies, such as encryption [64, 57, 49]. One notable development was the definition of "Grand Challenges" for achieving user-centred security in 2005 by Zurko [67], stating that for those developing secure systems:

> "The body of experience testing the usability of security both in the lab and in context will define the techniques and tools we need and can use. It will also generate a body of best practice we can begin to systematize in checklists and expert evaluations."

In this paper, we offer such a systematisation by producing a set of principles for user studies of both security and privacy. The principles complement each other, and are interdependent. Such a set of principles can support the development of robust study outcomes, comparison of results across user studies, and composition into a meaningful body of evidence centred around the users' security technology experience. We review a concise set of experiments studying user security technology use. The result of the review is five *principles* which can be reviewed in advance of performing a study or to help guide evaluation of past studies.

Zurko [67] states that explicit security mechanisms that are incomprehensible to users and which are not integrated with the task are not effective. There is then a need to capture end-user understanding of security and how security fits with their activities. Researchers who consider these principles will have a language to express assurance that their study is applicable, consistent, reliable, and should be believed. Furthermore, the principles assist in establishing relationships among outcomes of different studies on elements such as: identifying user needs, user risk profiles, impact of using particular technologies, and the impact of certain more-or-less controlled conditions of use. Findings can then be collated within specialised frameworks, where efforts are already underway in the research community – these include the "human in the loop" framework developed by Cranor in 2008 [13], and a repository of behavioural science findings as relate to IT-security [44].

The paper is arranged as follows: Section 2 describes background on general rules of research validity; Section 3 describes the process followed to derive the principles; Section 4 details the recommended study principles for examining the usability of security and privacy technologies. Discussion follows in Section 5, followed by Conclusions.

2 Background

Hatleback and Spring [22] identify and explain four desirable experiment design features by analogy with biology that should apply to experiments in computing and computing security, analogous to principles proposed specifically in malware research [51]. The four principles are:

Internal validity: The experiment is of "suitable scope to achieve the reported results" and is not "susceptible to systematic error" [22, p. 451]

External validity: The result of the experiment "is not solely an artifact of the laboratory setting" [22, p. 451]

Containment: No "confounds" in the results, and no experimental "effects are a threat to safety" of the participants, the environment, or society generally [22, p. 452]

Transparency: "there are no explanatory gaps in the experimental mechanism" and the explanatory "diagram for the experimental mechanism is complete" in that it covers all relevant entities and activities. [22, p. 452]

These four terms come from a background of experimental and quantitative research. In considering robust experimental principles in the junction between IT-security practice and behavioural sciences research, Pfleeger and Caputo [44] suggest that steps be taken to reduce confounding variables and biases, to then support transferability. Qualitative and case-study based research methods have analogous principles, although they are derived and ensured differently. The qualitative research methods tradition roughly makes the following translations. Validity [43] usually refers specifically to internal validity, whereas transferability (or generalisablity) [30] maps to external validity, trustworthiness [40] and credibility [29] map to transparency, and containment is expressed as ethics in research design and execution [11].

A famous principle in philosophy of science is falsifiability: the idea that good theories are those whose truth can be tested and hypothetically fail [45]. For Popper, a good experiment is taken to be one that tests an existing theory. Successful theories pass more tests than others. However, one must immediately ask how to design such an experiment, and what features it should have. Further, when two results conflict, we must evaluate the strength of evidence provided by each experimental result [39, p. 146]. Our principles provide positive, constructive guidance for both these processes that the falsifiability concept does not directly supply. That is, these principles provide something of the ”how” to experiment design.

3 Methodology

Our methodology for selecting which principles to discuss followed two phases. First, we oriented our search in broad strokes by the categories of desirable features introduced in the Section 2. These categories are derived from scientific investigation more broadly and enjoy general support across multiple disciplines, and thus provide a reasonable starting point. They are also explicitly very general, and so require more detailed specification for challenges common to our particular field of interest, user studies in security and privacy.

The second phase of our methodology was to evaluate and select principles based on our expertise and experience. We thought it important that studies should be realistic, bias as little as possible and use precise language and we structured these high-level goals into principles. The process of specifying them was as follows. First priority was that we had personal empirical experience designing studies that meet these principles and overcome the underlying challenge. Thus we prioritised principles for which we feel we have an adequate and accurate formulation. We also selected principles based on our perception of the importance to a high-quality study, where high-quality means one can explain adequately how it meets the four high-level rules of internal validity, external validity, containment, and transparency. These properties are not commensurate, that is they are not measured in equivalent units and thus are not directly comparable. Therefore we employed expert judgement to decide whether a certain drop in containment from one common challenge is more or less damaging to study quality than some certain drop in external validity, for example. We did not consider a rubric or other counting exercise to adequately help our principle selection process; the qualities are too contextual and rich to so easily put into bins.

Thus, one may reasonably disagree with our choice of principles. We would like to stress that the five principles we propose are what we have found important in our research and they might not be applicable to all types of studies and all research areas within usable security and privacy. As part of future work, our proposed principles should enable quantitative, empirical measurement of study outcome differences to further improve study design understanding. This experience-first strategy is more likely to produce an incomplete list, but our recommendations are more likely to be accurate. Secondly, we emphasise our set of principles is not meant to be exhaustive, but only accurate and high-value recommendations. From this point of view, additional principles of equally high quality are welcome from the rest of the community based on others' study design experience. At present, we prefer to be somewhat conservative in our coverage and confident the principles we recommend are accurate and useful.

4 Results: Principles

Our principles come from one of two angles, roughly those from security studies and those from user studies. More specifically, our principles arise from the following two challenges, (1) subject-matter-specific problems common to security generally interacting for the first time with techniques for exploring user experience, and (2) general challenges of human experience studies that have particularly pervasive or damaging impact when they arise in security usability. We propose five principles for robust experiment design; the first three are security subject-matter issues, and the second two are general user experience study problems with particular impact for the applicability of behavioural research outcomes in IT security [44].

- Give participants a **primary task**
- Ensure participants experience **realistic risk**
- **Avoid priming** the participants
- Perform experiments **double blind** whenever possible
- **Define** these elements precisely: threat model; security; privacy; usability

These five principles for robust studies in usable security are best viewed as subject-matter specific elaborations of desirable experiment-design elements for robust study design from across the sciences. Although in some important ways these principles are not new because they are based on existing ideas, in other important ways these recommendations are unique because they have been tailored to the specific challenges common to user studies in security and privacy.

The genesis of our principles is learning from experiments we have participated in or have read about; in parallel, the target impact of our proposed principles is to better learn from experiment. We have generated the principles by learning from shortcomings in our own experiment designs and those of others. However, in order to best learn from experiments in the future it would be prudent to follow the principles as recommendations or heuristics for overcoming some of the most common errors that arise in usable security studies; thus our principles are open to future revision, addition, and amendment as warranted.

We use past work as a series of case studies to elicit these principles, and use analogy to existing literature in other fields as evidence that our conclusions are robust and not mere idiosyncrasies of the cases used. Usable security makes use of methods from both qualitative and quantitative research disciplines. A measurement study may be used on a subset of participants to examine the extent to which the values reported by the humans match objective values of behaviour captured by a sensor. For example, privacy studies have repeatedly identified a discrepancy between reported preferences and actual behaviours (e.g., [60, 28]). Therefore we make use of the study design principles from both traditions, where appropriate. Subsequently, we describe how each of the five principles relates to widely-accepted generic principles of good study design.

4.1 Primary task

By giving participants a primary task in a study we make sure they are put in a realistic situation. In real life, people use computers in order to accomplish some task, be it to send an email, make a purchase or search for information, and so security as a task is secondary to a main purpose.

In usable security research, Brostoff and Sasse [10] were the first to have a primary task in their study. In a 3-month field trial, 34 students used an authentication mechanism called Passfaces to access their course materials. Although users were positive about the idea behind Passfaces when asked about it, a 3-month trial of participants using Passfaces and passwords in practice painted a different picture. The results show that the frequency of logging in to the system dropped when Passfaces guarded access to the system; participants logged in with one third of the frequency when they authenticating using a grid of Passfaces rather than passwords. Since logging in using Passfaces took longer, more recent research of security behaviours (e.g., [63]) would in retrospect imply that participants decided that it was not worth their time to spend a minute logging in only for a few minutes of work on a system.

Giving participants a primary task while we study their security behaviour is related to two important features in usable security. First, users in the real world have a primary task which is interrupted by performing security tasks. Including a primary task makes sure the experiment simulates the real world accurately enough to be meaningful; this is a form of ensuring external validity or transferability. A primary task adds external validity in another way, namely we know from psychology that human attention and other mental resources are bounded [59], where such bounds can impact security [3]. Further to this, users would rather achieve their goal than be distracted by secondary tasks that divert their attention from the primary task [63]. Herley [19] urges security designers to be mindful of how much security-related effort is demanded of users, and to use what is available to them wisely. Having the full mental resources of a participant available for a security task in a study setting does not necessarily translate to that person wanting – or being able – to devote their mental resources solely to security in a more realistic setting.

Giving participants a primary task in a study is not always appropriate. For example, user testing of a new authentication mechanism is a multi-stage process from requirements gathering to evaluation post-adoption. At one of the early stages, it is advisable to conduct a performance study with users to assess if a security-related task is achievable. For example, it would be confirmed whether it is possible to read and enter the digits from an RSA token into an entry field within the defined expiry window for a generated series of digits. In research on CAPTCHAs (Completely Automated Public Turing test to tell Computers and Humans Apart), many studies have focused on establishing if a paricular type of CAPTCHA is decipherable (e.g., [20]). Ideally such performance evaluations would be complemented with studies to assess user acceptance and suitability in real-life interactions to assess whether solutions are viable in genuine user populations.

Creating a primary task is difficult, and requires time and effort. Preibusch [47] provided a guide on how to study consumers' privacy choices in the tradition of behavioural economics and advocated using real-world shopping scenarios as a main approach. Researchers should create real shopping scenarios to study privacy choices; in this instance real means the participant can browse in an online shop, buy and pay for goods, and receive them. Studies in privacy have used primary tasks including purchasing gourmet goods [46], DVDs [5] and cinema tickets [31]. Some examples of primary tasks used in security studies include asking participants to buy goods from online retailers (e.g., [4]) and evaluate a tool for summarising academic articles [36]. While all experiments mentioned above were confined to a university laboratory, researchers are also increasingly conducting security experiments in the wild. A notable example is a field experiment by Felt et al. [17] which tested six proposed SSL warnings in Google Chrome and recorded 130,754 user reactions. The research has superior methodology since the behaviour of actual users is recorded as they are going about their daily online activities, and there is arguably no better primary task than the actual one that a user naturally chooses to do themselves. However using such superior methodology is not within reach of most academics and more collaborations between industry and academia are necessary to make such studies possible.

4.2 Realistic risk

Like the importance of a primary task, a realistic risk is part of the design principle of providing a realistic task environment for study participants, because the potential for real consequences is part of a realistic experience of security. Enumeration of the risks of usability failure is also important to the design of secure systems [67]. Participants are under 'realistic risk' when they perceive there is a threat to the security and privacy of their own information.

An experiment should introduce realistic risk to participants because people behave differently if they know a situation is a simulation. Lack of a realistic risk threatens external validity, where this threat stems from the fact that participants' perception of risk is one of the things (implicitly) being tested, and it changes if participants know a situation is a simulation. In this sense, lack of realistic risk causes the experimental results to be solely an artifact of the laboratory setting, with no adequate analog in the real world, and so transferability or external validity is undermined. Participant risk perception variability also represents a threat to the internal validity of a study when participants are exposed to different perceived risks without measuring, controlling, or monitoring those differences during the study.

Studies have introduced realistic risk to participants in different ways. Schechter et al. [54] (described in more detail below) asked a group of their participants to log in with their actual credentials to online banking. In a study by Beresford et al. [5], participants were purchasing DVDs and entered their own details to complete the purchase and have the products shipped.

The impact of using participants' actual credentials has been tested directly. Schechter et al. [54] tasked their participants with performing different online banking tasks, and manipulated a range of different website authentication measures such as HTTPS indicators and website authentication images. A group of participants in their experiment used their actual credentials while others role-played with simulated credentials. The researchers found that those participants who used their own credentials in the experiment behaved more securely than those using credentials provided for them.

In a study by Krol et al. [36], participants brought their own laptops to the laboratory and if they downloaded a file despite a security warning, it could have potentially infected their own computer with a virus. In interviews afterwards, a few participants stressed that if they have to download something from an untrusted source, they would do it on a public shared computer in order not to put their own machine in jeopardy. However, owning the laptop is not the only element of realism perceived by participants as 29 out of 120 participants said they considered the laboratory a trusted environment and assumed that the researchers checked the files beforehand and would not let them download something malicious. This fact highlights the need for continued assessment of users' perceptions of risk, both before and after studies, to improve researchers' interpretation of results and understanding of user attitudes.

Obviously inserting a realistic risk into a study protocol causes an interesting trade-off with containment, however, that must be addressed through the institutional review board (IRB) process. If the IRB is unfamiliar with the relevant technologies, the Menlo report [14] provides a framework for deciding whether the study poses too much of a threat to prospective participants. The Menlo report elaborates four principles for information and computer technology (ICT) research: respect for persons, beneficence, justice, and respect for law and public interest. These principles are based on ethics in biomedical studies but are thoroughly adapted for the ICT context.

The challenge of creating an ethically sound study with a realistic risk may lead a researcher to opt for sacrificing the external validity and ignoring this principle. Usually when a study sacrifices external validity it is to gain internal validity, losing representativeness in order to more carefully control the effects being studied. When internal and external validity are exchanged in this way, it immediately suggests a family of studies, some with strict controls and some descriptions of the real world and a gradient of more or less controlled studies in between, that could be synthesised in order to provide appropriate explanation for the phenomena. Relaxing the principle of a realistic risk to the participant does not provide such an exchange; if anything it negatively impacts both internal and external validity. The research into security usability has up to this point done a great deal of work in identifying factors which influence security behaviour – increasingly research is finding that the properties or severity of these factors can encourage a particular response to security. Individual utility of security can be influenced by factors personal to them; for example, the complexity and number of passwords that a person must manage can – once it reaches a perceived 'limit' – encourage the reuse of passwords or a reliance on recall aids such as written notes [2, 24].

4.3 Avoidance of priming

Priming is exposing participants to some information that might influence how they subsequently behave in the experiment. Non-priming the participant helps avoid biases such as demand characteristics where the participant gives answers based on what they believe the experimenter expects of them. Non-priming is an issue of internal validity, but also containment if the researcher comes into possession of personal or otherwise sensitive information. Non-priming can be achieved by simply not telling participants much about the purpose of the study, it can range from keeping the study description general to actively telling lies to participants. A common way to avoid priming is to deceive participants about the actual purpose of the study. Deception has been used in our field of research; Egelman et al. [15] advocate deception for user studies in security and privacy to produce generalisable findings. Krol et al. [36] told their participants they were examining a summary tool for academic papers where in reality they studied participants' reactions to download warnings.

Again ethical questions arise from the fact that participants are lied to. Psychology has traditionally dealt with this dilemma by requiring researchers to debrief participants at the end of the study and tell them what the actual purpose of the research was. However researchers have warned about potential negative consequences that might arise from deception. Horton et al. [27] emphasise that using deception can make participants distrust researchers in future studies. Researchers in the field of economics tend to avoid deception altogether as this could falsify the research results [26].

4.4 Double blind

In a double blind experiment, both the participant and the person executing the experiment do not know details of the study – this limits the capacity for either party to influence the study outcomes through knowledge of the study design itself. Traditionally used in medicine [50], the person executing the experiment would not be informed as to whether a patient is receiving an active medicine or a placebo. In this way, the designers of a medical trial hope to avoid a situation where an experimenter administering medicine treats the subject differently or influences the results in any other way. Double blind experiment design can improve internal validity and containment by preventing accidental transmission of biases, research goals, or sensitive information between the researcher and the participants.

To the best of our knowledge, experimental procedures using double blind have been used only once so far in usable security and privacy research. Malheiros et al. [41] studied individuals' willingness to disclose information in a credit card application. They employed three undergraduate psychology students to conduct experimental sessions. The students were told that the study was exploring individuals' willingness to disclose different types of information on a loan application for an actual loan provider. In reality, the study was looking at participants' privacy perceptions.

As previously, there are ethical considerations with not telling the entire truth not only to participants but also the person executing the experiment. Running a useful double-blind experiment introduces challenges to the experiment design. For example, if the person who executes the experiment is unaware of the purpose of the study, they for example cannot ask specific questions in response to the participant's behaviour, which may be valuable to adequately interpret the participant's behaviour in situ. Debriefing at the end of the study session might not be possible as the experimenter is not adequately prepared to do so themselves, and another researcher would need to be present to debrief the participant. Such an approach would further require that the experimenter be debriefed at the end of the study.

4.5 Define: threat model, security, privacy, usability

There are two important ways in which the researcher must carefully attend to how meaning is assigned to terms during explanation and during execution. Firstly, terms must have precise and well-defined meanings when articulating the design, protocol and results of an experiment to colleagues; secondly and more subtly, the researcher should be careful not to bias participants by priming them with definitions provided during the course of a study. In the first case, being clear and consistent with definitions during experiment design and execution improves internal validity. This reduces the chances for error or imprecision that would lead to systemic design flaws, as would result from confusion of similar concepts that are actually distinct at the detail level of experimental examination. Clear definitions improve transparency, trustworthiness, and credibility when describing and explaining an experiment. In the more subtle case, it is generally desirable that the researcher not provide any definitions of terms to the subject participants, to avoid biasing the participants' answers. This sense of attention to definitions overlaps heavily with the avoidance of priming, discussed in more detail in Section 4.3.

The terms we find to be most commonly impacted by definitional problems are *threat model*, *security*, *privacy*, and *usability*. These words are central to all research in the field, so it is both unsurprising and troubling that the terms are hard to define. Definitional disputes about the term information continue in information science, for example [66]. The difficulty is unsurprising because research in any field can be interpreted as wrestling with creating precise agreement for defining the terms and relations among them that adequately describe the mechanisms under study. The lack of definition is simultaneously troubling because lack of specificity prevents a genuine discussion about the merits of competing definitions to capture the mechanism adequately and instead hides behind ambiguity. Researchers should consider and contrast different terms in forming their own understanding to promote their ability to support study participants in articulating their own perspectives.

When articulating the design, protocol, and results of studies, researchers should take as a starting point the most widely agreed upon definitions. Shared definitions are critically important to a well-functioning research culture and community because without shared definitions we cannot genuinely compare results among studies. Appropriate international standards bodies include IETF (Internet Engineering Task Force), IAB (Internet Architecture Board), ISO (International Organization for Standardization), and IEEE (Institute of Electrical and Electronics Engineers). If these starting points are insufficient, then the researcher has a firm point of departure to explain why this is so; however, redefinition must be careful and must ensure usage does not slide between different definitions of a term without noting so doing. The security glossary from the IETF is an informational document that provides an excellent starting point [58]. Departures from definitions should be clear and justified. Therefore it is worth excerpting from the RFC for each of the terms to discuss common departure points.

Threat is "a potential for violation of security, which exists when there is an entity, circumstance, capability, action, or event that could cause harm" [58, p. 303]. Note that this does not define *threat model*, which is the set of threats and countermeasures considered relevant to the system at hand. Considering the relevant set of threats is essential for the external validity of a study.

Security is "a system condition that results from the establishment and maintenance of measures to protect the system," where the measures taken are suggested as deterrence, avoidance, prevention, detection, recovery, and correction [58, p. 263]. Studies often must contribute to a specific aspect of security, as it covers a broad range of activities. Authors would do well to specify which measures or aspects of the system condition of security on which their study focuses.

Privacy is "the right of an entity (normally a person), acting in its own behalf, to determine the degree to which it will interact with its environment" [58, p. 231]. This term has a particularly rich history of being difficult to define cleanly. For a comprehensive overview of the different definitions of privacy see the work of Gürses [21].

Usability is not directly defined by the IETF, however it is referenced as one of the two requirements for the availability pillar of the classic confidentiality-integrity-availability triad. Availability is "the property of a system or a system resource being... usable... upon demand..." [58, p. 29]. This supports the idea that, if availability is a requirement, an unusable system cannot be secure. Meanwhile, the failure of the standards to have even an informational definition of usability while giving it such a prominent position serves to highlight the importance of research in usable security. The usable security community cannot contribute to filling this gap for the Internet community as a whole if we are not clear about our own definitions.

A definition of *usability* is provided in the ISO 9241-11 standard for "office work with visual display terminals", as the *"Extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use"*. In studies, the inclusion of a primary task then provides an approximation of the context of use, against which to measure these qualities.

Although researchers need to be clear when communicating their definitions to peers, while conducting studies the researcher should not provide definitions to the participants when participant perceptions of these terms are being studied. Providing or sanctioning responses threatens the study because it injects a systemic error in the form of the researcher's pre-conceived definitions, threatening internal validity. Methods for avoiding even accidental transfer of ideas from the researcher to the participant are discussed in Section 4.3 and Section 4.4 on avoiding priming and performing double blind experiments, respectively.

We are often studying the definitions, because security and privacy mean different things to different people. The gap between security architect and user definitions of security is demonstrated by an example study on CAPTCHAs. Krol et al. [37] asked participants to make purchases on a ticket-selling website and part of the check-out process was to solve a CAPTCHA. After the purchase, participants were interviewed about their experience. In the security community, the security of CAPTCHAs is considered in terms of them being solvable by humans and not by robots. This is to protect the system from automated attacks leading to for example unavailability of the service to actual users. In the study, when participants mentioned security they did not speak about how well the CAPTCHAs protected the services but worried about the security of their own accounts and personal data.

Defining basic terms can be difficult as their meaning is often contextual. The challenge is to be precise, consistent and open to discussion with others.

4.6 Additional considerations

There are two areas which we treat as additional considerations, which are important but for which we have no direct recommendations: sampling bias, and the impact of current events on participants' perception and comprehension of security.

4.6.1 Sampling bias

An important consideration that we did not include in the principles is *sampling bias*. Sampling bias is, roughly, when the sample studied in an experiment is not relevantly representative of the population to which the researcher generalises their results. Therefore it is a type of threat to external validity. Sampling bias is a common scientific problem which has been studied in both psychology and information security, and thus it should not be surprising that user studies in security and privacy also contend with sampling bias.

In the larger fields of psychology and security, sampling bias has been studied in different ways. Many psychology research studies have had undergraduate students as participants, where a study with a mean participant age of 19 is not uncommon. Further, studies in psychology often rely on participants drawn from Western, educated, industrialised, rich and democratic (WEIRD) societies. Heinrich et al. [23] showed that these participants are not representative of all humans and are often outliers. In psychology, it often appears to be the richness or complexity of human individuals or systemic cultural differences that drives sample bias concerns. In information security, sampling bias is more often treated as an artifact of the sensor choices or as an artifact intentionally inserted by the adversaries being studied [42]. Sampling bias in information security may be assessed by technical measures with individual components which compare the whole sample to the population of available properties, such as total viable IP addresses [62]. However, like psychology, the argument for what qualifies a sample as sufficiently or relevantly unbiased must be made on a case-by-case basis.

In the field of usable security, sampling bias has already been discussed in at least two ways. Firstly, there has been a discussion as to whether samples drawn from crowd-sourcing platforms are representative of the wider population [32, 56]. Secondly, some studies have focused on the security and privacy of hitherto understudied populations. For example, Elliott and Sinclair Brody [16] studied the security and privacy needs of Afro-American New Yorkers. Bonneau and Xu [7] studied how character encoding can influence the choice of passwords for English, Chinese, Hebrew and Spanish speakers.

In our own studies, we have used pre-screening to maximise diversity of samples to include users of different age groups, gender and educational backgrounds (e.g., [38]). Pre-screening is our best effort to match our study sample to the population who uses the technology we are studying. In many cases, we do not know what subset of the whole human population is actually our target population of users, which makes targeting the correct sample particularly difficult. We have not had participants from non-WEIRD population groups in our own studies, thus we feel less qualified to talk about how to address this. How usable security is different from other disciplines in this respect requires further investigation, which is why we have not listed formal recommendations in this regard as we have for our five principles. Our general advice is to be mindful of the bias of one's research sample and not to frame any results as if they were universally applicable, but rather to frame results as applicable only to the population group(s) studied.

4.6.2 Current events

Participants often need to be considered in their social context, rather than as isolated individuals. While this tension is common between psychology and sociology, it is particularly important in security and privacy user studies because the social backdrop of information security has been changing so rapidly over the last two decades.

In Section 4.5, we elaborated on definitions of threat model, security, privacy and usability as a means to support dialogue with study participants as to what these concepts mean to them. Similarly, participants will develop an understanding of these concepts from the environment around them. Rader and Wash [48] found that people develop knowledge of security from both incidental and informal sources. Those who proactively learnt about security would learn – through sources such as news articles – about how to protect themselves from a range of attacks; others would learn incidentally by way of stories from others, sharing ideas primarily about the kinds of people who might attack them. Participants' perspectives about security may then be shaped by current events as they are documented and discussed with others. This may then require researchers to in some way be mindful of the current events around the time of a study – participants' perceptions may not be stable over time (and distinct studies) as the outside world changes.

5 Discussion

In this section, we discuss how researchers and practitioners could apply these principles. The principles do not replace researcher acumen or experience, however they provide a valuable service for facilitating evaluation, guiding younger researchers and students, and marking a baseline common language for discussing further improvements.

Our list of recommendations does not trivialise the difficulty of the research to be done. There is no replacement for the researcher's experience and skill; yet best principles to check for when designing, executing, or evaluating an experiment help in other critical ways. In weighing the advantages and disadvantages of checklists as a component of repeated procedures, Klein [35] notes that checklists should not supplant expertise but can be used to break complex procedures into repeatable steps. Most surgeries use checklists, but this improves patient outcomes only when hospital staff are properly trained and understand the checklist [12], unsurprisingly. Similarly, our mere process of listing these principles is not sufficient to improve research outcomes.

It may be that researchers in the field of usable security and privacy combine experiment tools to respond to the principles. Bravo-Lillo et al. [8] for instance have developed a reusable research ethics framework. Ferreira et al. [18] use a formal modelling technique to define technical and social threats as a precursor to designing and running experiments which involve human participants – such an approach may be applied to define the *threat model* for a study.

Security practitioners and developers of automated IT systems may want to account for the user when building security mechanisms that require human interaction – research that considers the principles can be more readily applied within a repeatable framework as advocated by Cranor [13]. Study of security alongside a primary task can identify communication impediments; realistic risk can characterise personal variables; clearly-articulated threat models can convey how behaviour and security mechanisms under evaluation respond to anticipated attacks.

6 Conclusions and future work

The five principles presented here provide an excellent example of learning from past experiments in order to produce incrementally better experimental designs going forward. Although we do not claim that the principles are exhaustive, they provide a fruitful starting point for reflecting on experimental design principles within the specific subfield of usable security and privacy research. The principles of primary tasks, realistic risks, avoiding priming, conducting double blind studies, and defining terms are reasonably intuitive from surveying the literature and have demonstrated benefits.

We recommend that anyone designing an experiment in usable security and privacy considers these principles carefully. If, after consideration, the researchers decide one or more principles do not apply to their study design, we simply recommend that they explain why when reporting their studies. This also serves to more concretely define the validity of subsequent study findings relative to the work of others in the field and in the wider world of security practice.

The work of describing principles that are important to experiments and other structured observations within a field is never done. The process is iterative; as helpful principles are applied more widely in new studies, new challenges will arise as old best principles are mastered. To facilitate such advancement of the field, future work should continually analyse the trade-offs between internal validity and external validity and the challenges of providing transparency and containment. With an eye keen to these potential problems, we can catalogue both further study designs and their impacts upon the capacity to capture user experiences of security technologies.

Acknowledgments

Kat Krol is supported by a Secure Usability Fellowship from the Open Technology Fund and Simply Secure. Jonathan M. Spring is supported by University College London's Overseas Research Scholarship & Graduate Research Scholarship. Simon Parkin is supported by UK EPSRC, grant nr. EP/K006517/1 ("Productive Security"). The authors also wish to thank the LASER Workshop's Program Committee, their paper shepherd and attendees of the workshop for their support in preparing the paper. Steve Dodier-Lazaro and Sören Preibusch provided useful feedback on earlier drafts of the paper.

References

[1] ADAMS, A., AND SASSE, M. A. Users are not the enemy. *Communications of the ACM 42* (1999), 40–46.

[2] BEAUTEMENT, A., SASSE, M. A., AND WONHAM, M. The compliance budget: Managing security behaviour in organisations. In *Proceedings of the 2008 workshop on New security paradigms* (2009), ACM, pp. 47–58.

[3] BENENSON, Z., LENZINI, G., OLIVEIRA, D., PARKIN, S., AND UEBELACKER, S. Maybe poor Johnny really cannot encrypt – The case for a complexity theory for usable security. In *New Security Paradigms Workshop (NSPW'15)* (2015).

[4] BERENDT, B., GÜNTHER, O., AND SPIEKERMANN, S. Privacy in e-commerce: stated preferences vs. actual behavior. *Communications of the ACM 48*, 4 (2005), 101–106.

[5] BERESFORD, A. R., KÜBLER, D., AND PREIBUSCH, S. Unwillingness to pay for privacy: A field experiment. *Economics Letters 117*, 1 (2012), 25–27.

[6] BONNEAU, J., AND SCHECHTER, S. Towards reliable storage of 56-bit secrets in human memory. In *Proc. USENIX Security* (2014).

[7] BONNEAU, J., AND XU, R. Of contraseñas, sysmawt, and mìmǎ: Character encoding issues for web passwords. In *Web 2.0 Security & Privacy* (May 2012).

[8] BRAVO-LILLO, C., EGELMAN, S., HERLEY, C., SCHECHTER, S., AND TSAI, J. You needn't build that: Reusable ethics-compliance infrastructure for human subjects research. In *Cybersecurity Research Ethics Dialog & Strategy Workshop* (2013).

[9] BROSTOFF, S., INGLESANT, P., AND SASSE, M. A. Evaluating the usability and security of a graphical one-time PIN system. In *BCS Interaction Specialist Group Conference* (2010), British Computer Society, pp. 88–97.

[10] BROSTOFF, S., AND SASSE, M. A. Are Passfaces more usable than passwords? A field trial investigation. *People and Computers* (2000), 405–424.

[11] CHEEK, J. Research design. In *The Sage Encyclopedia of Qualitative Research Methods*, L. M. Given, Ed. SAGE Publications, 2008, pp. 762–764.

[12] CONLEY, D. M., SINGER, S. J., EDMONDSON, L., BERRY, W. R., AND GAWANDE, A. A. Effective surgical safety checklist implementation. *Journal of the American College of Surgeons 212*, 5 (2011), 873–879.

[13] CRANOR, L. F. A framework for reasoning about the human in the loop. *UPSEC 8* (2008), 1–15.

[14] DITTRICH, D., AND KENNEALLY, E. The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research. *US Department of Homeland Security* (December 2011).

[15] EGELMAN, S., TSAI, J. Y., AND CRANOR, L. F. Tell me lies: A methodology for scientifically rigorous security user studies. In *Workshop on Studying Online Behaviour at CHI'10* (2010), ACM.

[16] ELLIOTT, A., AND BRODY SINCLAIR, S. S. Design Implications of Lived Surveillance in New York. In *Workshop on Everyday Surveillance at CHI'16* (2016).

[17] FELT, A. P., REEDER, R. W., ALMUHIMEDI, H., AND CONSOLVO, S. Experimenting at scale with Google Chrome's SSL warning. In *SIGCHI Conference on Human Factors in Computing Systems* (2014), pp. 2667–2670.

[18] FERREIRA, A., HUYNEN, J.-L., KOENIG, V., AND LENZINI, G. A conceptual framework to study socio-technical security. In *Human Aspects of Information Security, Privacy, and Trust*. Springer, 2014, pp. 318–329.

[19] FLORÊNCIO, D., HERLEY, C., AND VAN OORSCHOT, P. C. Password portfolios and the finite-effort user: Sustainably managing large numbers of accounts. In *USENIX Security* (2014).

[20] FUJITA, M., IKEYA, Y., KANI, J., AND NISHIGAKI, M. Chimera captcha: A proposal of captcha using strangeness in merged objects. In *Human Aspects of Information Security, Privacy, and Trust (HAS 2015), HCI International 2015*. Springer, 2015, pp. 48–58.

[21] GÜRSES, S. *Multilateral Privacy Requirements Analysis in Online Social Network Services*. PhD thesis, K.U. Leuven, 2010.

[22] HATLEBACK, E., AND SPRING, J. M. Exploring a mechanistic approach to experimentation in computing. *Philosophy & Technology 27*, 3 (2014), 441–459.

[23] HENRICH, J., HEINE, S. J., AND NORENZAYAN, A. The weirdest people in the world? *Behavioral and brain sciences 33*, 2-3 (2010), 61–83.

[24] HERLEY, C. So long, and no thanks for the externalities: The rational rejection of security advice by users. In *New Security Paradigms Workshop (NSPW'09)* (2009), ACM, pp. 133–144.

[25] HERLEY, C. More is not the answer. *IEEE Security & Privacy 12*, 1 (2014), 14–19.

[26] HERTWIG, R., AND ORTMANN, A. Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences 24*, 03 (2001), 383–403.

[27] HORTON, J. J., RAND, D. G., AND ZECKHAUSER, R. J. The online laboratory: Conducting experiments in a real labor market. *Experimental Economics 14*, 3 (2011), 399–425.

[28] JENSEN, C., POTTS, C., AND JENSEN, C. Privacy practices of internet users: Self-reports versus observed behavior. *International Journal of Human-Computer Studies 63*, 1 (2005), 203–227.

[29] JENSEN, D. Credibility. In *The Sage Encyclopedia of Qualitative Research Methods*, L. M. Given, Ed. SAGE Publications, 2008, pp. 139–140.

[30] JENSEN, D. Transferability. In *The Sage Encyclopedia of Qualitative Research Methods*, L. M. Given, Ed. SAGE Publications, 2008, p. 887.

[31] JENTZSCH, N., PREIBUSCH, S., AND HARASSER, A. Study on monetising privacy: An economic model for pricing personal information. *ENISA, Feb* (2012).

[32] KANG, R., BROWN, S., DABBISH, L., AND KIESLER, S. B. Privacy attitudes of mechanical turk workers and the us public. In *SOUPS* (2014), pp. 37–49.

[33] KIRLAPPOS, I., BEAUTEMENT, A., AND SASSE, M. A. "comply or die" is dead: Long live security-aware principal agents. In *Financial Cryptography and Data Security*. Springer, 2013, pp. 70–82.

[34] KIRLAPPOS, I., PARKIN, S., AND SASSE, M. A. Learning from "shadow security": Why understanding non-compliance provides the basis for effective security. In *USEC 2014: NDSS Workshop on Usable Security* (2014).

[35] KLEIN, G. *Streetlights and shadows: Searching for the keys to adaptive decision making*. MIT Press, 2009.

[36] KROL, K., MOROZ, M., AND SASSE, M. A. Don't work. Can't work? Why it's time to rethink security warnings. In *International Conference on Risk and Security of Internet and Systems (CRiSIS'12)* (2012), IEEE, pp. 1–8.

[37] KROL, K., PARKIN, S., AND SASSE, M. A. Better the devil you know: A user study of two CAPTCHAs and a possible replacement technology. In *USEC 2016: NDSS Workshop on Usable Security* (2016).

[38] Krol, K., Rahman, M. S., Parkin, S., De Cristofaro, E., and Vasserman, E. Y. An Exploratory Study of User Perceptions of Payment Methods in the UK and the US. In *USEC 2016: NDSS Workshop on Usable Security* (2016).

[39] Kuhn, T. S. *The structure of scientific revolutions*, fourth ed. University of Chicago press, 2012.

[40] Lisa M. Given, K. S. Trustworthiness. In *The Sage Encyclopedia of Qualitative Research Methods*, L. M. Given, Ed. SAGE Publications, 2008, pp. 896–897.

[41] Malheiros, M., Brostoff, S., Jennett, C., and Sasse, M. A. Would you sell your mother's data? Personal data disclosure in a simulated credit card application. In *The Economics of Information Security and Privacy*. Springer, 2013, pp. 237–261.

[42] Metcalf, L. B., and Spring, J. M. Blacklist ecosystem analysis: Spanning Jan 2012 to Jun 2014. In *The 2nd ACM Workshop on Information Sharing and Collaborative Security* (Denver, Oct 2015), pp. 13–22.

[43] Miller, P. Validity. In *The Sage Encyclopedia of Qualitative Research Methods*, L. M. Given, Ed. SAGE Publications, 2008, pp. 910–911.

[44] Pfleeger, S. L., and Caputo, D. D. Leveraging behavioral science to mitigate cyber security risk. *computers & security 31*, 4 (2012), 597–611.

[45] Popper, K. R. *The logic of scientific discovery*. 1959.

[46] Preibusch, S. Economic aspects of privacy negotiations. Master's thesis, Technische Universität Berlin, 2008.

[47] Preibusch, S. How to explore consumers' privacy choices with behavioral economics. In *Privacy in a Digital, Networked World*. Springer, 2015, pp. 313–341.

[48] Rader, E., and Wash, R. Identifying patterns in informal sources of security information. *Journal of Cybersecurity 1*, 1 (2015), 121–144.

[49] Renaud, K., Volkamer, M., and Renkema-Padmos, A. Why doesnt jane protect her privacy? In *Privacy Enhancing Technologies* (2014), Springer, pp. 244–262.

[50] Rivers, W., and Webber, H. The action of caffeine on the capacity for muscular work. *The Journal of physiology 36*, 1 (1907), 33–47.

[51] Rossow, C., Dietrich, C. J., Grier, C., Kreibich, C., Paxson, V., Pohlmann, N., Bos, H., and Van Steen, M. Prudent practices for designing malware experiments: Status quo and outlook. In *Security and Privacy (S&P), IEEE Symposium on* (2012), pp. 65–79.

[52] Sasse, A. Scaring and bullying people into security won't work. *IEEE Security & Privacy*, 3 (2015), 80–83.

[53] Sasse, M. A., Brostoff, S., and Weirich, D. Transforming the 'weakest link' – a human/computer interaction approach to usable and effective security. *BT Technology Journal 19*, 3 (2001), 122–131.

[54] Schechter, S. E., Dhamija, R., Ozment, A., and Fischer, I. The emperor's new security indicators. In *Security and Privacy, 2007. SP'07. IEEE Symposium on* (2007), IEEE, pp. 51–65.

[55] Schneier, B. Secrets and lies: Security in a digital world, 2000.

[56] Schnorf, S., Sedley, A., Ortlieb, M., and Woodruff, A. A comparison of six sample providers regarding online privacy benchmarks. In *SOUPS Workshop on Privacy Personas and Segmentation* (2014).

[57] Sheng, S., Broderick, L., Koranda, C. A., and Hyland, J. J. Why Johnny still can't encrypt: Evaluating the usability of email encryption software. In *Symposium On Usable Privacy and Security* (2006).

[58] Shirey, R. Internet Security Glossary, Version 2. RFC 4949 (Informational), Aug. 2007.

[59] Simon, H. A. Rational choice and the structure of the environment. *Psychological Review; Psychological Review 63*, 2 (1956), 129.

[60] Spiekermann, S., Grossklags, J., and Berendt, B. E-privacy in 2nd generation e-commerce: privacy preferences versus actual behavior. In *Proceedings of the 3rd ACM conference on Electronic Commerce* (2001), ACM, pp. 38–47.

[61] Spring, J. M. Toward realistic modeling criteria of games in internet security. *Journal of Cyber Security & Information Systems 2*, 2 (2014), 2–11.

[62] Spring, J. M., Metcalf, L. B., and Stoner, E. Correlating domain registrations and dns first activity in general and for malware. In *Securing and Trusting Internet Names: SATIN* (Teddington, UK, March 2011).

[63] Steves, M., Chisnell, D., Sasse, M. A., Krol, K., Theofanos, M., and Wald, H. Report: Authentication Diary Study. National Institute of Standards and Technology (NISTIR) 7983, 2014.

[64] Whitten, A., and Tyger, J. D. Why Johnny Can't Encrypt: A Usability Case Study of PGP 5.0. In *USENIX Security* (1999), pp. 169–184.

[65] Yee, K.-P. Aligning security and usability. *IEEE Security & Privacy*, 5 (2004), 48–55.

[66] Zins, C. Conceptual approaches for defining data, information, and knowledge. *Journal of the American society for information science and technology 58*, 4 (2007), 479–493.

[67] Zurko, M. E. User-centered security: Stepping up to the grand challenge. In *Computer Security Applications Conference, 21st Annual* (2005), IEEE, pp. 14–pp.