A Summary of Survey Methodology Best Practices for Security and Privacy Researchers

Elissa M. Redmiles (eredmiles@cs.umd.edu), Yasemin Acar, Sascha Fahl, and Michelle L. Mazurek

"Given a choice between dancing pigs and security, users will pick dancing pigs every time," warns an oft-cited quote from well-known security researcher Bruce Schneier [132]. This issue of understanding how to make security tools and mechanisms work better for humans (often categorized as *usability*, broadly construed) has become increasingly important over the past 17 years [7], [159], as illustrated by the growing body of research. Usable security and privacy research has improved our understanding of how to help users stay safe from phishing attacks [12], [62], [77], [105], [109], [129], [138], create strong passwords [39], [73], [130], [152], and control access to their accounts [16], [33], [93], [139], as just three examples.

One key technique for understanding and improving how human decision making affects security is the gathering of self-reported data from users. This data is typically gathered via survey and interview studies, and serves to inform the broader security and privacy community about user needs, behaviors, and beliefs. The quality of this data, and the validity of subsequent research results, depends on the choices researchers make when designing their experiments.

Contained here is a set of essential guidelines for conducting self-report usability studies distilled from prior work in survey methodology and related fields. Other fields that rely on self-report data, such as the health and social sciences, have established guidelines and recommendations for collecting high quality self-report data [10], [42], [55], [57], [70], [82], [98], [103], [119], [136], [148], [149].

I. Best-Practices & Pitfalls: Learning From Other Disciplines

The guidelines below are distilled from more than 100 relevant books and research articles from the survey methodology, psychology, and sociology fields to distill suggestions for conducting self-report interview and survey studies. In the following, an overview is provided of these best practices in the context of studying usability for security and privacy. These findings can be largely grouped around questionnaire writing (Section II), sampling (Section III), and pre-testing (Section IV).

II. Questionnaire Writing

The process of responding to a given questionnaire or interview item can be modeled in four steps (Figure 1) [32]. Respondents must first comprehend the question, then second attempt to retrieve the information necessary to answer it. Third, the respondent must judge whether this information is sufficient to answer the question, whether they wish to provide that information, and which answer choice most closely matches their desired response. Fourth, the respondent actually reports their answer.

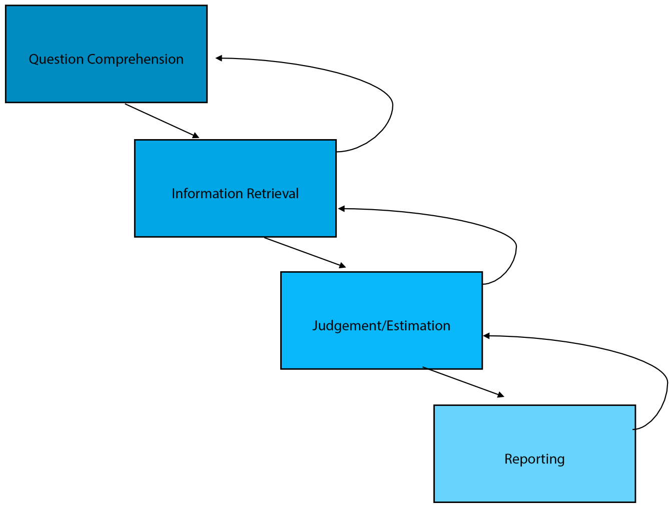

Fig. 1. Figure from Cannell et al. [32]: cognitive process of survey response.

In order to ensure that respondents can successfully complete this process, best practices with regard to word choice, question phrasing and ordering, question context, sensitive and demographic questions, survey length, Likert scales, and specific study modes (e.g. online vs. face-to-face) must all be considered. Each of these considerations is addressed below.

A. Choose Wording Carefully

Many possible problems can arise with word choice. The two largest such problems are different respondents having inconsistent understanding of terms in a survey and question writers using technical or abstract words that the respondent does not understand. Prior work in survey methodology has shown that even common words may be understood in multitude of different ways by respondents—for example, the word 'usually' was interpreted in 24 different ways by one study's participants [20], [68]. Such variations in understanding can creat data that cannot easily be compared [68]. Pre-testing (addressed in Section IV) is a key mechanism for ensuring that respondents understand terms consistently and in the way that you expect.

The problem of terms that respondents do not understand is particularly severe for security and privacy studies, which are riddled with technical concepts and terms. “The problem of unfamiliar or technical terminology is all the more serious,” notes a meta-analysis of survey literature on this problem, “because evidence indicates that respondents often ignore written definitions provided with the question” [148]. How then should we solve this problem? One option is to identify and use terms with which respondents are more comfortable, rather than technical or formal terms; Bradburn et al. found that this approach (e.g., using “sex” rather than “sexual intercourse”) resulted in more accurate responses [26]. Identifying such terms may sometimes be challenging, but focus groups can be used to successfully identify the language respondents use to discuss a concept [148]. When asking about a topic with which respondents are unfamiliar, definitions cannot be avoided, but it is recommended to thoroughly pre-test the questionnaire to ensure respondents consistently read and understand the definition.

To help with refining question wording, researchers can apply a tool such as QUAID [2] to automatically check for several common wording pitfalls, including technical and ambiguous terms.

Select Appropriate Likert Scales. One special case of the need for careful question wording is Likert scales, which are widely used to collect participants’ agreement with statements or another nuanced opinion. Prior work shows that providing too few or too many points on the scale may negatively impact the validity and interpretability of the results. Scale lengths between five and ten have been shown to optimal: fewer options can create bias by requiring the respondent to pick something that doesn’t quite fit, while too many options might render differences in responses meaningless. For example, when considering the difficulty of remembering a password on a scale from 1-100, what is the difference between responses of ‘72’ and ‘76’ [69], [82], [111], [123]? Thus, researchers should take care to select a Likert scale of appropriate length.

In addition, researchers should consider whether to choose an odd or even number of scale points. Both odd and even scales should be balanced; that is, they should contain an even number of options on each side of neutral (e.g., Very Bad, Somewhat Bad, Neither Good nor Bad, Somewhat Good, Very good) [78], [92]. Odd scales will thus have an explicit neutral option (Neither Good Nor Bad), while even scales will not. Even scales can therefore elicit stronger responses from respondents [9], [51].

Finally, word choice in Likert-scale options is also very important. For example, research has shown that the word “often” is very ambiguous and may be interpreted in a range of ways by respondents, leading to measurement error [20], [68]. Before creating a new scale, first consider whether an existing, validated scale can be applied. Vagias provides a good list of validated Likert scales [153]. If a new scale must be made, care should be taken to select unambiguous words and to pre-test the scale.

Avoid Double-Barreled Questions. In addition to taking care in their word choice, researchers must be careful to avoid double-barreled questions that implicitly ask more than one question. For example, asking a question such as, “Do you believe that your employer should require you to update your computer and change your password every six months?” (Yes, No, I Don’t Know) requires respondents to provide a single answer about both a requirement to update their computer and a requirement to change their passwords, even though they may agree with one requirement but disagree with the other. Prior work shows that such questions can negatively impact the validity of responses and may cause respondent frustration, leading them to drop out of the survey [42], [55], [69], [82], [148].

B. *Consider Question Context*

It is important to consider not only the text of the questions themselves, but also the context of these questions. Prior work shows strong evidence that the context in which a question is shown, including the questions asked previously, any advertisements shown in a web survey, and the sponsor and topic provided for the survey, may greatly affect responses [44], [46], [56], [69], [74], [133], [134], [148]–[150].

The effect of prior questions can be mitigated by asking questions about a number of different topics, diffusing the impact of any particular question on later questions. For example, Schwarz found that asking a question about marital happiness had a “pronounced effect on answers to a subsequent question about general life satisfaction when respondents marriages were the only specific life domain that the earlier questions asked about. When the earlier questions also asked about leisure time and jobs, the effect of marital happiness item on answers to the question on general life satisfaction was significantly reduced” [134]. A second, complementary way to mitigate such effects is to randomize the order in which questions are asked [82], [95]—while this does not remove order bias, it ensures that the bias is randomly distributed [119].

To ensure a closed-answer question includes all potentially necessary answer choices, the literature recommends focus-group or interview studies that solicit open-ended answers [103], [148]. Alternatively, if an open-answer question is used in a structured survey or interview, the responses should be systematically coded using open coding to achieve a concrete set of answers [103], [133]. For semi-structured interviews, on the other hand, it is recommended to begin with an open-answer question followed by prompts to elicit deeper answers from the participants or remind them of potential answers they may have forgotten to discuss [108].

Further, the sponsor, topic, and title of a study can all influence both who responds (*response bias*) and how they respond. It can be beneficial to emphasize that a study is being conducted by a university, as university sponsors (compared to unknown or corporate sponsors) elicit higher response rates [56]. However, it may be helpful to be opaque when stating the survey topic or title, as the topic may not only create response bias but also influence respondents to answer in ways they think the survey provider wants or expects [74], [107], [140]. For example, asking a question about software updating in a survey titled "Improving Computer Security" may lead to overreporting of the importance of updates.

Finally, to avoid bias from unknown images or ads that may also be displayed on the survey page, use an online survey platform that includes only your survey on a standardized template (such as Qualtrics) [44], [46].

Our experience suggests that question context can be critical for self-report studies in the domain of security and privacy, but there has been unfortunately little prior research on this specific topic. Here, are some hypotheses drawn from loosely related prior work and our own anecdotal research experiences in the security and privacy field. It is possible that the issue of prior questions may be especially salient for privacy, as asking personal questions first might increase respondents' reported privacy concern. Relatedly, in behavioral studies, administering a pre-experiment questionnaire querying respondents' privacy preferences or requesting personal information may subsequently impact respondents' security and privacy behavior by changing the context in which they are behaving. Framing about how the data will be used also can affect disclosure of sensitive data [8].

It is also possible that providing a trustworthy sponsor for your study (e.g. listing your university name) may be especially important for security and privacy studies in which security-sensitive and privacy-sensitive respondents may be hesitant to participate for fear of having their information compromised by an untrustworthy entity [131]. On the other hand, a trustworthy sponsor that leads a participant to feel safe may result in higher willingness to engage in behavior that might feel unsafe in another context [38].

C. Carefully Order Response Choices

The order in which answer choices are presented may also impact the survey results [75], [87], [95]. Prior work has found that in written (or online) surveys, the first answer choice is selected most often, while in phone and face-to-face surveys the last answer choice is more likely to be selected [19], [124]. To mitigate this effect, it is important to randomize the order of the answer choices, evenly distributing the occurrence of the the order bias [82]. Likert scales experience similar effects: for example, a Likert-scale item with the left-most answer option being positive (e.g., Strongly Agree (5)) is more likely to result in a higher overall set of responses [21], [37], [85], [157]. To address this, survey designers may wish to reverse the scale direction for a randomly selected half of respondents.

D. Be Aware of Sensitive Questions

While it is important to consider the phrasing and placement of all questions, extra consideration should be given to sensitive questions. A plethora of prior research summarized by Tourangeau and Yan [150] has shown that respondents tend to under-report socially-undesirable behaviors such as drug- and alcohol-use, bankruptcy, energy use, criminal behavior, and racist attitudes [59], [63], [102], [104], [150], while over-reporting desirable behaviors such as church attendance, eco-friendliness, exercise, library use, wearing a seatbelt, and voting [18], [72], [117], [144], [150]. Given the number of domains in which the effect of this *desirability bias* has been proven, it seems clear that such bias will be an issue for questions regarding security and privacy behaviors and beliefs, as well.

Desirability bias can be mitigated through a number of different methods [150]; here are three of the easiest. First, sensitive questions can be asked not about the respondent directly (e.g., "Do you behave securely?"), but rather indirectly ("Do you think your friends behave securely") [76]. Second, questions can be softened with "forgiving wording," such as an introductory sentence that makes clear to the respondent that all answer choices are acceptable (e.g., "People have many different reasons for choosing to update or not to update their phones. Which of the following best describes what you do when an update is available for your phone? . . . ") [69], [145], [150]. Third, questions should be balanced: e.g., "If there is a serious fuel shortage this winter, do you think there should **or should not** be a law requiring people to lower the heat in their homes?" where the **bold** text is used to balance the question [135].

A number of more advanced techniques for ensuring that respondents answer sensitive questions honestly also exist [103], [150]. One example is the use of a list experiment (also known as the unmatched count technique) [24], [43], [48], [52], [79], [150], which involves the following procedure: (1) divide respondents into a control and treatment group, (2) ask control respondents to report the number of items on a list that they wish to answer 'yes' to, rather than asking them to answer yes/no about each item individually, (3) ask treatment respondents to do the same, but add the sensitive question of interest to the list, and (4) compute the proportion of respondents who answered yes to the sensitive item of interest by comparing the answers of the two groups. This technique was recently employed to study the prevalence of snooping on other people's mobile phones [110]. Other techniques include the randomized response technique [155] and the bogus pipeline [150]. Given the power of these techniques to improve measurement validity for sensitive security and privacy questions, it is strongly recommended that researchers consider increased exploration and application of both simple and more advanced techniques for reducing desirability bias.

Finally, before we can mitigate desirability bias for sensitive questions, we must identify which items are sensitive. Thankfully, prior work shows that researchers are typically very accurate in their assessment of whether or not a question is sensitive [27]. If the researcher is unsure of whether a question is sensitive, a brief pilot test in which respondents are asked to rate question sensitivity on a Likert scale can be helpful [150].

Desirability bias is especially important for security and privacy studies, as respondents may feel pressure to over-report knowledge [88] or behaviors they perceive to be "good" such as patching. This may be especially true in the case of a lab or telephone study, if they know the interviewer to be someone who cares about security. There is some preliminary evidence of issues of over- and mis-reporting privacy behaviors [142] due to desirability bias. This issue may also play a small role in the "privacy paradox" [15] in which respondents have a tendency to report more privacy sensitivity than their activities warrant. Thus, security and privacy researchers should take great care in designing questions that ask about security and privacy behaviors and preferences, items for which respondents may be inclined to over- or mis-report.

Demographic Questions. Demographic questions are a special subset of sensitive questions. They are not only sensitive, but may be associated with significant stereotype threat that can affect survey measurements [116]. For example, asking demographic questions at the beginning of the survey frames the context of that survey in terms of respondent demographics: in one experiment, women performed worse on a math-related task when asked to provide their gender first than when asked to provide it afterward [116]. Thus, prior work suggests avoiding asking demographic questions at the beginning of a survey containing questions with "right or wrong" answers, such as math problems, and for surveys containing questions about sensitive topics or topics with socially-desirable answers [41], [56], [64], [91], [128], [143], [150], [158]. For questionnaires regarding less sensitive topics with no discernible socially-desirable or "correct" answer (e.g., "How many gallons of milk did you purchase this week?"), some research suggests that demographic item placement may not negatively impact on survey answers, and may increase response rates to demographic items (in surveys where respondents drop off before finishing the survey) [58], [71], [147].

More specifically, research on the U.S. Census has found that placing the Hispanic ethnicity item (e.g., "Are you of Hispanic origin?") prior to a question about race results in lower non-response for both items [113]. It is always recommended to ask ethnicity and racial identity as separate questions, but in the case of severe space constraints where only one question can be asked, the question should be formatted as a multiple response item, where respondents can select, for example, both Hispanic and White as their answer to the singular race/ethnicity question [115].

In security and privacy research, it is often tempting to ask demographic questions near the beginning of a survey, in order to use up time while some computational task (e.g., processing a participant's social media posts in order to ask privacy or access-control questions) runs in the background. However, given that security and privacy behaviors are potentially sensitive topics (similar to health behavior), it is recommended that researchers follow the literature-supported best practice of placing demographic questions at the end of the survey [41], [56], [64], [91], [128], [143], [150], [158].

E. Keep Surveys to a Reasonable Length

Since recruiting participants can be difficult and time-consuming, it is often tempting to bombard each participant with as many questions as possible at once. However, extending the time it takes to participate in the survey may lead to a lower response (or completion) rate [54], [107], [137].

The inclusion of a completion bar has been shown to increase completion rates for closed-answer surveys [47]; contrastingly, misrepresenting the length of a survey (e.g., advertising a five-minute survey that really takes 10 minutes to complete) will significantly reduce the completion rate [50].

F. Consider Your Survey Mode

The mode (online, telephone, face-to-face) that you employ to conduct your study may have additional considerations. For example, telephone and face-to-face surveys and interviews may be impacted by interviewer effects, discussed below; online surveys, which enable researchers to easily require that respondents answer questions, have different considerations.

Interviews and Interviewers. One key consideration for interview studies is interview style. Semi-structured or conversational interviewing, which allows the interviewer to slightly modify questions or supply clarification and definitions, has been shown to result in more accurate measurements for studies about sensitive topics [146]. However, these approaches are also more subject to error from stereotype threat and social-desirability biases [70].

In interview studies and telephone or face-to-face surveys the gender, race, and age of an interviewer may influence responses [86], [96], [150]. Additionally, the quality of the interviewing, and the prompts the interviewer provides can also significantly affect the survey measurements [89].

There are three main mechanisms to mitigate these effects. The first is to thoroughly train interviewers to provide neutral prompts and read the interview or survey protocol as it is written [67], [89]. Fowler and Mangione found that interviewers who had less than one day of training were significantly worse at reading the questions correctly and prompting respondents when they gave incomplete or ambiguous answers [70]. Second, participants should be randomly assigned to different interviewers, preferably with differing demographic characteristics, in order to diffuse bias. The third method is to calculate and report the statistical interviewer effect metric [82].

Behave Ethically When Requiring Responses. Requiring respondents to provide answers to every question on a survey, without offering a "don't know" (DK) or "prefer not to answer" option, is considered unethical: it is in direct violation of the American Association for Public Opinion Research code of ethics [6], and is explicitly disallowed by many Institutional Review Boards [5]. It can also affect the quality of responses offered, and may lead to respondent frustration and survey drop-offs [42], [57], [82], [103]. This problem is most prevalent in web surveys, as interviewers are not typically able to force respondents to answer questions (although respondents who refuse to answer certain questions may be removed from a study). While there are legitimate reasons to want to require some answer to each question (e.g., being charged per response, wanting better data quality), it is highly recommended that a DK or "prefer not to answer" option be offered in all cases, including for demographic questions [107]. There is unfortunately little research into the trade-offs between offer a DK answer choice, a "prefer not to answer," choice, and simply making the question optional, so researchers must (at least for now) use their best judgment.

G. *Reuse Existing, Validated Questions When Possible*

Using survey questions and scales that have previously been validated and tested by other researchers can improve the soundness of study results, provide a baseline of results from prior work to compare against, and save researchers significant time and energy that would otherwise be spent on carefully framing and then pre-testing these questions. While formally validated scales are most authoritative, questions previously pre-tested and used by other researchers also provide comparability with prior work and can save time and effort. While it is rare to be able to rely entirely on previously existing questions, including as many of them as possible (especially when dealing with topics that arise frequently in many security and privacy studies) can be highly beneficial. This is a commonly supported best practice in survey design [65], [82].

Some experimentally validated measures that may be useful for security and privacy researchers include scales for assessing internet skill, privacy inclination, security behavior intention, and usability [28], [61], [83], [120]. Pew surveys and the Roper database provide large banks of previously written and pre-tested items [35], [36]. Note that these existing scales and questions were developed and tested in particular contexts; researchers should carefully consider their strengths and weaknesses when deciding whether re-use of a particular item is appropriate.

III. Sampling

Representative samples ensure that the opinions of one subset of the population (those who are over-represented) are not inaccurately magnified while ignoring or under-reporting the opinions of another subset (those who are under-represented, or not represented at all) [82], [100], [103]. The classic illustration of the importance of sampling is the 1936 Literary Digest poll on the U.S. presidential election: Using a 2.4-million-respondent sample, the poll predicted Alfred Landon would win with 57% of the vote, when in fact Roosevelt won with 62%. The poll was so wrong largely because of its severely biased sample: Literary Digest subscribers, who tended to be older, wealthier, and more politically conservative than the general population [13], [29], [53], [141]. As shown by this example, even large samples cannot make up for sampling bias. In fact, in the same year, a much smaller Gallup poll (50,000 respondents) with a significantly better sampling strategy successfully predicted the election results [53].

In usable security and privacy research, different sampling methods may be more or less appropriate for different studies. Below, is an outline of four prominent sampling methods and literature-based suggestions for when they should be used. The majority of our discussion centers around quantitative (e.g., survey) sampling methods and recommendations, rather than on qualitative (e.g., interview) sampling methods, as the time-intensive nature of qualitative studies restricts the feasibility of many sampling strategies.

A. *Convenience and Snowball Samples*

Convenience sampling typically involves sampling from the most accessible participants (e.g., university students, social-network contacts) [70], [82], [112]. This sampling method is very low cost, but "may result in poor quality data and lacks intellectual credibility," as the recruited participants tend not to be very representative of the researcher's target population [112]. Snowball sampling involves using a somewhat more rigorous approach to recruiting initial participants (e.g., using demographic quotas) and then recruiting the friends of those initial participants as additional respondents in the sample [23]. Snowball sampling is similarly low cost, with potentially low data quality [23], [81].

Qualitative Sampling. While convenience and snowball samples can be largely rejected outright for quantitative surveys, the majority of qualitative studies must rely on some form of convenience sampling due to time, travel, and cost considerations that restrict qualitative sample sizes [49], [112], [118]. These approaches may also be necessary when the target population is very difficult to reach, such as users with very specific security experiences or visually impaired users. To mitigate the inherent drawbacks of these sampling strategies and maximize result validity, prior work suggests that researchers should construct a theoretical model of the factors that will impact the results of their study and then attempt to screen and select respondents who represent a diverse sample of these factors [112]. For security and privacy research, factors such as demographics, internet skill and beliefs about security may often be relevant [125], [126], [156].

B. *Crowdsourced and Social Media Samples*

Another commonly used sampling method is sampling from crowdsourcing platforms such as Amazon Mechanical Turk (MTurk). These platforms will provide more representative and less biased samples than most convenience samples (e.g., as compared to selecting only local university undergraduates) [30], [34], but they still suffer from sample bias [90]. For example, MTurk users tend to be more highly educated, younger, and more technically-savvy than the general population [90], [97], and they may have different values and personality characteristics [80]. This sample bias may have important implications for the results of security and privacy studies [97], [99]. Similarly, social media platforms, and specifically Twitter, are increasingly used as sources of data. This data may represent a biased sample population, as only 13% of Internet users use Twitter, for example, and the users of social media tend to be non-representative on lines of class, gender, and education [60], [84], [151].

This is not to say, however, that these samples should not be used. MTurk samples provide significantly better sample diversity than convenience samples [30], [34] and may be particularly useful for studying young, educated populations.

Further, MTurk (and newer competitors, such as Prolific [1]), allow behavioral tasks beyond the simple answering of survey questions (e.g., testing out a new security tool and providing feedback); crowdsourcing allows for far larger sample sizes for these behavioral experiments than could be reasonably be obtained in a laboratory setting [34]. MTurk samples also offer researchers relatively easy access to respondents from a variety of countries, a population that was previously challenging to reach [106]. Social media data, on the other hand, provides the opportunity for researchers to observe users' online behavior and language without any intervention; this type of ecological validity and observation is impossible to obtain with a traditional representative survey. Moreover, most social media data is available publicly, for free, consequently lowering the barrier of entry for research. For some studies, these benefits will outweigh the sampling drawbacks; for others the lack of representativeness may be a critical problem.

C. Online Census-Representative Samples

An alternative to convenience, crowdsourced, and social-media samples is to use quota sampling from an online panel to achieve a census-representative sample distribution. Online survey samples are typically obtained from panels, which are put together by companies such as Qualtrics [2] and Forsa [3] [14]. These companies may recruit participants from all over the world via paper mailings, airline frequent flyer programs, mailing lists, and other methods. These online panels are made up of thousands of potential respondents, who are compensated for their participation when they take a survey [14], [45]. Researchers can submit requests to have their survey distributed to a set number of these panel respondents, based on demographic criteria. For example, researchers can request a 500-person sample that is census-representative of their country with regard to age, income, education, gender, race/ethnicity, and household size. Response prices for such samples begin at around $3 per response, depending on the panel and criteria selected, but can be much higher.

While these panel samples are more representative than crowd-sourced, social media, and convenience samples, they are still subject to sample bias. Of panel members who are invited to take a given survey, over 90% chose not to respond. This high rate of non-response, may lead to response bias, the extent and effects of which is not yet fully understood [14]. Thus, while online census-based quota panel samples provide a more representative set of respondents, researchers should be careful to avoid overclaiming about generalizability to the entire population on the basis of results from these samples.

D. Probabilistic Samples

Probabilistic samples—that is, samples in which every person in the given population (e.g., the U.S.) had a non-zero probability of taking a given survey—are the gold-standard of samples [94], [101], [111], [148]. Probabilistic samples allow researchers to use statistical weighting techniques to estimate the true prevalence of their results in the entire sample population. Probabilistic samples are rarely used in usable security and privacy [125], [127], [142], but hold significant promise for providing findings regarding true prevalence of behaviors and beliefs in the population and for studying populations that are typically under-represented in online panel, convenience, and crowd-source samples (e.g., low-SES or low-internet-skill users). However, such samples can be extremely expensive, with individual responses typically costing $12 to $30.

While probabilistic samples are typically created by contacting potential respondents via mail or via telephone, two mechanisms exist for conducting nearly-probability-based online surveys: Google Consumer Surveys [4] and KnowledgePanel [5]. These methods use IP and internet behavior patterns to discern users' demographics, cross-reference this information with local census information, and apply stratified weighting techniques accordingly to create a representative sample [14], [114]. Given the inferences that must be made, these methods are not fully probabilistic and thus cannot be used to infer prevalence to the entire sample population. Further, KnowledgePanel can be nearly as expensive as traditional probabilistic samples, while Google Consumer Surveys allows researchers to ask a maximum of 10 items per survey, including demographic items, which limits its applicability for many research tasks. However, with these limitations in mind, preliminary research suggests that these nearly-probabilistic methods may provide results that are as representative as probabilistic methods [114].

E. Choosing a Sampling Method

All of the methods discussed above have both benefits and drawbacks. Researchers must often make tradeoffs that balance the desire for the highest-possible-quality data with resource and feasibility constraints. Thus, researchers should carefully consider how the properties of different sampling methods apply in the context of their specific research question: while representativeness is always important, reaching specific under-represented populations may be more important for some questions than for others. Perhaps most critically, researchers should always clearly describe why their sampling approach was selected and how its limitations affect the scope and generalizability of their results [106]. As a specific limitation in the context of privacy, it also seems likely that individuals with very high privacy concerns are less likely to participate in research studies at all and may therefore be underrepresented [131]; researchers should take care to consider this possibility when drawing conclusions.

IV. Pre-testing

As mentioned in Section II, respondents frequently interpret words in different ways, and may hesitate to report answers due to social-desirability bias. Additionally, while writing questionnaires we may inadvertently miss key answer choices

[1] https://www.prolific.ac/

[2] https://www.qualtrics.com/online-sample/

[3] https://www.forsa.de/

[4] https://www.google.com/analytics/surveys/

[5] http://join.knpanel.com/about.html

or accidentally include technical words that our respondents do not understand. Pre-testing surveys and interview protocols can help prevent these and other measurement errors and ensure that self-report survey and interview measurements are as accurate as possible [57], [66], [82], [103], [122], [148]. Further, best-practices literature also recommends reporting pre-testing results, in addition to using those results to form better surveys, as the results of such pre-tests can aid future researchers [122], [154]. Below, these three different methods—which are optimally used together—are presented for pre-testing self-report user studies.

A. Piloting

Field tests or pilots are perhaps the most common form of survey pre-testing, and have been used in the social sciences since the 1930s [20], [122]. Piloting surveys involves running the survey or interview protocol on a small set of respondents, and then examining the data and feedback from the interviewers to identify potential problems [122]. Piloting is useful for identifying technical issues, consistent misunderstandings, and problems for interviewer implementation (in the case of an interview study or face-to-face/phone survey). However, traditional piloting can do little to identify issues of respondent misinterpretation, missing answer choices, or even stress and discomfort, as respondents are not aware that the survey or interview protocol is being evaluated [20], [122], [160].

B. Cognitive Interviewing

Cognitive interviews—which involve asking respondents to think aloud as they complete a survey as well as asking them questions about each survey item—can help to identify more subtle measurement errors. Per-item questions might include "How did you feel answering that question?" or "What does [a particular term] mean to you?" Cognitive interviewing is broadly recommended as a necessary and highly-important pre-testing measure in the survey literature [11], [17], [40], [103], [122], [160]. The required sample size for these interviews is small, with only 10 participants typically illuminating more than 50% of potential survey problems [25]. Such pre-testing has been shown to significantly reduce measurement error [22], [31], [121].

C. Expert Reviewing

Cognitive interviewing and piloting may not catch the most basic of best-practice errors, such as missing DK options and too-short Likert scales (Section II. To identify these errors before deployment, prior work [31], [46], [82], [122] suggests that it may be helpful for researchers to solicit expert reviews from colleagues with expertise in survey methodology or human-computer interaction, or from their campus statistical and survey consulting department (e.g., [1], [3], [4]), if available.

References

[1] Princeton university data and statistical services.

[2] Question understanding aid (quaid) tool.

[3] University of maryland university libraries statistical consulting.

[4] University of michigan survey research center.

[5] AAPOR. Irb faqs for survey researchers.

[6] AAPOR. AAPOR code of ethics, 2015.

[7] A. Adams and M. A. Sasse. Users are not the enemy. *Communications of the ACM*, 42(12):40–46, 1999.

[8] I. Adjerid, A. Acquisti, L. Brandimarte, and G. Loewenstein. Sleights of privacy: Framing, disclosures, and the limits of transparency. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, page 9. ACM, 2013.

[9] G. Albaum. The likert scale revisited: an alternate version. *Journal of the Market Research Society*, 39(2):331–332, 1997.

[10] L. Andres. *Designing and doing survey research.* Sage Publications Ltd, London, 2012.

[11] D. Andrews, B. Nonnecke, and J. Preece. Electronic survey methodology: A case study in reaching hard-to-involve internet users. *International journal of human-computer interaction*, 16(2):185–210, 2003.

[12] N. A. G. Arachchilage and S. Love. A game design framework for avoiding phishing attacks. *Comput. Hum. Behav.*, 2013.

[13] J. S. Armstrong and T. S. Overton. Estimating nonresponse bias in mail surveys. *Journal of marketing research*, pages 396–402, 1977.

[14] R. Baker, S. J. Blumberg, J. M. Brick, M. P. Couper, M. Courtright, J. M. Dennis, D. Dillman, M. R. Frankel, P. Garland, R. M. Groves, et al. Research synthesis aapor report on online panels. *Public Opinion Quarterly*, 74(4):711–781, 2010.

[15] S. B. Barnes. A privacy paradox: Social networking in the united states. *First Monday*, 11(9), 2006.

[16] L. Bauer, L. F. Cranor, R. W. Reeder, M. K. Reiter, and K. Vaniea. A user study of policy creation in a flexible access-control system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 543–552. ACM, 2008.

[17] P. C. Beatty and G. B. Willis. Research synthesis: The practice of cognitive interviewing. *Public Opinion Quarterly*, 71(2):287–311, 2007.

[18] R. F. Belli, M. W. Traugott, and M. N. Beckmann. What leads to voting overreports? contrasts of overreporters to validated voters and admitted nonvoters in the american national election studies. *Journal of Official Statistics*, 17(4):479, 2001.

[19] W. A. Belson. Effects of reversing presentation order of verbal rating scales. *Journal of Advertising Research*, 6(4):30–37, 1966.

[20] W. A. Belson. *The design and understanding of survey questions.* Gower Aldershot, 1981.

[21] L. Betts and J. Hartley. The effects of changes in the order of verbal labels and numerical values on childrens scores on attitude and rating scales. *British Educational Research Journal*, 38(2):319–331, 2012.

[22] P. P. Biemer, R. M. Groves, L. E. Lyberg, N. A. Mathiowetz, and S. Sudman. *Measurement errors in surveys*, volume 173. John Wiley & Sons, 2011.

[23] P. Biernacki and D. Waldorf. Snowball sampling: Problems and techniques of chain referral sampling. *Sociological methods & research*, 10(2):141–163, 1981.

[24] G. Blair, K. Imai, and Y.-Y. Zhou. Design and analysis of the randomized response technique. *Journal of the American Statistical Association*, 110(511):1304–1319, 2015.

[25] J. Blair and F. G. Conrad. Sample size for cognitive interview pretesting. *Public Opinion Quarterly*, 75(4):636–658, 2011.

[26] N. Bradburn and S. Sudman. Associates (1979). *Improving Interviewing Methods and Questionnaire Design: Response Effects to Threatening Questions in Survey Research*, 85.

[27] N. M. Bradburn and C. Miles. Vague quantifiers. *Public Opinion Quarterly*, 43(1):92–101, 1979.

[28] J. Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

[29] M. C. Bryson. The literary digest poll: Making of a statistical myth. *The American Statistician*, 30(4):184–185, 1976.

[30] M. Buhrmester, T. Kwang, and S. D. Gosling. Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, 6(1):3–5, 2011.

[31] P. Campanelli. Testing survey questions: New directions in cognitive interviewing. *Bulletin de Methodologie Sociologique*, 55(1):5–17, 1997.

[32] C. F. Cannell, K. H. Marquis, A. Laurent, et al. *A summary of studies of interviewing methodology.* Number 69. US Government Printing Office, Washington, DC 20402, 1977.

[33] X. Cao and L. Iverson. Intentional access management: Making access control usable for end-users. In *Proceedings of the second symposium on Usable privacy and security*, pages 20–31. ACM, 2006.

[34] K. Casler, L. Bickel, and E. Hackett. Separate but equal? a comparison of participants and data gathered via amazons mturk, social media, and face-to-face behavioral testing. *Computers in Human Behavior*, 29(6):2156–2160, 2013.

[35] P. R. Center. Pew question search.

[36] T. R. D. Center. ipoll search.

[37] J. C. Chan. *Response-order effect in Likert-type scales*. ERIC Clearinghouse, 1990.

[38] N. Christin, S. Egelman, T. Vidas, and J. Grossklags. It's All about the Benjamins: An Empirical Study on Incentivizing Users to Ignore Security Advice. In *Financial Cryptography*, pages 16–30, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[39] M. Ciampa. A comparison of password feedback mechanisms and their impact on password entropy. *Information Management & Computer Security*, 2013.

[40] D. Collins. Pretesting survey instruments: an overview of cognitive methods. *Quality of life research*, 12(3):229–238, 2003.

[41] D. Colton and R. W. Covert. *Designing and constructing instruments for social research and evaluation*. John Wiley & Sons, 2007.

[42] J. M. Converse and S. Presser. *Survey questions: Handcrafting the standardized questionnaire*. Number 63. Sage, 1986.

[43] D. Corstange. Sensitive questions, truthful answers? modeling the list experiment with listit. *Political Analysis*, 17(1):45–63, 2009.

[44] M. P. Couper, F. G. Conrad, and R. Tourangeau. Visual context effects in web surveys. *Public Opinion Quarterly*, 71(4):623–634, 2007.

[45] M. P. Couper and P. V. Miller. Web survey methods introduction. *Public Opinion Quarterly*, 72(5):831–835, 2008.

[46] M. P. Couper, R. Tourangeau, and K. Kenyon. Picture this! exploring visual effects in web surveys. *Public Opinion Quarterly*, 68(2):255–266, 2004.

[47] M. P. Couper, M. W. Traugott, and M. J. Lamias. Web survey design and administration. *Public opinion quarterly*, 65(2):230–253, 2001.

[48] E. Coutts and B. Jann. Sensitive questions in online surveys: Experimental results for the randomized response technique (rrt) and the unmatched count technique (uct). *Sociological Methods & Research*, 40(1):169–193, 2011.

[49] I. T. Coyne. Sampling in qualitative research. purposeful and theoretical sampling; merging or clear boundaries? *Journal of advanced nursing*, 26(3):623–630, 1997.

[50] S. D. Crawford, M. P. Couper, and M. J. Lamias. Web surveys perceptions of burden. *Social science computer review*, 19(2):146–162, 2001.

[51] R. A. Cummins and E. Gullone. Why we should not use 5-point likert scales: The case for subjective quality of life measurement. In *Proceedings, second international conference on quality of life in cities*, pages 74–93, 2000.

[52] D. R. Dalton, J. C. Wimbush, and C. M. Daily. Using the unmatched count technique (uct) to estimate base r. *Personnel Psychology*, 47(4):817, 1994.

[53] D. DeTurk. Case study i: The 1936 literary digest poll.

[54] E. Deutskens, K. De Ruyter, M. Wetzels, and P. Oosterveld. Response rate and response quality of internet-based surveys: An experimental study. *Marketing letters*, 15(1):21–36, 2004.

[55] D. A. Dillman. *Mail and telephone surveys*, volume 3. Wiley Interscience, 1978.

[56] D. A. Dillman. *Mail and Internet surveys: The tailored design method–2007 Update with new Internet, visual, and mixed-mode guide*. John Wiley & Sons, 2011.

[57] D. A. Dillman, R. D. Tortora, and D. Bowker. Principles for constructing web surveys. In *Joint Meetings of the American Statistical Association*, 1998.

[58] F. J. Drummond, L. Sharp, A.-E. Carsin, T. Kelleher, and H. Comber. Questionnaire order significantly increased response to a postal survey sent to primary care physicians. *Journal of clinical epidemiology*, 61(2):177–185, 2008.

[59] J. C. Duffy and J. J. Waterto. Under-reporting of alcohol consumption in sample surveys: The effect of computer interviewing in fieldwork. *British journal of addiction*, 79(4):303–308, 1984.

[60] M. Duggan and J. Brenner. *The demographics of social media users, 2012*, volume 14. Pew Research Center's Internet & American Life Project Washington, DC, 2013.

[61] S. Egelman and E. Peer. 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*, pages 2873–2882. ACM, 2015.

[62] S. Fahl, M. Harbach, T. Muders, M. Smith, and U. Sander. Helping johnny 2.0 to encrypt his facebook conversations. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, page 11. ACM, 2012.

[63] M. Fendrich and C. M. Vaughn. Diminished lifetime substance use over time: An inquiry into differential underreporting. *Public Opinion Quarterly*, 58(1):96–123, 1994.

[64] A. Fink. *The survey kit: How to conduct self-administered and mail surveys*. Sage, 2003.

[65] J. Floyd J. Fowler. *Survey Research Methods (4th ed.)*. SAGE Publications, Inc., 2009.

[66] B. Forsyth, J. M. Rothgeb, and G. B. Willis. Does pretesting make a difference? an experimental test. *Methods for testing and evaluating survey questionnaires*, pages 525–546, 2004.

[67] F. J. Fowler. Reducing interviewer-related error through interviewer training, supervision, and other means. *Measurement errors in surveys*, pages 259–278, 1991.

[68] F. J. Fowler. How unclear terms affect survey data. *Public Opinion Quarterly*, 56(2):218–231, 1992.

[69] F. J. Fowler. *Improving survey questions: Design and evaluation*, volume 38. Sage, 1995.

[70] F. J. Fowler Jr and T. W. Mangione. *Standardized survey interviewing: Minimizing interviewer-related error*, volume 18. Sage, 1990.

[71] A. Frick, M.-T. Bächtiger, and U.-D. Reips. Financial incentives, personal information and drop-out rate in online studies. *Dimensions of Internet science*, pages 209–219, 1999.

[72] E. T. Fuj, M. Hennessy, and J. Mak. An evaluation of the validity and reliability of survey response data on household electricity conservation. *Evaluation Review*, 9(1):93–104, 1985.

[73] M. Fujita, M. Yamada, S. Arimura, Y. Ikeya, and M. Nishigaki. An attempt to memorize strong passwords while playing games. In *NBIS*, 2015.

[74] M. Galesic and R. Tourangeau. What is sexual harassment? it depends on who asks! framing effects on survey responses. *Applied cognitive psychology*, 21(2):189–202, 2007.

[75] M. Galesic, R. Tourangeau, M. P. Couper, and F. G. Conrad. Eye-tracking data new insights on response order effects and other cognitive shortcuts in survey responding. *Public Opinion Quarterly*, 72(5):892–913, 2008.

[76] S. Garfinkel and H. R. Lipford. *Usable Security: History, Themes, and Challenges*. Synthesis Lectures on Information Security, Privacy, and Trust. Morgan & Claypool Publishers, 2014.

[77] V. Garg, L. J. Camp, K. Connelly, and L. Lorenzen-Huber. Risk communication design: Video vs. text. In *PETS*, 2012.

[78] R. Garland. The mid-point on a rating scale: Is it desirable. *Marketing bulletin*, 2(1):66–70, 1991.

[79] A. N. Glynn. What can we learn with statistical truth serum? design and analysis of the list experiment. *Public Opinion Quarterly*, 77(S1):159–172, 2013.

[80] J. K. Goodman, C. E. Cryder, and A. Cheema. Data collection in a flat world: The strengths and weaknesses of mechanical turk samples. *Journal of Behavioral Decision Making*, 26(3):213–224, 2013.

[81] L. A. Goodman. Snowball sampling. *The annals of mathematical statistics*, pages 148–170, 1961.

[82] R. M. Groves, F. J. Fowler Jr, M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau. *Survey methodology*, volume 561. John Wiley & Sons, 2009.

[83] E. Hargittai. Survey measures of web-oriented digital literacy. *Social science computer review*, 23(3):371–379, 2005.

[84] E. Hargittai. Whose space? differences among users and non-users of social network sites. *Journal of Computer-Mediated Communication*, 13(1):276–297, 2007.

[85] J. Hartley. Some thoughts on likert-type scales. *International Journal of Clinical and Health Psychology*, 14(1):83–86, 2014.

[86] S. Hatchett and H. Schuman. White respondents and race-of-interviewer effects. *The Public Opinion Quarterly*, 39(4):523–528, 1975.

[87] A. L. Holbrook, J. A. Krosnick, D. Moore, and R. Tourangeau. Response order effects in dichotomous categorical questions presented

orally the impact of question and respondent attributes. *Public Opinion Quarterly*, 71(3):325–348, 2007.

[88] A. E. Howe, I. Ray, M. Roberts, M. Urbanska, and Z. Byrne. The psychology of security for the home computer user. In *2012 IEEE Symposium on Security and Privacy*, pages 209–223. IEEE, 2012.

[89] A. Hughes, J. Chromy, K. Giacoletti, and D. Odom. Impact of interviewer experience on respondent reports of substance use. *Redesigning an Ongoing National Household Survey*, pages 161–84, 2002.

[90] P. G. Ipeirotis. Demographics of mechanical turk. 2010.

[91] S. L. Jackson. *Research methods and statistics: A critical thinking approach*. Cengage Learning, 2015.

[92] R. Johns. Likert items and scales. *Survey Question Bank: Methods Fact Sheet*, 1, 2010.

[93] M. Johnson, S. Egelman, and S. M. Bellovin. Facebook and privacy. In *the Eighth Symposium*, page 1, New York, New York, USA, 2012. ACM Press.

[94] G. Kalton and D. Kasprzyk. *Treatment of missing survey data*. Department of Biostatistics, University of Michigan, 1986.

[95] G. Kalton and H. Schuman. The effect of the question on survey responses: A review. *Journal of the Royal Statistical Society. Series A (General)*, pages 42–73, 1982.

[96] E. W. Kane and L. J. Macaulay. Interviewer gender and gender attitudes. *Public opinion quarterly*, 57(1):1–28, 1993.

[97] R. Kang, S. Brown, L. Dabbish, and S. B. Kiesler. Privacy attitudes of mechanical turk workers and the us public. In *SOUPS*, pages 37–49, 2014.

[98] K. Kelley. Good practice in the conduct and reporting of survey research. *International Journal for Quality in Health Care*, 15(3):261–266, 2003.

[99] P. G. Kelley. Conducting usable privacy & security studies with amazons mechanical turk. In *Symposium on Usable Privacy and Security (SOUPS)(Redmond, WA*, 2010.

[100] L. Kish. Survey sampling. 1965.

[101] B. Kitchenham and S. L. Pfleeger. Principles of survey research: part 5: populations and samples. *ACM SIGSOFT Software Engineering Notes*, 27(5):17–20, 2002.

[102] F. Kreuter, S. Presser, and R. Tourangeau. Social desirability bias in cati, ivr, and web surveys the effects of mode and question sensitivity. *Public Opinion Quarterly*, 72(5):847–865, 2008.

[103] J. A. Krosnick. Survey research. *Annual review of psychology*, 50(1):537–567, 1999.

[104] M. Krysan. Privacy and the expression of white racial attitudes: A comparison across three contexts. *Public Opinion Quarterly*, pages 506–544, 1998.

[105] P. Kumaraguru, S. Sheng, A. Acquisti, L. F. Cranor, and J. Hong. Teaching johnny not to fall for phish. *ACM Trans. Internet Technol.*, 2010.

[106] R. N. Landers and T. S. Behrend. An inconvenient truth: Arbitrary distinctions between organizational, mechanical turk, and other convenience samples. *Industrial and Organizational Psychology*, 8(02):142–164, 2015.

[107] J. Lazar, J. H. Feng, and H. Hochheiser. *Research Methods in Human-Computer Interaction*. Wiley Publishing, 2010.

[108] B. L. Leech. Asking questions: techniques for semistructured interviews. *Political Science & Politics*, 35(04):665–668, 2002.

[109] E. Lin, S. Greenberg, E. Trotter, D. Ma, and J. Aycock. Does domain highlighting help people identify phishing sites? In *CHI*, 2011.

[110] D. Marques, I. Muslukhov, T. Guerreiro, L. Carriço, and K. Beznosov. Snooping on mobile phones: Prevalence and trends. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, 2016.

[111] P. V. Marsden and J. D. Wright, editors. *Handbook of survey research*. Emerald, Bingley, UK, 2 edition, 2010.

[112] M. N. Marshall. Sampling for qualitative research. *Family practice*, 13(6):522–526, 1996.

[113] E. Martin, T. J. DeMaio, and P. C. Campanelli. Context effects for census measures of race and hispanic origin. *Public Opinion Quarterly*, 54(4):551–566, 1990.

[114] P. McDonald, M. Mohebbi, and B. Slatkin. Comparing google consumer surveys to existing probability and non-probability based internet surveys. *Google Whitepaper*, 2012.

[115] N. R. McKenney and C. E. Bennett. Issues regarding data on race and ethnicity: the census bureau experience. *Public health reports*, 109(1):16, 1994.

[116] B. Mirel. Usability and hardcopy manuals: Evaluating research designs and methods. *SIGDOC Asterisk J. Comput. Doc.*, 14(4):69–77, Sept. 1990.

[117] H. J. Parry and H. M. Crossley. Validity of responses to survey questions. *Public Opinion Quarterly*, 14(1):61–80, 1950.

[118] M. Q. Patton. *Qualitative research*. Wiley Online Library, 2005.

[119] Pew Research Methods. Questionnaire design.

[120] S. Preibusch. Guide to measuring privacy concern: Review of survey and observational instruments. *International Journal of Human-Computer Studies*, 71(12):1133–1143, 2013.

[121] S. Presser and J. Blair. Survey pretesting: Do different methods produce different results. *Sociological methodology*, 24(1):73–104, 1994.

[122] S. Presser, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, J. M. Rothgeb, and E. Singer. Methods for testing and evaluating survey questions. *Public opinion quarterly*, 68(1):109–130, 2004.

[123] C. C. Preston and A. M. Colman. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences. *Acta psychologica*, 104(1):1–15, 2000.

[124] S. B. Quinn and W. A. Belson. *The effects of reversing the order of presentation of verbal rating scales in survey interviews*. Survey Research Centre, the London School of Economics and Political Science, 1969.

[125] L. Rainie, S. Kiesler, R. Kang, M. Madden, M. Duggan, S. Brown, and L. Dabbish. Anonymity, privacy, and security online. *Pew Research Center*, 5, 2013.

[126] E. M. Redmiles, S. Kross, and M. L. Mazurek. How i learned to be secure: a census-representative survey of security advice sources and behavior. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 666–677. ACM, 2016.

[127] E. M. Redmiles, S. Kross, and M. L. Mazurek. Where is the digital divide? examining the impact of socioeconomics on security and privacy outcomes. 2016.

[128] M. T. Roberson and E. Sundstrom. Questionnaire design, return rates, and response favorableness in an employee attitude questionnaire. *Journal of Applied Psychology*, 75(3):354, 1990.

[129] S. A. Robila and J. W. Ragucci. Don't be a phish: Steps in user education. In *SIGCSE*, 2006.

[130] S. Schechter and J. Bonneau. Learning assigned secrets for unlocking mobile devices. In *SOUPS*, 2015.

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

[132] B. Schneier. Security in the real world: How to evaluate security technology. *Computer security journal*, 15:1–14, 1999.

[133] H. Schuman and S. Presser. *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage, 1996.

[134] N. Schwarz. Self-reports: how the questions shape the answers. *American psychologist*, 54(2):93, 1999.

[135] E. M. Shaeffer, J. A. Krosnick, G. E. Langer, and D. M. Merkle. Comparing the quality of data obtained by minimally balanced and fully balanced attitude questions. *Public Opinion Quarterly*, 69(3):417–428, 2005.

[136] G. M. Shapiro. *Sample Size*, pages 782–784. Sage Publications, Inc., 0 edition, 2008.

[137] K. B. Sheehan. E-mail survey response rates: A review. *Journal of Computer-Mediated Communication*, 6(2):0–0, 2001.

[138] S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. Downs. Who falls for phish?: A demographic analysis of phishing susceptibility and effectiveness of interventions. In *CHI*, 2010.

[139] D. K. Smetters and N. Good. How users use access control. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, page 15. ACM, 2009.

[140] A. J. Søgaard, R. Selmer, E. Bjertness, and D. Thelle. The oslo health study: The impact of self-selection in a large, population-based survey. *International journal for equity in health*, 3(1):1, 2004.

[141] P. Squire. Why the 1936 literary digest poll failed. *Public Opinion Quarterly*, 52(1):125–133, 1988.

[142] J. Staddon, D. Huffaker, L. Brown, and A. Sedley. Are privacy concerns a turn-off?: engagement and privacy in social networks. In *Proceedings of the eighth symposium on usable privacy and security*, page 10. ACM, 2012.

[143] J. Stoutenbourgh. Demographic measures. *Encyclopedia of survey research methods*, 1:185–186, 2008.

[144] J. V. Stulginskas, R. Verreault, and I. B. Pless. A comparison of observed and reported restraint use by children and adults. *Accident Analysis & Prevention*, 17(5):381–386, 1985.

[145] S. Sudman and N. M. Bradburn. Asking questions: a practical guide to questionnaire design. 1982.

[146] J. Tarnai and D. L. Moore. Measuring and improving telephone interviewer performance and productivity. *De Leeuw, L. Japec, PJ Lavrakas, MW Link, and RL Sangster (Eds.), Advances in Telephone Survey Methodology*, pages 359–384, 2008.

[147] R. Teclaw, M. C. Price, and K. Osatuke. Demographic question placement: Effect on item response rates and means of a veterans health administration survey. *Journal of Business and Psychology*, 27(3):281–290, 2012.

[148] R. Tourangeau, L. J. Rips, and K. Rasinski. *The psychology of survey response*. Cambridge University Press, 2000.

[149] R. Tourangeau and T. W. Smith. Asking sensitive questions the impact of data collection mode, question format, and question context. *Public opinion quarterly*, 60(2):275–304, 1996.

[150] R. Tourangeau and T. Yan. Sensitive questions in surveys. *Psychological bulletin*, 133(5):859, 2007.

[151] Z. Tufekci. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *arXiv preprint arXiv:1403.7400*, 2014.

[152] B. Ur, P. G. Kelley, S. Komanduri, J. Lee, M. Maass, M. L. Mazurek, T. Passaro, R. Shay, T. Vidas, L. Bauer, N. Christin, and L. F. Cranor. How does your password measure up? the effect of strength meters on password creation. In *USENIX Sec.*, 2012.

[153] W. M. Vagias. Likert-type scale response anchors. *Clemson International Institute for Tourism & Research Development, Department of Parks, Recreation and Tourism Management. Clemson University*, 2006.

[154] E. R. Van Teijlingen, A.-M. Rennie, V. Hundley, and W. Graham. The importance of conducting and reporting pilot studies: the example of the scottish births survey. *Journal of advanced nursing*, 34(3):289–295, 2001.

[155] S. L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

[156] R. Wash and E. Rader. Too much knowledge? security beliefs and protective behaviors among united states internet users. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*, pages 309–325, 2015.

[157] L.-J. Weng and C.-P. Cheng. Effects of response order on likert-type scales. *Educational and psychological measurement*, 60(6):908–924, 2000.

[158] B. E. Whitley, M. E. Kite, and H. L. Adams. *Principles of research in behavioral science*. Routledge, 2012.

[159] A. Whitten and J. D. Tygar. Why johnny can't encrypt: A usability evaluation of pgp 5.0. In *Usenix Security*, volume 1999, 1999.

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