I never reuse passwords! Development and Validation of a Security and Privacy Social Desirability Scale (SP-SDS) for end users without a background in computer science

Laura Marie Abels
University of Bonn

Matthew Smith
University of Bonn, Fraunhofer FKIE

Anna-Marie Ortloff
University of Bonn

Abstract

Social desirability bias can be a problem in human-subjects-research, if participants give answers they believe researchers want to hear, instead of their true opinion. This is especially concerning for sensitive topics, which are prevalent in Usable Security and Privacy (USP) research, e.g. when asking users about their security habits, experiences of digital abuse or opinions on surveillance. While validated scales measuring general social desirability bias exist, it is unclear how applicable they are in USP. Besides the jarring context switch, it is uncertain how well social desirability of security and privacy related behavior matches general social desirability. To address this, we developed and validated a 13-item security and privacy-specific social desirability scale (SP-SDS), (total N=1167). A correlation of $\tau = .43$ between SP-SDS and the established Marlowe-Crowne SDS confirms that social desirability bias in USP is related to, but distinct from, general social desirability bias. Based on our validated scale we conducted a study with a representative US-sample (N=867) for participants without a CS-background, to measure the perception of social desirability for the behaviors contained in the SP-SDS and to create a baseline for comparison with other samples. Finally, we make recommendations for using SP-SDS in USP studies.

1 Introduction

Biases in research are systematic errors that can occur in all phases of the research process [65]. They exist in all research [84] and can occur intentionally or unintentionally [83]. Biases affect the validity and reliability of study results, and misinterpretation of data can have significant consequences for practice [84]. Biases in studies cannot be completely avoided, but measures can be taken to reduce biases, e.g. random sampling to minimize selection bias [83]. Some biases which have been investigated and discussed in Usable Security and Privacy (USP) research are sampling bias from online sampling [40, 71] or the focus on Western, Educated, Industrialized, Rich, and Democratic (WEIRD) demographics in USP studies [34], and biases and differences in different study environments [85], as well as self-reporting and social desirability bias [72]. Many studies in USP rely on self-reported data from participants, and especially for large-scale studies, data is often collected through surveys. This means that our conclusions and findings may be susceptible to social desirability bias, especially where studies measure sensitive behavior, such as experiencing online abuse [1, 55, 95, 102] or personal security practices [23, 37, 82] or collect opinions on sensitive topics, such as surveillance [18, 21, 28, 29, 76]. Social desirability bias is the tendency to give answers in surveys that do not reflect true opinions, but rather socially desirable behavior. The resulting overestimation of socially desirable behavior and underestimation of undesirable behavior can lead to a bias in study results. It is therefore important to recognize situations in which social desirability bias can occur and to determine the extent of the bias so that it can then be taken into account and corrected accordingly [43]. There are several general validated scales to measure social desirability bias [15, 66]. However, since these scales ask about participants' behavior in areas not at all related to USP, it is not clear whether social desirability bias in general contexts is transferable to more specific contexts [22], such as USP. In addition, including such scales can represent a significant context-switch for participants.

Therefore, we developed and validated a security and privacy social desirability scale (SP-SDS)[1] with a specific focus

USENIX Symposium on Usable Privacy and Security (SOUPS) 2025.
August 10–12, 2025, Seattle, WA, United States.

[1] the finalized scale can be found in an easy to grab format on the following website: https://www.besec.uni-bonn.de/for-researchers/sp-sds

on end users, as they are the largest demographic interacting with privacy and security topics and still frequently studied demographic in USP. We followed recommended scale development processes such as described in Votipka et al. [101] and Boateng et al. [9].

First, we developed a pool of items and tested them for comprehensibility (N=16) and suitability (N=30). We based the structure of our scale on the classic Marlowe-Crowne social desirability scale (M-C SDS), which is the most widely used scale in information system research [48]. We then conducted a principal component analysis (N=297) for item reduction, identifying five components relating to different types of socially desirable security and privacy behavior, and confirmed the five-factor structure using a confirmatory factor analysis (N=824). We found that responses to SP-SDS exhibit a medium correlation (Kendall's τ=.43) with M-C SDS and smaller correlations with a different general Social Desirability Scale (SDS) (τ=.26) and a context-specific SDS (τ=.17). This shows that social desirability bias in USP is related to more general types of social desirability bias, but the concept is nevertheless distinct and we recommend using SP-SDS in USP studies. Our finalized scale includes 13 security and privacy items and allows USP researchers to measure how prone their participants are to the social desirability bias to aid in interpreting their data.

We further conducted a calibration study to elicit the proportions of people who classify the behaviors covered by the SP-SDS as socially (un)desirable with a representative US-sample for participants without CS background (N=867). The resulting proportions can be used as weights when calculating weighted social desirability scores which take into account that not all behaviors are perceived as equally socially (un)desirable. We provide guidance on interpreting the resulting SP-SDS scores by providing a baseline distribution of social desirability scores, to which researchers can compare their sample. Since our score was developed for and only validated with participants without a CS background, researchers using it should take care that their samples fit this criterion, e.g. by asking about participants' experience and background with respect to computer science.

2 Related Work

In the following we discuss biases, specifically social desirability bias and introduce prior work on general and context-specific SDS in other domains. Finally, we present prior research on biases in USP.

2.1 Biases

Biases are systematic errors, which can be present throughout the research process, from planning, through data collection and analysis, to reporting and publishing research [65]. They may be introduced by the researcher, e.g. experimenter bias [56, 103] regarding behavior [75] or inherent characteristics of the experimenter [17, 19, 51, 104], or by the participant, e.g. social desirability bias [19, 30, 48, 49] or self-reporting bias [64, 72], occur in the data collection process, e.g. sampling bias [34, 71] or environmental factors bias [23, 85], analysis process, e.g. confirmation bias [39, 52, 89], or as publication bias in the community as a whole [79, 80].

Social desirability bias often occurs when asking about sensitive topics on self-report measures [26] and is considered one of the most prominent response biases in survey research [48]. Sensitive survey questions generally lead to comparatively higher non-response rates or greater measurement errors in the responses than questions on other topics [98]. Misreporting on sensitive topics occurs quite frequently and is largely situation-dependent, with the extent of misreporting depending on whether the respondent has something embarrassing to report and on the design features of the survey [98]. It is a more or less motivated process in which respondents adapt their statements in order not to embarrass themselves in front of the interviewer or to avoid repercussions from third parties [98]. Several studies compared methods of data collection on sensitive topics, e.g. [13, 44, 97].

In an evaluation of different methods to collect survey data on sensitive topics, the methods did not differ in response rates, but influenced the extent to which sensitive behaviors were reported: Computerized self-administration increased respondents' willingness to make potentially embarrassing admissions in surveys [97]. In another comparison of three different survey administering strategies, Kreuter et al. found that each performed best on a different error metric of unit non-response, item non-response and reporting accuracy, so the choice of survey mode could depend on which source of error is most important in a survey [44]. Coutts and Jann tested two techniques trying to provide anonymity to respondents' to reduce socially desirable responses: the randomized response technique, where a randomizing device is used to determine whether the respondent answers the sensitive question truthfully (the result is only known to the respondent) and the unmatched count technique, where the participants are randomly divided into two groups (with only one group being asked the sensitive question) and only indicate the number of behaviors which apply to them [13]. While the randomized response techniques were problematic, the unmatched count technique was a more promising approach for a self-administered setting [13]. However, social desirability bias still occurs to an extent even in anonymous online surveys, and meta-analyses have found that the influence of data collection method is negligent [20, 69]. Kwak et al. conducted a literature review and found that, social desirability bias is rarely addressed in information system studies that use self-reported measures and where it has been mentioned, it was not adequately addressed [48].

2.2 Social Desirability Scales

Social desirability scales can assess the extent of social desirability bias and can be used to control this bias [49]. They are easily compatible with online surveys, in contrast to other alternative strategies for avoiding social desirability bias [49], such as using EEG data [5], and techniques to provide more anonymity to participants, like the randomized response or unmatched count technique [13].

The M-C SDS is one of the most used scales for measuring general social desirability bias [6, 68, 74]. It consists of 33 items, which describe behaviors that are either socially desirable and improbable of occurrence, or socially undesirable but frequently occurring [15]. Respondents answer "true" or "false" whether the behaviors apply to them personally [15]. Since the original questionnaire is lengthy, various short-forms have been developed [74, 87]. Other general scales used to detect social desirability bias are the Social Desirability Scale–17 [88], the Self-Deception Questionnaire [77] and the Balanced Inventory of Desirable Responding (BIDR) [67]. Some, like BIDR consist of two constructs: self-deceptive positivity and impression management [67], while others, like the M-C SDS or the Social Desirability Scale–17 [15,88] have only a single factor. To measure social desirability bias in specific domains, various context-specific scales for measuring social desirability have been developed over time. Some examples are the Environmental desirability responding scale (EDRS) [22], and scales for the food safety area [38], in business contexts [54] and for children [14].

2.3 Biases in USP

Biases have also been observed in USP studies, including research focused specifically on biases. Prior work detected differences due to the study environment, e.g. between online and laboratory studies [23, 85]. (Environmental factors bias)

The influence of data collection methods and sampling strategies have also been investigated, focusing on the representativeness of online crowd sourcing platforms, with the degree of representativeness of MTurk workers' responses changing over time [40, 71, 94]. Differences were also identified between recruitment channels for software developer participants [41, 92]. Most participants in USP studies belong to the WEIRD demographic group [34]. (Sampling bias)

Priming participants can cause bias and by not priming the participants, biases such as demand characteristics can be avoided [45]. Studies with students, freelancers and company developers, who were either primed to securely implement password storage or did not receive this security priming, showed that priming can have a large influence on performance [58–61]. On the other hand, using deception in the study design did not change the outcome of a study on password storage [16]. (Experimental procedure bias)

Redmiles et al. compared survey data to field measurements on software updates to investigate social desirability bias and found that they differed systematically, with respondents reporting faster update speeds for themselves in the survey than were measured in the field data, and recommending even faster update speeds [72]. Given that USP research encompasses various sensitive topics, we believe that social desirability bias is relevant to our community. As the currently available SDSs measure social desirability bias in a more general context [15, 67] or in different contexts than USP [22, 38, 54], we believe a USP specific SDS is beneficial.

3 Study Design

USP research encompasses studies with different types of users, among them marginalized groups [1, 102] and specialists like software developers [58,91] or administrators [78,96]. Different groups vary in their perceptions and behaviors related to privacy and security, e.g. by their expertise in the area of security and privacy [11, 12, 37, 46], and thus may also differ with respect to their perception of social desirability of such behaviors. We wanted to develop a scale that was applicable to a large demographic and chose to focus the design and validation process of the SP-SDS on end users without a computer science (CS) background as the target population.

We followed recommended scale development processes, such as described in Boateng et al. [9], which is outlined in recent efforts in scale development in USP [10,24,33,81]. We first generated an initial item pool and evaluated content validity using an expert focus group. We used two pilot studies (N=16, N=30) to test the comprehensibility of the items. For all further studies, we used the representative sampling feature on the crowd-sourcing platform Prolific [2], based on data from the US-census bureau on age, gender and simplified ethnicity. However, we screened out participants with a background in CS, since they were not our target population. We refer to these samples as non-CS US samples. Participants were compensated with £9/hr based on duration estimates. Participants with CS background received partial compensation for their time taken to finish the screening questions.

The demographic information on our study samples are in Table 2. In the next step, we conducted a **scale development study** (N=297) to determine the factor structure of our scale and remove inconsistent items. In a **validation study**, with a new sample (N=824), we confirmed the factor structure from the scale developing study and evaluated the SP-SDS' reliability and validity. Finally we conducted an additional **calibration study** (N=867) to determine population levels of social desirability for the behaviors included in our final scale. This allows future users of the SP-SDS to weight responses by their social desirability. We provide the measurement instruments used in our studies on OSF[3].

[2] https://www.prolific.com/

[3] https://osf.io/v5rph/?view_only=429ec281ad504b5499206e5af323c1bf

3.1 Ethics

Our project was approved by the ethics review board at one of our institutions, and we complied with the General Data Protection Regulation. Data was collected anonymously, and we informed our participants about the study and our data collection process before the start of their participation. As the aim of the study is to measure social desirability bias, participants in the scale development and validation study were only informed that we were asking questions about computer-related and general behavior, but not about the full aim of the study. This was necessary as otherwise the answers to our survey questions could have been influenced. Participants were fully informed at the end of the study.

4 Item Generation

We based the questions on the M-C SDS. This means that our items represent either socially desirable and rarely occurring behaviors or socially undesirable, but frequently occurring behaviors. We refer to this condition as item requirements in the following. Items that contain socially desirable, but rare behaviors are coded as true, meaning that the social desirable answer to this would be "Yes, I do this behavior". Conversely, items with frequent but undesirable behaviors are coded as false. For analyses throughout this paper, we recoded those items coded as false before further analysis.

4.1 Initial Item Generation

We started generating items by adjusting general items from M-C SDS to a more security or privacy related context, e.g. the M-C SDS item “If I could get into a movie without paying and be sure I was not seen I would probably do it” was modified to “If I knew I wouldn’t get caught, I would watch movies illegally”. To supplement this item pool, we asked CS and IT Security students, who were enrolled in an empirical methods course and had already studied social desirability bias to come up with additional security and privacy related socially desirable behaviors. This led to a total initial pool of 373 items. We removed duplicates, adapted the phrasing and supplemented this item pool by additional items we generated based on the topics of interest we identified in the initial item pool, leading to 95 filtered items. The items covered the following topics: *AI*, *backups*, *encryption*, *illegal behavior*, *installation*, *network*, *passwords*, *phishing*, *privacy policies*, *programming*, and *updates*. There were 18 items which do not fit into any of the topics above. We removed 18 items not suitable for end users, most of which belonged to the categories of *programming* and *network*.

The item pool was then separately presented to two experts. The first person is familiar with biases and study design from their experimental work and the second person has previously done work on social desirability bias specifically. After considering their comments, an additional 17 items were removed and phrasing was further improved. Reasons for the removal of items were ambiguous wording, insufficient item requirements and lack of applicability to the target group.

4.2 Expert Review

The remaining 60 items were discussed in a brainstorming meeting with six researchers with experience in conducting human subject studies in USP, i.e. those who we expect to use the scale in their studies. The item pool was presented to the researchers, and they were then asked to discuss each item in terms of wording and suitability for measuring social desirability. As a result, 23 items were excluded, some items were rephrased and two new items were added.

4.3 Piloting

To ensure that the items were understandable for the target group, 16 participants that did not have a CS background were recruited from different age groups (20-30 years old, 30-40 years old, 40-50 years old, 50+ years old) for our first pilot study. Starting with the youngest age group, participants were asked to rate every item according to how socially desirable they think a given behavior is, and how many users they think exhibit this behavior. A 7-point Likert scale from "Very undesirable" (1) to "Very desirable" (7) was used for the first assessment, with an additional response option being "I don’t understand the statement". A slider from 0 to 100% of users was used to measure the estimated percentage of end users exhibiting these behaviors. The items presented in the pilot studies are in Table 7 in the Appendix. The survey was conducted via Zoom or in person, depending on participant preferences, to clarify any ambiguities immediately. Participants completed the survey for themselves and were encouraged to ask questions as soon they had trouble understanding anything. Researchers were available for direct feedback. After each participant, we used the feedback to improve the items. As soon as at least two people in an age group had no more suggestions for improvement, the next age group was surveyed. Four items were adapted by explaining the (technical) terms they contained.

We then recruited 48 additional participants on Prolific. We screened out 15 participants based on failed comprehension checks, 3 based on failed attention checks and 5 participants with missing values. To identify participants with CS background we asked the following questions:

- Are you working or have you ever worked in an area related to computer science?
- Are you or have you studied something related to computer science?
- Do you have hobbies related to computer science?

We count participants who answered "Yes" to at least one of these questions as participants with CS background. 22 out of 41 participants had a CS background. We found that participants with a CS background found the behaviors more desirable (Mdn=6) than those without a CS background (Mdn=5) (r=.15) but also believed they were more frequent in the general population (Mdn_{CS}=30% vs $Mdn_{non\text{-}CS}$=25%, r=.15). For individual behaviors, such as watching movies illegally (Mdn_{CS}=5.5 vs $Mdn_{non\text{-}CS}$=3) and reading terms and conditions (Mdn_{CS}=6 vs $Mdn_{non\text{-}CS}$=4), differences in median social desirability were as high as 2 or 2.5, which can represent a perception flip from *desirable* to *neutral* on our measuring instrument. Given that prior work also showed differences in reported behavior and mental models between experts and non-expert in IT security [12,27,37,46], we decided to screen out participants with a CS background from further studies. To do this, we provided a definition of computer science [4], examples of computers science related jobs [5] and used the questions presented above.

We recruited 71 participants on Prolific for our second pilot study. Prolific does not contain pre-screeners specifically based on CS background, but allows screening within the surveys deployed on their platform. We screened out three participants based on failed attention checks, 28 who indicated a CS background meaning they did not fit our target group and excluded 2 with missing values, resulting in 38 participants for our analysis.

The participants were asked to complete a similar survey as in the first pilot study. However, mirroring the question format of the M-C SDS [15], these questions were binary and asked for participants' judgment on whether the behavior in the statement was socially desirable, and whether the behavior in the statement was realistic. We still provided an option to indicate problems understanding the statement. As we identified some difficulties in understanding the term "social desirability" in the first part of our pilot study, we included a validated item from the M-C SDS as a comprehension check and excluded participants from the analysis who classified the socially desirable item "I'm always willing to admit it when I make a mistake" as undesirable since we cannot check whether the participants' opinion on this item simply differs from the general public or whether the term "social desirability" was misunderstood. To consolidate this decision, we compared the responses of the two groups (number of items rated as socially desirable) using a Wilcoxon Rank Sum test, assuming that participants who failed our comprehension check would answer inconsistently and differently. We found that participants who rated the M-C SDS item as undesirable (Mdn=11.5) differed in their assessment of social desirability from those who rated it as desirable (Mdn=32), z=-3.039, p=.0024; r=.493. We therefore removed 8 participants and used the remaining 30 for our further analysis.

A total of four items were marked as hard to understand once each but since the participants did not provide a reasoning in the following free text field, we examined these items, but ultimately did not adjust their phrasing. Based on this study, we retained items where no more than 30% of respondents rated the behavior as undesirable or as realistic,

abbreviation	socially undesirable	realistic	decision
AI forbidden	**16.67**	**26.67**	accept
annoyed	**23.33**	33.33	reject
back up	46.67	**13.33**	reject
check AI	**10.00**	**23.33**	accept
check HTTPS	43.33	**23.33**	reject
check backup	36.67	**13.33**	reject
check leaks	**26.67**	**13.33**	accept
check program	**6.67**	43.33	reject
clicked link	**23.33**	33.33	reject
data collected	**30.00**	**20.00**	accept
different passwords	**10.00**	**23.33**	accept
disclose AI	33.33	**16.67**	reject
encrypt mail	36.67	**6.67**	reject
ignore update	**26.67**	36.67	accept*
ignore warnings	**20.00**	46.67	accept*
illegal movies	**23.33**	**16.67**	accept
install updates	40.00	**30.00**	reject
laughed	**23.33**	40.00	reject
lock device	**16.67**	50.00	reject
log out	**6.67**	63.33	reject
looked screen	43.33	**10.00**	reject
mail attachment	**23.33**	**26.67**	accept
personal password	50.00	**20.00**	reject
pirated software	**16.67**	**23.33**	accept
policy access	**20.00**	**20.00**	accept
polite online	**10.00**	**20.00**	accept
random passwords	53.33	**13.33**	reject
read messages	**30.00**	**26.67**	accept
read policy	**30.00**	**16.67**	accept
read terms	**16.67**	**20.00**	accept
required cookies	**30.00**	46.67	reject
reuse passwords	**20.00**	**23.33**	accept
secure passwords	**16.67**	**26.67**	accept
share passwords	40.00	**30.00**	reject
smashing computer	**16.67**	**26.67**	accept
troll comment	**30.00**	**20.00**	accept
turnoff location	**26.67**	**20.00**	accept
two fa	**23.33**	40.00	reject
write password	46.67	**20.00**	reject

Table 1: Pilot study (N=30) results in percent. Bold values mean that our inclusion criteria are fulfilled. * means we overrode these decisions.

[4] Computer Science is the study of computers and computational systems. Computer scientists deal mostly with software and software systems; this includes their theory, design, development, and application. Principal areas of study within Computer Science include artificial intelligence, computer systems and networks, security, database systems, human computer interaction, vision and graphics, numerical analysis, programming languages, software engineering, bioinformatics and theory of computing. (source: https://undergrad.cs.umd.edu/what-computer-science)

[5] Examples of computer science related jobs are: Data scientist, network administrator, software developer, IT Project Manager

to ensure that our items could measure social desirability. A total of 19 items were removed, 11 due to the social desirability assessment. We kept the items *ignore warning* and *ignore update* despite their rating as realistic by 46.67% and 36.67% of participants.

Prior work shows that security warnings are in fact often ignored by users [73, 90, 99], with click-through-rates as high as e.g. 70.2% for google Chrome's SSL warning in 2013 [3] and 62.4% in 2018 [73], making it likely that users have at least once ignored a security warning. Similarly, ensuring timely update behavior has been investigated for multiple user groups [96], including end users [57, 72, 100], as delayed updates are a considerate problem for IT security [70, 93].

This research shows that despite a fairly high proportion of our pilot participants believing this behavior is realistic, it is actually not, leaving the items as suitable for our scale. The exact proportions for all items are shown in Table 1.

5 Refining the Scale - the Scale Development Study

In order to reduce the number of items on our scale and evaluate its factor structure, we recruited more participants and performed a Principal Components Analysis (PCA). This section describes the methodology and results of the PCA. The participants saw the remaining 20 items, which we retained after the two pilot studies, and indicated with true/false whether the statements applied to them personally or not. To avoid biasing participants, participants were told that we wanted to gain insights into computer-based behaviors in our survey and were told the actual purpose of the study at the end.

5.1 Participants

We recruited 516 participants. We screened out 5 participants who took part in the survey twice, 212 participants with CS background and excluded 2 with missing values in our 20 true/false items. After removing these participants, we retained 297 participants in our analysis. As per guidelines used in other scale development studies, our final sample included at least 10 participants per scale item included in the study [24, 33]. Participants were compensated with 0.75 USD.

Item	Scale Development	Scale Validation	Calibration
total N	297	824	867
Gender			
Woman	52.86%	57.16%	52.02%
Man	46.13%	41.62%	38.75%
Non-binary	1.01%	0.85%	1.85%
Self-described	0%	0.12%	0.46%
No Answer	0%	0.24%	6.92%
Age			
Minimum	18	18	18
Maximum	89	86	84
Mean	47.50	47.71	47.26
Standard deviation	15.87	15.1	15.60
No answer	3.37%	0.02%	8.65%
Ethnicity			
White	70.03%	70.27%	67.36%
Black	11.45%	8.86%	10.38%
Mixed	7.07%	6.92%	6.34%
Asian	6.40%	7.77%	5.88%
Other	4.04%	4.61%	2.65%
No Answer	1.01%	1.58%	7.38%
Education			
Professional degree	2.02%	2.31%	1.61%
Doctoral degree	2.36%	1.33%	1.85%
Master's degree	11.78%	12.62%	14.65%
Bachelor's degree	34.01%	35.80%	32.87%
Associate degree	10.44%	12.50%	9.34%
Some college (no degree)	18.18%	20.15%	18.34%
Technical certification	4.71%	2.31%	2.42%
High school (including GED)	16.16%	12.14%	11.76%
Less than High School	0.34%	0.85%	0.35%
No answer	0%	0%	6.81%
Occupation			
Employed full-time	45.79%	48.06%	47.87%
Employed part-time	15.82%	17.96%	17.07%
Contract or temporary	1.35%	0%	0%
Retired	13.80%	12.86%	13.49%
Unable to work	3.03%	2.67%	1.04%
Unemployed	9.43%	9.83%	8.42%
Other	10.10%	8.00%	4.84%
No answer	0.67%	0.61%	7.27%

Table 2: Demographic information (Percentage) for the participants in the scale development, validation and calibration studies.

5.2 Results

A PCA was conducted to examine the factors of our scale. We calculated the polychoric correlation matrix. The items *AI forbidden*, *check AI*, *check leaks*, *mail attachment*, *read messages*, *smashing computer* and *turnoff location* were removed because they only had polychoric correlations below .3. Since we intend to measure behavior related to social desirability, if individual items have too low correlations with the rest, this might not be the case for them, so we excluded them. A PCA was conducted on the 13 remaining items with oblique rotation (oblimin). The Kaiser-Meyer-Olkin measure verified the sampling adequacy for the analysis KMO = .74 ("good" [25]), and all KMO values for individual items were >.62, which is above the acceptable limit of .5 [25]. Barlett's test of sphericity, $\chi^2(78)=1091.49$, $p<.001$, indicated that correlations between items were sufficiently large for PCA. An initial analysis was run to obtain eigenvalues for each component in the data. Five components had eigenvalues over Kaiser's criterion of 1 and in combination explained 68.2% of the variance. The scree plot was ambiguous so Kaiser's

criterion was used, and five components were retained in the final analysis. Table 3 shows the factor loadings after rotation and table 4 the structure matrix. The items that cluster on the same components suggest that component 1 represents data protection, component 2 passwords, component 3 illegal behavior, component 4 online behavior and component 5 ignore (security) advice.

To measure the internal consistency reliability, we used the Kuder-Richardson formula 20 (KR-20), which is intended for dichotomous data [53, 63]. The internal consistency coefficient according to KR-20 is .76 and is considered acceptable (>0.7) [53, 63, 86]. We calculated the item-total correlations of our scale, which indicates that reliability could not be substantially improved by removing an item.

6 Finalizing the Scale - the Validation Study

We recruited another sample to validate our scale. The participants were presented the 13 items remaining after the PCA and again had to indicate with true/false whether the statements applied to them. To test validity, we added three further social desirability scales. These were the M-C SDS [15], the BIDR [66] and the Environmentally Desirable Response Scale (EDRS) [22]. In addition, we asked the participants directly whether they were susceptible to social desirability. The first of three such questions asked directly how likely they were to admit to negative behavior in the survey. The two others were based on prior work suggesting two sub-constructs of social desirability: *self-deceptive positivity* and *image management* [66, 67].

6.1 Participants

We recruited 1398 participants. We screened out 23 participants that who took part in the survey more than once, 509 participants with a CS background, and removed 4 participants, who did not give consent, 23 participants that did not complete our survey and 6 participants that answered our survey in under 5 minutes (with mean completion time: 14.7 minutes, Q1=9.5 minutes, Q3=16.1 minutes).

We additionally removed participants with missing answers to the SP-SDS items. After removing these 9 participants, we had 824 participants for our analysis. For some parts of the analysis we excluded another 19 participants with missing values in other scale items. Participants were compensated with 2.25 USD.

6.2 Results

To confirm our identified components from the PCA, we conducted a Confirmatory Factor Analysis (CFA). We calculated the following fit indices: the Root Mean Square Error of Approximation (RMSEA), the Standardized Root Mean Square Residual (SRMR), the Comparative Fit Index (CFI), and the Tucker-Lewis Index (TLI) [9]. We did not consider the chi-square goodness-of-fit test which is unreliable for a large sample size, such as the one in our study [24, 33, 36]. All indices indicate a good fit. CFI=.978 and TLI=.969, which is above the recommendation of .95. The RMSEA was .039, which is below 0.05, indicating a close, good fit and SRMR was .037, also indicating a good fit, as it is below 0.08 [9]. The KR-20 reliability is .77 and is considered acceptable [53, 63, 86].

To measure the convergent validity of the SP-SDS we calculated correlations (Kendall's τ) between SP-SDS and other social desirability scales. Our scale has significant positive correlations with all other tested scales. We found a medium positive correlation (r_τ=.431, p<0.001) between the SP-SDS and the M-C SDS indicating that our scale measures the same concept as the M-C SDS and yet they still differ from each other. We also found positive, but smaller correlations between SP-SDS and BIDR (r_τ=.262 , p<0.001) and SP-SDS and EDRS (r_τ=.174, p<0.001) indicating that our scale shares some conceptual overlap with the others. Given that we started our item generation by adapting items from M-C SDS, the higher correlation that with the other general SDS, BIDR, makes sense. The correlation with the EDRS, which measures social desirability bias in the, different, environmental context, is the smallest. We conclude that our scale measures the concept of social desirability, as these other scales do, but the larger correlation coefficients with the general SDS indicate that social desirability bias differs by context, and our scale is focused on the domain of security and privacy related behavior.

We also investigated the relationship between our three ad-hoc questions directly asking whether participants believed themselves to be susceptible to social desirability bias. Since the first question was formulated as a Likert item, we used Spearman's rank-order correlation to determine the relationship between the likelihood of admitting to behaviors that could be seen negatively in surveys and the SP-SDS score. The correlation was negligible, but negative (ρ=-.07, p=.051). We conducted a Wilcoxon Rank Sum test for the other two questions. Participants who stated that they gave answers that seem more acceptable to others rather than their true opinions, have lower SP-SDS scores (Mdn=5) than participants who answered that they did not give answers that seem more acceptable to others rather than their true opinions (Mdn=6), but the difference was not statistically significant, z=-1.44, p=.149; r=-.05 Participants who stated that their answers were influenced by the way they want to see themselves have higher SP-SDS scores (Mdn=6.5) than those who stated their answers were not influenced by the way they want to see themselves (Mdn=6), z=-2.19, p=.029; r=-.08.The relationship between the three ad-hoc questions and the SP-SDS score is weak and somewhat contradictory. For the first and second question, participants who answered this question in a socially desirable way, also had higher scores on the SP-SDS, but the effect sizes were small in both cases. For the third question, it

item	data protection	passwords	illegal behavior	ignore (security) advice	online behavior
read policy	**0.91**	0.00	-0.04	-0.02	-0.05
read terms	**0.87**	-0.01	0.07	0.01	-0.01
policy access	**0.84**	0.00	0.06	-0.08	0.07
data collected	**0.66**	0.04	-0.13	0.17	0.03
reuse passwords	0.00	**0.87**	0.01	-0.06	0.03
different passwords	-0.02	**0.85**	-0.01	0.01	0.07
secure passwords	0.09	**0.52**	0.05	0.24	-0.32
pirated software	-0.04	0.01	**0.91**	0.00	-0.02
illegal movies	0.05	0.00	**0.87**	0.02	0.03
ignore warnings	-0.03	-0.01	0.04	**0.84**	-0.04
ignore update	0.04	0.00	-0.01	**0.75**	0.13
troll comment	-0.01	0.05	0.02	-0.05	**0.83**
polite online	0.04	0.01	0.02	0.15	**0.75**
Eigenvalues	2.77	1.78	1.64	1.43	1.39
% of variance	21.0	13.5	12.5	10.6	10.6

Table 3: Oblimin rotated factor loadings (N=297)

item	data protection	passwords	illegal behavior	ignore (security) advice	online behavior
read policy	**0.89**	0.14	0.19	0.21	0.04
read terms	**0.88**	0.16	0.29	0.25	0.09
policy access	**0.84**	0.15	0.27	0.16	0.15
data collected	**0.69**	0.17	0.09	0.34	0.10
reuse passwords	0.14	**0.86**	0.19	0.12	0.14
different passwords	0.14	**0.86**	0.18	0.19	0.17
secure passwords	0.23	**0.56**	0.18	0.35	-0.21
pirated software	0.20	0.19	**0.90**	0.14	0.12
illegal movies	0.29	0.20	**0.89**	0.18	0.17
ignore warnings	0.19	0.16	0.16	**0.83**	0.05
ignore update	0.25	0.18	0.15	**0.77**	0.21
troll comment	0.08	0.15	0.14	0.05	**0.83**
polite online	0.16	0.15	0.17	0.25	**0.78**

Table 4: Structure matrix (N=297)

was the other way around, where participants that admitted to giving socially desirable answers in this question, received slightly higher scores for being susceptible to social desirability bias, although the difference between the two groups was small. The ad-hoc questions asked for the specific behavior participants may have just partaken in, and were not validated otherwise, consequently, this may have led to complex interactions between social desirability bias and feeling put in the spotlight about acting this way. As such, we believe the contradictory results from this further analysis do not detract from the validity of the SP-SDS.

7 Using the scale - the Calibration Study

We wanted to consider the fact that not all items are considered socially desirable by all people. For people who do not consider a behavior to be socially desirable, their answer does not depend on their susceptibility to social desirability bias, and they will admit to a certain behavior regardless, since they do not view it as undesirable. When calculating a score by counting how many social desirable responses were given and interpreting how much the person is affected by the social desirability bias based on this, it can lead to incorrect conclusions. To illustrate this, we provide an extreme example. Assume that all items are coded as true (i.e. they are socially desirable but rarely true) and person A finds the first half of the items socially desirable and the second half neutral (i.e. they find it neither socially desirable nor undesirable). Person B, on the other hand, finds the first half of the items neutral and the second half desirable. Also assume that both A and B are completely susceptible to social desirability bias, and thus respond that they do the behaviors they consider to be socially desirable. However, they freely report their actual behavior (i.e. that they are not exhibiting the behavior) in answering the

other items, as their stance towards these behaviors is neutral and they do not feel social pressure in answering these items. When calculating scores by summing the number of socially desirable responses, both A and B would have a medium score, despite being completely susceptible to the bias. If A and B were representative of the whole population, we would need to adjust our scale accordingly. To take this into account, we conducted an additional calibration study to measure the levels of social desirability for each behavior in our scale for our population of interest. We then use the degree of social desirability for each item as weights when calculating our score. For our calibration study, the participants were shown the 13 items of the SP-SDS and were asked whether the behavior in the statement was socially desirable, and whether the behavior in the statement was realistic with the binary answer options "yes" and "no".

7.1 Participants

We recruited 1746 participants. We screened out 20 participants who took part in the survey multiple times, 711 participants with a CS background and 144 who failed our comprehension check question. We also excluded 1 participant that did not complete our survey and 1 straightliner, who always answered "no", regardless of the behavior in question.

We additionally removed 2 participants with missing values, resulting in 867 participants. Participants were compensated with 0.90 USD.

7.2 Results

The proportion of participants rating the behavior in the SP-SDS items as socially desirable is shown in Table 5. These proportions range from 0.68 (for item *read policy*) to 0.84 (for item *polite online*), mean = 0.76. 42% of the participants found all 13 items socially desirable, but the number of behaviors perceived as socially desirable differed widely (M=9.83, SD=3.86)

To calculate an SP-SDS score, taking into account the different levels of social desirability from the calibration study, each item response is weighted by the proportion of participants in our sample, who think it is socially desirable. Since the maximum possible score is then 9.83, which makes interpretation unintuitive, we normalize the score, so the maximum possible value of the SP-SDS is 1 and the minimum is 0. To calculate the social desirability score of participant j, use the following formula with the weights from Table 5:

$$\text{Score(j)} = \frac{\sum_{i=1}^{13} score_{ji} * p_i}{9.83}$$

$$\text{where } score_{ji} = \begin{cases} 1 & \text{if j answers item i socially desirably} \\ 0 & \text{otherwise} \end{cases}$$

and $p_i \in [0,1]$ proportion of people who consider i socially (un)desirable. Scores can range from 0 to 1 with high scores indicating probably biased responses.

7.3 Exploratory Comparisons for Demographic Subgroups

Prior work indicates that susceptibility to social desirability bias may be different depending on demographic factors like gender [7,8,35], age [4] or education level [31]. We conducted an exploratory multiple regression to determine whether SP-SDS scores may have been influenced by these demographic factors. The baseline group in this regression was an Asian man of average age (48 years), with less than a Bachelor's degree.

Age significantly predicted social desirability scores; β_{age}=0.005, p<.001, 95% CI [0.004,0.006]. Black and Mixed ethnicity also significantly predicted social desirability scores; β_{Black}=0.22, p<.001, 95% CI [0.14,0.30] and β_{Mixed}=0.13, p=.002, 95% CI [0.05,0.21].The other computed standardized regression coefficients were not as strongly significant and ranged from β=-0.37 to β=0.1. The demographic regression accounted together for R^2=.14, 95% CI [.1,.19]. The results of the regression are in Table 6.

This suggests that older people and people of Black, Mixed or self-described ethnicity had higher SP-SDS scores and were thus more susceptible to social desirability bias. A 45 year age difference would represent a similar difference in mean SP-SDS scores as the difference in between Asian and Black participants, a change in SP-SDS scores of about 0.23. People with higher education levels than a Bachelor's degree, and people with a Bachelor's degree had lower SP-SDS scores (a change of -0.06 and -0.04 respectively, compared to participants education lower than a Bachelor's degree) and were thus slightly less susceptible to social desirability bias.

Scores are additionally influenced by participants' perception of social desirability of the behaviors included in the scale, even though we weighted them in our score calculation. As such, we believe that these effects are not practically relevant enough to justify changing experimental procedures for these demographic groups and that the statistical significance we observe is due to our large sample size. Furthermore, our sample was not balanced for all groups, for example non-binary people and people preferring to self-describe their gender were only a very small part of our sample, so even though differences for these groups are larger than between women and men, due to the smaller sample size, the effect cannot be reliably estimated. However, we recommend that researchers using our scale monitor differences in demographic characteristics to evaluate our recommendation further.

abbrevation	statement	Socially desirable
data collected	I am always aware of what personal data is collected by the platforms I use and how it is used	0.75
different passwords	I use different passwords for all my accounts	0.78
ignore update	I never ignore software update reminders	0.73
ignore warnings	I never ignore security warnings on my computer	0.79
illegal movies*	If I knew I wouldn't get caught, I would watch movies illegally	0.79
pirated software*	If I knew I wouldn't get caught, I would use pirated software[1] [1] use of software that is not properly licensed	0.83
policy access	I always read the privacy policy before giving an app/application access to my personal data	0.72
polite online	I am always just as polite online as I am in the real world	0.84
read policy	I always read the privacy policy completely before I agree to them	0.68
read terms	I always read the terms and conditions completely before I agree to them	0.69
reuse passwords	I never reuse passwords	0.69
secure passwords	I only use secure passwords (passwords that are long and complex)	0.80
troll comment	I have never considered posting a troll-comment[1] [1] comments that deliberately try to provoke an argument, conflict or emotional reaction	0.74

Table 5: Ratings of the SP-SDS for end users without a computer science background in the calibration study (N=867). Items coded "false" (marked with *) were recoded before analysis.

Predictor	β	95% CI
Intercept	0.19***	[0.11,0.27]
Age	0.005***	[0.004,0.006]
Gender (Woman)	0.02	[-0.01,0.05]
Gender (Non-binary)	-0.1	[-0.27,0.07]
Gender (Prefer to self-describe)	-0.37	[-0.81,0.08]
Education (Bachelor's degree)	-0.04*	[-0.07,-0.0003]
Education (higher than Bachelor)	-0.06*	[-0.10,-0.01]
Ethnicity (Black)	0.22***	[0.14,0.30]
Ethnicity (Mixed)	0.13**	[0.05,0.21]
Ethnicity (Prefer to self-describe)	0.1*	[0.005,0.19]
Ethnicity (White)	0.05	[-0.01,0.11]

Table 6: Demographic Regression with weighted social desirability score as the outcome variable. * $p < .05$, ** $p < .01$, *** $p < .001$

8 Limitations

Our target population for the SP-SDS was end users without a computer science background. In the scale development, validation and calibration studies, we used Prolific to recruit a representative US-sample based on census-level data on gender, age and ethnicity. Due to our screening questions filtering out people with a background in computer science, regardless of their other demographic characteristics, the remaining participants may not be representative of the US population anymore. Prior research suggests that social desirability bias varies by gender [7, 8, 35] with women being more prone to social desirability bias [7, 8]. There are also differences in age and education: older people are more likely to give socially desirable answers [4], more educated people are less like to give socially desirable answers [31]. In our exploratory regression analysis, we identified statistically significant differences in age, education level and ethnicity. However, we do not judge them to be practically relevant in the sense that researchers should adjust their study protocols to account for the difference.

Furthermore, the SP-SDS is currently only validated based on a US-sample, but the extent and patterns of social desirability bias varies depending on the cultural background [7, 42]. Since many USP studies are also conducted with US-samples, and our recruitment platform Prolific supports representative sampling for some demographic characteristics for this population, we chose to develop our scale based on this sample. Future work needs to examine the validity of the SP-SDS in other cultural contexts.

We also excluded people with a computer science background from our studies, as our target population was end users without a computer science background. Some of the behaviors included in our scale may have a different degree of social desirability for people with more experience in computer science, specifically in IT security. For example, these people might be more capable to determine whether a security warning is a false positive, which has been shown to be prevalent for browser certificate warnings security warnings in the past [2]. As such never *ignore warnings* may not be socially desirable for experts. On the other hand, the realism of behaviors may also be affected. If participants use a password manager to generate complex passwords and because they

do not have to remember them, they can afford to have only *secure passwords* and *different passwords* for every account. The validity of SP-SDS for other populations of interest in USP research, like software developers or other people with a computer science background, needs to be determined, but since end users are still an important demographic of interest at USP, we chose to first develop the SP-SDS for this population. When using the SP-SDS, researchers should ensure that their sample consists of end users for which the scale is validated, e.g. by asking participants about their background with respect to computer science.

The M-C SDS, which we used as the base of the SP-SDS has several limitations, such as its length, outdated wording and low reliability [6, 32]. It has also been criticized for representing social desirability as a uni-dimensional construct, measuring need for approval, while other studies on measuring social desirability have identified more dimensions, i.e. self-deceptive positivity and impression management [67]. In addition, a recent meta-analysis tested the validity of SDSs, with studies in their sample including, among others, M-C SDS and BIDR, by investigating the relationship between exhibited socially desirable behavior in economic games and socially desirable responses [50]. They did not find consistent links between behavior and responses [50]. However, since we adapted and developed new items for SP-SDS, and the factor structure differs in that our five factors represent different groups of behaviors related to IT security and privacy for which the perceived social desirability and behavior can vary, these limitations do not directly transfer to SP-SDS. Nevertheless, to further test the validity of SP-SDS, we recommend conducting a study where behaviors referenced in SP-SDS are observed and measured in an objective way, while simultaneously applying the SP-SDS to be able to compare responses to behavior directly, as has been done by Redmiles et al. [72].

9 Discussion and Recommendations for Using the SP-SDS

We developed a security and privacy specific scale to measure social desirability bias. We conducted four studies to develop our scale: two pilot studies to test for comprehensibility and suitability of our developed items, a scale development study, where we reduced the included items and identified a five-factor structure for SP-SDS and finally a validation study to check the previously found factor structure. The resulting scale demonstrated desirable psychometric properties such as high fit indices for the factor structure, acceptable reliability and convergent validity. We found that SP-SDS had a medium size correlation with general M-C SDS and smaller correlations with BIDR and a SDS from a completely different context. This shows that our scale measures the concept of social desirability, but specifically for the context of IT security and privacy. We conducted an additional calibration

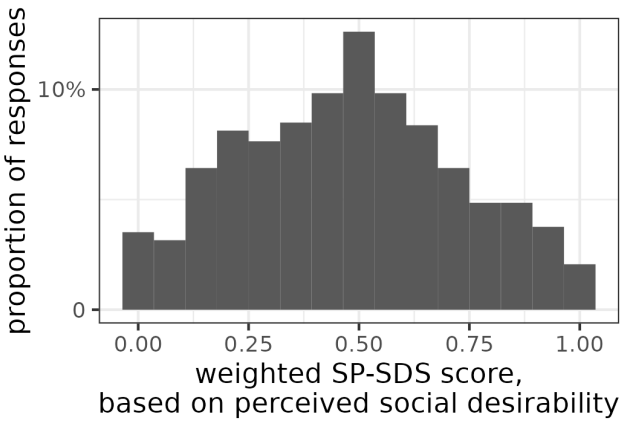

Figure 1: Distribution of the weighted scores calculated in the validation study

study and suggest to calculate weighted scores to account for different levels of social desirability in the items in our scale. High values on the SP-SDS can be a warning sign for researchers, that their data can be subject to social desirability bias.

In the following we describe how researchers using the SP-SDS, can interpret the resulting scores. To provide a baseline for interpretation, we calculated the SP-SDS scores for the participants in the validation study based on our weights and formula described above. The distribution of the weighted scores is shown in Figure 1. Researchers applying the SP-SDS in their work can either compare the distribution of their sample or the scores of individual participants to the distribution in Figure 1. A distribution that is more skewed towards 0 than in our validation study sample indicates that social desirability bias is lower than in this baseline, while a skew towards 1 indicates that social desirability bias is higher. When comparing an individual participant's score to the distribution, researchers can assess the percentile that the participant's score is in, again with higher scores suggesting higher social desirability bias. We provide percentiles for direct comparison in Table 8 in the appendix.

Our exploratory comparison of social desirability bias between different groups suggested an influence of age and ethnicity and prior work indicates that susceptibility to social desirability bias may vary based on demographic factors like gender [7, 8, 35], age [4] or education level [31]. Even though we believe these differences are not large enough to warrant changes to experimental design to compensate, we present distributions of SP-SDS scores for different demographic groups, so other researchers can decide for themselves whether they want to account for this. The distributions for different gender are in Figure 2, for three age groups in Figure 3, for the three groups of education level in Figure 4 and for different ethnicity in Figure 5. These distributions and proportions can be used for samples with specific characteristics in the same way

as the general proportions and distributions. Consistent with our exploratory regression analysis, the distribution for the higher age bracket is skewed a little more towards 1, while the distribution for the lower age bracket is skewed more towards 0 than the middle age bracket. Similarly the distributions for participants with Black and Mixed ethnicity are also skewed more towards 1 and the distribution for Asian participants more towards 0, although the smaller sample sizes make this relationship less clear.

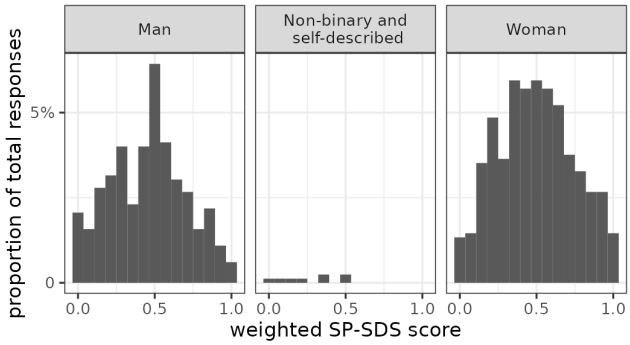

Figure 2: Distribution of the weighted scores in the validation study, grouped by gender

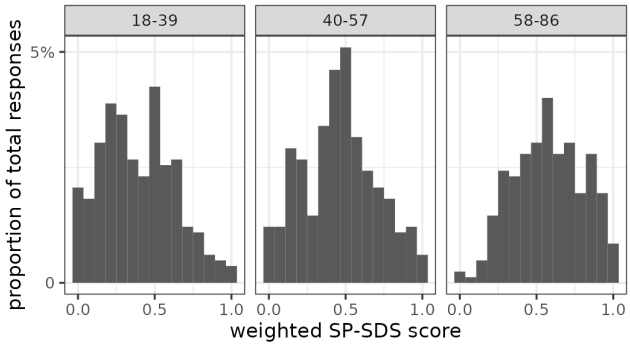

Figure 3: Distribution of the weighted scores in the validation study, grouped by age

Interpreting scores from a SDS comes with some difficulties. Even though we filtered our initial item pool to only include behaviors which we and the additional experts involved in our brainstorming session, believe to be rare, participants may actually be able to behave in such a way. Without additional behavioral measures, it is hard to tell, which is the case. However, measuring actual prevalence of behaviors, similarly to Redmiles et al.'s work on updates [72] for some of the items in the SP-SDS could help clarify this. Social desirability bias is also not the only type of response bias affecting data quality. Acquiescence bias is somewhat related to social desirability bias but refers to the general tendency to respond positively, with "yes" answers being selected more often than "no" answers [47]. Memory bias [62] means that how well users remember what they did in the past, regardless of whether

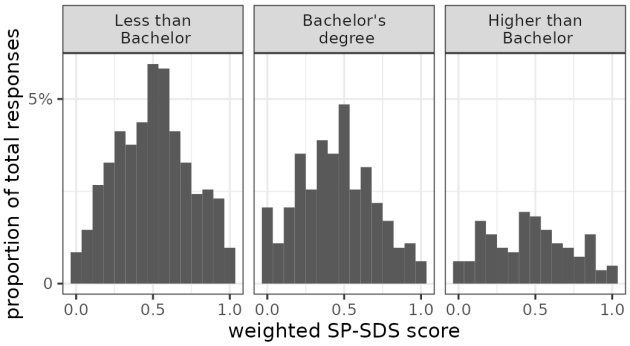

Figure 4: Distribution of the weighted scores in the validation study, grouped by education levels

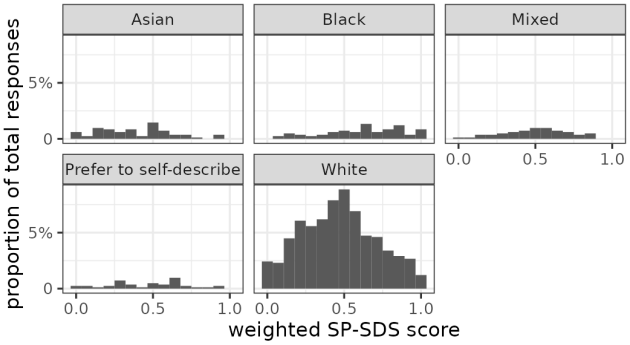

Figure 5: Distribution of the weighted scores in the validation study, grouped by ethnicity

the behavior they are reporting on is socially desirable or not, influences their responses.

10 Conclusion

When conducting human factors research on security and privacy topics that utilize self-reported data, social desirability bias is a common concern. In this paper, we develop the SP-SDS to help to assess the potential impact. The scale demonstrated high internal consistency and convergent validity. It enables USP researchers to estimate the extent of social desirability bias with a focus specifically on security and privacy.We additionally provide tools to compare the extent of social desirability bias for end-user samples without a computer science background to a US-baseline for different demographic subgroups.

References

[1] Tanisha Afnan, Yixin Zou, Maryam Mustafa, and Florian Schaub. Aunties, Strangers, and the FBI: Online Privacy Concerns and Experiences of Muslim-American Women. In *Proceedings of the Eighteenth Symposium on Usable Privacy and Security*

(SOUPS 2022, SOUPS'22, page 21, Boston, USA, 2022. USENIX Association.

[2] Devdatta Akhawe, Johanna Amann, Matthias Vallentin, and Robin Sommer. Here's my cert, so trust me, maybe? understanding TLS errors on the web. In *Proceedings of the 22nd international conference on World Wide Web*, WWW '13, pages 59–70, New York, NY, USA, May 2013. Association for Computing Machinery.

[3] Devdatta Akhawe and Adrienne Porter Felt. Alice in Warningland: A Large-Scale Field Study of Browser Security Warning Effectiveness. In *Proceedings of the 22nd USENIX Security Symposium*. USENIX Association, 2013.

[4] Liisi Ausmees, Christian Kandler, Anu Realo, Jüri Allik, Peter Borkenau, Martina Hřebíčková, and René Mõttus. Age differences in personality traits and social desirability: A multi-rater multi-sample study. *Journal of Research in Personality*, 99:104245, August 2022.

[5] Hermann Baumgartl, Philipp Roessler, Daniel Sauter, and Ricardo Buettner. Measuring Social Desirability Using a Novel Machine Learning Approach Based on EEG Data. In *PACIS 2020 Proceedings*, page 12, Dubai, UAE, June 2020.

[6] S. Natasha Beretvas, Jason L. Meyers, and Walter L. Leite. A Reliability Generalization Study of the Marlowe-Crowne Social Desirability Scale. *Educational and Psychological Measurement*, 62(4):570–589, August 2002. Publisher: SAGE Publications Inc.

[7] Richard A. Bernardi. Associations between Hofstede's Cultural Constructs and Social Desirability Response Bias. *Journal of Business Ethics*, 65(1):43–53, April 2006.

[8] Richard A. Bernardi and Steven T. Guptill. Social Desirability Response Bias, Gender, and Factors Influencing Organizational Commitment: An International Study. *Journal of Business Ethics*, 81(4):797–809, September 2008.

[9] Godfred O. Boateng, Torsten B. Neilands, Edward A. Frongillo, Hugo R. Melgar-Quiñonez, and Sera L. Young. Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. *Frontiers in Public Health*, 6:149, June 2018.

[10] Nele Borgert, Oliver D. Reithmaier, Luisa Jansen, Larina Hillemann, Ian Hussey, and Malte Elson. Home Is Where the Smart Is: Development and Validation of the Cybersecurity Self-Efficacy in Smart Homes (CySESH) Scale. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, page 15, Hamburg Germany, April 2023. ACM.

[11] Cristian Bravo-Lillo, Lorrie Faith Cranor, Julie Downs, and Saranga Komanduri. Bridging the Gap in Computer Security Warnings: A Mental Model Approach. *IEEE Security & Privacy*, 9(2):18–26, March 2011. Conference Name: IEEE Security & Privacy.

[12] Karoline Busse, Julia Schäfer, and Matthew Smith. Replication: No one can hack my mind revisiting a study on expert and non-expert security practices and advice. In *Proceedings of the fifteenth symposium on usable privacy and security (SOUPS 2019)*, SOUPS'19, page 21, Santa Clara, CA, August 2019. USENIX Association.

[13] Elisabeth Coutts and Ben 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, February 2011.

[14] Virginia Crandall, Vaughn Crandall, and Walter Katkovsky. A Children's Social Desirability Questionnaire. *Journal of consulting psychology*, 29:27–36, February 1965.

[15] Douglas P. Crowne and David Marlowe. A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 24(4):349–354, February 1960. ISBN: 0095-8891.

[16] Anastasia Danilova, Alena Naiakshina, Johanna Deuter, and Matthew Smith. Replication: On the Ecological Validity of Online Security Developer Studies: Exploring Deception in a {Password-Storage} Study with Freelancers. In *Proceedings of the Sixteenth Symposium on Usable Privacy and Security*, SOUPS'20, pages 165–183. USENIX Association, 2020.

[17] Darren W. Davis and Brian D. Silver. Stereotype Threat and Race of Interviewer Effects in a Survey on Political Knowledge. *American Journal of Political Science*, 47(1):33–45, 2003. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1540-5907.00003.

[18] Elmira Deldari, Parth Thakkar, and Yaxing Yao. Users' Perceptions of Online Child Abuse Detection Mechanisms. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):147:1–147:26, April 2024.

[19] Nicola Dell, Vidya Vaidyanathan, Indrani Medhi, Edward Cutrell, and William Thies. "Yours is better!": participant response bias in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing*

Systems, CHI'12, pages 1321–1330, Austin, USA, May 2012. ACM.

[20] D. Dodou and J. C. F. de Winter. Social desirability is the same in offline, online, and paper surveys: A meta-analysis. *Computers in Human Behavior*, 36:487–495, July 2014.

[21] ECPAT. What do EU Citizens think of the balance between online privacy and child protection? Polling Research, ECPAT, 2021.

[22] Alan Ewert and Graeme Galloway. Socially desirable responding in an environmental context: Development of a domain specific scale. *Environmental Education Research - ENVIRON EDUC RES*, 15:55–70, February 2009.

[23] Sascha Fahl, Marian Harbach, Yasemin Acar, and Matthew Smith. On the ecological validity of a password study. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, page 13, Newcastle United Kingdom, July 2013. ACM.

[24] Habiba Farzand, Karola Marky, and Mohamed Khamis. Out-of-Device Privacy Unveiled: Designing and Validating the Out-of-Device Privacy Scale (ODPS). In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, page 15, Honolulu HI USA, May 2024. ACM.

[25] Andy Field, Jeremy Miles, and Zoe Field. *Discovering Statistics Using R*. SAGE Publications Ltd, London ; Thousand Oaks, Calif, 1. edition edition, April 2012.

[26] Robert J. Fisher. Social Desirability Bias and the Validity of Indirect Questioning. *Journal of Consumer Research*, 20(2):303–315, September 1993.

[27] Kevin Gallagher, Sameer Patil, and Nasir Memon. New Me: Understanding Expert and Non-Expert Perceptions and Usage of the Tor Anonymity Network. In *Proceedings of the Thirteenth Symposium on Usable Privacy and Security*, SOUPS'17, Santa Clara, USA, 2017. USENIX Association.

[28] Lisa Geierhaas, Florin Martius, Arthi Arumugam, and Matthew Smith. "Not the Right Question?" A Study on Attitudes Toward Client-Side Scanning with Security and Privacy Researchers and a U.S. Population Sample. In *2025 IEEE Symposium on Security and Privacy (SP)*, SP'25, pages 86–86, San Francisco, USA, 2025. IEEE Computer Society. ISSN: 2375-1207.

[29] Lisa Geierhaas, Fabian Otto, Maximilian Häring, and Matthew Smith. Attitudes towards Client-Side Scanning for CSAM, Terrorism, Drug Trafficking, Drug Use and Tax Evasion in Germany. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 217–233, San Francisco, USA, May 2023. IEEE Computer Society. ISSN: 2375-1207.

[30] Marton Gergely and V. Srinivasan Rao. Social desirability bias in software piracy research: Evidence from pilot studies. In *2016 12th International Conference on Innovations in Information Technology (IIT)*, pages 1–4, November 2016.

[31] Katja Haberecht, Inga Schnuerer, Beate Gaertner, Ulrich John, and Jennis Freyer-Adam. The Stability of Social Desirability: A Latent Change Analysis. *Journal of Personality*, 83(4):404–412, 2015. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jopy.12112.

[32] Claire M. Hart, Timothy D. Ritchie, Erica G. Hepper, and Jochen E. Gebauer. The Balanced Inventory of Desirable Responding Short Form (BIDR-16). *Sage Open*, 5(4):2158244015621113, October 2015. Publisher: SAGE Publications.

[33] Rakibul Hasan, Rebecca Weil, Rudolf Siegel, and Katharina Krombholz. A Psychometric Scale to Measure Individuals' Value of Other People's Privacy (VOPP). In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, page 14, New York, NY, USA, April 2023. Association for Computing Machinery.

[34] Ayako A. Hasegawa, Daisuke Inoue, and Mitsuaki Akiyama. How WEIRD is Usable Privacy and Security Research? In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 3241–3258, Philadelphia, USA, 2024. USENIX Association.

[35] James R. Hebert, Yunsheng Ma, Lynn Clemow, Ira S. Ockene, Gordon Saperia, Edward J. Stanek, III, Philip A. Merriam, and Judith K. Ockene. Gender Differences in Social Desirability and Social Approval Bias in Dietary Self-report. *American Journal of Epidemiology*, 146(12):1046–1055, December 1997.

[36] Daire Hooper, Joseph Coughlan, and Michael Mullen. Structural Equation Modeling: Guidelines for Determining Model Fit. *The Electronic Journal of Business Research Methods*, 6, November 2007.

[37] Iulia Ion, Rob Reeder, and Sunny Consolvo. "...No one can hack my mind": Comparing expert and non-expert security practices. In *Proceedings of the Eleventh symposium on usable privacy and security (SOUPS 2015)*, pages 327–346, Ottawa, July 2015. USENIX Association.

[38] Lone Jespersen, Tanya MacLaurin, and Peter Vlerick. Development and validation of a scale to capture social desirability in food safety culture. *Food Control*, 82:42–47, December 2017.

[39] Alex Kale, Matthew Kay, and Jessica Hullman. Decision-Making Under Uncertainty in Research Synthesis: Designing for the Garden of Forking Paths. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI'19, page 14, Glasgow Scotland Uk, May 2019. ACM.

[40] Ruogu Kang, Stephanie Brown, Laura Dabbish, and Sara Kiesler. Privacy Attitudes of Mechanical Turk Workers and the {U.S}. Public. In *10th Symposium On Usable Privacy and Security (SOUPS 2014)*, SOUPS'14, pages 37–49, Menlo Park, USA, 2014. USENIX Association.

[41] Harjot Kaur, Sabrina Klivan, Daniel Votipka, Yasemin Acar, and Sascha Fahl. Where to Recruit for Security Development Studies: Comparing Six Software Developer Samples. In *Proceedings of the 31st USENIX Security Symposium*, pages 4041–4058, Boston, USA, 2022. USENIX Association.

[42] Seung Hyun Kim and Sangmook Kim. National Culture and Social Desirability Bias in Measuring Public Service Motivation. *Administration & Society*, 48(4):444–476, May 2016. Publisher: SAGE Publications Inc.

[43] Maryon F. King and Gordon C. Bruner. Social desirability bias: A neglected aspect of validity testing. *Psychology & Marketing*, 17(2):79–103, 2000.

[44] Frauke Kreuter, Stanley Presser, and Roger Tourangeau. Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity. *Public Opinion Quarterly*, 72(5):847–865, December 2008.

[45] Kat Krol, Jonathan M. Spring, Simon Parkin, and M. Angela Sasse. Towards Robust Experimental Design for User Studies in Security and Privacy. In *The LASER Workshop: Learning from Authoritative Security Experiment Results (LASER 2016)*, pages 21–31, 2016.

[46] Katharina Krombholz, Karoline Busse, Katharina Pfeffer, Matthew Smith, and Emanuel von Zezschwitz. "If HTTPS Were Secure, I Wouldn't Need 2FA" - End User and Administrator Mental Models of HTTPS. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 246–263, San Francisco, CA, USA, May 2019. IEEE.

[47] Ozan Kuru and Josh Pasek. Improving social media measurement in surveys: Avoiding acquiescence bias in Facebook research. *Computers in Human Behavior*, 57:82–92, 2016.

[48] Dong-Heon Kwak, Philipp Holtkamp, and Sung S. Kim. Measuring and Controlling Social Desirability Bias: Applications in Information Systems Research. *Journal of the Association for Information Systems*, 20(4):317–345, 2019.

[49] Dong-Heon (Austin) Kwak, Xiao Ma, and Sumin Kim. When does social desirability become a problem? Detection and reduction of social desirability bias in information systems research. *Information & Management*, 58(7):103500, November 2021.

[50] Lukas Lanz, Isabel Thielmann, and Fabiola H. Gerpott. Are social desirability scales desirable? A meta-analytic test of the validity of social desirability scales in the context of prosocial behavior. *Journal of Personality*, 90(2):203–221, April 2022.

[51] Mingnan Liu and Yichen Wang. Race-of-Interviewer Effect in the Computer-Assisted Self-Interview Module in a Face-to-Face Survey. *International Journal of Public Opinion Research*, 28(2):292–305, June 2016.

[52] Yang Liu, Tim Althoff, and Jeffrey Heer. Paths Explored, Paths Omitted, Paths Obscured: Decision Points & Selective Reporting in End-to-End Data Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, page 14, Honolulu, USA, April 2020. ACM.

[53] Idowu Longe and Aneshkumar Maharaj. Investigating Students' Understanding of Complex Number and Its Relation to Algebraic Group Using and APOS Theory. *Journal of Medives : Journal of Mathematics Education IKIP Veteran Semarang*, 7:117, January 2023.

[54] Kenneth Manning, William Bearden, and Kelly Tian. Development and validation of the Agents' Socially Desirable Responding (ASDR) scale. *Marketing Letters*, 20:31–44, March 2009.

[55] Tara Matthews, Kathleen O'Leary, Anna Turner, Manya Sleeper, Jill Palzkill Woelfer, Martin Shelton, Cori Manthorne, Elizabeth F. Churchill, and Sunny Consolvo. Stories from Survivors: Privacy & Security Practices when Coping with Intimate Partner Abuse. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2189–2201, Denver Colorado USA, May 2017. Association for Computing Machinery.

[56] Jim McCambridge, John Witton, and Diana R. Elbourne. Systematic review of the Hawthorne effect: New concepts are needed to study research participation effects. *Journal of Clinical Epidemiology*, 67(3):267–277, March 2014.

[57] Andreas Möller, Florian Michahelles, Stefan Diewald, Luis Roalter, and Matthias Kranz. Update Behavior in App Markets and Security Implications: A Case Study in Google Play. In *Research in the LARGE : Proceedings of the 3rd International Workshop. Held in Conjunction with Mobile HCI*, pages 3–6, September 2012.

[58] Alena Naiakshina, Anastasia Danilova, Eva Gerlitz, and Matthew Smith. On Conducting Security Developer Studies with CS Students: Examining a Password-Storage Study with CS Students, Freelancers, and Company Developers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, page 13, Honolulu HI USA, April 2020. ACM.

[59] Alena Naiakshina, Anastasia Danilova, Eva Gerlitz, Emanuel von Zezschwitz, and Matthew Smith. "If you want, I can store the encrypted password": A Password-Storage Field Study with Freelance Developers. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–12, Glasgow, Scotland, Uk, May 2019. ACM.

[60] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, Marco Herzog, Sergej Dechand, and Matthew Smith. Why Do Developers Get Password Storage Wrong? A Qualitative Usability Study. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, pages 311–328, New York, NY, USA, October 2017. Association for Computing Machinery.

[61] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, and Matthew Smith. Deception Task Design in Developer Password Studies: Exploring a Student Sample. In *Proceedings of the Fourteenth Symposium on Usable Privacy and Security*, SOUPS'18, pages 297–313, Baltimore, USA, 2018. USENIX Association.

[62] Pawarat Nontasil and Stephen J. Payne. Emotional Utility and Recall of the Facebook News Feed. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 9, Glasgow Scotland Uk, May 2019. ACM.

[63] Simon Ntumi, Sheilla Agbenyo, and Tapela Bulala. Estimating the Psychometric Properties (Item Difficulty, Discrimination and Reliability Indices) of Test Items using Kuder-Richardson Approach (KR-20). *Shanlax International Journal of Education*, 11(3):18–28, June 2023.

[64] Jakob Ohme, Theo Araujo, Claes H. de Vreese, and Jessica Taylor Piotrowski. Mobile data donations: Assessing self-report accuracy and sample biases with the iOS Screen Time function. *Mobile Media & Communication*, 9(2):293–313, May 2021.

[65] Christopher J. Pannucci and Edwin G. Wilkins. Identifying and Avoiding Bias in Research:. *Plastic and Reconstructive Surgery*, 126(2):619–625, August 2010.

[66] Delroy Paulhus. Two-Component Models of Socially Desirable Responding. *Journal of Personality and Social Psychology*, 46:598–609, 1984.

[67] Delroy L. Paulhus. Measurement and Control of Response Bias. *Measurement of Personality and Social Psychological Attitudes*, 1, 1991.

[68] Enrico Perinelli and Paola Gremigni. Use of Social Desirability Scales in Clinical Psychology: A Systematic Review. *Journal of Clinical Psychology*, 72(6):534–551, 2016.

[69] Mikael Persson and Maria Solevid. Measuring Political Participation—Testing Social Desirability Bias in a Web-Survey Experiment. *International Journal of Public Opinion Research*, 26(1):98–112, 2014.

[70] Prashanth Rajivan, Efrat Aharonov-Majar, and Cleotilde Gonzalez. Update now or later? Effects of experience, cost, and risk preference on update decisions. *Journal of Cybersecurity*, 6(1):tyaa002, January 2020.

[71] Elissa M. Redmiles, Sean Kross, and Michelle L. Mazurek. How Well Do My Results Generalize? Comparing Security and Privacy Survey Results from MTurk, Web, and Telephone Samples. In *2019 IEEE Symposium on Security and Privacy (SP)*, SP'19, pages 1326–1343, May 2019. ISSN: 2375-1207.

[72] Elissa M. Redmiles, Ziyun Zhu, Sean Kross, Dhruv Kuchhal, Tudor Dumitras, and Michelle L. Mazurek. Asking for a Friend: Evaluating Response Biases in Security User Studies. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 1238–1255, Toronto Canada, October 2018. ACM.

[73] Robert W. Reeder, Adrienne Porter Felt, Sunny Consolvo, Nathan Malkin, Christopher Thompson, and Serge Egelman. An Experience Sampling Study of User Reactions to Browser Warnings in the Field. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–13, Montreal QC Canada, April 2018. ACM.

[74] William Reynolds. Development of reliable and valid short forms of the marlow–crowne social desirability scale. *Journal of Clinical Psychology*, 38:119–125, 01 1982.

[75] Simone Romano, Davide Fucci, Giuseppe Scanniello, Maria Teresa Baldassarre, Burak Turhan, and Natalia Juristo. On researcher bias in Software Engineering experiments. *Journal of Systems and Software*, 182:111068, December 2021.

[76] Scott Ruoti, Mark O'Neill, Daniel Zappala, and Kent Seamons. User attitudes toward the inspection of encrypted traffic. In *Proceedings of the Twelfth Symposium on Usable Privacy and Security*, SOUPS'16, pages 131–146, Denver, USA, June 2016. USENIX Association.

[77] Harold A. Sackeim and Ruben C. Gur. Self-Deception, Self-Confrontation, and Consciousness. In Gary E. Schwartz and David Shapiro, editors, *Consciousness and Self-Regulation: Advances in Research and Theory Volume 2*, pages 139–197. Springer US, Boston, MA, 1978.

[78] Sena Sahin, Suood Al Roomi, Tara Poteat, and Frank Li. Investigating the Password Policy Practices of Website Administrators. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 552–569, San Francisco, USA, May 2023. ISSN: 2375-1207.

[79] Anne M. Scheel, Mitchell R. M. J. Schijen, and Daniël Lakens. An Excess of Positive Results: Comparing the Standard Psychology Literature With Registered Reports. *Advances in Methods and Practices in Psychological Science*, 4(2):25152459211007467, April 2021. Publisher: SAGE Publications Inc.

[80] Roberta W. Scherer, Joerg J. Meerpohl, Nadine Pfeifer, Christine Schmucker, Guido Schwarzer, and Erik von Elm. Full publication of results initially presented in abstracts. *Cochrane Database of Systematic Reviews*, (11), 2018. Publisher: John Wiley & Sons, Ltd.

[81] Matthias Schmidmaier, Jonathan Rupp, Darina Cvetanova, and Sven Mayer. Perceived Empathy of Technology Scale (PETS): Measuring Empathy of Systems Toward the User. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, page 18, Honolulu HI USA, May 2024. ACM.

[82] Camelia Simoiu, Joseph Bonneau, Christopher Gates, and Sharad Goel. "I was told to buy a software or lose my computer. I ignored it": A study of ransomware. In *Fifteenth symposium on usable privacy and security (SOUPS 2019)*, pages 155–174, Santa Clara, CA, August 2019. USENIX Association.

[83] Ana-Maria Simundic. Bias in research. *Biochemia Medica*, 23(1):12–15, February 2013. Publisher: Hrvatsko društvo za medicinsku biokemiju i laboratorijsku medicinu.

[84] Joanna Smith and Helen Noble. Bias in research. *Evidence-Based Nursing*, 17(4):100–101, October 2014. Publisher: Royal College of Nursing Section: Research made simple.

[85] Andreas Sotirakopoulos, Kirstie Hawkey, and Konstantin Beznosov. "I did it because I trusted you": Challenges with the Study Environment Biasing Participant Behaviours. In *SOUPS Usable Security Experiment Reports (USER) Workshop*, SOUPS'10, Redmond, USA, 2010. USENIX Association.

[86] Mythily Srinivasan. Psychometric Characteristics of Oral Pathology Test Items in the Dental Hygiene Curriculum—A Longitudinal Analysis. *Dentistry Journal*, 9:56, May 2021.

[87] Robert Strahan and Kathleen Carrese Gerbasi. Short, homogeneous versions of the Marlow-Crowne Social Desirability Scale. *Journal of Clinical Psychology*, 28(2):191–193, 1972. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/1097-4679%28197204%2928%3A2%3C191%3A%3AAID-JCLP2270280220%3E3.0.CO%3B2-G.

[88] Joachim Stöber. Die Soziale-Erwünschtheits-Skala-17 (SES-17): Entwicklung und erste Befunde zu Reliabilität und Validität. *Diagnostica*, 45(4):173–177, 1999. ISBN: 0012-1924.

[89] Poorna Talkad Sukumar and Ronald Metoyer. Towards Designing Unbiased Replication Studies in Information Visualization. In *2018 IEEE Evaluation and Beyond - Methodological Approaches for Visualization (BELIV)*, pages 93–101, Berlin, Germany, October 2018. IEEE.

[90] Joshua Sunshine, Serge Egelman, Hazim Almuhimedi, Neha Atri, and Lorrie Faith Cranor. Crying Wolf: An Empirical Study of SSL Warning Effectiveness. In *18th USENIX Security Symposium (USENIX Security 09)*, page 34. USENIX Association, 2009.

[91] Mohammad Tahaei and Kami Vaniea. A Survey on Developer-Centred Security. In *2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pages 129–138, June 2019.

[92] Mohammad Tahaei and Kami Vaniea. Recruiting Participants With Programming Skills: A Comparison of Four Crowdsourcing Platforms and a CS Student Mailing List. In *Proceedings of the 2022 CHI Conference*

on Human Factors in Computing Systems, CHI '22, pages 1–15, New York, NY, USA, April 2022. Association for Computing Machinery.

[93] Mahzabin Tamanna, Joseph D. Stephens, and Mohd Anwar. Security Implications of User Non-compliance Behavior to Software Updates: A Risk Assessment Study, November 2024. arXiv:2411.06262 [cs].

[94] Jenny Tang, Eleanor Birrell, and Ada Lerner. Replication: How Well Do My Results Generalize Now? The External Validity of Online Privacy and Security Surveys. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, SOUPS'22, pages 367–385, Boston, USA, 2022. USENIX Association.

[95] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, Damon McCoy, Sarah Meiklejohn, Thomas Ristenpart, and Gianluca Stringhini. SoK: Hate, Harassment, and the Changing Landscape of Online Abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*, SP'21, pages 247–267, San Francisco, CA, USA, May 2021. IEEE.

[96] Christian Tiefenau, Maximilian Häring, Katharina Krombholz, and Emanuel von Zezschwitz. Security, Availability, and Multiple Information Sources: Exploring Update Behavior of System Administrators. In *Proceedings of the Sixteenth Symposium on Usable Privacy and Security*, SOUPS'20, pages 239–258. USENIX Association, 2020.

[97] Roger Tourangeau and Tom W. Smith. Asking Sensitive Questions: The Impact of Data Collection Mode, Question Format, and Question Context. *Public Opinion Quarterly*, 60(2):275, 1996.

[98] Roger Tourangeau and Ting Yan. Sensitive questions in surveys. *Psychological Bulletin*, 133(5):859–883, 2007. Place: US Publisher: American Psychological Association.

[99] Anthony Vance, David Eargle, Jeffrey L. Jenkins, C. Brock Kirwan, and Bonnie Brinton Anderson. The Fog of Warnings: How Non-essential Notifications Blur with Security Warnings. In *Proceedings of the Fifteenth Symposium on Usable Privacy and Security*, pages 407–420, 2019.

[100] Kami Vaniea, Emilee Rader, and Rick Wash. Betrayed by updates: how negative experiences affect future security. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI'14, pages 2671–2674, Toronto Ontario Canada, April 2014. ACM.

[101] Daniel Votipka, Desiree Abrokwa, and Michelle L. Mazurek. Building and Validating a Scale for Secure Software Development Self-Efficacy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–20, New York, NY, USA, April 2020. Association for Computing Machinery.

[102] Leijie Wang, Ruotong Wang, Sterling Williams-Ceci, and Amy X Zhang. "Is Reporting Worth the Sacrifice of Revealing What I've Sent?": Privacy Considerations When Reporting on End-to-End Encrypted Platforms. In *Proceedings of the Nineteenth Symposium on Usable Privacy and Security*, SOUPS'23, Anaheim, USA, 2023. USENIX Association.

[103] Senuri Wijenayake, Niels van Berkel, and Jorge Goncalves. Bots for Research: Minimising the Experimenter Effect. In *Proceedings of the CHI 2020 Workshop on Detection and Design for Cognitive Biases in People and Computing Systems*, page 9, Honolulu, USA, 2020. ACM.

[104] Sandra R. Wilson, Nancy L. Brown, Carolina Mejia, and Philip W. Lavori. Effects of Interviewer Characteristics on Reported Sexual Behavior of California Latino Couples. *Hispanic Journal of Behavioral Sciences*, 24(1):38–62, February 2002.

A Items

Items of the SP-SDS from the first pilot survey to the final scale. The items marked with x in the SV column form the final scale.

abbreviation	t/f	Statements	SD	SV	CA
AI forbidden	f	I use AI for my work, even if it's not allowed	x		
annoyed	t	I am never annoyed if people ask me to help them with their computers			
back up	t	I back up my data every day			
check AI	t	When I use an AI, I always double-check if the generated results are correct	x		
check HTTPS	t	I always check whether the website uses HTTPS[1] [1] protocol with which data can be encrypted and thus protected from eavesdropping, URL begins with https if it's used			
check backup	t	I check that all my backups work			
check leaks	t	I regularly check whether my passwords and e-mail addresses have been found in data leaks	x		
check program	t	I always check if the program is trustworthy before installing software			
clicked link	t	I have never clicked a link in a mail without checking if it's safe			
data collected	t	I am always aware of what personal data is collected by the platforms I use and how it is used	x	x	x
different passwords	t	I use different passwords for all my accounts	x	x	x
disclose AI	t	I always disclose when I have used AI			
encrypt mail	t	I always encrypt[1] e-mails containing personal data [1] convert information in a way so that only			
ignore update	t	I never ignore software update reminders	x	x	x
ignore warnings	t	I never ignore security warnings on my computer	x	x	x
illegal movies	f	If I knew I wouldn't get caught, I would watch movies illegally	x	x	x
install updates	t	I always install non-automatic (security) updates right away			
laughed	t	I have never laughed about somebody's lack of knowledge about computers			
lock device	t	I always lock all my devices if someone else could access them			
log out	t	I always log out of my accounts after using someone else's device			
looked screen	t	I have never looked at someone else's screen without their permission			
mail attachment	t	I have never opened a mail attachment without checking if it's safe	x		
personal password	t	I have never used a password that contains personal data (e.g. birthday, pet name)			
pirated software	f	If I knew I wouldn't get caught, I would use pirated software[1] [1] use of software that is not properly licensed	x	x	x
policy access	t	I always read the privacy policy before giving an app/application access to my personal data	x	x	x
polite online	t	I am always just as polite online as I am in the real world	x	x	x
random passwords	t	I only use randomly generated passwords			
read messages	t	I have never read somebody else's messages without explicit permission	x		
read policy	t	I always read the privacy policy completely before I agree to them	x	x	x
read terms	t	I always read the terms and conditions completely before I agree to them	x	x	x
required cookies	t	I always only agree to required cookies			
reuse passwords	t	I never reuse passwords	x	x	x
secure passwords	t	I only use secure passwords (passwords that are long and complex)	x	x	x
share passwords	t	I have never in my life given my password to another person			
smashing computer	f	There have been occasions when I felt like smashing a computer	x		
troll comment	t	I have never considered posting a troll-comment[1] [1] comments that deliberately try to provoke an argument, conflict or emotional reaction	x	x	x
turnoff location	t	I always turn off location services when I don't need them	x		
two fa	t	I use two-factor authentication (e.g. password and code from authentication app)			
write password	t	I have never written down my password (except within password managers or other encrypted files)			

Table 7: Pilotitems with codings true(t)/false(f) and changes from pilot in (blue). Items used for Scale Development (SD), Scale Validation (SV) and Calibration (CA) are marked with x

B Response distribution percentiles

Percentiles for direct comparison:

Quantile	Scores
0%	0
5%	0.08
10%	0.16
15%	0.23
20%	0.25
25%	0.31
30%	0.32
35%	0.38
40%	0.40
45%	0.41
50%	0.47
55%	0.48
60%	0.54
65%	0.56
70%	0.61
75%	0.63
80%	0.70
85%	0.77
90%	0.84
95%	0.92
100%	1

Table 8: Percentiles of weighted SP-SDS scores in the validation study