On recruiting and retaining users for security-sensitive longitudinal measurement panels

Akira Yamada[1,3*], Kyle Crichton[2*], Yukiko Sawaya[1], Jin-Dong Dong[2],
Sarah Pearman[2], Ayumu Kubota[1], and Nicolas Christin[2]
[1] *KDDI Research, Inc.*
[2] *Carnegie Mellon University*
[3] *National Institute of Information and Communications Technology*

Abstract

Many recent studies have turned to longitudinal measurement panels to characterize how people use their computing devices under realistic conditions. In these studies, participants' devices are instrumented, and their behavior is closely monitored over long time intervals. Because such monitoring can be highly intrusive, researchers face substantial challenges recruiting and retaining participants.

We present three case studies using medium- to large-scale longitudinal panels, which all collect privacy- and security-sensitive data. In evaluating factors related to recruitment, retention, and data collection, we provide a foundation to inform the design of future long-term panel studies.

Through these studies, we observe that monetary and non-monetary incentives can be effective in recruiting panel participants, although each presents trade-offs and potential biases. Contrary to our initial expectations, we find that users do not behave any differently in their first few weeks of participation than in the remainder of their time in the study. In terms of retention, we note that personalized enrollment follow-ups can lower initial dropout rates, but they are challenging and costly to scale. Communication, including following up with inactive users, is vital to retention. However, finding the right balance of communication is equally important. Interfering with a participant's everyday device use is a sure way to lose users. Finally, we present several findings, based on practical experience, to help inform the design of the data collection process in observational panels.

1 Introduction

Many recent studies have attempted to characterize how people use their computing devices under realistic conditions. Because of the limitations of user surveys and lab experiments, researchers have increasingly turned to longitudinal measurement panels, in which participant devices are instrumented, and their behavior extensively monitored over long time intervals [7, 14, 15, 20, 26, 29, 31, 33, 49, 57]. While these panels provide rich insights into real-world user behavior, they are difficult to conduct due to technical complexity, cost, and logistical challenges. As such, longitudinal panels remain relatively rare in the field despite the advantages they afford.

Central to the problem researchers face is the highly intrusive nature of longitudinal measurement studies. As users increasingly rely on computing devices—in particular smartphones—for all aspects of their life, measurements of device use become more and more privacy-invasive. This requires special attention be paid to data collection and storage security, further complicating cost and logistics. Equally important is that the privacy and security risks be properly communicated to potential participants. However, in presenting this information users may understandably be reluctant to participate. This leads to the fundamental challenge for researchers in conducting security-sensitive longitudinal measurement panels: recruitment and retention.

To better understand these challenges we present three case studies of recent large-scale longitudinal panels, featuring approximately 2 million, 2,000, and 600 users, respectively, and running for periods ranging from two to over four years. These studies were conducted in diverse geographical (Japan and the United States) and computing (personal computers and mobile devices) environments, using very different recruitment and retention techniques. For instance, one study used monetary incentives to recruit users, while another adopted a popular animation character; and the third study provided additional security functionality—in the form of an anti-phishing toolbar. Likewise, one of the studies features frequent interactions between the research team and the participants, while others only rely on minimal communication.

We aim to synthesize recommendations for recruiting and retaining participants in future privacy-intrusive panel studies. We selected these three studies because we were collectively

[*] Both authors contributed equally.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2022.
August 7–9, 2022, Boston, MA, United States.

involved in various aspects of the design, conduct, and analysis of the research. Thus, we had direct access to the data, participants, and other researchers involved in each project. Our goal is not to provide a meta-analysis, but to assess recruitment and retention issues, based on (usually publicly unavailable) retention data and first-hand accounts. While our findings can apply to a broader set of studies relying on longitudinal panels, such as clinical health studies, we focus on security-sensitive panels where data collected are privacy-invasive and used to study security and privacy behavior.

We acknowledge that the differences between studies make direct, quantitative comparisons difficult, as does the relatively limited number of the panels considered. However, given the rarity of large-scale longitudinal measurement panels, we believe that there is great value in drawing what lessons can be learned from the few studies available. Acknowledging the aforementioned limitations, we employ a case study approach to qualitatively assess the three panel studies, supporting observations and findings with an appropriate level of quantitative evidence. We use a combination of measurements, research logs, surveys, and practical experience to compile a set of lessons learned regarding recruitment, retention, and data collection in long-term observational panels.

Overall, we find that both monetary and non-monetary incentives are effective in recruiting participants, although each may introduce its own potential bias. Contrary to our expectations, newly recruited users do not behave differently in their first few weeks than they do later on. As for participant retention, personalized enrollment and follow-ups can lower initial dropout rates, but are challenging and costly to scale. Communication, including following up with inactive users, is vital to retention, but finding the right balance of communication is equally important. Interfering with a participant's everyday device use is a sure way to lose users. Finally, we highlight the importance of monitoring for sensor outages and user dropouts, maintaining the order of observed events, establishing good measures for active user engagement, and handling multi-user devices and multi-device users.

2 Related work

We next discuss related studies by grouping them into three sets: recent user behavior measurement panels, work on participant retention in longitudinal studies, and inquiries in recruitment, motivation, and bias.

2.1 Measurement panels

Panels of personal computer users have been recruited to study a variety of behaviors related to human-computer interaction. These studies, which instrument the participant's computer with sensors, enable researchers to observe detailed information about the user's behavior over long periods of time. One major area of research using these panels has been to study how users browse the internet and how that behavior changes over time [7,29,33,49,57].

In addition, numerous studies have used longitudinal panels to examine certain user security and privacy behaviors (e.g., password creation [35] or private browsing use [17]). Other work has examined behavior leading up, and in response, to encountering security threats such as cross-site scripting attacks and related scams [34] or drive-by-downloads [27,28]. Some research has leveraged user behavior gleaned from these panels to predict exposure risk to malicious content [6,25,26,42]. Besides characterizing user responses, several studies have used longitudinal panels to examine how users maintain their machines [38] and how accurately users perceive their own maintenance and security behavior [15,51].

With users spending an increasing amount of time on their smartphones and tablets, researchers have recently taken to collecting data on mobile device use. Several early smartphone panels were created to enable researchers to deploy experiments related to smartphone use [20,31]. These panels were used to compare a user's security intention to their actual behavior [8] and to develop a measure of users' information security awareness [4]. Other recent smartphone panels include investigations of smartphone lock use [50], and of how users evaluate requested permissions [53].

2.2 Recruitment motivations and bias

Previous work on recruitment incentives—predominantly focused on survey studies—has demonstrated that offering monetary incentives to participants improves recruitment rates and decreases non-response rates [23,44,46,58]. Specific reward methods, such as lotteries, attract participants with psychologically-biased personalities and are highly effective in certain tasks [18]. Prior research on the use of non-monetary rewards suggests a similar, yet possibly weaker, effect [3,58]. Alternatively, in volunteer-based platforms [1,2,37], the participants' motivation types highly affect attentions and dropouts [21]. However, relatively few studies have compared the effects of various recruitment incentives on sample composition or the quality of data collected [46]. What evidence exists suggests that monetary and non-monetary rewards do not equally appeal to all participants [58]. As a result, the use of different incentives can result in under- or over-representation of various demographic groups, especially related to education and income level [36,40,45]. Yet, previous studies have shown that incentives generally have no statistically significant effect on question non-response [43,55].

2.3 Retention in longitudinal studies

Researchers conducting a measurement panel study must also retain user participation throughout a (often long) study. Maintaining contact with participants, recontacting participants who do not respond or show up, and using incentives have been found to be key factors in user retention [52]. In their systematic review of 88 clinical studies, Robinson et al. identified 985 retention strategies and found a positive correlation between the number employed and retention rate. However, most clinical studies examined were descriptive, with only six of them designed to directly compare between strategies [39]. Of these studies, three found that cash payments and higher compensation led to higher retention [11,12,54], two reported higher retention rates for participants who received more contact and reminders from the research team [10, 13], and one found that small non-monetary rewards had no effect [5].

3 Methods

We next give an overview of the three measurement panel studies used in our analysis: the Security Behavior Observatory (SBO, [14, 15, 17, 35]), a Security Toolbar's trace data, and a Mobile Security Behavior Observatory (mSBO, [56]). We close with a discussion of the ethical review process and copyright licensing.

3.1 Security Behavior Observatory

The SBO was a longitudinal study of home computer use conducted between May 2015 and July 2019. As a part of the study, participants consented to have their home computers instrumented with a variety of sensors that collected, encrypted, and then transmitted data back to a central repository, in exchange for monthly payments. The study was limited to Windows desktop and laptop computers that were primarily used at home. The study received Institutional Review Board approval from Carnegie Mellon University.

Recruitment Over four years, the SBO project recruited a total of 623 participants who on average stayed in the study for just under two years ($\mu = 1.76, \sigma = 1.05$). Participants were predominantly recruited from one major U.S. metropolitan area, using a university research recruitment service as the primary recruitment source along with several secondary sources. Participants completed a pre-enrollment survey to confirm eligibility and provide consent, after which they received a phone call from a research team member to step them through the enrollment process in which consent was reconfirmed audibly. Individuals received $30 upon enrollment and $10 for each month they stayed in the study. If a participant encountered technical issues or data stopped being sent for an extended period of time, a member of the SBO research team would directly contact the participant via phone or email. Participants could discontinue their participation at any time.

Data collection The SBO was designed using a client-server architecture with several client-side sensors to collect different data types from participants' machines. Information including the state of the user's machine, installed software, current processes, user interactions, and web browsing were sent whenever the participant's computer was powered on. We refer to Forget et al. [14] for a thorough discussion of the SBO architecture. Participants who reported issues with the sensors interfering with their daily use received a lightweight version of the sensor that only collected browsing data.

Upon completion of the study, participants were asked to complete an exit survey, described in Appendix B. The survey was distributed to the SBO email list to participants who had been in the study at any point. The survey was run on the Qualtrics online survey platform, where 203 responses were recorded. Those who completed the survey received a $15 Amazon gift card as additional compensation.

3.2 Security Toolbar trace data

The second panel we look at is derived from data provided by a Japanese security company. This company offers a security tool to its customers which, as a part of its service, and with explicit customer agreement, collects web browsing information from the customer device.[1] This dataset contains more than four years of browsing data, ranging from December 2016 to February 2021. The data is limited to Microsoft Windows Internet Explorer (IE) users. However, this is less of a limitation than it may seem, as many Japanese administrations and businesses required IE until recently [30].

Recruitment The Security Toolbar is used as part of a specific type of web service used primarily in Japan. The web service partners distribute the toolbar on behalf of the security company as part of their services' security enhancement. Users can use the toolbar as long as they continue to subscribe to the web service and have the toolbar installed on their device. The data we have access to features over 2 million participants, with between 50,000–300,000 daily active users. Since Microsoft stopped IE support, the number of installations has declined over time. Prior to downloading the software, users are provided information about the data collected through the security tool, and are prompted to provide consent to continue. We obtained this data under a research agreement with the security company and the sharing of the data was approved by the Institutional Review Board at Carnegie Mellon University.

Data collection Data collection has been ongoing since December 2016. The collection software is installed as an add-on to the IE browser and sends encrypted data back to the company's servers. The data provided to us has been anonymized and does not include any demographic information. As such, we are unable to compare the sample composition with that of the other panel studies.

[1] Due to a non-disclosure agreement with the company providing the tool and data, we cannot refer to the tool by name.

3.3 Mobile Security Behavior Observatory

The mSBO is an ongoing research project inspired by the SBO to observe user security behavior on mobile devices, and compare it to that of personal computer users. The application, which is free to download from the Google Play Store, collects data on how users interact with their mobile devices and periodically transmits the data to a central server when an Internet connection is available. Through a chat interface built in the app, users can report spam, phishing schemes, and malicious websites they encounter. Included in the application is a gamified animation character that appears on the user's home screen. Using "experience points" accumulated from interacting with the app and filling out periodic questionnaires (also provided in the app), users can customize the character's color, emotes, and vocabulary. Further details about the mSBO application, the system architecture, and the animation character can be found in Appendix C.

Recruitment The mSBO application was first distributed via the Google Play Store (Japan only) on March 16, 2020. The IARC generic rating was set to 18+ to prevent participants under the age of 18 from participating in the experiment. Upon downloading the app, users are asked to read and understand the terms and conditions to install. During installation, users are informed of the research project and are presented with information about the data collected through the app. Participants must separately consent to each type of data collected before they can start using the application. Participation can be discontinued at any time by uninstalling the application. In addition, users can withdraw consent at any time using a one-click option that leads to the deletion of all data collected from their device.

Coinciding with the launch of the app, recruitment was advertised on seven of our organization's websites and through our organization's Twitter accounts. An additional two-week Twitter recruitment campaign was run in June 2020. As of May 2021, 2,031 participants had installed the app, with approximately 400 daily active users.

Data collection Similar to the SBO, the mSBO relies on a client-server architecture. The mSBO application monitors the use of all other applications on the smartphone device as a background app. The sensor collects data on other installed applications, the use of those applications, web browsing, and network information. Within the app, a local heuristic filter purges email addresses, phone numbers, credit cards, SNS account names, and passwords from the collected data. In addition, the mSBO captures fuzzy hashes [24] of SMS messages that contain URLs, along with the plain text URL, to check for spam and malicious content. The data is then encrypted and sent back to a central server when the user's device has access to the Internet. Further details about the application architecture can be found in Appendix C.

Through the app, users can report security incidents and potential threats through a chat-based interface. In addition, short questionnaires are distributed twice a week which users can complete in exchange for experience points. The contents of the questionnaires vary widely and include topics such as security, information technology, and artificial intelligence.

Lastly, we distributed a 36-question survey through the mSBO application starting in December 2020. The survey asked users about their experience with prior research studies, their security behavior, and general demographic information. Included in the survey is a modified version of the 16-question Security Behavior Intentions Scale (SeBIS) developed by Egelman and Peer [9]. Since the survey was distributed to Japanese-speaking users, we utilized the revised RSeBIS scale which is more robust to language translation [41]. Because the SeBIS scale is geared toward personal computer users, we made slight modifications to several questions as follows. First, we replaced the phrase "computer screen" with "smartphone screen." Second, we combined two questions about device locking (F3 and F4) as they became essentially identical on smartphones. Third, we added a question about biometric authentication to better capture locking and unlocking behavior. Fourth, we removed a question regarding "mouse-over" use prior to clicking a link (F10) as that functionality does not exist on a mobile device. The full list of survey questions, including the modified SeBIS scale can be found in Appendix A. We will refer to this mobile-friendly version of the SeBIS instrument as the mRSeBIS scale. In total, we received 318 valid responses to the survey.

3.4 Ethics and copyright

Ethical review Data from the mSBO study and the Security Toolbar was collected in Japan by Japanese companies. In lieu of an academic Institutional Review Board (IRB), these studies were approved by an external ethics board which included privacy, legal, and ethics experts. All of the data collected as a part of these two studies was used for academic research purposes only and was not monetized in any way. U.S. researchers on the team did not collect any data related to these two studies, but received IRB approval from Carnegie Mellon University to receive and analyze it. The SBO study, which was conducted in the United States, received IRB approval from Carnegie Mellon University.

Copyright licensing To implement the mSBO mobile application, we adopted characters from a famous science fiction animated series. We obtained an official educational license from the copyright owner. The Android application is available on Google Play.We submitted additional license documents to Google for limited use of the characters when registering the app on Google Play. Users residing in Japan can download and install this smartphone application during the license period (currently ending in 2025).

4 Demographics

The key demographics from both SBO and mSBO studies are summarized in Table 1. Demographic information was not collected as a part of the Security Toolbar dataset. While the demographics in both samples are skewed in comparison to the general population, we find that the SBO sample is less representative. Most notably, participants in the SBO study had generally disproportionately lower incomes than the overall U.S. population. In the United States, 17.1% of the population have an income lower than $24,000 [48], while as many as 32.1% of SBO participants reported an annual income lower than $24,000. On the other hand, we do not observe significant income bias in the mSBO sample, which roughly aligns with the income distribution in Japan [47].

In addition, we observe a bi-modal age distribution in the SBO sample, skewed towards participants under 30 and over 60. This may be related to the income skew as the two largest subgroups in the SBO sample consist of university students and retirees, both which tend to have lower levels of income. Again, the mSBO sample does not present the same bias. However, the mSBO sample is strongly skewed toward men. We hypothesize this is because the sci-fi animation character in the mSBO app is based on Seinen manga, Japanese animation targeted toward younger adult men.

We do not observe substantial bias in the sample's education levels. The mSBO sample slightly over-represents those with a high school degree or less, however this can plausibly relate to the animation character attracting younger male participants. On the other hand, the SBO sample is over-representative of participants with higher education.

Table 1: **Demographics from SBO and mSBO studies**

	Demographic	**mSBO**	**SBO**
Gender	Male	69.5%	40.2%
	Female	26.7%	59.3%
	Other/No response	3.5%	0.5%
Age	18-21	2.2%	5.3%
	22–30	10.4%	43.9%
	31–40	23.6%	16.0%
	40–50	36.8%	9.4%
	50–60	22.0%	8.9%
	Over 61	2.5%	16.0%
	No response	2.5%	0.5%
Education	No High School GED	3.8%	0.3%
	High School GED	28.0%	9.2%
	Some College	4.1%	24.4%
	Trade School Degree	18.9%	1.9%
	Bachelor's Degree	29.9%	39.9%
	Master's Degree	8.5%	20.1%
	Doctoral Degree	2.2%	4.2%
	Other/No response	4.7%	0.0%
Income	<2.5M JPY / <25K USD	21.1%	32.1%
	2.5–5M JPY / 25–50K USD	32.1%	22.0%
	5–7.5M JPY / 50–75K USD	19.8%	13.6%
	7.5–10M JPY / 75–100K USD	9.4%	8.2%
	10–15M JPY/100–200K USD	2.2%	8.7%
	>15M JPY / >200K USD	0.3%	2.1%
	No response	15.1%	13.2%
Occupation	Student	2.2%	35.9%
	Company employee	64.2%	40.2%
	Self-employed	5.3%	0.3%
	Public servant	8.2%	–
	Part-time job	6.3%	–
	Unemployed/Retired	–	22.0%
	Housewives and husbands	5.3%	0.5%
	Other/No response	8.5%	1.1%

5 Findings

We next present our findings and observations from the three panel studies. First, we examine various aspects of participant recruitment across the three studies. Second, we assess participant retention to identify factors that had positive and negative effects. Third, we draw upon these experiences to identify important practices for data collection and analysis.

5.1 Participant recruitment

Across the three panels we observe a range of different recruitment strategies, particularly in regards to the incentives offered to participants. We find that both monetary and non-monetary incentives are effective at recruiting panel participants. While we cannot draw causal conclusions about the effect of the incentives, based on survey responses from two of the panels we do observe key descriptive differences in participants' motivation to join the study, privacy concerns, and security behavior. Despite these differences, and contrary to our own hypothesis, we do not find evidence to support the theory of a more acute Hawthorne effect for users immediately after they are recruited into either study. Participants' behavior and device use did not change between the period immediately following recruitment and the remainder of their time in the study.

5.1.1 Monetary and non-monetary incentives

Despite the use of a variety of incentives across the three studies, we observe that all of the incentives offered, both monetary and non-monetary, were effective at recruiting participants. Monetary incentives, like those offered in the SBO study, are a well-established form of compensation in research studies. In contrast, non-monetary incentives are infrequently used by the research community. However, longitudinal panels require incentives that can retain user participation over an often long period of time. This can be an expensive undertaking using monetary incentives. Looking towards alternative methods, the mSBO and Security Toolbar studies offered participants a non-monetary incentive. The Security Toolbar, appropriately named, incentivized users by providing a security service as they browsed the web. In the mSBO study, users were offered a gamified, customizable in-app character from a popular sci-fi animation series.

The Security Toolbar, whose recruitment and distribution was done through a software company, was able to recruit and maintain several hundred thousand participants. The SBO and mSBO studies, whose recruitment channels were similar to that of a typical research study, both were able to recruit hundreds of participants and maintain over 300 daily active users despite very different incentives being offered. In fact, recruitment for the SBO study using monetary incentives was arguably more difficult, required advertisement through multiple channels, and took a longer period of time to ramp up to the same number of users as the mSBO study.

Although we find monetary and non-monetary incentives to work effectively, there are several tradeoffs for researchers to consider and potential bias, discussed in the following sections, to be aware of. First, experimental design can be simpler when using financial rewards as there are fewer variables and design decisions involved compared to using non-monetary incentives. In the case of monetary rewards, only the amount of time the user has to spend and the amount of the reward are considered. On the other hand, the types of non-monetary motivations are "boredom," "comparison," "fun," "science," and "self-learning," which affect the attributes and behaviors of the participants [21]. Second, while non-monetary incentives can lower the direct costs of recruitment, the indirect costs stemming from the design and maintenance of the non-monetary reward should be considered. Third, while both sets of studies compete with other platforms for a limited pool of participants, the incentive design can affect the type of competing platform. With monetary incentives, we find that participants have often used a variety of crowd-sourcing platforms that compete for their time and attention. Although research projects must compete with these other platforms, simply offering higher monetary rewards is generally enough. On the other hand, with non-monetary incentives, researchers cannot easily control the many intangible factors that lead to the widespread adoption of some free apps but not others.

5.1.2 Research participation and motivation

From the surveys in Appendix A and B, we found that SBO participants had more prior experience with research and survey platforms, signed up for research studies more frequently, and were more financially motivated to participate in research than their mSBO counterparts. Two-thirds of SBO participants reported having used at least one crowd-working or survey platform outside of the university recruitment service the SBO study used. In fact, 23% of SBO participants had signed up for research studies at least once a month over the previous year. Conversely, less than 10% of mSBO participants had used a crowd-working service, and less than 30% had used a survey platform service. Fewer than 5% of participants had signed up for research studies at least once a month over the previous year.

Furthermore, when asked to select among eight factors that were important when deciding to participate in a study, SBO participants reported they would prioritize how much they will be paid (76%) and the amount of work required (67%). In contrast, mSBO participants reported that the study purpose (77%) and the security and privacy of the data collected (65%) were most important. Payment amount (16%) and the amount of work required (48%) ranked among the least important factors for mSBO participants. The full prioritized lists of user motivations are shown in Appendix D.

5.1.3 Privacy concerns

mSBO participants were more concerned about how their data was being collected and by whom than SBO participants. mSBO participants rated the "security or privacy of data collected in the study" (65%) as the second most important motivating factor for participation out of a total of eight. "Who is conducting the study" (57%), an indicator of trust and reputation, was the third highest-rated. However, in the SBO study, security and privacy (37%) rated fifth, and who is running the study (26%) rated sixth. While not definitive, these differences could also be related to the incentive being offered, as previous work has shown that people are willing to sell their privacy for minimal amounts of money [16].

5.1.4 Security behavior

Similar to the self-reported privacy concerns, mSBO participants also reported having greater security concerns than their SBO counterparts. The participants' security concerns were measured using the SeBIS, RSeBIS, and mobile RSeBIS (described in Section 3.3) scales in the SBO and mSBO studies. In addition, because we cannot survey users of the Security Toolbar, we instead compare mSBO and SBO results to those obtained in the original RSeBIS work, that targeted Japanese PC users [41], which is the closest proxy for our Security Toolbar users we could find in the literature. The distribution of the SeBIS scores of participants in these three studies is reported in Table 2. Participants in the mSBO reported the highest level of security concerns, followed by SBO participants and then Security Toolbar participants. The difference in the distribution of scores between all three studies was statistically significant at the 95% confidence interval ($p < 0.001$).

To validate our comparison among different versions of the SeBIS scale, we evaluated the mobile-friendly version of SeBIS (mRSeBIS) using the same methodology in the original SeBIS [9] and the revised RSeBIS [41] papers. This method relies on confirmatory factor analysis (CFA) and Cronbach's α to evaluate the validity and reliability of the proposed instru-

Table 2: **Distribution of SeBIS scores across PC users [41], mSBO, and SBO studies.** Scores are normalized by the number of questions (RSeBIS: 16, mRSeBIS: 15, SeBIS: 16)

	PC users [41]	mSBO	SBO
Scale	RSeBIS	mRSeBIS	SeBIS
Responses	500	318	399
Mean	2.572	3.739	3.406
Standard Deviation	0.931	0.763	0.523
Minimum	1.067	1.667	2.250
Maximum	5.000	5.000	5.000

ment. Confirmatory factor analysis measures the alignment between the scales' items and a set of hypothesized latent factors, which, in this case, include proactive awareness, password selection, device locking, and software updating. A high level of alignment indicates that the scale measures the factors we expect them to measure, i.e., the scale is valid. Cronbach's α measures the scale's reliability; in other words, the items are measuring the same construct. This is important, as an unreliable scale cannot be valid. Our results in Table 3 show that the mRSeBIS scale has high reliability and a good fit, roughly equivalent to that of the original SeBIS scale.

5.1.5 Influence of monitoring on initial behavior

In analyzing usage data from the SBO and mSBO studies, we did not find any differences in behavior during the period immediately following user recruitment and their long-term behavior. This ran contrary to our hypothesis that users would change their behavior during their first few weeks in the study in response to being more aware that their device was being monitored. In other words, we expected the Hawthorne effect to be more acute during this initial period since participants were repeatedly made aware of the data collection and monitoring procedures during on-boarding. In particular, we expected that participants might use their devices less initially and would refrain from engaging in privacy-sensitive activities like viewing pornography or visiting video streaming sites that frequently contain pirated content.

Table 3: **mRSeBIS scale validation.**

Scale	mRSeBIS (JP)	Recommended
N	318	
Cronbach's α	0.818	>0.60 [9]
RMSEA	0.055	<0.06 [19]
SRMR	0.058	<0.08 [19]
CFI	0.954	>0.90 [32]
TLI	0.942	>0.90 [32]

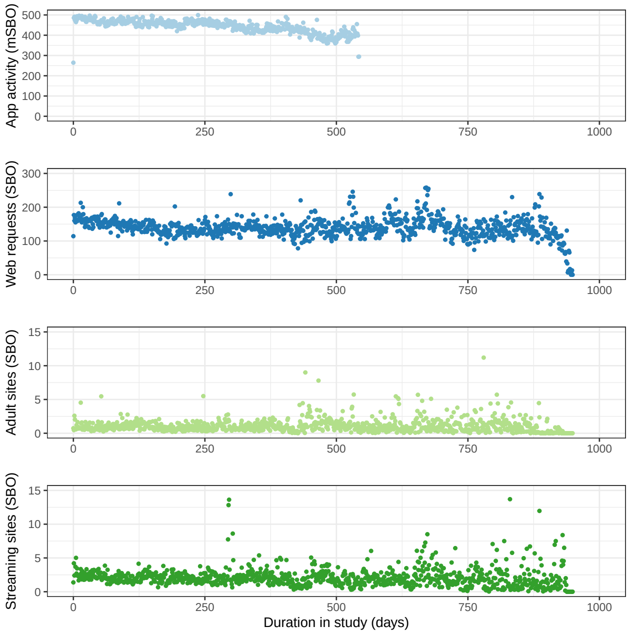


Figure 1: **Activity over time.** From top to bottom, average (1) foreground application records(mSBO), (2) user-initiated web requests(SBO), (3) visits to adult websites(SBO), and (4) visits to streaming websites(SBO), per user by the number of days elapsed since they joined the study.

Figure 1 shows, relative to the number of days participants were in the study, the average application use for the mSBO; and user-initiated web requests, visits to adult websites, and visits to streaming sites for the SBO. As the figure shows, device use remained relatively constant regardless of the length of time a participant was in the study. We also observe SBO users visit adult and streaming websites from day zero onward. Thus, participants do *not* behave differently in an initial ramp-up period before reverting to usual device and browsing patterns. In other words, observed behavior in the period immediately following recruitment appears representative of true behavior. This is particularly important for short-term observational studies, which are much more common than longitudinal research panels.

5.1.6 Lessons learned on participant recruitment

- Both monetary and non-monetary incentives work effectively for recruiting panel participants.
- Indirect costs stemming from the design and maintenance of the non-monetary reward should be considered.
- Researchers compete for a limited pool of participants; incentives affect which platforms one is competing with.

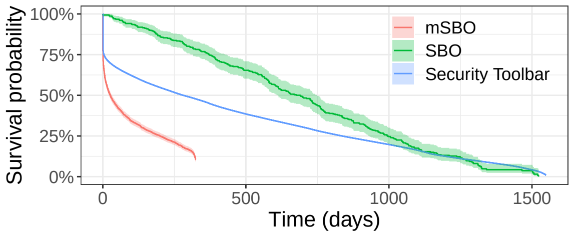


Figure 2: **Kaplan-Meier survival curves for all user panels.** Each point is the probability that a user participating at a time $t = 0$ will still participate at time $t = x$. The shaded area denotes the 95% confidence interval.

- Potential bias related to incentives should be considered, particularly related to privacy and security concerns.
- Newly recruited participants do not behavior differently in their first few days or weeks, than they do throughout the remainder of their time in the study.

5.2 Participant retention

Between the three studies, we observe markedly different retention rates among participants. Figure 2 shows the results of a Kaplan-Meier survival analysis [22] which illustrates the probability of a participant remaining in the study after a certain number of days. As shown, the survival curve for the SBO is relatively linear, with half of the participants dropping out after about 700 days in the study. In contrast, the Security Toolbar and mSBO study have high initial dropout rates, with participation stabilizing for users who stay in the study for at least a month. In fact, after a month, Security Toolbar users are more likely to maintain their participation compared to the SBO and mSBO, as indicated by the flatter downslope of the curve. The mSBO study suffers the highest initial dropout rate, losing about 60% of participants over the first month. After stabilizing, participants drop out at a rate similar (slightly steeper) as in the SBO study. We next identify four factors that help to explain the differences we observe between studies.

5.2.1 Minimizing interference

The first factor influencing retention is the stability of the sensor software and its interference with the participant's use of the device. The mSBO application was tested prior to the initial roll out, however we could not cover the entire spectrum of possible Android devices and versions of the Android operating system that could run the mSBO. Our testing focused on the functionality and usability of the app, such as not interfering with the participant's normal use of their device. Unfortunately, unanticipated compatibility issues and software bugs led to application instability and unexpected crashes during the first four months of the app's release.

The initial version of the app also continuously displayed the character icon on the home screen and when using other applications. This was designed to remind participants of the app's monitoring. However, we received feedback that this display feature severely interfered with user activities. The top panel of Figure 3, which shows the number of new and active mSBO users, demonstrates that during the period from March 2020 to July 2020, users left the study at a high rate. The second graph represents new installation, and the spike in late June 2020 reflects an additional Twitter recruitment campaign. The third graph illustrates crash encountering users per active users. After several bug fixes, a stable version of the app that disabled the constant character visibility was released on July 9, 2020. The fourth graph shows application version history. We released bug fixes and new features ten times during the first year. After the bug fixes released with version 1.0.5, the number of daily active users stabilized.

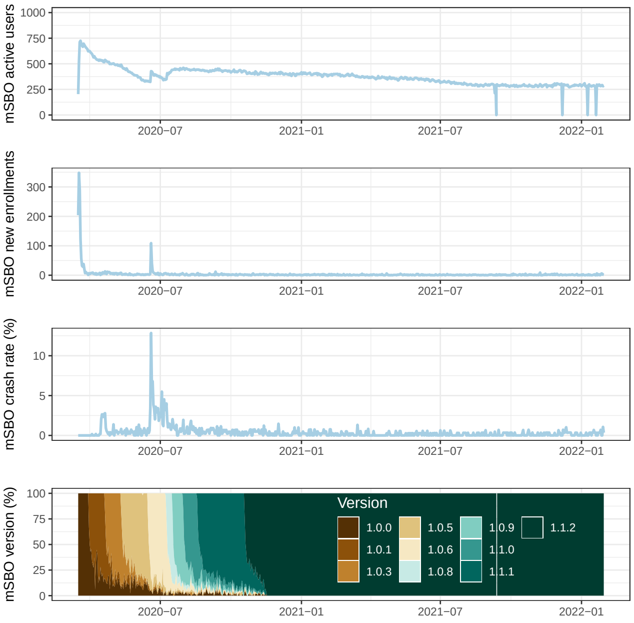


Figure 3: **mSBO panel evolution over time.** Each point is a computed over a one-day window.

Similarly, one of the main complaints from participants in the early part of the SBO study was that the sensor software noticeably slowed down their device. The bottom plot in Figure 4 shows that opt-outs early in the study were primarily due to performance issues. This feedback led to the development of a lightweight version of the sensors. This was initially deployed only for impacted users before being rolled out to a broader set of users in December 2017. As the study continued, performance-related dropouts subsided, and a distinct decline in all dropouts occurred after the December 2017 deployment.

5.2.2 Communication balance

The second important retention factor is striking the right balance of communication with participants. In the mSBO study, we hypothesized that regular notifications would increase users' engagement with the app. Initially, users received three notifications per week with messages or questions for them to answer. However, it became clear that users found this level of communication too high, and many uninstalled the app. In the stable version of the app released in July 2020, we turned off the notifications and made them optional. Combined with fixing the bugs mentioned in Section 5.2.1, stopping the notifications helped to stabilize the number of active daily users. Neither the SBO nor the Security Toolbar studies employed notifications.

5.2.3 Following up with participants

As a corollary to communication, the third retention factor is the level of follow ups with participants. In the SBO study, new participants received an initial enrollment call from a member of the research team when they signed up. A member of the research team would also call participants to follow up whenever the participant had stopped sending data for an extended period of time. Figure 4 shows the number of active SBO users, new enrollments and calls, inactive users and follow-up calls, and opt-outs over the course of the study. Unlike the other panels, the SBO study maintained a positive increase in active daily users during the early phase of the study and relatively linear survival curve throughout. For comparison, of the 1,502 users who installed the mSBO app, during the first week, 25% effectively dropped out by either not opening the app, failing to provide consent, refusing the required permissions, or configuring the app settings incorrectly. We believe that the SBO enrollment calls helped alleviate this problem by addressing participant concerns upfront and resolving initial technical issues. In addition, the follow-up calls to inactive users likely helped achieve a lower attrition rate compared to the other panels: actual opt-outs were low.

5.2.4 Tangible benefits to participants

The fourth and final factor is providing a tangible benefit to participants. This effect is primarily observed among Security Toolbar users who, after an initial drop in participation, were the most likely to remain in the study long-term. While the mSBO app offers some utility through its reporting mechanism, the Security Toolbar provides everyday security benefits by helping prevent social engineering attempts. This benefit makes the toolbar quite popular among IE users and helps to explain the high retention rate.

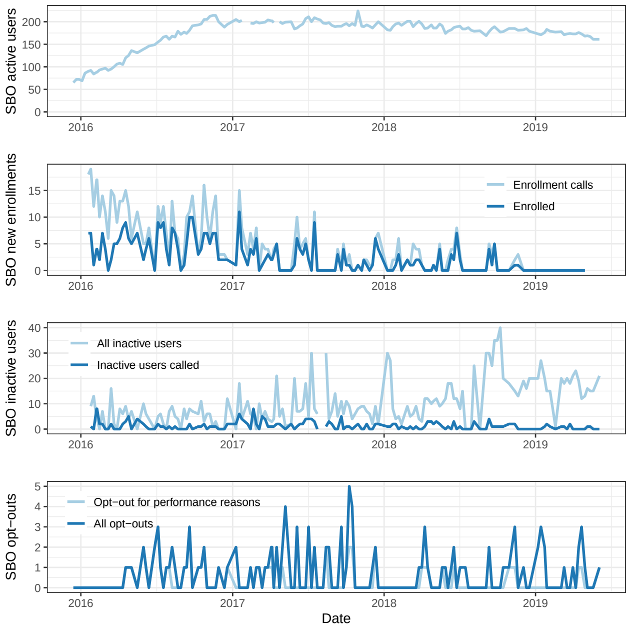


Figure 4: **SBO panel evolution over time.** Each point is a computed over a one-week window.

5.2.5 Device use and retention

One additional area of interest was the relationship between different types of users and their likelihood of remaining in the observational panel. In particular, we theorized that the frequency of device use might impact retention and, as a secondary effect, bias the sample. Based on an analysis of the mSBO and SBO studies, we find a mix of evidence. Table 4 shows the results of a series of regressions comparing the relationship between several metrics of device use and the length of time participant's remaining in the panel. The metrics for device use were log-transformed to obtain normal distributions and heteroskedastic robust standard errors were employed where Breusch-Pagan tests indicated heteroskedasticity. For additional details, see Appendix E.

While we observe a substantial amount of noise, we find a statistically significant positive relationship between how frequently a participant uses their device and how long they stay in the mSBO study. In the SBO study, we only find a significant relationship between average web requests and the duration in the study. It is possible that participants who used their device more frequently were more motivated to stay in the mSBO study due to the gamification of the animation character. The primary means that users leveled up their character, thereby unlocking additional features, was by filling out weekly surveys and reporting security issues they encountered. However, neither of these factors were strongly correlated with average web requests (surveys: $\rho = 0.287$, reports: $\rho = 0.171$) or average app use (surveys: $\rho = 0.356$, reports: $\rho = 0.250$).

Table 4: **Results of regression models**. These compare several metrics of average device use (independent variables) and the number of days in the study (dependent variable).

Study	*n*	Independent Variable	Coefficient	Intercept	*p*-value
mSBO	2229	Average web requests per active day (log)	11.526	112.879	0.020
		Average app use per active day (log)	64.359	-36.656	<0.001
		Network connections per active day (log)	126.923	-28.276	<0.001
SBO	307	Average web requests per active day (log)	35.227	239.782	0.0.031
		Average user-initiated web requests per active day (log)	31.571	326.858	0.081
		Average tab use per active day (log)	14.006	397.662	0.362

5.2.6 Lessons learned on participant retention

- Researchers should test the usability of the sensor software, which should not interfere with normal device use.
- The stability of the sensor software is vital for retention.
- Finding the right balance of communication between researchers and participants is critical.
- Researchers should monitor technical difficulties and follow up with participants quickly.
- Providing a tangible benefit to participants contributes to long-term retention.

5.3 Data collection

Planning for and designing the data collection process in longitudinal studies can be quite challenging. Often, unanticipated events arise during the course of the study that were not accounted for initially. This is particularly true for panels studies that span multiple years. In the following sections, we identify several data collection challenges and useful design decisions based on practical experience that can aid future researchers in creating observational panel studies.

5.3.1 Data collected per user

To give researchers a sense of how much data they can expect to collect, we analyzed the average amount of data collected per user in the SBO and mSBO studies. We find that on average researchers can expect to collect between 550–600 web requests on personal computers and between 50–100 web requests on mobile devices per user per day. Of the web requests made using personal computers, only about 12% of those are initiated in response to user-initiated navigation (e.g., link, bookmark, search, etc.). The remaining 88% were automatically generated by the browser or the web page. In addition, we find that personal computers users interact with browser tabs (e.g. create, switch, or close) about 120 times per day on average. On mobile devices, users interact with and switch between different apps about 500 times per day on average.

OS limitations make it difficult to observe all web requests on mobile devices. VPN or web proxies could help alleviate this issue, but may degrade the user experience. Directly observing the URLs displayed in the web browser navigation bar is also challenging, as different smartphones frequently use different default web browsers (manufacturers often pre-install their own fork of, e.g., Chrome), with their own navigation bar. This, in turn, increases the complexity of the mobile sensor. Even more importantly, users spend more time on other applications than web browsers, and those applications may rely on internal browsers—using system HTML-rendering libraries, but with a different layout.

5.3.2 Identifying dropouts and technical difficulties

As mentioned in Section 5.2.3, following up with inactive users played an important role in retaining users. Therefore, the monitoring software should identify user inactivity. One way to accomplish this, which was used in several panels, was to create automated alerts or regular reports for users whose devices had stopped sending data for an extended period of time. In the SBO study, regular reports were used to follow up with participants manually. In the mSBO study, a "forget me" button was deployed so that participants could clearly signal their intention to dropout of the experiment.

However, a user device might also stop sending data due to a technical problem rather than the user intending to drop out of the study.[2] When it comes to data analysis, these kinds of gaps can be difficult to account for. The observation software used in the SBO study, which was designed with two sets of independent sensors, encountered many cases where one set of sensors would temporarily go down while the other would continue to send data. As a result, a large amount of analysis work was applied to detecting these gaps, and a substantial portion of the data collected had to be thrown out. One solution to this problem is to install an independent

[2] Often this was a result of a sensor failing until the device was restarted or the installation of other software that conflicted with the sensor software.

heartbeat sensor that regularly pings the home server. This can alert the research team when sensors go down as opposed to when a user is simply inactive.

5.3.3 Timestamps and order of events

For observational studies where data is being collected from a user device, timestamps alone are often not sufficient in maintaining the order of observed events. Many computational events can occur within the same (milli)second. In the SBO study, this led to significant post-hoc analysis to recreate the proper sequence of events, and in some ambiguous cases, data had to be discarded. A simple sequence counter enumerated by the sensor software would have alleviated this problem.

However, an ordering mechanism would not have fully solved all timestamping issues. Multiple studies observed skew in the timestamps recorded on participants' devices. This can occur if the user's device is not synchronized to a global time source, the user's device is defective, or the user manipulates their device's internal time intentionally. Furthermore, relying on the timestamp of data arrival at the server is insufficient. Users go offline often, even if they are connected to a mobile network, which delays the upload of sensor data. While not perfect, we find capturing the order of events, the number of seconds that have elapsed between events, and a combination of client and server timestamps to be most effective for data cleansing and analysis. Careful consideration and storage of a user's time zone, which may change throughout the course of the study, is also recommended.

5.3.4 Defining active user engagement

One limitation in using sensor data is that it provides the perspective of the device and only indirectly that of the participant. This can prove challenging when attempting to determine how long a user is actively interacting with their device, an important metric for many applications. For example, when a user navigates to a new web page, that information is logged by the sensors. However, if the user does not interact with their device for an extended period of time, it is unclear whether they are still engaged with that page or if they have left their device on but unattended.

In these studies, user engagement was roughly time-boxed using other recorded events, such as when the user navigated to the next web page or switched browser tabs. The mSBO study also used the foreground application history. However, some mobile apps, e.g., calendar and weather, always occupy the foreground of the screen, which makes it difficult to determine whether the user is active. The SBO study also relied on log in/out, power on/off, lock/unlock, and application change events. To refine this further, mouse movements were overlaid with the activity trace to determine active periods of user interaction. However, even this method is imperfect, as users could still be passively engaged with their device, like when watching a movie, even if they are not actively using their mouse. We recommend that future studies explicitly capture events that indicate the end of a user interaction (e.g., closing a web page or application) if available. While the use of audio and video could provide precise measurements, they also raise substantial privacy concerns, and were avoided in these studies. Alternative measures of active engagement like mouse movements, keyboard use, touchscreen interaction, and resource usage are likely a better, albeit less precise, method.

5.3.5 Multiple users and devices

Over the course of a longitudinal panel study, there likely will be instances of multiple users sharing a given device (more so for personal computers than smartphones), and individual users with multiple devices. In the SBO study, in several cases more than one person was using a single personal computer. It would have been very useful to differentiate users, either by requiring separate logins or using some other identifier. In addition, over the course of the five-year study, most participants upgraded or replaced their computers at some point. A process for handling these cases was not originally in place, resulting in several substantial gaps in data coverage as users switched from one device to another. In the mSBO study, the multi-device problem typically occurred when one person owned more than one smartphone. The use of the primary smartphone differed greatly from that of secondary devices.

5.3.6 Survey distribution

In the mSBO study, the platform was designed such that researchers could distribute surveys to participants directly through the application. This provided a quick and easy way for researchers to interact with participants and gather supplemental data. Using this feature, researchers sent weekly questionnaires to which participants responded at an average rate of 30% of active users. In the SBO study, researchers had to distribute surveys to participants by email. Having to coordinate with participants, often individually, created significant overhead, so that surveys were distributed very infrequently. However, communicating by email rather than through the platform also enabled the SBO study to survey users who had previously participated in the study but had since left.

5.3.7 Lessons learned on data collection

- The monitoring software should identify user inactivity.
- Software should accurately record timestamps and event order, consider clock skew, and network disconnections.
- Researchers should capture metrics to define active user engagement.
- Monitoring should be designed with multiple users and devices in mind.

- Distributing surveys through the sensor software provides greater ease of use and flexibility.

6 Discussion

In the following sections we discuss the limitations of our work, and the implications our findings have for future studies.

6.1 Limitations

Since these studies were run independently, the main limitation of our analysis is that we cannot draw causal inference from any comparison across studies. These studies were also run in two different countries, each with distinct cultures, which may account for some of the observed differences. In addition, the sensors in these studies were device- and platform-specific. The mSBO application was limited to Android smartphone devices, the SBO platform was only available to PC users running Microsoft Windows, and the Security Toolbar was specific to Microsoft's Internet Explorer browser. These factors limit the generalizability of our findings.

6.2 Recommendations for future panel studies

In general, we find that retaining users in measurement panel studies is challenging, especially with new users. Personalized enrollment follow-ups can lower initial drop out rates but are also demanding and costly to scale. Communication, including following up with inactive users, is vitally important to retention. However, finding the right balance of communication is equally important and likely depends on the context of the study. Ideally, communication with participants should be enough to engage users without annoying them. In practice, making the sensors as invisible as possible may be best as interfering with everyday use of a device is a sure way to lose users. Conducting user testing early, with a variety of hardware and devices, is highly recommended.

There is no evidence of a ramp-up period for new users. Hence, participant data collected on initial device use are not as biased as we had originally hypothesized. This helps alleviate concerns over the results of short-term observation studies, and justifies including initial observations alongside long-term observations.

Foresight in designing data collection methods for long-term observation is quite difficult and unanticipated challenges are almost guaranteed. We propose five recommendations to help assuage these issues. First, design mechanisms within the observational software to identify user dropouts and sensor outages. Second, use a variety of sequences, time deltas, and timestamps to maintain the correct order and timing of observed events. Third, collect data that can help to clearly define when users are actively using the device such as start and end events, mouse movements, keyboard and touchscreen interactions, and device resource utilization. Fourth, create a proactive process to handle cases where multiple participants use the same device and multiple devices are used by the same participant. Fifth, build in a mechanism, preferably within the observation platform, to easily follow up with participants, solicit feedback, and distribute surveys.

6.3 Future research

Our results indicate that monetary and non-monetary incentives provide viable means of recruiting participants for longitudinal measurement studies. However, both types of incentives have tradeoffs to consider and potential bias that they introduce. In presenting these three case studies, we are unable to draw causal conclusions about the effects of the various incentives offered. Further research, in a controlled setting, is needed to understand these effects with particular focus on participants' privacy concerns, security behavior, and motivation for participation. In addition, differences in privacy and security concerns may provide an opportunity for researchers to appeal to participants in recruitment and retention. Future work examining such methods would greatly benefit work on longitudinal panels.

7 Conclusions

This paper provides the first evaluation of factors that influence recruitment, retention, and data collection in longitudinal, security- and privacy-sensitive measurement panels. While substantial related work has been done in the context of surveys and clinical studies, privacy/security measurement panels are unique in the intrusive nature of the data collected. These types of studies are relatively rare, although are increasingly being used to observe behavior in a variety of research related to human-computer interaction.

We examined three medium- to large-scale panel studies, which all primarily collect privacy- and security-sensitive data (notably web browsing data). The three studies differed in origin (Japan vs. United States), recruitment incentives (monetary, gamification, added functionality), devices studied (personal computer vs. mobile), degree of interaction, and monitoring software visibility.

Our work provides new insight into recruitment efforts for longitudinal panels, including the effectiveness of monetary and non-monetary incentives, and into participant motivations, privacy concerns and security behavior. We show evidence that users do not act differently during their initial time in the study compared to their long-term behavior, alleviating concerns of potential bias. We identify key factors that affect user retention, including device interference, communication, follow-ups with potential dropouts, and tangible participant benefits. Finally, we derive recommendations to inform the design of the data collection process in future panel studies.

Acknowledgments

This work was partially funded by the National Security Agency (NSA) Science of Security Lablet at Carnegie Mellon University (contract #H9823014C0140), and through a Carnegie Bosch Institute (CBI) Fellowship. This work was also partially supported by WarpDrive: Web-based Attack Response with Practical and Deployable Research InitiatiVE, the commissioned Research of National Institute of Information and Communications Technology (NICT), Japan.

References

[1] Lab in the Wild. http://www.labinthewild.org/.

[2] Project Implicit. https://www.projectimplicit.net/.

[3] Johannes Abeler and Daniele Nosenzo. Self-selection into laboratory experiments: pro-social motives versus monetary incentives. *Experimental Economics*, 18(2):195–214, 06 2015.

[4] Ron Bitton, Kobi Boymgold, Rami Puzis, and Asaf Shabtai. Evaluating the Information Security Awareness of Smartphone Users. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, pages 1–13, April 2020.

[5] Deborah Bowen, Mark Thornquist, Gary Goodman, Gilbert S. Omenn, Karen Anderson, Matt Barnett, and Barbara Valanis. Effects of incentive items on participation in a randomized chemoprevention trial. *Journal of Health Psychology*, 5(1):109–115, 2000.

[6] Davide Canali, Leyla Bilge, and Davide Balzarotti. On the effectiveness of risk prediction based on users browsing behavior. In *Proceedings of the 9th ACM symposium on Information, computer and communications security*, ASIA CCS '14, pages 171–182, Kyoto, Japan, June 2014.

[7] Lara D. Catledge and James E. Pitkow. Characterizing browsing strategies in the world-wide web. *Computer Networks and ISDN Systems*, 27(6):1065 – 1073, 1995. Proceedings of the Third International World-Wide Web Conference.

[8] Serge Egelman, Marian Harbach, and Eyal Peer. Behavior Ever Follows Intention? A Validation of the Security Behavior Intentions Scale (SeBIS). In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 5257–5261, 2016.

[9] Serge Egelman and Eyal 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*, CHI '15, page 2873–2882, 2015.

[10] M. Florencia Etcheverry, Jennifer L. Evans, Emilia Sanchez, Eva Mendez-Arancibia, Mercé Meroño, José M. Gatell, Kimberly Page, and Joan Joseph. Enhanced retention strategies and willingness to participate among hard-to-reach female sex workers in barcelona for hiv prevention and vaccine trials. *Human Vaccines & Immunotherapeutics*, 9(2):420–429, 2013.

[11] David S. Festinger, Douglas B. Marlowe, Jason R. Croft, Karen L. Dugosh, Nicole K. Mastro, Patricia A. Lee, David S. DeMatteo, and Nicholas S. Patapis. Do research payments precipitate drug use or coerce participation? *Drug and Alcohol Dependence*, 78(3):275–281, 2005.

[12] David S. Festinger, Douglas B. Marlowe, Karen L. Dugosh, Jason R. Croft, and Patricia L. Arabia. Higher magnitude cash payments improve research follow-up rates without increasing drug use or perceived coercion. *Drug and Alcohol Dependence*, 96(1):128–135, 2008.

[13] Marvella E. Ford, Suzanne Havstad, Sally W. Vernon, Shawna D. Davis, David Kroll, Lois Lamerato, and G. Marie Swanson. Enhancing Adherence Among Older African American Men Enrolled in a Longitudinal Cancer Screening Trial. *The Gerontologist*, 46(4):545–550, 08 2006.

[14] Alain Forget, Saranga Komanduri, Alessandro Acquisti, Nicolas Christin, Lorrie Cranor, and Rahul Telang. Security Behavior Observatory: Infrastructure for Long-term Monitoring of Client Machines (CMU-CyLab-14-009). Jul 2014.

[15] Alain Forget, Sarah Pearman, Jeremy Thomas, Alessandro Acquisti, Nicolas Christin, Lorrie Cranor, Serge Egelman, Marian Harbach, and Rahul Telang. Do or do not, there is no try: User engagement may not improve security outcomes. In *Proceedings of the Tenth Symposium on Usable Privacy and Security (SOUPS'16)*, Denver, CO, July 2016.

[16] Jens Grossklags and Alesssandro Acquisti. When 25 cents is too much: an experiment on willingness-to-sell and willingness-to-protect personal information. In *Proceedings (online) of the Sixth Worskhop on Economics of Information Security (WEIS'07)*, Pittsburgh, PA, 2007.

[17] Hana Habib, Jessica Colnago, Vidya Gopalakrishnan, Sarah Pearman, Jeremy Thomas, Alessandro Acquisti, Nicolas Christin, and Lorrie Cranor. Away from prying eyes: Analyzing usage and understanding of private browsing. In *Proceedings of the Twelfth Symposium on*

Usable Privacy and Security (SOUPS'18), Baltimore, MD, August 2018.

[18] Gary Hsieh and Rafał Kocielnik. You Get Who You Pay for: The Impact of Incentives on Participation Bias. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, CSCW '16, pages 823–835, February 2016.

[19] Li-tze Hu and Peter M. Bentler. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1):1–55, January 1999.

[20] Kasthuri Jayarajah, Rajesh Krishna Balan, Meera Radhakrishnan, Archan Misra, and Youngki Lee. LiveLabs: Building In-Situ Mobile Sensing &amp; Behavioural Experimentation TestBeds. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services*, MobiSys '16, pages 1–15, June 2016.

[21] Eunice Jun, Gary Hsieh, and Katharina Reinecke. Types of Motivation Affect Study Selection, Attention, and Dropouts in Online Experiments. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW):56:1–56:15, December 2017.

[22] E. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958.

[23] Ngo Manh Khoi, Sven Casteleyn, M. Mehdi Moradi, and Edzer Pebesma. Do Monetary Incentives Influence Users' Behavior in Participatory Sensing? *Sensors (Basel, Switzerland)*, 18(5), May 2018.

[24] Jesse Kornblum and Tsukasa Oi. Ssdeep – Fuzzy hashing program, Apr 2018. https://ssdeep-project.github.io/ssdeep/index.html.

[25] Fanny Lalonde Lévesque, Sonia Chiasson, Anil Somayaji, and José M. Fernandez. Technological and Human Factors of Malware Attacks: A Computer Security Clinical Trial Approach. *ACM Transactions on Privacy and Security*, 21(4):18:1–18:30, July 2018.

[26] Fanny Lalonde Lévesque, José M. Fernandez, and Anil Somayaji. Risk prediction of malware victimization based on user behavior. In *2014 9th International Conference on Malicious and Unwanted Software: The Americas (MALWARE)*, pages 128–134, October 2014.

[27] Takashi Matsunaka, Junpei Urakawa, and Ayumu Kubota. Detecting and Preventing Drive-By Download Attack via Participative Monitoring of the Web. In *2013 Eighth Asia Joint Conference on Information Security*, pages 48–55, 2013.

[28] Takashi Matsunaka, Akira Yamada, Ayumu Kubota, and Takahiro Kasama. A User-participating Framework for Monitoring the Web with Privacy Guaranteed. *IPSJ Journal*, 57(12):2682–2695, 2016.

[29] B. McKenzie and A. Cockburn. An empirical analysis of web page revisitation. In *Proceedings of the 34th Annual Hawaii International Conference on System Sciences*, 2001.

[30] Shusuke Murai. Japan sticks with Internet Explorer as microsoft ends support for old versions. Japan Times, January 2016. https://www.japantimes.co.jp/news/2016/01/12/business/tech/japan-sticks-internet-explorer-microsoft-ends-support-old-versions/.

[31] Anandatirtha Nandugudi, Anudipa Maiti, Taeyeon Ki, Fatih Bulut, Murat Demirbas, Tevfik Kosar, Chunming Qiao, Steven Y. Ko, and Geoffrey Challen. PhoneLab: A Large Programmable Smartphone Testbed. In *Proceedings of First International Workshop on Sensing and Big Data Mining - SENSEMINE'13*, pages 1–6, Roma, Italy, 2013. ACM Press.

[32] Richard G. Netemeyer, William O. Bearden, and Subhash Sharma. *Scaling Procedures: Issues and Applications*. SAGE Publications, Inc, Thousand Oaks, Calif, 1st edition edition, March 2003.

[33] Hartmut Obendorf, Harald Weinreich, Eelco Herder, and Matthias Mayer. Web page revisitation revisited: Implications of a long-term click-stream study of browser usage. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, page 597–606, 2007.

[34] Kaan Onarlioglu, Utku Yilmaz, Engin Kirda, and Davide Balzarotti. Insights into User Behavior in Dealing with Internet Attacks. May 2012.

[35] Sarah Pearman, Jeremy Thomas, Pardis Emami Naeini, Hana Habib, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Serge Egelman, and Alain Forget. Let's Go in for a Closer Look: Observing Passwords in Their Natural Habitat. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, CCS '17, pages 295–310.

[36] Daniel Petrolia and Sanjoy Bhattacharjee. Revisiting incentive effects: Evidence from a random-sample mail survey on consumer preferences for fuel ethanol. *Public Opinion Quarterly*, 73, 08 2009.

[37] Many Brain Project. TestMyBrain. https://www.testmybrain.org.

[38] 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*, CCS '18, page 1238–1255, 2018.

[39] Karen A. Robinson, Victor D. Dinglas, Vineeth Sukrithan, Ramakrishna Yalamanchilli, Pedro A. Mendez-Tellez, Cheryl Dennison-Himmelfarb, and Dale M. Needham. Updated systematic review identifies substantial number of retention strategies: using more strategies retains more study participants. *Journal of Clinical Epidemiology*, 68(12):1481–1487, 2015.

[40] Erica Ryu, Mick P. Couper, and Robert W. Marans. Survey Incentives: Cash vs. In-Kind; Face-to-Face vs. Mail; Response Rate vs. Nonresponse Error. *International Journal of Public Opinion Research*, 18(1):89–106, 07 2005.

[41] Yukiko Sawaya, Mahmood Sharif, Nicolas Christin, Ayumu Kubota, Akihiro Nakarai, and Akira Yamada. Self-Confidence Trumps Knowledge: A Cross-Cultural Study of Security Behavior. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 2202–2214, May 2017.

[42] Mahmood Sharif, Jumpei Urakawa, Nicolas Christin, Ayumu Kubota, and Akira Yamada. Predicting Impending Exposure to Malicious Content from User Behavior. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, CCS '18, pages 1487–1501, Toronto, Canada, January 2018.

[43] Eleanor Singer, Robert M. Groves, and Amy D. Corning. Differential incentives: Beliefs about practices, perceptions of equity, and effects on survey participation. *The Public Opinion Quarterly*, 63(2):251–260, 1999.

[44] Eleanor Singer and Richard Kulka. Paying respondents for survey participation. *Studies of welfare populations: Data Collection and Research Issues*, 01 2002.

[45] Eleanor Singer, John van Hoewyk, and Mary P. Maher. Experiments with incentives in telephone surveys. *The Public Opinion Quarterly*, 64(2):171–188, 2000.

[46] Eleanor Singer and Cong Ye. The use and effects of incentives in surveys. *The ANNALS of the American Academy of Political and Social Science*, 645(1):112–141, 2013.

[47] Statista. Distribution of annual household income in japan in 2019. https://www.statista.com/statistics/614245/distribution-of-annual-household-income-japan/.

[48] Statista. Percentage distribution of household income in the u.s. in 2019. https://www.statista.com/statistics/203183/percentage-distribution-of-household-income-in-the-us/.

[49] Linda Tauscher and Saul Greenberg. Revisitation patterns in world wide web navigation. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, CHI '97, page 399–406, 1997.

[50] Dirk Van Bruggen, Shu Liu, Mitch Kajzer, Aaron Striegel, Charles R. Crowell, and John D'Arcy. Modifying smartphone user locking behavior. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, SOUPS '13, pages 1–14, Newcastle, United Kingdom, July 2013.

[51] Rick Wash, Emilee Rader, and Chris Fennell. Can People Self-Report Security Accurately? Agreement Between Self-Report and Behavioral Measures. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 2228–2232, May 2017.

[52] Nicole Watson, Eva Leissou, Heidi Guyer, and Mark Wooden. *Best Practices for Panel Maintenance and Retention*, chapter 29, pages 597–622. John Wiley & Sons, Ltd, 2018.

[53] Primal Wijesekera, Arjun Baokar, Ashkan Hosseini, Serge Egelman, David Wagner, and Konstantin Beznosov. Android permissions remystified: A field study on contextual integrity. In *24th USENIX Security Symposium (USENIX Security 15)*, pages 499–514, Washington, D.C., August 2015.

[54] Claire E. Wilcox, Michael P. Bogenschutz, Masato Nakazawa, and George E. Woody. Compensation effects on clinical trial data collection in opioid-dependent young adults. *The American Journal of Drug and Alcohol Abuse*, 38(1):81–86, 2012.

[55] Diane K. Willimack, Howard Schuman, Beth-Ellen Pennell, and James M. Lepkowski. Effects of a prepaid nonmonetary incentive on response rates and response quality in a face-to-face survey. *The Public Opinion Quarterly*, 59(1):78–92, 1995.

[56] Akira Yamada, Shoma Tanaka, Yukiko Sawaya, Ayumu Kubota, So Matsuda, Reo Matsumura, Shun Umemoto, Jun Nakajima, Kyle Crichton, Jin-Dong Dong, and Nicolas Christin. Mobile security behavior observatory: Long-term monitoring of mobile user behavior. In *Proceedings of USENIX Symposium on Usable Privacy and Security (SOUPS)*, August 2020. Poster abstract.

[57] Haimo Zhang and Shengdong Zhao. Measuring web page revisitation in tabbed browsing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, page 1831–1834, 2011.

[58] Guili Zheng, Sona Oksuzyan, Shelly Hsu, Jennifer Cloud, Mirna Ponce Jewell, Nirvi Shah, Lisa V. Smith, Douglas Frye, and Tony Kuo. Self-Reported Interest to Participate in a Health Survey if Different Amounts of Cash or Non-Monetary Incentive Types Were Offered. *Journal of Urban Health: Bulletin of the New York Academy of Medicine*, 95(6):837–849, December 2018.

A mSBO survey

(Note: The questions below are translated into English from the original Japanese survey.)

Demographics

1. Which gender do you most identify with?
   - man
   - woman
   - non-binary
   - self-describe: [free text form]
2. What is your age?
   - I would prefer not to respond
   - 18–21 years old
   - 22–30 years old
   - 31–40 years old
   - 41–50 years old
   - 50–60 years old
   - Over 61 years old
3. What is your highest level of education?
   - I would prefer not to respond
   - No High School GED
   - High School GED
   - Some College/Current College Student
   - Trade or Technical School Degree
   - Bachelor's Degree
   - Master's Degree
   - Doctoral Degree or Equivalent
4. What is your occupation?
   - I would prefer not to respond
   - student (esp. a university student)
   - company employee
   - self-employed
   - public servant
   - part-time job
   - Housewife/husband
   - Other:
5. What is your income level?
   - I would prefer not to respond
   - Less than 2,500,000 yen
   - 2,500,000–5,000,000 yen
   - 5,000,000–7,500,000 yen
   - 7,500,000–10,000,000 yen
   - 10,000,000–15,000,000 yen
   - More than 15,000,000 yen

Modified SeBIS (5-point Likert scale; from "never" to "always")

- F3‡: I manually lock my smartphone screen when I step away from it.
- F4‡: I set my smartphone screen to automatically lock if I don't use it for a prolonged period of time.
- F5: I use a PIN or passcode to unlock my mobile phone.
- F6‡: I use biometrics (fingerprint scanner, facial recognition) to unlock my mobile phone
- F12†: I change my passwords even if it is not needed.
- F13: I use different passwords for different accounts that I have.
- F14†: I include special characters in my password even if it's not required.
- F15: When I create a new online account, I try to use a password that goes beyond the site's minimum requirements.
- F8': When someone sends me a link, I open it only after verifying where it goes.
- F11†: I know what website I'm visiting by looking at the URL bar, rather than by the website's look and feel.
- F16†: I verify that information will be sent securely (e.g., SSL, "https://", a lock icon) before I submit it to websites.
- F7†: If I discover a security problem, I fix or report it rather than assuming somebody else will.

- F1: When I'm prompted about a software update, I install it right away.
- F2: I try to make sure that the programs I use are up-to-date.
- F9: I verify that my anti-virus software has been regularly updating itself.

The dagger (†) symbol represents questions modified in RSeBIS [41] from the original SeBIS [8, 9]. The double-dagger (‡) symbol denotes questions modified from RSeBIS [41]. F6‡ is introduced instead of F6 because F6 and F5 become identical in the smartphone context. A question related to using mouse-over as a strategy (F12 in the original SeBIS) is removed because smartphones do not offer this functionality.

Do you have any complaints about using the app? (5-point Likert scale; from "strongly disagree" to "strongly agree")

1. The application slows down my device.
2. The application drains the battery on my phone.
3. The app shuts down unexpectedly.
4. I receive too many messages from the app.
5. The application interferes with the normal use of my phone.
6. I am concerned about the privacy of the data collected.
7. Other: [free text form]

What is your level of satisfaction? (5-point Likert scale; from "very dissatisfied" to "very satisfied")

1. The character on the home screen
2. Experience / Level
3. Changing the emote color
4. Periodic questionnaire
5. Phishing/spam report
6. Profile (screen time)
7. Protocol (install/consent process)

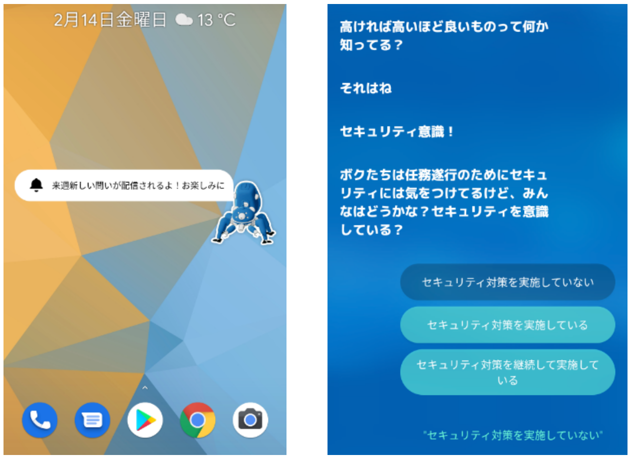
About the app

1. Where did you hear about this app?
    - Press Release
    - news site
    - Twitter
    - Other social networking sites (e.g. Facebook)
    - Friend/Colleague
    - Other: [free text form]
2. Over the past year, how frequently have you signed up for a new research study? (Not including this study)?
    - Never
    - Less than one per month
    - About one per month
    - About one per week
    - Several times a week
    - Multiple times a day
3. What factors are important to you when deciding what studies to participate in?
    - How much I will will be compensated for participating
    - Amount of effort or work
    - Whether I can participate at home / online (versus going somewhere to participate in person)
    - Purpose or topic of the study
    - Security or privacy of data collected in the study
    - Who is conducting the study
    - How quickly I will be compensated get paid
    - The study's consent form
    - Other: [free text form]
4. Would you be more likely to continue participating if you received a small recurring payment or if we continued to add new customizations and features to the character?
    - Small recurring payment
    - The character's new customizations and features
    - Neither

B SBO exit survey

1. Have you participated in any other research besides this study? ("Research," includes academic research, like this study, or marketing research, like surveys for companies.)
    (a) Yes, both in person and remotely (e.g., online, by phone or mail).
    (b) Yes, in person only.
    (c) Yes, remotely only (e.g. online, by phone or mail).
    (d) No.
    (e) Not sure.

2. What research platforms have you used to sign up for studies? (Please select all that apply.)
   (a) [University 1 platform]
   (b) Other [University 1] platform (please describe).
   (c) [University 2 platform]
   (d) Other [University 2] platform (please describe).
   (e) Amazon Mechanical Turk (MTurk).
   (f) Qualtrics.
   (g) Prolific.
   (h) Other (please describe).
   (i) None of the above.
3. Over the past year, how frequently have you signed up for a new research study? (Not including this study.)
   (a) Never.
   (b) Less than one per month.
   (c) About one per month.
   (d) About one per week.
   (e) Several times a week.
   (f) Multiple times a day.
4. What factors are important to you when deciding what studies to participate in? (You may select multiple factors.)
   (a) How much I will get paid.
   (b) Amount of effort or work.
   (c) Whether I can participate at home / online (versus going somewhere to participate in person).
   (d) Purpose or topic of the study.
   (e) Security or privacy of data collected in the study.
   (f) Who is conducting the study.
   (g) How quickly I will get paid.
   (h) The study's consent form.
   (i) Other.
5. Have you participated in other studies that collected data about how you used computer(s), smartphone(s), or other internet-connected devices?
   (a) Yes.
   (b) No.
   (c) Not sure.
6. How was your experience participating in this study, overall?
   (a) Positive
   (b) Negative
   (c) Not sure
   (d) Other
7. What did you like about the study, if anything?
8. What did you dislike about the study, if anything?
9. Did you have any concerns or reservations about enrolling in this study?
   (a) Yes.
   (b) No.
   (c) I don't remember.
10. What were those concerns, and what caused you to enroll anyway?
11. Would you participate in a study that used software to collect data about your computer usage again?
    (a) Yes.
    (b) No.
    (c) Not sure / Depends.
12. If you wish, you may elaborate on why you would or would not participate in this type of study again.
13. How did this study's payments compare to other studies you have participated in?
    (a) Less generous.
    (b) More generous.
    (c) About the same.
14. Do you have any feedback about the payments for this study? This could include payment amounts, the payment method (Amazon gift cards), payment timing, or other payment details.
15. Do you think you used your computer differently than you normally would due to our research software being installed on it?
    (a) Yes.
    (b) No.
16. What caused you to use your computer differently when our research software was installed?
17. What was different about your computer usage while our software was installed?
18. Do you have any other feedback or comments about this study that you would like for us to know?



2月14日金曜日 13 °C

来週新しい問いが配信されるよ！お楽しみに

高ければ高いほど良いものって何か知ってる？

それはね

セキュリティ意識！

ボクたちは任務遂行のためにセキュリティには気をつけてるけど、みんなはどうかな？セキュリティを意識している？

セキュリティ対策を実施していない
セキュリティ対策を実施している
セキュリティ対策を継続して実施している

"セキュリティ対策を実施していない"

©Shirow Masamune, Production I.G/KODANSHA

(a) The character is displayed on the homescreen of users' smartphone anytime, and informs messages to users

(b) When a user taps the character, the screen transitions to the chat interface, where the user responds to questions

Figure 5: **mSBO user interface** with (a) the animation character on the users' home screen and (b) the chat interface used for reporting potential security threats and answering short surveys.

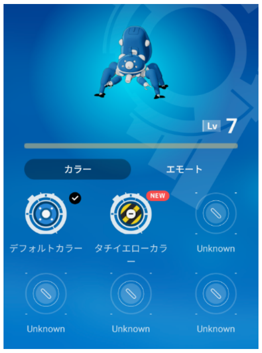
C mSBO app description

The mSBO app consists of a cartoon character agent on the smartphone's home screen and a chat-type interface. Fig. 5 shows a screenshot of the app. The animation character appears not just in the home screen but in any application. Tapping on the character icon launches the app and brings the user to the chat-type interface to interactively talk. We can send any message to the users as a pop-up attached to the character. We delivered questionnaire invitations through this pop-up.

We implement the character using Android's screen overlay functionality, enabling us to overlay Tachikoma on other applications. We do not implement the automatic location feature, so the icon and pop-up possibly cover other app displays or buttons. The users have to move the icon before tapping on something else. Although we use this design to clearly notify the users they were being observed, having the character constantly in the foreground admittedly may interfere with regular phone usage. Tapping on the character invokes the mSBO app, even if the participant is using another app.

The mSBO app provides experience points and stage levels to incentivize users. Figure 6 illustrates the color selection scene. Participants can earn points by answering regular questionnaires and reporting spam/phishing. As participants collect these points, they reach higher experience levels and can in turn further decorate the home screen character.

Figure 7 shows the mSBO sensor app configuration. The sensor app extensively relies on Android's Accessibility Service, which is designed to provide alternative navigation feedback to applications installed on Android devices. For example, the Accessibility Service can be used to convert text to speech, or to warn of malicious web sites in addition to other tools (e.g., Google Safe Browsing). Most apps (e.g., Chrome, SMS, ...) fire `AccessibilityEvents` to communicate UI changes to the Accessibility Service.



©Shirow Masamune, Production I.G/KODANSHA

Figure 6: **Color selection to decorate the home screen character**. This is based on the experience level reached.

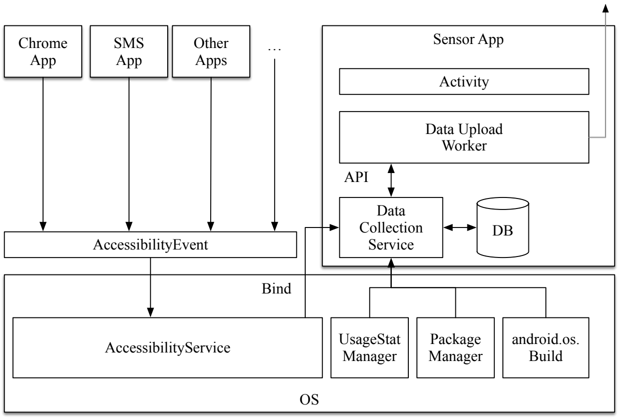


Figure 7: mSBO app configuration.

The mSBO app binds its own Data Collection Service to the Accessibility Service. That way, as long as the user grants Accessibility Service permission to the mSBO app, the Data Collection Service can capture whatever text is displayed in the app the user is running; e.g., the URL in the navigation bar, any anchor text in the browser, or any URL in an SMS.

The second major component of the mSBO is a `DataUpload` Worker. This worker, under Android's `WorkerManager`, uploads collected data as a background service. These uploads are scheduled, deferrable, asynchronous tasks, and are resilient to app crashes or device restarts.

D Participant motivations

In Section 5.1.2, we describe user motivations for engaging with the SBO or the mSBO. Below we present the full lists of motivations, ordered by decreasing priority, for both types of participants in Tables 5 and 6. SBO participants reported prioritizing how much they would be paid (76%) and the amount of work required (67%). In contrast, mSBO participants reported that the study purpose (77%) and the security and privacy of the data collected (65%) were most important. Payment amount (16%) and the amount of work required (48%) ranked among the least important factors for mSBO participants.

Table 5: **Prioritized motivation list (SBO)**

No.	Motivation	Rate
1	How much I will be paid	75.8%
2	Amount of effort or work	66.7%
3	Whether I can participate at home / online (versus going somewhere to participate in person)	64.1%
4	Purpose or topic of the study	52.0%
5	Security or privacy of data collected in the study	36.9%
6	Who is conducting the study	25.8%
7	How quickly I will get paid	20.7%
8	The study's consent form	13.1%
9	Other	4.0%

Table 6: **Prioritized motivation list (mSBO)**

No.	Motivation	Rate
1	Purpose or topic of the study	77.1%
2	Security or privacy of data collected in the study	64.9%
3	Who is conducting the study	56.7%
4	Amount of effort or work	48.0%
5	Whether I can participate at home / online (versus going somewhere to participate in person)	45.8%
6	The study's consent form	44.8%
7	How much I will will be compensated for participating	16.0%
8	How quickly I will be paid	4.1%
9	Other	3.1%

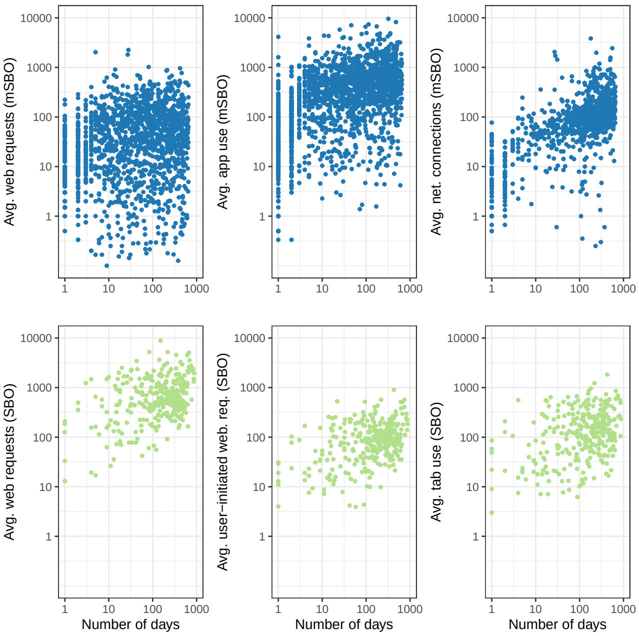


Figure 8: **Device use and retention (scatter plots).**

E Device use and retention graph

In Section 5.2.5, we ran several linear regressions to evaluate the relationship between device use and the length of participation in the study. As discussed, the dependent variables in the model were log-transformed to obtain normal distributions, an underlying assumption required for linear regressions. We tested for heteroskedasticity using Breusch-Pagan tests and found evidence of it in the mSBO sample. As such, we applied heteroskedastic robust standard errors in those regressions. All three metrics in the mSBO sample and the number of web requests in the SBO sample were found to have a statistically significant positive relationship with participants' duration in the study. We visually represent the relationship between the various metrics of device use and retention in both samples, in Figure 8. The figure presents scatter plots where the *x*-axis is the number of days users participated in the study, and the *y*-axes are the corresponding use of the device, according to various metrics. These scatter plots indicate that while a small positive relationship sometimes exists, the data are quite noisy.