Can People Self-Report Security Accurately? Agreement Between Self-Report and Behavioral Measures

Rick Wash
Michigan State University
East Lansing, MI USA
wash@msu.edu

Emilee Rader
Michigan State University
East Lansing, MI USA
emilee@msu.edu

Chris Fennell
Michigan State University
East Lansing, MI USA
cfennell@msu.edu

ABSTRACT
It is common for researchers to use self-report measures (e.g. surveys) to measure people's security behaviors. In the computer security community, we don't know what behaviors people understand well enough to self-report accurately, or how well those self-reports correlate with what people actually do. In a six week field study, we collected both behavior data and survey responses from 122 subjects. We found that a relatively small number of behaviors – mostly related to tasks that require users to take a specific, regular action – have non-zero correlations. Since security is almost never a user's primary task for everyday computer users, several important security behaviors that we directly measured were not self-reported accurately. These results suggest that security research based on self-report is only reliable for certain behaviors. Additionally, a number of important security behaviors are not sufficiently salient to users that they can self-report accurately.

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords
security; self-report; intentions

INTRODUCTION
Users must make choices and then take action to protect their computing devices from malicious actors [15]. These actions include choosing strong and unique passwords, installing software updates for operating systems and third party software, and many others. Researchers often ask users to answer self-report questions about what security-related actions they took in the past, or intend to take in the future (e.g. [18, 12, 5]). The responses are assumed to be reasonably accurate, and are used as a basis for drawing conclusions about security-related outcomes. Survey methods are used because measuring many security-related behaviors in situ is difficult; security tasks and therefore behaviors occur intermittently and are rarely the primary focus of attention [19], and behavioral measures that cover the same range of tasks as survey questions are more labor-intensive to develop, collect and analyze.

CHI 2017, May 06 - 11, 2017, Denver, CO, USA

ISBN 978-1-4503-4655-9/17/05…$15.00

DOI: http://dx.doi.org/10.1145/3025453.3025911

However, Sheeran's [13] recent meta-meta-analysis of research across a wide variety of non-security situations found that self-reported intentions only account for about 28% of the variance in behavior. Situational characteristics can vary widely; correlations between self-report and behavior ranged from 0.40 to 0.82. People don't remember everything they do; they often substitute typical cases for specific memories [3]. They form intentions despite constraints that prevent them from acting on those intentions [2], and let their intentions alter their memories [3]. Their answers to self-report questions are often biased, for example, by over-reporting behaviors that are more socially desirable [11]. There are some behaviors that people can likely self-report accurately, while other behaviors are more difficult.

There is a growing body of computer security research comparing user self-reports about security-related perceptions, intentions and actions with evidence of their behaviors. Users report password complexity and reuse moderately accurately [4, 17]. Installing a single MacOS update within three weeks of release was associated with higher Updating sub-scale scores on the Security Behavior Intentions Scale (SeBIS) [4]. But, descriptions of Windows update settings often do not match what the computer is actually configured to do [18], and the overall security state of a computer is unrelated to the way non-expert users self-describe their level of engagement with computer security [7].

Understanding to what extent and under what circumstances self-reported security agrees with behavioral measures is important, because security researchers who use survey measures need to know when it is valid to do so. Also, understanding which security-related behaviors people can self-report accurately will help to identify types of actions that are harder for users to be aware of or remember, which could be associated with poor security decisions. Finally, knowing more about how aware users are of the security state of their computers and how that connects with their own security-related actions could lead to new opportunities for the design of security tools and interventions.

Our work builds on prior studies by measuring self-report of security-related behaviors using a survey, and collecting evidence of users' security behaviors over a six-week period of actual real-world use, rather than in an experimental setting.

We found that while some behavioral measures are moderately correlated with survey responses (approximately a 0.3 correlation), other behaviors considered to be important by experts [10] had virtually no correlation with the self-reported answers. Security behaviors that involve choices that have visible effects are self-reported moderately accurately, while behaviors that are passive are much less likely to be self-reported accurately.

METHOD

Our study combined two surveys with a log data collection tool designed to collect evidence of wide-ranging security-relevant behaviors. We began the study by giving subjects a survey (the "pre-survey") that asked a number of questions about their past behaviors and their future intentions. At the end of the pre-survey, we provided instructions for installing a custom-written tool that collected log data from both Microsoft Windows and from subjects' web browsers (Google Chrome and Mozilla Firefox). The subjects ran the software on their computers for at least six weeks, after which they were asked to take a second survey (the "post-survey") and given instructions for uninstalling the log data collector.

We recruited subjects in the spring of 2015 from a large midwestern university by asking the registrar to email a random sample of students (both undergraduate and graduate). Students in computer science and engineering were excluded. Out of 15,000 emails, we had approximately 247 students respond (1.6% response rate). About 180 were eligible to participate in the study: they had a personal computer running Windows 7 or Windows 8, which they said they used regularly; they used either Google Chrome or Mozilla Firefox as their main web browser; and they responded to our instruction emails. They were also required to have the ability to install software on the computer, and be the only user of the computer.

We received usable data from 122 of the eligible subjects (0.8% usable response rate). The remaining subjects were excluded due to unforeseen bugs in the data collection software or because they did not use their computer enough (e.g. had more than 7 consecutive days without using the computer, not counting spring break). Two subjects' computers had hardware problems that caused them to withdraw, and two others withdrew without explanation. The sample is fairly representative of the student population of the university. Almost all subjects were in the 18-29 age range, 52% were female, and 76% were white. Approximately 72% of the subjects were undergraduates, and the remaining were graduate students.

All subjects provided informed consent to the data collection. Subjects were compensated $70 for their participation; those who withdrew early received partial compensation. They were able to turn off the data collection software at any time using a control panel that we provided, or by entering private browsing mode. Our study was approved by our institution's IRB.

MEASURES

Our self-report measures are individual survey questions or scales composed of a set of questions. Most use a 5-point Likert scale ('Strongly Disagree' to 'Strongly Agree', or 'Never' to 'Always'). The text of all questions is presented in Table 1. The questions are from two survey instruments published in 2015, Egelman and Peer [5], and Wash and Rader [16]. We use subjects' answers to pre-survey questions in our analysis, except the SeBIS items asked only in the post-survey. Survey questions were included verbatim to better understand the validity of prior research. The survey measures we used are:

- *How often do you? Block popups* (M=4.1, SD=0.83), *Update patches regularly* (M=2.8, SD=1.2), and *Use good passwords* (M=4.1, SD=0.91) [16].
- Password Generation sub-scale questions *F13* (M=3.0, SD=0.96) and *F14* (reversed, M=3.1, SD=1.2) [5].
- *Updating Sub-Scale* (M=3.33, SD=0.78, Chronbach's Alpha=0.72), and question *F2* (M=3.6, SD=0.91) [5].

The behavioral measures were calculated using log data collected by our software, which contains a record of many security-related actions from each subject over the six-week study period. We calculated per-subject log data variables in several ways; for example, counts of the number of times an action occurred (e.g., number of passwords that include a special character), percentages indicating how frequently a state was true (e.g., percentage of days that iTunes was up-to-date), or binary variables representing whether a state was true at all during the study (e.g., was a 3rd party ad-block extension installed). Two password measures were special cases: password entropy measured in bits, and the website-to-password ratio measuring password reuse (higher values mean more reuse). The log data measures we used are:

- 3rd party ad-block installed (Yes=71)
- 3rd party ad-block running (On=51, Changed=17, Off=3)
- Avg. 3rd party softw. Days up-to-date (M=35%, SD=17%)
- iTunes Days up-to-date (M=37%, SD=42%)
- Firefox Days up-to-date (M=27%, SD=30%)
- Chrome Days up-to-date (M=32%, SD=15%)
- Avg. Windows Days-to-Patch (M=5.6, SD=4.2)
- Windows updates installed (M=42, SD=29)
- Password w/special character (Yes=69)
- Average Password Entropy (M=48, SD=8.8)
- Website-to-password ratio (M=3.4, SD=2; 1=no reuse)

We focused on comparing survey and log data measures for software updates, passwords, and ad-blocking browser extensions. Keeping software up to date is a security activity that both expert and non-expert computer users believe is important, as is creating strong passwords [10]. Web advertisements can distribute malware, and many users find them to be a nuisance [9]. All of these measures involve user choices that occur at irregular time intervals, and the focus of users' attention only briefly. For example, notices about updates for 3rd party software often interrupt users [14].

RESULTS

We calculated correlations between self-report responses and related log data measures. Table 1 shows the main results. We roughly grouped our comparisons: correlations 0.20 or higher we believe indicate that users can self-report moderately accurately. Correlations less than 0.10 indicate effectively no relationship; these are behaviors where the survey answers show little to no correspondence to what the logs indicate hap-

Table 1. Correlations between survey measures and log data measures. Each row is a comparison between one survey question and one log data measure. The number of subjects included in each correlation varies depending on details like whether particular software was installed (e.g., iTunes, Firefox, ad blocking extension), or how many subjects did not answer a survey question. WR refers to Wash & Rader [16]; EP refers to Egelman & Peer [5]. The question in row 6. below is reverse-coded.

Source	Survey Question	Log Measure (Units)	Correlation		N
1. WR	How often do you? Block popups	Install 3rd party ad-block (binary)	0.37	***	122
2. WR	How often do you? Use good passwords	Average Password Entropy (bits)	0.31	***	122
3. EP	I try to make sure that the programs I use are up-to-date (F2).	iTunes Days up-to-date (percent)	0.29	**	84
4. EP	I try to make sure that the programs I use are up-to-date (F2).	Average 3rd party software Days up-to-date (percent)	0.23	*	111
5. EP	Updating Subscale	Average Windows Days-to-Patch (count)	0.17	•	101
6. EP	I do not include special characters in my password if it's not required (F14, reversed).	Password w/ special character (binary)	0.095		112
7. WR	How often do you? Update patches regularly	Windows updates installed (percent)	-0.078		121
8. EP	I use different passwords for different accounts that I have (F13).	Website-to-password ratio	-0.073		111
9. EP	I try to make sure that the programs I use are up-to-date (F2).	Firefox Days up-to-date (percent)	0.023		73
10. EP	I try to make sure that the programs I use are up-to-date (F2).	Chrome Days up-to-date (percent)	0.022		106
11. WR	How often do you? Update patches regularly	Average Windows Days-to-Patch (days)	-0.017		109
12. EP	Updating Subscale	Windows updates installed (percent)	-0.017		111
13. WR	How often do you? Block popups	Run 3rd party ad-block (binary)	0.015		71

$p < 0.001$: '***' ; $p < 0.01$: '**' ; $p < 0.05$: '*'; $p < 0.10$: '•'

pened on the computer. Correlations between 0.10 and 0.20 are inconclusive; they might indicate an important relationship, but they also might simply be noise.

We conducted a Pearson product moment hypothesis test that compares each of the correlations with 0; a statistically significant result suggests that we can rule out a 0 correlation. However, most of these measurements include a fair amount of measurement error, and it is known that measurement error reduces correlations and inflates standard errors, which increases the likelihood of finding a non-significant result even when there is a large, real correlation [13]. We considered using a Bonferroni correction for multiple comparisons. However, all corrections for multiple comparisons significantly reduce the power of the statistical tests, and would make it very unlikely to find true effects if they are present. Additionally, such a correction would dramatically increase the type M error rate and would mean that any effect that was statistically significant was likely to be a large over-estimate [8]. We believe that focusing on the effect size—the size of the correlation—is a more statistically valid way of drawing conclusions.

Blocking Popups: We begin by looking at two ways of measuring whether people block popups while web browsing. The survey question, from Wash and Rader [16], asks "How often do you? Block popups". We compared subjects' answers on this question to two log measures for whether people block popups: whether they have a 3rd party browser extension installed that blocks ads, and whether they have a 3rd party browser extension for blocking ads activated and running.

There is a positive correlation between the answer to the survey question and whether the subject's computer has an ad-blocking extension installed. Indeed, this is the largest correlation we found (r=0.37, Row 1 in Table 1). However, there is virtually no correlation between the survey question and whether that extension was actually activated and running (r=0.02, Row 13). Twenty of the 71 subjects who had an ad blocking extension installed had disabled it at some point during the six weeks of the study.

Both installing and enabling/disabling the extension require a decision from the user. When the extension is running, there are visible artifacts, such as missing ads, on almost every webpage visited. But the decision to install an ad-blocker may be more easily recalled when filling out a survey than decisions to change the state of an already-installed ad blocker. Also, as early as the Spring of 2015, websites were blocking browsers with ad-blocking extensions turned on [1]. Some subjects may have turned ad-blocking off to visit certain websites, and this variability may be why the survey measure is not correlated with whether the ad-blocking extension is running.

3rd Party Software Updates: Next, we examine the update status of 3rd party (non operating system) software on subjects' computers. The survey question, from Egelman and Peer [5], asks people to "strongly agree" or "strongly disagree" with the statement "I try to make sure that the programs I use are up-to-date." We collected information about the dates that updates for common 3rd party software programs were released. We compared the version recorded in the log data with current version information for each date during the study period to calculate the percentage of days that Apple's iTunes and Google's Chrome, two commonly-used pieces of 3rd party software, were up-to-date.

The self-reported agreement with keeping software up-to-date is correlated with whether iTunes was up to date (r=0.29, Row

3 in Table 1, the third strongest correlation we measured). However, subjects' answers to the same survey question were uncorrelated with whether Google Chrome was up-to-date (r=0.02, Row 10). The correlation with Mozilla's Firefox was also very low (r=0.02, Row 9). iTunes has a highly visible update system that requires user interaction when an update is released. Chrome and Firefox, on the other hand, handle updates in the background; they automatically download and install updates without any user interaction. We suspect that the visible nature of iTunes' updates means that users have a better understanding of whether that software is up-to-date.

Operating System Software Updates: We can also analyze how well survey questions about software updates correlate with operating system updates. Rather than using that one question, we instead draw comparisons against the SeBIS Updating sub-scale, which asked multiple questions about keeping systems up-to-date [5]. We calculated two measures of how up-to-date the operating system was: the percentage of Windows Updates patches that were successfully installed, and for those installed, the average number of days from the time the patch was released to the time it was finally installed.

Average days-to-patch and scores on the SeBIS Updating sub-scale are positively correlated, although slightly below our threshold for a moderate correlation (r=0.17, Row 5 in Table 1). However, there is almost zero correlation between the SeBIS Updates sub-scale and the percentage of Windows Updates that were successfully installed (r=-0.017, Row 12). On average, our subjects' computers successfully installed only 42% of updates, but 19 computers had installed every update successfully.

Days-to-patch is somewhat under the control of the user. The default setting for Windows Update is to download all updates and then prompt the user when a reboot is needed [18]; users have the option of delaying the install. However, whether an update is installed successfully or ends with an error is not really under the user's control. When users self-report, they seem to self-report based on updating actions that are under their control as opposed to actions that the computer takes on their behalf.

Passwords: Finally, we compare password composition and reuse measures with subjects' self-report answers. This has been examined in detail in past research (e.g., [17]), and our findings are similar. However, here we highlight one particular finding that is new. We asked subjects the SeBIS question, "I do not include special characters in my password if it's not required." [5]. Following Florêncio and Herley [6], we examined the entropy subjects' passwords to determine if any of them included a special character.

Of all of our pairs of log and survey measures, we believe this comparison has the strongest surface correspondence between the measures. However, there is a fairly low correlation between the survey question and the log data (r=0.10, Row 6 in Table 1). Also, other evidence of password choices, such as the average entropy, are more strongly correlated with self-reported survey answers (r=0.31, Row 2). This indicates that users are generally aware of their choices about passwords [17], but may be unsure whether specific passwords use special characters or not.

SeBIS Sub-scales: Finally, we can partially validate the SeBIS Password Generation and Updating sub-scales, which do appear to be correlated with some log data measures. However, they are not correlated with all of the relevant log-data measurements (see below).

Sub-scale	Log Measure	Correlation	
Password	Average Password Entropy	0.18	•
Password	Password Re-use	-0.11	
Updates	3rd Party Software	0.23	*
Updates	Windows Days-to-Patch	0.17	•
Updates	Windows % updates installed	-0.02	

$p < 0.001$: '***' ; $p < 0.01$: '**' ; $p < 0.05$: '*'; $p < 0.10$: '•'

Similar to the patterns we found in the other comparisons, these two SeBIS sub-scales were more highly correlated with visible actions of the user (such as days-to-patch 3rd party software, r=0.23) and less correlated with actions that the user has less control over or require everyday attention to security details (Windows percent updates installed, r=-0.02).

DISCUSSION

Across a number of important security activities, we found that people can self-report some types of behavior relatively accurately, but other types of behaviors are not correlated with self-report survey responses. People seem to be able to accurately self-report behaviors that are related to choices that are more salient, either because they are made proactively, such as installing an ad-blocking extension, or because the computer prompts them to do something (iTunes updates). If a behavior involves awareness rather than action (such as whether the ad-blocking extension is running), or if a behavior isn't visible (such as automatic Chrome updates), then people are less able to accurately answer questions about it.

This finding has important implications for how we interpret survey-based security research. There are many user decisions that people do not self-report accurately. When studying these decisions, it is important to measure actual behaviors rather than relying on self-reports. Additionally, these findings suggest that when users are considering their own past behaviors, they likely remember actions they initiated better than states that results from their actions. Their perceptions of their own security are likely biased by their explicit actions, and discount awareness behaviors and less visible behaviors.

ACKNOWLEDGEMENTS

We thank Kami Vaniea, Ruthie Berman, Zac Wellmer, Tyler Olson, Nick Saxton, Nathan Klein, Raymond Heldt, Ruchira Ramani, Jallal Elhazzat, Tim Hasselbeck, Shiwani Bisht, Robert Plant Pinto Santos, Meghan Huynh, and Simone Merendi for assistance in developing the software and analyzing the data. This material is based upon work supported by the National Science Foundation under Grant Nos. CNS-1116544 and CNS-1115926.

REFERENCES

1. 2015. Block Shock. *The Economist* (June 6 2015). http://www.economist.com/node/21653644/print
2. Icek Ajzen. 1991. The theory of planned behavior. *Organizational Behavior and Human Decision Processes* 50, 2 (Dec. 1991), 179–211. DOI: http://dx.doi.org/10.1016/0749-5978(91)90020-T
3. John R. Anderson. 2009. *Cognitive Psychology and Its Limitations* (7th ed.). Worth Publishers.
4. Serge Egelman, Marian Harbach, and Eyal Peer. 2016. 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*. 5257–5261. DOI: http://dx.doi.org/10.1145/2858036.2858265
5. Serge Egelman and Eyal Peer. 2015. Scaling the Security Wall: Developing a Security Behavior Intentions Scale. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. 2873–2882. DOI: http://dx.doi.org/10.1145/2702123.2702249
6. Dinei Florêncio and Cormac Herley. 2007. A large-scale study of web password habits. In *Proceedings of the 16th International Conference on World Wide Web (WWW)*. 657–666. DOI: http://dx.doi.org/10.1145/1242572.1242661
7. Alain Forget, Sarah Pearman, Jeremy Thomas, Alessandro Acquisti, Nicolas Christin, Lorrie Faith Cranor, Serge Egelman, Marian Harbach, and Rahul Telang. 2016. Do or Do Not, There Is No Try: User Engagement May Not Improve Security Outcomes. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*. 97–111.
8. Andrew Gelman and John Carlin. 2014. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. *Perspectives on Psychological Science* 9, 6 (2014), 641–651. DOI: http://dx.doi.org/10.1177/1745691614551642
9. Joel Hruska. 2016. Forbes forces readers to turn off ad blockers, promptly serves malware. (January 8 2016). http://www.extremetech.com/internet/220696-forbes-forces-readers-to-turn-off-ad-blockers-promptly-serves-malware
10. Iulia Ion, Rob Reeder, and Sunny Consolvo. 2015. "... no one can hack my mind": Comparing Expert and Non-Expert Security Practices. . . . *On Usable Privacy and Security (SOUPS . . .* (2015).
11. Ivar Krumpal. 2013. Determinants of social desirability bias in sensitive surveys: a literature review. *Quality & Quantity* 47, 4 (2013), 2025–2047. DOI: http://dx.doi.org/10.1007/s11135-011-9640-9
12. Robert LaRose, Nora J Rifon, and Richard Enbody. 2008. Promoting personal responsibility for internet safety. *Commun. ACM* 51, 3 (March 2008), 71–76. DOI: http://dx.doi.org/10.1145/1325555.1325569
13. Paschal Sheeran. 2011. Intention—Behavior Relations: A Conceptual and Empirical Review. *European Review of Social Psychology* 12, 1 (2011), 1–36. DOI: http://dx.doi.org/10.1080/14792772143000003
14. Kami Vaniea, Emilee Rader, and Rick Wash. 2014. Betrayed by Updates: How Negative Experiences Affect Future Security. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2671–2674. DOI: http://dx.doi.org/10.1145/2556288.2557275
15. Rick Wash. 2010. Folk Models of Home Computer Security. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*. Seattle, WA. DOI: http://dx.doi.org/10.1145/1837110.1837125
16. Rick Wash and Emilee Rader. 2015. Too Much Knowledge? Security Beliefs and Protective Behaviors Among US Internet Users. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*.
17. Rick Wash, Emilee Rader, Ruthie Berman, and Zac Wellmer. 2016. Understanding Password Choices: How Frequently Entered Passwords are Re-used Across Websites. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*. Denver, CO.
18. Rick Wash, Emilee Rader, Kami Vaniea, and Michelle Rizor. 2014. Out of the Loop: How Automated Software Updates Cause Unintended Security Consequences. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*. 89–104.
19. Ryan West. 2008. The Psychology of Security. *Commun. ACM* 51, 4 (2008), 34–40. DOI: http://dx.doi.org/10.1145/1330311.1330320