International Journal of Recent Trends in Engineering & Research (IJRTER)
Conference on Recent Trends in Electrical and Electronics Engineering (RTEEE '19)
Special Issue; March- 2019 [ISSN: 2455-1457]
DOI: 10.23883/IJRTER.CONF.20190304.009.7M9IX

Using Fingerprint Authentication to Reduce System Security: An Empirical Study

[1]S.T.Chithra, [2]P.A.Dayalin Abisha
[1,2] UG Students
Arunachala College of Engineering For Women

***Abstract*—Choosing the security architecture and policies for a system is a demanding task that must be informed by an understanding of user behavior. We investigate the hypothesis that adding visible security features to a system increases user confidence in the security of a system and thereby causes users to reduce how much effort they spend in other security areas. In our study, 96 volunteers each created a pair of accounts, one secured only by a password and one secured by both a password and a fingerprint reader. Our results strongly support our hypothesis—on average. When using the fingerprint reader, users created passwords that would take one three-thousandth as long to break, thereby *potentially* negating the advantage two-factor authentication could have offered.**

***Index Terms*—user study, security policy, risk compensation, two-factor authentication**

I. Introduction

Passwords remain the weakest component of many im-portant security systems, so there is a concerted push from many directions to supplant or supplement passwords with less-fragile security measures. While this push has had some effects, particularly in environments that require more security, it has failed to replace passwords—the vast majority of com-puter users still use passwords on a day-to-day basis. Since the security of passwords relies so heavily on user behavior, studies that empirically examine patterns of password creation and use remain important in the evaluation of security policies.

Most empirical studies of user password behavior provide baseline information about how users create, remember, or use passwords [1]–[3] or the kind of passwords users create [4]. Some offer practical results based on comparing user groups given different sets of advice [5] or by persuading users to use stronger passwords [6]. A larger number of user studies have been done on alternatives to passwords, such as graphical passwords [7] or biometric authentication [8], though most of these have studied usability or subjective preferences rather than the security-related behavior of the users. Our study fills a void in the examination of a multi-factor authentication scheme by reviewing how users behave in such a system compared to how they behave in similar systems that use only single-factor authentication.

When policymakers decide to add to or improve upon password authentication rather than replace it with a different authentication scheme, the default assumption is that the ad-ditional layer or layers of security are essentially independent of other layers; that adding a security mechanism to a system will not decrease the overall security of that system. From a software perspective, this assumption is generally true and can usually be verified. However, the security of a system is inextricably linked to and dependent upon the behavior of the authorized users of that system. If adding visible security features increases the level of confidence those users have in the security of the system, they may decide to expend less effort in behaving securely. This is an outcome predicted by risk compensation theory [9] or risk homeostasis theory [10], but has not been previously tested in the arena of computer security.

Our approach to partially address this question was to offer volunteers a monetary incentive to secure accounts they were told were going to be attacked. Each volunteer was asked to create two accounts with distinct passwords: one protected only by a password and one protected by both a password and a fingerprint reader—the most commonly used biometric for authentication. Users were given $5 to divide as they saw fit between the two accounts and told that they would receive whatever money was in the accounts so long as that account was not compromised by the end of the study.

In addition to noting the evidential measure of security each user assigned to their accounts (the ratio of money put in each account), we asked the users directly about their confidence in the security measures protecting the two accounts. Finally, we asked some demographic questions in order to assess potential differences between user groups. We also observed users to see how many wrote down their passwords, but these data are unpaired to user accounts, so we cannot draw conclusions from it.

In order to analyze the results of our study we measured the strength of user passwords using a hybrid metric that attempts to estimate the number of guesses needed to guess each password. We used entropy measurements in this metric, but derived it primarily from the performance of several different password-cracking techniques on each password. Using this measure of strength, we found a large variance for how different each participant's pair of passwords was. However, across our entire study, the passwords connected to fingerprint-using accounts were $\frac{1}{2980}^{th}$ the strength of passwords for the password-only accounts. The decrease in strength was even more marked for several subgroups within our study, and we found strong correlations between a decrease in password strength and an expectation of the fingerprint reader providing

DOI: 10.23883/IJRTER.CONF.20190304.009.7M9IX

good security.

II. Previous Work

As mentioned in the introduction, many studies have been done to establish the baseline behavior of users, including the outdated but early study presented by Zviran and Haga [11] and the extremely large study of over half a million users by Florencio and Herley [1]. Moving from those baselines to studying the effects of other policies on user behaviors, a number of empirical studies have determined factors that may improve the quality of user passwords, such as the study of 288 Cambridge Natural Science first-year students [5]. Other empirical studies have developed systems to improve the passwords users produce, such as the Persuasive Text Password system developed by and tested on 83 students of Carleton University in Ottawa [12].

However, the evaluation of policies designed to improve user passwords is significantly different from the purpose of our work, which aims to determine the unintentional and potentially negative consequences of policies and security architecture. Research has been done into unintended negative consequences of password policies [13], determining that attempts to increase security can have adverse affects [14]; this aligns with overcompensation which potentially reduces or eliminates the effectiveness of security improvements [15].

An excellent analysis of different combinations of authentication techniques is given by O'Gorman [16]; however, they model each technique in isolation and treat combinations of them as an independent combination. For this reason, our research raises warnings related to their concluding advice in the areas where they recommend two-factor authentication.

In the area user-focused security, there is limited data and analysis of user motivation regarding their security posture and behavior. Weirich and Sasse [17] report the results of a series of user interviews about password behavior and develops suggestions of how to phrase arguments to persuade users to adopt better security behavior. Stanton et al. [18] likewise interviewed users for the purpose of discovering the range of user security behavior in a wide variety of contexts—not just in password choice and use. The laboratory-based users studies that the authors are familiar with mostly attempt to test new security systems, such as the evaluation of D´ej`a Vu by Dhamija and Perrig [19] and one of the evaluations of the PassPoints system such as done by Wiedenbeck et al. [20].

On the other hand, our hypothesis is intricately tied to the theory of risk compensation in psychology, that when people feel less apparent risk they will be less cautious and expend less effort to ensure their own security. Taken to an extreme, risk homeostasis theory proposes that people will modify their risk-taking behavior to always achieve the same level of risk, regardless of the safety of the environment. These theories have been extensively studied in various areas of human behavior [15], [21]–[23]. These theories have also been applied to computer security as theoretical applications [24] and explanations [25] or through qualitative surveys [26]. We seek to extend this by documenting a well-controlled experiment to determine the presence and strength of this effect in password choice.

III. Methodology

Our experiment consisted of a user study conducted in a laboratory setting that asked users to create two accounts each; we recorded the two passwords each user created. We describe the design of the user study itself in Section III-A. For analysis of the collected data, adequately measuring the strength of passwords collected in this study was of primary importance. We discuss what such measurement means in Section III-B.

The methods that we decided to use in assessing the strength of the collected passwords included various modes of John the Ripper, discussed in Section III-D, and two analytic measures. Our use of John the Ripper required password dictionaries described in Section III-C; these password dictionaries were also used to establish a baseline expectation about the dis-tribution of password strengths and are therefore themselves relevant in our later analysis. The analytic measures we used to gauge password strength were class size estimation, discussed in Section III-E, and entropy, discussed in Section III-F.

A. Experimental Procedure

In order to determine how our subjects chose passwords in different security environments, we wrote a desktop application that guided each volunteer through an account creation process, similar to creating an account on a website. Each volunteer was asked to create two accounts: a "standard" account secured only by a password and a "biometric" account secured by both a password and a fingerprint. The order that volunteers created accounts was randomized between subjects. After they had created both accounts, they were asked questions to assess how secure they felt the two accounts were, as well as questions regarding their security background.

1) Recruitment: After we received approval from our Institute Review Board (IRB) to run a psychological study with human participants, we began recruiting students. Our volunteers were recruited primarily through fliers posted around our college campus offering $5 to participate in an experiment to test hackers. Some volunteers heard about our experiment through word of mouth, and toward the end of the experiment we set up a booth with a large poster outside the dining hall in the student union building.

2) Objective and threat: In order for our study to obtain realistic passwords, before creating the accounts, each volunteer was given instructions indicating the following:

- The researchers were going to put $5 in total into the accounts created by each volunteer.
- The accounts would be attacked by a group of hackers who would attempt to gain access to each.
- Any money in any account the hackers compromised would go to the hackers.
- Money remaining in a volunteer's account after one week would be given to the volunteer.
- Passwords, compromised or not, would be released at the end of the study, so users should avoid using an existing password.

DOI: 10.23883/IJRTER.CONF.20190304.009.7M9IX

3) Phase I: Our first group of 20 volunteers was given an account number of the form 12-3456. They were told that they would need to remember their account number to log in at the end of the experiment.

Observing the participants, 19 of the first 20 volunteers wrote down their password when they wrote down their account number. Since writing down passwords was not a behavior we had intended to cause, we revised the experiment.

4) Phase II: The remaining 74 volunteers were asked instead to create a username along with their password. Many users still wrote down their password and/or username, approximately 54%, but far fewer than in Phase I.

5) Process: The account creation process required that the users enter the same password into two text boxes (obscured by dots) and checked that they were identical before proceeding. The only password requirements were that the password box be non-empty and that the password consist exclusively of printable ASCII characters other than newlines or carriage returns. Additionally, the account that purported to use a fingerprint scanner required that the user swipe his or her finger along a provided fingerprint scanner. Although this is an unrealistic behavior, we displayed a scanned fingerprint on the screen as visual feedback when the user correctly scanned his or her fingerprint. All users that were asked assumed (as intended) that the displayed fingerprint was the scan of their own finger; in fact it was a randomly chosen scan of one of the researchers' fingerprints. No fingerprints were actually collected as part of this study.

6) Evaluation: After each user finished creating both accounts, he or she was asked to allocate $5 between the accounts in whatever way he or she preferred. This was intended to give us a practical measure of how confident each user felt about the security of each account. Next, the user was directly asked about the security of the accounts. Finally, the user was asked a few demographic questions intended to help us interpret group differences in the data.

7) Compensation: Users were asked to return to our lab at least one week after they set up accounts to see if their accounts had been compromised and get rewarded if the accounts were intact. They were also told that they would need to re-enter their password at this time to get paid. The only reason for imposing the delay was to disguise the fact that all volunteers would be paid no matter what (since there were no actual hackers). Most students never returned to collect their $5, indicating that the reward was not substantial for them and was not the primary motivator in participating. There were insufficient users who returned to collect their compensation to determine whether there was any significant difference in their security posture, as compared to those who did not return.

B. Password Rating

To evaluate differences in password choice, we developed a numerical metric for rating the strength of passwords. Rather than a single monolithic metric, we opted to use or model several cracking techniques. For each technique, we estimated how much time—relative to the computing power available— that technique would take to guess or break each password. Our final metric is formed by taking the minimum of all of these, as any of them individually could be used by an attacker.

The strength of a password can be regarded as a function of how many guesses a sophisticated attacker focusing only on the security of the password itself (ignoring any other aspect of the overall security system) would have to make before happening upon the correct password. There are two features which make this measure inherently variable; one is very predictable, the other is very unpredictable.

System design is much more predictable in the variability of password strength and reflects the degree to which the design of the authentication system itself restricts the number of passwords that can be tested per second. For example, system design includes the choice of password-hashing algorithm, the policy limiting login attempts, and how well password hash databases are secured. Ultimately, system design strengthens or weakens all passwords equally. Thus, any measure of password strength should take system design into consideration only as a constant factor. Since our analysis reflects comparisons between passwords used on an arbitrary system, we omit this feature.

Ordinality is much less predictable and reflects the order in which passwords are guessed by *any* hypothetical attacker. It may be efficient for a knowledgeable attacker to semi-manually guess a series of passwords related to names, contacts, dates, and interests known to be relevant to a target. In most cases, it is worth testing against known password lists, sorted by frequency of known usage. Even if neither of these work, there are several independent methods of brute-forcing passwords that result in different password orderings.

The difference that different password guessing orders can make is significant and unpredictable. The password "`j_IGc5}7vTky(wQr`" could be an exceptionally good password, but by virtue of being printed here and possible to index, it could end up in a wordlist less than one hundred million words long, making it crackable in a very short amount of time.

This principle also extends to other features commonly associated with the strength of a password; for example, a password that a user uses on multiple systems is inherently weaker in each system than it would be otherwise, because there is a potential risk of a first compromised account affecting a second. However, this weakness through password reuse is really an example of a knowledgeable attacker being able to lower the ordinality of that password within his or her guess sequence.

Despite the obvious impossibility of knowing in which order an arbitrary attacker will guess passwords, ordinality almost exclusively represents the strength of a password, and any password-strength metric is an attempt to estimate the ordinality of a potential password in a hypothetical but representative list that an attacker might go through. We chose to directly assess the index of passwords in various sequences that an attacker might try and choose the minimum of all of

DOI: 10.23883/IJRTER.CONF.20190304.009.7M9IX

these as a reasonably likely ordinal drawn from the distribution of likely guess sequences. This metric has the advantage that it should be no more than a small constant factor smaller than the actual number of guesses an attacker would take. It could be much larger, especially if the attacker has special knowledge about a victim's password, but should otherwise be representative of real-world cracking attempts.

1) Online versus offline attacks: Our analysis implicitly assumes that passwords created are vulnerable to offline attacks, where attackers have, for example, access to the hashes of the passwords they wish to crack. Most security schemes are designed to prevent such an attack, including the kind of bank service we were mimicking, but assuming that passwords are only vulnerable to online attacks also assumes perfect implementation of additional security safeguards. In additional, we should not assume only online attacks because many publicly available dictionaries of passwords and password hashes were originally obtained by compromising web servers. With so many out there, we are forced to the conclude that online services are quite often vulnerable to offline attack.

C. Password dictionaries

The first and easiest technique for most attackers is to try testing known passwords, as drawn from large-scale account leaks. Based on the demographics of the pool of volunteers our study drew from, we used a dictionary of English words (approximately 900,000 words) and combinations of English words (48,400,000,000), but also obtained and used password lists from famous phishing or SQL–injection related leaks, including from MySpace, Hotmail, VKontakte and RockYou.

The MySpace list was originally published in 2006, after an attacker created a fake MySpace login page and phished tens of thousands of users. The attacker or attackers left the server they were collecting passwords on unprotected, and their list was compromised and later publicly posted. Our version of this list contains 49,711 passwords, although not all of these passwords actually belonged to users, since some people recognized the phishing attempt and entered fake data. Like other researchers [27], we make no attempt to distinguish between the fake and the real passwords.

The Hotmail list was similarly collected from a phishing attempt in 2009. The original list contains username and password pairs from usernames starting with 'a' or 'b' and has 9,856 entries. Also in 2009, the social site RockYou.com suffered a data breach. Since the site operators stored passwords in plain text, the passwords for 32,603,387 users were exposed. Lastly, 55,766 unique usernames and passwords were collected from the Russian social networking site VKontakte.ru by the trojan Trojan.Win32.VkHost.an, which redirected users of an infected host to a phishing site whenever they attempted to access the real VKontakte.ru.

We used these four lists to determine the strength of passwords collected by our study and as a baseline for our expectations of password strength. For the former purpose, we also used the list published by Openwall [28], consisting of more than 3,400,000 unique passwords, sorted by the expected frequency of use.

D. John the Ripper

John the Ripper [29] is probably the preeminent password cracking software. We employed three typical modes of using John the Ripper. We spent approximately 24 hours running each mode on the passwords we collected, and recorded the number of guesses taken to get to each password in our set.

1) Mangling: John the Ripper has a "mangling" mode where it takes passwords from an existing list and mangles them according to a set of rules. This can be very useful for specific modifications, such as adding "1" or "!" to the end of each password, or replacing the letter "o" with "0". Using all the mangling rules dramatically expands the input password list.

2) Iterative: John the Ripper's "iterative" mode is capable of brute-forcing all passwords of up to 8 characters drawn from the set of non-whitespace printable ASCII characters. Under normal operation, this is the default mode after the wordlist provided to John has been exhausted. In our study, we used it essentially in parallel with all the other modes. Although all possible passwords are examined, the examination order is somewhat unorthodox. For instance, the first five passwords tested are (in order) *1952*, *sammy*, *stark*, *start*, and *stack*, which are by no means the most commonly used passwords.

3) Markov: There is now a "Markov" mode plugin for John the Ripper that uses character digram frequencies[1] to build a finite-state machine able to generate passwords according to the Markov-chain estimate of their likelihood. Using the Markov mode, John the Ripper can generate passwords of any length, but the complexity of the password is bounded by the frequency with which the digrams in the password appear in the Markov-mode's database, and passwords above a stated complexity threshold will be skipped. This is theoretically similar to the iterative mode, but more sophisticated and able to generate more realistic and longer passwords.

Limiting the maximum password length of the Markov mode or changing the maximum complexity of generated passwords can result in a dramatically different password generation order. So, for our purposes, it is worth taking the index that results for each password, when the Markov mode is run for that specific length, and with a maximum complexity that exactly matches the password. This results in the smallest set that is guaranteed to contain the password, and thus results in the smallest reasonable estimate.

The `mkvcalcproba` utility, provided with the Markov plugin, directly computes the complexity of each password (as measured by its own generator), even for passwords with complexity higher than the Markov mode can generate. For these passwords that lie above the complexity that the Markov mode of John the Ripper will generate, we take half the theoretical number of such passwords as a reasonable estimate of how many passwords the Markov utility would go through before

[1] A digram is two characters in sequence. Digram frequencies measure how often a character follows the previous character in a corpus.

DOI: 10.23883/IJRTER.CONF.20190304.009.7M9IX

getting the correct password. The number of passwords of less or equal complexity is generated by the `genmkvpwd` utility, but for large passwords the number it computes overflows the 64-bit integers it uses, so it is inapplicable to long or very complex passwords.

E. Class size estimation

As a fallback analysis, we also provide a strictly analytical measure that still represents a potential method of brute-forcing passwords, but does not correspond to an actual industry technique of password cracking.

For each password, we count the number of members it has in each of five classes: lowercase alphabetic characters, uppercase alphabetic characters, numerals, common symbols[2], and uncommon symbols[3]. If $c_i(p)$ is the count of characters from class i for password p and s_i is the size of that class, our artificial measure is

$$m(p) = 2^{\sum_{i=1}^{5} c_i(p)} \prod_{j=0}^{c_i(p)} s_i - j.$$

This results in a measure that is always the same size or smaller than the more naïve estimation of $s^{|p|}$ where s is the sum of the size of all the character classes that p uses. Essentially, our analytic measure distinguishes between an alphabetic password with a single digit repeated and an arbitrary alphanumeric password of the same length. As an example, the difficult password mentioned above in Section III-B, `p = j_IGc5}7vTky(wQr`, has measure

$$3.8\times10^{28} = 2^{1} \cdot 26^{7} \cdot 26^{4} \cdot 4 \cdot 10^{2} \cdot 2 \cdot 7^{1} \cdot 1 \cdot 26^{2} \cdot 2^{16}.$$

The measure is more intuitive when explained as follows: it is the number of passwords formed by the characters from one set in any order, multiplied by the number of ways to insert all the characters from the second class, multiplied by the number of ways to insert all the characters from the next class, and so on. Our intention is to model the smallest class that would be reasonably constructed and would contain the password, and we assume that enumeration within that class would happen on average halfway through.

F. Entropy

While all password entropy measures are estimates and must be measured based on comparison of a password to a larger sample space, we also found it useful to measure the entropy of each password. The standard measure for entropy is

$$H(X) = -\sum_{x\in X} f(x) \log_2 f(x),$$

where f is the probability mass function of X. This is what is referred to as entropy and is measuring over a random variable—an entire space, rather than points within that space.

[2] those with a frequency more than 0.05% in our database of leaked passwords; these are `.`, `!`, `_`, `-`, `*`, space, and `@`

[3] other printable ASCII characters, 26 total

TABLE I
PARTICIPANT AGE

Age	Number	Percent
18–24	74	78.7
25–34	10	10.6
35–44	2	2.1
44–54	2	2.1
55+	1	1.1
No answer	5	5.3

In the context of password strength, the information content $\log_2 \frac{1}{f(x)}$ referred to as the entropy of that password. We continue to refer to the information content of a password as its entropy for the sake of consistency.

In the context of a password strength measure, we do not know and can only estimate the value of f for any given password. One memoryless approximation is to measure individual character frequencies within a corpus. If it is known that a character c occurs with frequency f_c in a corpus, each appearance carries $\log_2 \frac{1}{f_c}$ bits of information, neglecting conditioning upon other characters. Thus, by measuring the frequencies at which each potential password character appears, we can estimate the entropy of a new password as

$$l(p) = \sum_{c\in p} \log_2 \frac{1}{f_c}.$$

Thus, the number of potential passwords with entropy less than or equal to p is $2^{l(p)}$, which is just the reciprocal of the product of the individual character frequencies.

Entropy, at least by itself, is a poor measure of password strength, as demonstrated by Weir et al. [30]. However, entropy, especially when accounting for specific character frequencies, is better than nothing. For passwords that we are currently unable to crack it still gives us some information about which passwords will be cracked sooner than others if we continued to employ our cracking techniques.

IV. Results

Ninety-six individuals participated in our experiment. Two of these samples were rejected because the data our program gathered was incomplete. Since subjects were not closely monitored, we suspect that these two individuals did not finish going through the program and closed it or left before the end.

A. Sample demographics

Our 94 samples consist mostly of undergraduate university students, disproportionally drawn from the computer science department. Thus, they are overwhelmingly young: 79% were under 25 and 89% were under 35. Users were similarly well educated: 89% had at least some college experience and we expect that the vast majority of the 78% that reported "some college" were pursuing an undergraduate degree. The answers to the demographic questions are in Tables I, II and III.

DOI: 10.23883/IJRTER.CONF.20190304.009.7M9IX

TABLE II
PARTICIPANT EDUCATION

Education	Number	Percent
No high school degree or GED	1	1.1
High school degree only	7	7.4
Some college	73	77.7
Bachelor's degree	9	9.6
Master's degree or higher	2	2.1
No answer	2	2.1

TABLE III
PARTICIPANT COMPUTER & SECURITY BACKGROUND

Background	Number	Percent
Do not often use computers	0	0.0
Use computers for Internet and email only	14	14.9
Heavily use computers, no coding experience	52	55.3
Design or build software, not in security	12	12.8
Familiar with internals of computer security	13	13.8
No answer	3	3.2

B. Analysis decisions

During the analysis of the data for this paper, we adopted several conventions for the way we examined data. For example, we found histograms to very poorly represent the distribution of password strengths. Instead, we used kernel density estimates to approximate a probability distribution of the strength of passwords over a given group. In our testing, we found that simple Gaussian kernels worked very well, so we use Gaussian kernels throughout this paper. Furthermore, we found that using canonical kernel bandwidths resulted in satisfactory distributions that demonstrated neither over- nor under-smoothing, so we avoided bandwidth tuning for each of the distributions we looked at.

For many of the questions we sought to answer, we used parametric hypothesis tests that impose restrictions on the data they evaluate. In particular, we compared several data sets with t-tests, which require that the data approximate a normal distribution.

We earlier defined password strength to be the number of guesses we expect an attacker to take to break a password. For passwords successfully broken by any of our cracking techniques, this number is exact in some sense—this is simply how many guesses it took for us to break the password. For passwords long enough not be cracked, this represents our best guess as to how long it would have taken us, based on all the information we have about the password.

Since password strength grows exponentially in terms of password length, if password lengths were normally distributed, password strengths would be log-normally distributed. Although we expected such a distribution, we found that the logarithm of password strength was still positively skewed, as indicated in Figure 1.

The very large numbers used to describe such strengths is awkward, so for the remainder of the paper we refer

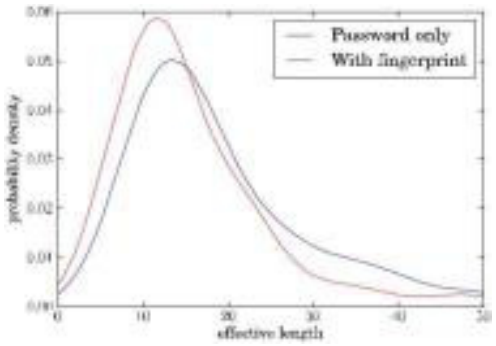

Fig. 1. Password strength of control and experimental groups

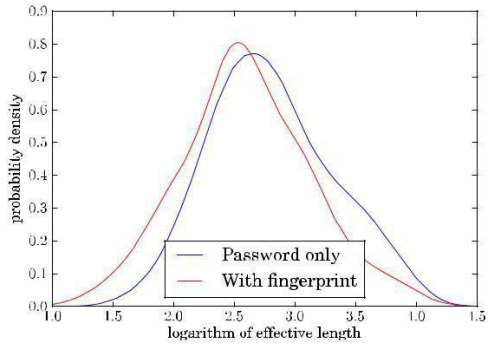

Fig. 2. Logarithmic strength of control and experimental groups

to "effective length". This is the base-10 logarithm of the expected number of guesses needed to correctly guess the password. This is a more easily interpreted metric; a password rating of 7.5 takes approximately thirty million guesses (a small number) to crack, while a password rated at 19.6 is well out of the reach of the attacks we rated the passwords against. This can also be intuitively thought of as representing a perfectly random base-10 password of this length.

However, for the purposes of quantitative comparison between groups of passwords, we additionally take the natural logarithm of effective length, resulting in distributions like those in Figure 2. These distributions are not normal either, but they are reasonably well balanced; the Anderson-Darling test for normality gives p-values of 0.4851 and 0.1920 respectively for the two distributions in Figure 2, so we have no strong reason to treat them as non-normal. Many of the graphs in this paper use this more balanced metric; their axes are labeled as the logarithm of effective length.

For the remainder of this paper, p-values reported for testing when two groups are drawn from the same distribution are computed using Welch's Two Sample t-test, using the $\ln \circ \log_{10}$ form of measuring strength. Intra-subject compar-isons use a Paired t-test, but over the $\log_{10}$ of the strength, since intra-subject differences in that measure do appear to be symmetric and approximately normally distributed. For the

DOI: 10.23883/IJRTER.CONF.20190304.009.7M9IX

TABLE IV
EXAMPLE PASSWORDS

Percentile	Study example	MySpace example
10	cab1007	jordan
25	d3rp217857	melch.
50	1Yes27Miss77	lillius4
75	anamishaizfalous9594	hunter345
90	_3_@_m!nEr@Ls_b(uEImperitor1	

hypotheses we examine in this paper we found a wide range of p-values, from $p > 0.4$ to $p < 10^{-15}$. Note that while lower p-values indicate greater statistical significance, and in particular $p < 0.05$ is often used as a cutoff for deciding whether or not to accept a hypothesis, p-values scale to both the sample size and the effect strength, both of which are different among our different hypotheses. Thus, there are some hypotheses we expect are true but we cannot accept given our sample sizes and hypotheses with very low p-values that are associated with a small enough effect that they should not be taken to represent important findings. In other words, alternate hypotheses that we appear to support with $p = 0.001$ (as the only alternate hypothesis) should not be taken to be ten times the significance of other alternate hypotheses we appear support with $p = 0.01$.

To mitigate concern surrounding the transforms we used on the data and their distribution, we evaluated our findings with both parametric and non-parametric statistical tests. In all cases, the resulting p-values were not meaningfully different— they did not change any conclusions. Because the non-parametric tests cannot assist in building confidence intervals, we usually omitted mention of them except in cases where reviewers specifically asked about alternate tests.

C. Password strength

The passwords collected during our study differ dramatically from passwords sampled from other sources, e.g., the large MySpace leak. For instance, we rate the passwords in Table IV as the 10^{th}, 25^{th}, 50^{th}, 75^{th} and 90^{th} percentiles from our study and the famous MySpace phishing leak respectively.

It should be evident that the passwords our study col-lected are dramatically stronger than what we expected to collect. More descriptively, Figure 3 overlays estimations of the probability distributions of the password strength over the Hotmail, MySpace and RockYou datasets, comparing them to the estimated probability distribution of password strength found in our study. As in Figure 2, these plots use the natural logarithm of effective password length.

It is also worthwhile to note that even the weaker passwords associated with the fingerprint-using accounts almost always (95% of the time) still exceed the minimum strength recommended by Florˆencio et al. [31] for web passwords. In many environments (such as general web usage), passwords stronger than this are superfluous, and add an unnecessary burden to users.

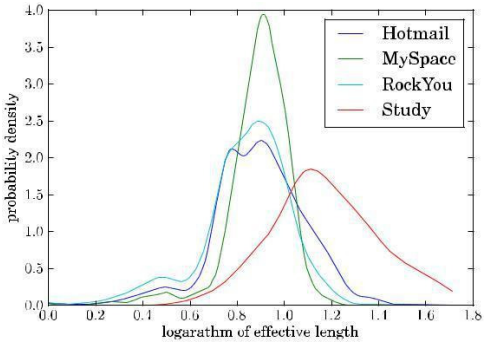

Fig. 3. Password strength from different datasets

There is really no possibility that the passwords in our study could have been drawn from one of the other distributions by coincidence; in all cases, the p-value for that hypothesis is less than 10^{-15}. The variance is dramatically larger, and the average difference in strength is incredible. The closest dataset is the Vkontakte set, which has an average rating "only" 7.14 less than the passwords in our study—a nearly fourteen million fold difference, whose relative smallness may be best explained by the fact that it is a primarily non-English set of passwords that are less easily cracked by the tools we used. The other datasets compare as 8.65, 8.97 and 9.62 points weaker than the dataset collected in our study for Hotmail, MySpace, and RockYou respectively. The latter can perhaps be explained in being somewhat larger because, unlike the other sources, RockYou did not require both letters and digits in the passwords of their users (Hotmail and MySpace did prior to their famous leaks).

Although we were unable to conclusively demonstrate that any of these factors was the cause of the difficult passwords we observed, we believe that there are a few relevant factors at hand.

1) Demographics: Our users were by and large more computer savvy and more familiar with computer security than a random sample of the Internet-using population, at least as compared to our password sets and reports from other password studies. This is especially true in comparison to some of our test sets, such as the MySpace list, which contains only passwords of people who fell prey to a phishing attack. As such, it is not surprising to see better-constructed passwords that are longer, more complex and less vulnerable to guessing than we would see in an unbiased sample.

However, using passwords too long or too complex to remember is *not* a hallmark of good password habits. Since so many of the passwords we collected are too long and contain too many special characters to remember or type in easily, we believe that demographics play a smaller role than the other factors as will be discussed next.

2) An obvious threat: Adams and Sasse [32] propose that the fundamental problem with password security is that security is an obstacle to achieving tasks. In standard use

Special Issue; March- 2019 [ISSN: 2455-1457]
DOI: 10.23883/IJRTER.CONF.20190304.009.7M9IX

cases, passwords are seen as an impediment to productivity; unless users have a very strong drive to protect the specific resource they believe they are protecting (a MySpace user page, for instance) and a strong belief that it is or will be under attack, users are disinclined to use any password more difficult than the minimum required. These results are confirmed by Florencio and Herley [1] and backed up by an economic analysis by Herley [14].

Our experiment did not mirror this presumption; users were told that they were protecting a bank account from a specific, concerted attack by a group of hackers. Security was their primary objective, not an ancillary obstacle. More over, instead of having to log into an account many times over an extended period of time, users were only required to remember their passwords for one week and log in once at the end of the period.

We believe this is the most significant factor and explains why a significant majority of the users chose a password they did not expect to remember—we noted that 68% of users visibly recorded their password in some way: some wrote on their hands or arm, some on paper, and most in their phones. A small number of users had lost the password they had written down by the end of the week and none of these were able to log back into their accounts; if these users had intended to memorize their passwords, they failed to do so before logging back into their accounts.

3) Competition: In addition to having an obvious threat, the experiment provided users with a well-defined competition. Each user believed he or she was directly competing against a group of hackers. Discussions we overheard from users that had just finished creating their accounts indicated that many decided they were also in competition with the other volunteers of the study; we heard sentiments such as "I bet my password is better than yours!" and "There's no way the hackers are going to break into my account!" from several users. The $5 we offered users is a paltry sum that failed to convince many users to volunteer, but the competition itself seemed to be a strong motivator once they had agreed to take part.

These final two observations, if true, provide interesting material for potentially convincing users to create and use stronger passwords. Regardless, we think that our general results are valid because we compare subjects in the same environment given the same objective to each other, rather than comparing the passwords that our subjects created to those in our test sets. In this comparison, we still find striking and statistically significant differences.

D. Group comparisons

Treating the set of passwords collected from volunteers as two separate groups, a control (the password-only accounts) and an experimental group (the accounts using a fingerprint), rather than as paired samples, we can determine substantial differences. Recall that users were required to use different passwords. One significant aspect of this is that we can use only the results of the first of each pair of account creations, avoiding a potential bias that might be introduced through different behavior on the second account creation. Thus, we compare the 44 volunteers who were randomly chosen to first create a fingerprint-using account against the 50 who first created a password-only account. The difference in password strength, which is a function of length, between their first account created is visible in Figure 1, although the difference in means between the two groups may seem minimal. In fact, we can assert that the distributions are different with $p < 0.003$ (against the null hypothesis of the means being equal).

The strongest position we can assert from a strict group comparison with $p < 0.05$ is that the mean of the experimental group is at least 0.09 below that of the control group. Since the examined data is the natural logarithm of our effective length, this corresponds to an effective length difference of at least 1.094, representing slightly more than a 10-fold reduction in the expected time to crack each password. We cannot assert from group data alone a small confidence interval on the true difference of means. Unfortunately, the variance in password strength between individuals is very large, so this is the limit of our between-subject results. For reliable conclusions on a sample size this small, we need to use paired comparisons.

To examine our results in the most "realistic" scenario, we can opt to exclude any of our abstract measures, and only examine the numbers from the passwords that we were able to crack in the single day of computer time that we allocated to the problem. In this analysis, only 16.8% of password-only accounts had passwords weak enough to crack, while 28.4% of the passwords from the fingerprint-using accounts were cracked. There was little overlap between these sets; only 2 (2.1%) of users had both their passwords cracked. Because so few passwords in our study were actually cracked by John the Ripper, we found the abstract measures quite important.

E. Paired comparisons

By blocking and ensuring that our samples are paired, we can reach much stronger and more specific conclusions. Examining paired data, we find that the differences in password length are significant with $p < 0.0000003$ (two-tailed test, $t=5.556$ with 93 degrees of freedom). More importantly, the 95% confidence interval of the true difference in means can be bounded to (2.23, 4.72), corresponding to a range of 170 to 52,000 in the expected factor by which passwords would be easier to crack. The difference in sample means is 3.47, corresponding to a nearly three-thousand fold decrease in expected time to crack. Non-parametric testing produced results just as dramatic; the Wilcoxon signed rank test has a p-value $< 10^{-7}$.

Furthermore, more than 77% of participants used a more difficult password for their password-only account than their fingerprint-using account, so the effect—that users used weaker passwords when also using a fingerprint reader— was not only large, it was widespread. Perhaps another 15% of users created passwords that were virtually identical in construction or difficulty; some but not all of these were given very similar strength scores by our metric. These passwords are displayed in Table V. One password is omitted because of

DOI: 10.23883/IJRTER.CONF.20190304.009.7M9IX

TABLE V
SIMILAR PASSWORD PAIRS FOR USERS THAT USED MORE DIFFICULT PASSWORDS IN THEIR FINGERPRINT-USING ACCOUNTS

Password-only account	Fingerprint-using account
SD1429	R714TP
2nd4cc	!st4cc
4721jeffdum	4721jeffguy
A8294e9c95...	~A8294e9c95...~
AlexPoli	AlexPoli2
M1n@B3v$	MnBv!2#4
!I91MovK	P?w2U8!y
cookies87	anyone87
mahasamatan	nyarlethotep
tS.k)4xsh3(m3.	Cr_(@d)8m7Xx_o.4
aple3%#8	w?,3EqQ7
4udr3yl1lly	p3anutbu77er

length; it was the same 34 character string, with a '#' character appended to the password of the second account created. This example illustrates why very similar passwords might not be given similar ratings, because an extra '#' character can be expected to increase the search space for a password by approximately 100-fold. While comparing the similarity of passwords is both somewhat subjective and faults from assigning a binary value to a continuous relationship, our analysis suggests that only 6 of 94 users created dissimilar password pairs where the easier-to-crack password was assigned to the password-only account.

F. Comparison within subgroups

Given that the overall difference between password strengths across all participants was so large, it is worth examining to see how that difference is distributed across the demographic subgroups that participated in our study. Unfortunately, our sample was not well distributed across the questions that we asked, so many of the demographics that we looked at do not present enough samples to be statistically significant. Regardless, we present all responses, albeit with caveats for each subgroup which is too small to draw conclusions from.

1) Age: In the researchers' experiences, other factors being equal, people who are young enough to have grown up with computers have a better understanding of security, even if they do not necessarily behave in a substantially more secure way. Figure 4 presents the observed differences. Note from Table I that we have too few volunteers 35 or older to draw serious conclusions. However, the differences between the 18–24 and 25–34 groups exceed the differences within those groups when using or not using the fingerprint reader, indicating that the magnitude of the demographic effect is smaller than the magnitude of the effect of adding a biometric.

In Figure 4 and similar figures later in the paper, we indicate the uncertainty of our estimation of the true mean through error bars representing a 95% confidence interval. These error bars

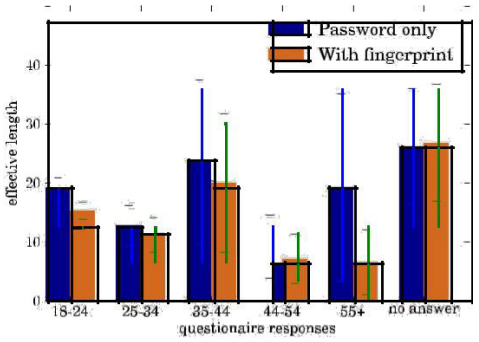

Fig. 4. Password strength by participant age

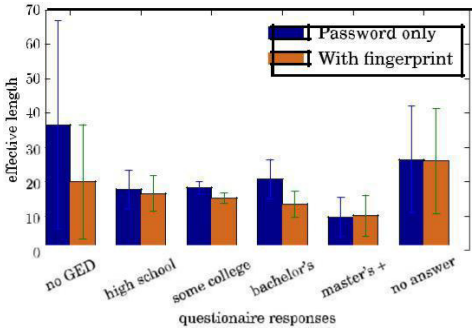

Fig. 5. Password strength by participant's educational level

would ideally be produced through bootstrapping, but since our sample sizes are often small enough to make bootstrapping produce optimistic (tight) confidence intervals, we use the Gaussian-based asymptotic approximation recommended by McGill et al. [33], but force a minimum interquartile range (R) of at least the minimum observed difference among *any* passwords of similar mean strength. Even after this, we feel that the error bars presented are overly optimistic for our smallest subgroups and likely under-represent the true variance.

2) Education: Like age, education is sometimes considered a factor in security awareness. Note that the large majority of our participants are currently undergraduate students in the "some college" category. There are enough students in the high school and bachelor's degree groups that it would be reason-able to draw conclusions, if the differences between groups were large, but in this case, no differences are significant at these sample sizes. For example, $p \approx 0.2$ for the hypothesis that the difference between the password-only and fingerprint-using accounts is larger for those with bachelor's degrees than those with high school degrees; $p \approx 0.4$ for the hypothesis that the password-only accounts of users with bachelor's degrees have stronger passwords than the password-only accounts of users with no college education.

DOI: 10.23883/IJRTER.CONF.20190304.009.7M9IX

3) Computer and security background: Participants were asked questions about their familiarity with software and computer security, as detailed in Table III. The answers to this question were much more evenly dispersed than answers to other demographic questions, enabling us to look seriously at the results, as all groups in Figure 6, save the "no answer" group, are large enough to be interesting. For Figure 6, we abbreviated the backgrounds listed in Table III to, in order, "non-user", "basic user", "standard user", "programmer", "security expert", and "no answer".

In this case, although the differences between groups are mostly smaller than the differences between the password-only and fingerprint-using accounts, the differences are still large. These samples sizes are still not quite large enough to support statistical significance; the least likely difference to have occurred by chance is that between the password-only account created by those who answered "I use computers for many tasks, but do not develop software" ("standard user" in Figure 6) and those who answered "I am familiar with the internals of computer security systems" ("security expert" in Figure 6), but this has a p-value of 0.07, so it misses standard criteria for statistical significance. However, by combining the samples of those who responded that they have a familiarity with programming with those who responded that they were familiar with the internals of computer security systems results in a p-value of 0.039, and this difference is quite large—a password rating of 4.42, representing a 26,000-fold increase in the time expected to crack the passwords of programers and security experts compared to those who merely use computers extensively. Because we are testing multiple demographic hypotheses, a p-value of 0.039 should not necessarily be thought of as significant. However, this result is enough to suggest further examination since the differences between groups are large—the only reason the p-values are high is the relatively small numbers in each sub-sample.

The computer science majors that participated in the experiment also did not appear to alter our results. While we cannot exactly classify participants based on the answers to this question, we can see that among all groups, password strength was noticeably decreased for the fingerprint-using account, and by relatively similar proportions.

4) Number of passwords used: As another measure into gaining insight into the security awareness and behaviors of our participants, we asked each user how many passwords they used on a regular basis. The choices we provided were overly broad, and most of our users fell into only two categories. The responses to that question are in Table VI.

As with several of the questions intended to distinguish group membership, only two of the groups are large enough to begin to draw conclusions from, but we present the results of all groups in Figure 7. Nearly all our volunteers (93%) reported using between 2 and 11 passwords on a regular basis, with the majority using between 2 and 5 (72% of all respondents). This matches expectations provided by Florencio and Herley [1], especially when conditioning those expectations upon the relative proportions of tech-savvy subjects in our studies.

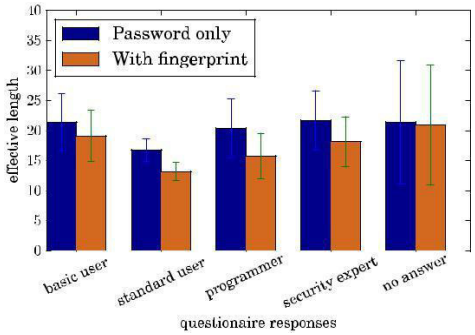

Fig. 6. Password strength by user's computer and security background

TABLE VI
NUMBER OF PASSWORDS USED ON A REGULAR BASIS

Number used	Count	Percentage
0-1	2	2.1
2-5	66	70.2
6-11	21	22.3
12-19	1	1.1
20+	2	2.1
No answer	2	2.1

The apparent difference between the 2–5 group and the 6–11 group is small but present for both the password-only and the fingerprint-using account. It is not, however, large enough to reach statistical significance ($p \approx 0.086$) even when comparing all passwords of the first account to all passwords of the second account. However, it is still reasonably likely that there is a true group difference.

5) Password managers: As a final metric to evaluate the security awareness and practices of our user groups, we asked each participant the question "Do you store or generate passwords so that you do not have to remember them all?", with the intention of determining whether they use some sort of password manager in order to use a larger number of passwords and/or use more complex passwords. The answers

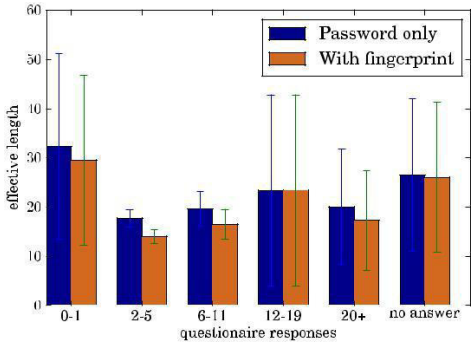

Fig. 7. Password strength by number of passwords typically used

DOI: 10.23883/IJRTER.CONF.20190304.009.7M9IX

TABLE VII
PASSWORD MANAGER USE

Used	Count	Percentage
No	54	57.4
Yes	31	33.0
No answer	9	9.6

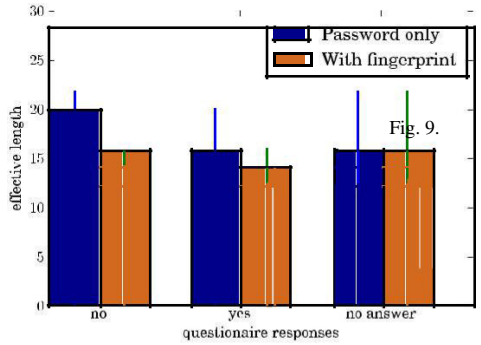

Fig. 8. Password strength by password manager use

to this question appear in Table VII.

We were impressed by the large number of users that claimed that they used some sort of password management system. However, any difference that this might indicate in security consciousness was not reflected in the passwords that our study collected ($p = 0.43$). In fact, as seen in Figure 8, it appears that the subgroup that uses a password management system produces slightly weaker passwords, a surprise to the researchers—prior to the experiment, we had a mild expectation that users who used password management systems were more security conscious and would create better passwords. It is possible that those users may be less experienced at coming up with good passwords on the spot, either because of or leading to their use of a password management system. However, we cannot authoritatively state that any observed difference was not a matter of chance.

G. Comparison by faith in biometrics

Our central hypothesis is that by giving users a second authentication mechanism, we risk inflating their sense of security in the overall system to the degree where they choose less secure passwords to fulfill their own part in securing that system. We can confirm this hypothesis insofar as, in the limited circumstances of our experiment, we found a remarkable difference in the strength of the passwords created by users expecting only those passwords to secure that account and users expecting a fingerprint to be also required to gain access to that account. Examining only the strength of passwords means that we can support the effect our hypothesis predicts, but not the actual mechanism of that effect. Supporting the cause we propose for the effect requires that we also somehow measure the degree to which each user actually had faith in the security of a fingerprint as an authenticator. In order to

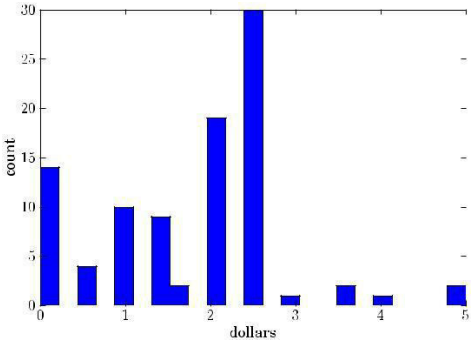

Fig. 9. Allocation of money to the password-only account

enable this sort of analysis, we asked several questions of each participant prior to asking the demographic questions.

1) Account allocation: The first question we asked participants, intended to be the most revealing, was when we informed each participant after they had already created their accounts that the $5 that was to be deposited in their accounts could be divided as they saw fit. They were given a slider control and asked to allocate $\$n$ to their password-only account and $\$(5 - n)$ to their fingerprint-using account.

Despite the fact that participants could choose any dollar and cents amount for this question, most chose round numbers such as $2.50 or $2.00, as can be seen in Figure 9. The mean allocation was $1.76 to the password-only account ($3.24 to the fingerprint-using account), seeming to indicate that even though passwords were weaker for the fingerprint-using account, participants had more faith in the combination of a fingerprint reader and a (usually weaker) password. This undershoots the prediction of risk homeostasis theory, since users still believe that they are more secure even after reducing the complexity of their passwords[4]. However, it matches most expectations of risk compensation theory that expect additional security measures to be sub-additive [10].

It is of some interest that participants overall had greater faith in the fingerprint-secured account, but it is of more interest that the more faith an individual participant had in the fingerprint-secured account, the weaker that participant's password for the fingerprint-using account was compared to their other password. Figure 10 shows the graphical breakdown. In this and future figures, a positive difference indicates that the password for the password-only account was stronger than the password for the fingerprint-using account. The correlation is not strikingly obvious (the best fit line is dashed in green), but Pearson's product-moment correlation for those two variables is -0.468, and the hypothesis that there is no correlation between them has a p-value of $p < 0.000002$, indicating that there is actually quite a strong degree of correlation between

[4] Users didn't make their passwords weak enough (according to their own subjective appraisals) to compensate for the security they expected the fingerprint reader to add; they judged the system as more secure overall, but homeostasis theory implies that it should be the same for both accounts.

DOI: 10.23883/IJRTER.CONF.20190304.009.7M9IX

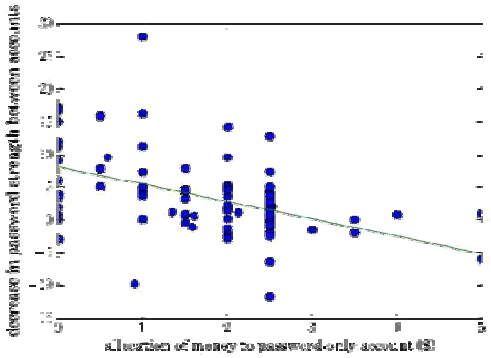

Fig. 10. Change in password strength by account allocation

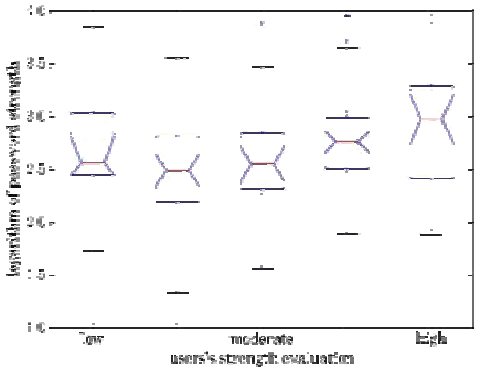

Fig. 11. User's evaluation of password strength by measured strength

the feeling of security users had and how much weaker their passwords were.

It may concern some readers that the data plotted in Figure 10 have significant outliers and that the Pearson product-moment correlation is too sensitive to outliers for this to be a trusted result. However, non-parametric tests reveal essentially the same result. For the hypothesis that they have no correlation: Spearman's rank correlation $\rho = -0.464$ with a p-value less than 0.000003 and Kendall's rank correlation $\tau = -0.347$, also with a p-value less than 0.000003.

2) User's security evaluation: After allocating money between the two accounts they had created, participants were asked to directly evaluate the security of the two accounts, as well as the components of each account. This evaluation was in the form of radio buttons that allowed participants to choose a response to the prompt "For each security measure, estimate how likely it is that the hackers will defeat it." for each of the following security measures:

1) the password used for the password-only account
2) the password used for the fingerprint-using account
3) the fingerprint used for the fingerprint-using account
4) all security on the fingerprint-using account

The results of these responses are in Table VIII. The participants responded with "Very unlikely", "Somewhat unlikely", "Neither likely nor unlikely", "Somewhat likely", or "Very likely". For the sake of clarity, in Table VIII, those answers have been replaced with "Very secure", "Somewhat secure", "Neither secure nor insecure", "Somewhat insecure", or "Very insecure", respectively.

The ability of participants to determine the quality of their passwords, directly, can perhaps be considered quite poor. Figure 11 demonstrates this through boxplots. If users correctly estimated the strength of their passwords (at least as measured by our ability to crack them), then Figure 11 would consist of medians, boxes and whiskers[5] increasing in value for increasing user security evaluations. While this appears, the correlation is very weak, and in fact testing for a correlation provides only p-values of 0.169 and 0.083 for the each group of passwords (password-only and fingerprint-using, respectively). Aggregating over both groups results in a sample Pearson's product-moment correlation of 0.172, and a p-value of 0.018 for the hypothesis that the two are not correlated, indicating that users are somewhat effective at discerning the strength of their passwords, but inconsistently.

Users proved much more able at the task of determining which of their passwords was the stronger. Figure 12 is similar to 11, but compares the difference in strength participants claimed against the differences in measured password strength. The clustering is not much better; the Pearson's product-moment correlation of the two data is 0.186, with insufficient statistical significance (p-value of 0.07). However, the actual measured strength difference is much less important than whether such a difference exists; comparing only the sign of the difference to the evaluation improves the correlation to 0.261 and the p-value to 0.01. Practically speaking, only 9.5% of users incorrectly gauged which of their passwords would be easier to break, although 55.8% of users marked their passwords as equally likely to be cracked by hackers. Thus their ability to gauge a numeric difference between their passwords was minimal, but their ability to correctly assess the sign of that difference was very good.

As before when measuring account allocation, we can attempt to use these answers to determine whether greater faith in the fingerprint-protected account was reflected by the choice of a weaker password for that account. Like Figure 10, Figure 13 compares a change in password strength to the difference of security evaluations participants gave for their password versus the entire security of the fingerprint-using account. The correlation, as expected, is much weaker, and is in fact too weak to be considered significant. Similar graphs and analyses appear when comparing measured strength decreases against the user's evaluations of the security of the fingerprint reader, as well as the increase in their rating of the security from fingerprint password to their rating of all the security of the

[5] Whiskers appear at the end of the dotted lines, and indicate 1.5 times the median to quartile difference. The top and bottom of each box represent the 25^{TH} and 75^{TH} quartiles, while the middle line is the median. The notches indicate a 95% confidence interval around the true median. Outliers appear as crosses.

Special Issue; March- 2019 [ISSN: 2455-1457]
DOI: 10.23883/IJRTER.CONF.20190304.009.7M9IX

TABLE VIII
USER EVALUATION OF SECURITY OF ACCOUNTS AND ACCOUNT PARTS (PERCENTAGES)

Security rating	Password used in password-only account	Password used in fingerprint account	Fingerprint used in fingerprint account	Aggregate security of fingerprint account
Very secure	23.4	13.8	21.3	31.9
Somewhat secure	35.1	38.3	37.2	34.0
Neither secure nor insecure	13.8	16.0	17.0	22.3
Somewhat insecure	21.3	23.4	16.0	7.4
Very insecure	6.4	8.5	8.5	4.3

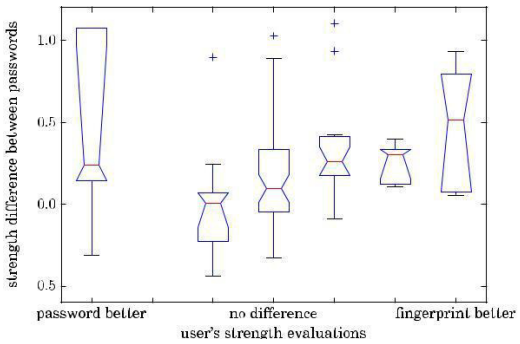

Fig. 12. Difference in user evaluation of password strength by measured strength differences

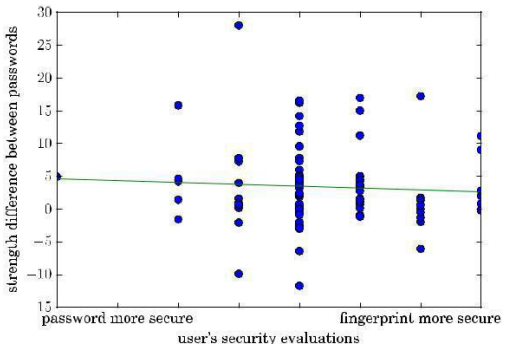

Fig. 13. Change in password strength by users' security evaluation

fingerprint account.

It is possible that a correlation fails to appear because of the discrete nature of the questions asked, but given that participants mostly chose a few discrete values when allocating money to their accounts, that seems unlikely to explain the difference. It seems more likely that the nature of the questions were less concrete; asking users to directly allocate their money resulted in a much more purposeful evaluation.

V. FUTURE WORK

The very nature of user studies opens them to a large degree of variability; since our study and results have not been heretofore tested, we can only make claims within the limited focus of our study. Similar tests in other circumstances are important to broaden the use of our conclusions. In particular, studies requiring users to log into accounts many times over an extended period of time and studies that do not stress the immediate security threats are probably necessary in order to extend our conclusions to general user behavior. In addition, long term studies that allow users to reuse passwords, modify passwords and opt out of either the password or fingerprint should be considered. Finally, a study similar to this one should be conducted with users who are less informed with respect to security.

In the same vein, our study only examined one kind of biometric authenticator in one operational mode. Other studies are warranted to examine other forms of security, including biometric and token-based systems. It is probably even worth studying the impact of "back-end" security that users do not have to directly interact with, such as the logos on websites that purport to assure users that the web form they are interacting with is secured in some way.

Our study only examined the strength of passwords, which measures only one small aspect of security that users are responsible for. There are many other areas in which users might act less or more securely; for instance, are users less reluctant to disclose their passwords if they expect a fingerprint to secure their account? Deeper investigations into these questions are likely to be more difficult to execute, but comprehensive studies that consider all of the interacting issues that arise from multi-factor authentication are necessary to inform administrators for effective security policy decisions. In our study, although there was a distinct difference in the strength of passwords between the password-only account and the fingerprint-using account, the vast majority of participants chose both their passwords to be strong enough to foil even a very dedicated adversary. We do not know whether this trend would have existed if there were limits on the size of the password, but it is likely that the attention users pay to security in other areas will end up more greatly impacting the security of a system than their password choice.

Additionally, our study suggests two potential ways to increase the strength of user passwords—through mention of a specific, imminent threat and through competition with other users. The former is already done in some instances and to some degree by system administrators that proactively crack passwords [34], although proactively cracking passwords

DOI: 10.23883/IJRTER.CONF.20190304.009.7M9IX

without the knowledge of the users should fail to achieve this psychological advantage. More broadly, many companies have training programs that warn users about the threats associated with weak security. Further study would be needed before we could identify the actual effects of either of these factors in increasing the strength of user passwords.

VI. Conclusion

Our experiment showed clearly that at least under controlled conditions users can be expected to behave less securely when interacting with systems that seem to them to be better secured. Although this is the predicted result of risk compensation theory, the implications of this finding are still potentially significant: if the results transfer to users and systems outside of a laboratory setting, system administrators should be wary of the kind of changes they make to secure a system, lest they inadvertently make the system less secure by unduly influencing user behavior.

Because of the much-better-than-expected strength of pass-words in our study, we do not expect changes in the actual strength of real user passwords to be well predicted by our findings, but it is possible that similar relative differences—around a three thousand fold decrease in time-to-crack—will be exhibited in the real-world. Given differences of this mag-nitude, using a password with a second required authentication mechanism is probably wise only if that mechanism can be trusted to secure the system at least well enough to make up for that difference. This advice, of course, becomes less and less relevant the less the mechanism under consideration resembles the environment of our experiment. Our results demonstrate a strong correlation between a perception of security and de-creased password strength. This implies that our results can be explained by risk compensation theory, in which case our most important result can be phrased as "do not introduce security mechanisms that appear to provide more security than they actually provide". Notably, if our fingerprint readers provided as much security as our participants expected they would, the resultant decrease in password strength with the addition of the fingerprint readers would not have significant impact on overall security. However, if our readers were connected to exploitable software or poorly designed, our participants could have been choosing weaker passwords without any compensation made by the use of the fingerprint reader.

Our personal feelings are that "sexy" security mechanisms that are heavily used in fiction and popular culture—including forms of biometric authentication, like fingerprint readers—are most liable to violate this policy. Security mechanisms like smart cards that provide good security but do not have a significant presence in spy movies may be less likely to be implicitly trusted. Of course, if a mechanism is a good enough security measure to trust it regardless of user behavior, then employing it will always be a good choice. We are, however, reluctant to classify any particular technology into a category where we trust that poor user practices will not affect it.

Lastly, although studying user behavior is seldom a primary focus in the field of computer security, it is significant. The better we do to secure systems, the more likely that the weakest component of the system will be the user. Thus anticipating user behavior becomes an important aspect of our job in designing secure systems, mandating a degree of confidence in what sorts of behaviors we can expect from users. While our contribution to this broad subject is focused and limited to only a few facets, we hope to directly assist the development of more secure systems and encourage further study of user behavior in a variety of security contexts.

References

[1] D. Florencio and C. Herley, "A large-scale study of web password habits," in *WWW '07: Proceedings of the 16th International Conference on World Wide Web*. Banff, Alberta, Canada: ACM, 2007, pp. 657–666.

[2] J. E. Webber, D. Guster, P. Safonov, and M. B. Schmidt, "Weak password security: An empirical study." *Information Security Journal: A Global Perspective*, vol. 17, no. 1, pp. 45–54, 2008.

[3] P. Hoonakker, N. Bornoe, and P. Carayon, "Password authentication from a human factors perspective: Results of a survey among end-users," *Human Factors and Ergonomics Society Annual Meeting Proceedings*, vol. 53, pp. 459–463(5), September 2009.

[4] M. Dell'Amico, P. Michiardi, and Y. Roudier, "Password strength: an empirical analysis," in *INFOCOM'10: Proceedings of the 29th Conference on Information Communications*. Piscataway, NJ, USA: IEEE Press, 2010, pp. 983–991.

[5] J. Yan, A. Blackwell, R. Anderson, and A. Grant, "Password memora-bility and security: empirical results," *Security & Privacy, IEEE*, vol. 2, no. 5, pp. 25–31, 2004.

[6] A. Forget, S. Chiasson, P. van Oorschot, and R. Biddle, "Persuasion for stronger passwords: Motivation and pilot study," in *Persuasive Technology*, ser. Lecture Notes in Computer Science, 2008, vol. 5033, pp. 140–150.

[7] D. Davis, F. Monrose, and M. K. Reiter, "On user choice in graphical password schemes," in *In 13th USENIX Security Symposium*. Berkeley, CA, USA: USENIX Association, 2004, pp. 151–164.

[8] L. Coventry, A. De Angeli, and G. Johnson, "Usability and biometric verification at the ATM interface," in *CHI '03: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Ft. Lauderdale, Florida, USA: ACM, 2003, pp. 153–160.

[9] A. Stewart, "On risk: perception and direction," *Computers & Security*, vol. 23, no. 5, pp. 362 – 370, 2004.

[10] R. M. Trimpop, "Risk homeostasis theory: roblems of the past and promises for the future," *Safety Science*, vol. 22, no. 1-3, pp. 119 – 130, 1996.

[11] M. Zviran and W. J. Haga, "Password security: an empirical study," *J. Manage. Inf. Syst.*, vol. 15, pp. 161–185, March 1999.

[12] K.-P. L. Vu, R. W. Proctor, A. Bhargav-Spantzel, B.-L. B. Tai, J. Cook, and E. E. Schultz, "Improving password security and memorability to protect personal and organizational information," *International Journal of Human-Computer Studies*, vol. 65, no. 8, pp. 744–757, 2007.

[13] P. G. Inglesant and M. A. Sasse, "The true cost of unusable password policies: password use in the wild," in *Proceedings of the 28th inter-national conference on Human factors in computing systems*, ser. CHI '10. Atlanta, Georgia, USA: ACM, 2010, pp. 383–392.

[14] C. Herley, "So long, and no thanks for the externalities: the rational rejection of security advice by users," in *Proceedings of the 2009 workshop on New security paradigms workshop*, ser. NSPW '09, 2009, pp. 133–144.

[15] S. Sagan, "The problem of redundancy problem: Why more nuclear security forces may produce less nuclear security," *Risk Analysis*, vol. 24, pp. 935–946, 2004.

[16] L. O'Gorman, "Comparing passwords, tokens, and biometrics for user authentication," *Proceedings of the IEEE*, vol. 91, no. 12, pp. 2021–2040, Dec. 2003.

[17] D. Weirich and M. A. Sasse, "Pretty good persuasion: a first step towards effective password security in the real world," in *Proceedings of the 2001 workshop on New security paradigms*, ser. NSPW '01. Cloudcroft, New Mexico: ACM, 2001, pp. 137–143.

[18] J. M. Stanton, K. R. Stam, P. Mastrangelo, and J. Jolton, "Analysis of end user security behaviors," *Computers & Security*, vol. 24, no. 2, pp. 124–133, 2005.

Special Issue; March- 2019 [ISSN: 2455-1457]
DOI: 10.23883/IJRTER.CONF.20190304.009.7M9IX

[19] R. Dhamija and A. Perrig, "D´ej`a vu: a user study using images for authentication," in *Proceedings of the 9th conference on USENIX Security Symposium - Volume 9*, 2000, pp. 4–4.

[20] S. Wiedenbeck, J. Waters, J.-C. Birget, A. Brodskiy, and N. Memon, "Authentication using graphical passwords: effects of tolerance and image choice," in *Proceedings of the 2005 symposium on Usable privacy and security*, ser. SOUPS '05. Pittsburgh, Pennsylvania: ACM, 2005, pp. 1–12.

[21] F. M. Streff and E. Geller, "An experimental test of risk compensation: Between-subject versus within-subject analyses," *Accident Analysis & Prevention*, vol. 20, no. 4, pp. 277 – 287, 1988.

[22] T. Assum, T. Bjrnskau, S. Fosser, and F. Sagberg, "Risk compensation—the case of road lighting," *Accident Analysis & Prevention*, vol. 31, no. 5, pp. 545 – 553, 1999.

[23] W. N. Evans and J. D. Graham, "Risk reduction or risk compensation? the case of mandatory safety-belt use laws," *Journal of Risk and Uncertainty*, vol. 4, pp. 61–73, 1991.

[24] D. Besnard and B. Arief, "Computer security impaired by legitimate users," *Computers & Security*, vol. 23, no. 3, pp. 253 – 264, 2004.

[25] R. West, "The psychology of security," *Commun. ACM*, vol. 51, pp. 34–40, April 2008.

[26] E. Albrechtsen, "A qualitative study of users' view on information security," *Computers & Security*, vol. 26, no. 4, pp. 276 – 289, 2007.

[27] M. Weir, S. Aggarwal, B. de Medeiros, and B. Glodek, "Password cracking using probabilistic context-free grammars," *Security and Privacy, IEEE Symposium on*, vol. 0, pp. 391 –405, May 2009.

[28] (2011, Mar.) Openwall wordlists collection. [Online]. Available: http://www.openwall.com/wordlists/

[29] (2010) John the ripper password cracker. [Online]. Available: http://www.openwall.com/john/

[30] M. Weir, S. Aggarwal, M. Collins, and H. Stern, "Testing metrics for password creation policies by attacking large sets of revealed passwords," in *Proceedings of the 17th ACM conference on Computer and communications security*, ser. CCS '10. Chicago, Illinois, USA: ACM, 2010, pp. 162–175.

[31] D. Florˆencio, C. Herley, and B. Coskun, "Do strong web passwords accomplish anything?" in *Proceedings of the 2nd USENIX workshop on Hot topics in security*. Berkeley, CA, USA: USENIX Association, 2007, pp. 10:1–10:6.

[32] A. Adams and M. A. Sasse, "Users are not the enemy," *Commun. ACM*, vol. 42, pp. 40–46, December 1999.

[33] R. McGill, J. W. Tukey, and W. A. Larsen, "Variations of Box Plots," *The American Statistician*, vol. 32, no. 1, pp. 12–16, 1978.

[34] M. Bishop and D. V. Klein, "Improving system security via proactive password checking," *Computers & Security*, vol. 14, no. 3, pp. 233–249, 1995.

