On the Memorability of System-generated PINs: Can Chunking Help?

Jun Ho Huh*
Honeywell ACS Labs
Golden Valley, MN USA
junho.huh@honeywell.com

Hyoungschick Kim
Sungkyunkwan University
Suwon, Korea
hyoung@skku.edu

Rakesh B. Bobba*
Oregon State University
Corvallis, OR USA
Rakesh.Bobba@oregonstate.edu

Masooda N. Bashir
University of Illinois
Urbana-Champaign, IL USA
mnb@illinois.edu

Konstantin Beznosov
University of British Columbia
Vancouver, BC Canada
beznosov@ece.ubc.ca

ABSTRACT

To ensure that users do not choose weak personal identification numbers (PINs), many banks give out *system-generated random PINs*. 4-digit is the most commonly used PIN length, but 6-digit system-generated PINs are also becoming popular. The increased security we get from using system-generated PINs, however, comes at the cost of memorability. And while banks are increasingly adopting system-generated PINs, the impact on memorability of such PINs has not been studied.

We conducted a large-scale online user study with 9,114 participants to investigate the impact of increased PIN length on the memorability of PINs, and whether number *chunking*[1] techniques (breaking a single number into multiple smaller numbers) can be applied to improve memorability for larger PIN lengths. As one would expect, our study shows that system-generated 4-digit PINs outperform 6-, 7-, and 8-digit PINs in long-term memorability. Interestingly, however, we find that there is no statistically significant difference in memorability between 6-, 7-, and 8-digit PINs, indicating that 7-, and 8-digit PINs should also be considered when looking to increase PIN length to 6-digits from currently common length of 4-digits for improved security.

By grouping all 6-, 7-, and 8-digit chunked PINs together, and comparing them against a group of all non-chunked PINs, we find that chunking, overall, improves memorability of system-generated PINs. To our surprise, however, none of the individual chunking policies (e.g., 0000-00-00) showed statistically significant improvement over their peer non-chunked policies (e.g., 00000000), indicating that chunking may only have a limited impact. Interestingly, the top performing 8-digit chunking policy did show noticeable and statistically significant improvement in memorability over shorter 7-digit PINs, indicating that while chunking has the potential to improve memorability, more studies are needed to understand the contexts in which that potential can be realized.

Categories and Subject Descriptors

D.4.6 [**Security and Protection**]: Authentication; H.1.2 [**User/Machine Systems**]: Human factors

General Terms

Experimentation, Human Factors, Measurement, Security

Keywords

Security, Usability, PINs, Passwords, Policy, Chunking

1. INTRODUCTION

A personal identification number (PIN) is a numeric password that is used to authenticate users. PINs are commonly used in banking systems and on handheld devices (e.g., mobile phones and tablets) that require quick and easy yet sufficiently secure access. Many banks use 4-digit PINs to authenticate debit card (and sometimes credit card) transactions. Mobile phones often require users to enter 4-digit PINs to authenticate and unlock the screen.

To strengthen PIN security, some banks and others have recently started using 6-digit PINs, to take advantage of the larger PIN space of 10^6 possible entries. That could provide a significant improvement in security. However, if users generate their own 6-digit PINs, the improvement in entropy will be marginal as there tends to be a small pool of commonly selected 6-digit PINs [20]. Also, people find it harder to remember 6-digit PINs.

To get around the problem of low entropy in user generated PINs, many banks are adopting *system-generated PINs*, asking users to remember randomly generated 4- or 6-digit PINs. Banks in Switzerland, for example, assign 6-8 digit PINs; Canadian banks use both 4- and 6-digit PINs. It is

*Part of this work was done while Dr. Huh and Dr. Bobba were at the University of Illinois.

[1]Note that our notion of chunking differs from the traditional notion in that we do not chunk numbers into semantically meaningful pieces.

Copyright is held by the author/owner. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee.
Symposium on Usable Privacy and Security (SOUPS) 2015, July 22–24, 2015, Ottawa, Canada.

common practice for banks to send physical mail to customers that contains a randomly generated PIN and instructions on how that PIN should be used and protected. System-generated PINs ensure that users do not use common PINs like "0000" and increase the entropy of PINs, making it more difficult for an attacker to guess them. Randomly generated PINs, when used together with an account lock-out policy (e.g., a user's account should be locked after 5 failed authentication attempts), can be highly effective against online brute-force attacks.

The biggest drawback with system-generated PINs (as with system-generated passwords [5]), however, is their memorability. Although many banks are moving to system-generated 6-digit PIN, the impact on memorability is not clearly known. Are the banks making the right decision in moving toward 6-digit PINs? Why should they not consider 7- or 8-digits?

We conducted a large-scale online user study[2], recruiting a total of 9,114 participants to understand the memorability of system-generated PINs of varying lengths, from 4 to 8 digits. As one would expect, our study shows that system-generated 4-digit PINs outperform 6-, 7-, and 8-digit PINs in long-term memorability. Interestingly, however, we find that there is no statistically significant difference between long-term memorability of 6-, 7-, and 8-digit PINs (see Section 4.4).

To investigate ways of improving memorability, we applied different "*chunking*" policies [14] on system-generated PINs, and studied their impact on memorability through the same online study. It is important to note that the notion of *chunking* used in this paper is different from the traditional notion of chunking. Traditional notion of *chunking* refers to the practice of breaking a single number into multiple smaller numbers that are *semantically meaningful*. Phone numbers are a good example of chunked numbers. In the United States a ten-digit phone number is chunked into smaller chunks of 3-3-4 (000–000–0000) that represent area code, exchange code and subscriber number respectively, to help people remember it easily. Based on what is already working well in the real world when using semantically meaningful chunks, we hypothesized that breaking longer system-generated PINs into smaller chunks, even if they did not have semantic meaning, would improve their long-term memorability.

We investigated a a variety of chunking combinations (referred to as *chunking policies*) to see how different arrangements of smaller chunks can affect memorability. In total, we investigated 12 different chunking policies. To the best of our knowledge, this is the first large-scale study on the impact of applying different variations of chunking techniques to randomly generated information and specifically to PINs; previous studies [9] often focused on showing that chunking is useful for information that has some meaning to a user.

A summary of the study findings is as follows:

- Our empirical evaluation of the relative memorability of 4-, 6-, 7-, and 8-digit system-generated PINs showed that 4-digit PINs outperform all other larger length PINs in long-term memorability as expected. Interestingly, however, the evaluation showed that there is no statistically significant difference between memorability of 6-, 7-, and 8-digit PINs.
- Our large-scale empirical evaluation of a wide variety of chunking policies for system-generated PINs did not find statistically significant improvement in long-term memorability when individual chunking policies (e.g., 0000-00-00) were compared against their non-chunked peers of same PIN length (e.g., 00000000).
- However, chunking policies did show statistically significant improvement in memoraiblity when grouped together and compared against a group of non-chunked polices.
- Further, we found an 8-digit chunking policy (0000-00-00) that noticeably outperformed shorter 7-digit PINs in long-term memorability with statistical significance.
- Differences in short-term memorability of system-generated PINs for different PIN lengths and chunking policies were found not to be statistically significant in most cases. Even in the cases where the differences were found to be statistically significant, the observed differences were relatively small ($< 3\%$).

Our findings suggest that, while chunking randomly generated PINs may not be universally effective in improving memorability, such chunking does have the potential to improve memorability in certain contexts. More studies are needed to understand the contexts in which chunking can help improve memorability of system generated PINs.

The next section discusses related work on PIN security and chunking techniques. Section 3 explains our hypotheses, methodology, and empirical study. Section 4 discusses the memorability results and participants' thoughts on the policies. Section 5 discusses our hypotheses in terms of the results. We discuss limitations, future directions and conclude in Section 6.

2. RELATED WORK

Over the years, many user authentication technologies have been designed and deployed on security-critical systems. Some of the popular technologies include passwords, PINs, digital certificates, physical tokens, one-time passwords, transaction profile scripts, and biometric identification. Among those, "what you know" forms of authentication, generally passwords or PINs, are still the dominant technology, mainly because of their familiarity and low implementation and deployment costs[16, 6]. However, to ensure good memorability, many users choose passwords or PINs that are easy to remember; such passwords/PINs can be guessed easily and are vulnerable to brute-force and dictionary attacks. Bonneau et al. [18] studied the difficulty of guessing human-selected 4-digit PINs, concluding that many users use their birth dates or other memorable dates as PINs, and an effective attack would involve brute-forcing PINs using dates.

To help users choose stronger PINs, PIN selection policies may be enforced. Such policies would specify, for instance, the combinations of numbers that can be used when a user is creating a PIN, or specify a list of PINs that cannot used. Even with those policies in place, however, users could still choose PINs that are weak but different from prohibited set of PINs, skewing the PIN distribution to the next set of popular, easy to remember PINs that are allowed by the policy.

[2]The study was approved by the Institutional Review Boards of both University of Illinois and Oregon State University.

Hence, the effectiveness and usability of those policies must be carefully evaluated before they are used. Kim et al. [20] analyzed the effectiveness of a few PIN selection policies, and found that a blacklist policy (e.g., forbidding the top 200 most popularly used PINs) can help users choose more secure 4- and 6-digit PINs that also have good memorability. Their study, however, looks at human-selected PINs.

As for passwords, there have also been active debates about the effectiveness of password selection policies. Several studies have already been carried out on understanding the relationship between password selection policies and the resulting passwords. Some of those studies were based on theoretical estimates [8, 26, 25]; some were based on small-scale laboratory studies [23, 7, 33, 13, 32]; and some were based on large-scale studies [29, 21]. Vu et al. [32] conducted a laboratory study that demonstrated that passwords chosen under strong selection policies are generally harder to compromise with automated password-cracking tools, but that they are also harder to generate and remember, affecting the overall usability. Kuo et al. [22] showed that automated tools were less effective against mnemonic passwords than against control passwords. Simulations performed by Shay et al. [26, 25] have shown that stringent-password selection policies can lead users to write down their passwords and thereby jeopardize their confidentiality. Shay et al. [29] examined users' behaviors and practices related to password creation under a new, more strict policy. Users were annoyed by the transition to a stricter password policy, but felt more secure under the new policy. Some users struggled to comply with the new policy, taking longer to create passwords and finding it harder to remember them. In a recent study, Komanduri et al. [21] tried to understand the relationship between password selection policies and the resulting passwords, and recommended some policies (e.g., a 16-character minimum with no additional requirements) that result in strong passwords without unduly burdening users. Shay et al. [27] particularly investigated the password selection policies for long passwords. Inglesant and Sasse [17] have shown that many users, despite knowing that repeated use of the same passwords is a bad security practice, rarely change their passwords.

Despite all those efforts in helping users select better passwords or PINs, advanced attackers have been effective in finding ways to crack them. Attackers combine dictionary attacks (with massive databases of dictionary words as well as commonly used passwords and PINs) with brute-forcing attacks, permutation attacks, rule-based attacks, or fingerprint attacks to crack just about anything that users create and remember [24, 15]. Hashcat [2] is a commonly used password-cracking tool that supports a combination of all of those attacks. With 4-digit PINs, it is even easier to perform those attacks, since the search space is much smaller. An alternative way to guarantee security is to use system-generated passwords or PINs, relying on a computer to generate a random password or PIN for you. System-generated PINs, although widely used by, for example, banks and the Department of Defense, have memorability issues [5]. To overcome such memorability weaknesses, we study the effects of using number-chunking techniques on system-generated PINs. In the past, several studies have shown the effectiveness of using chunking techniques as a memory tool for human brains. The hypothesis is based on the well-known process of chunking, in which primitive stimuli are grouped into larger conceptual groups such as letters into words [14]. Druzal et al. [11] claim that the use of chunking in working memory might be helpful for identification of faces. Thornton et al. [30] suggest that chunking is an effective mechanism for improving social working memory. Carstens et al. [9] show that human errors associated with password-based authentication can be significantly reduced through the use of passwords that are composed of data *meaningful* to the user.

Likewise, previous studies often looked at associating meaningful information with chunks (e.g., the first chunk of three digits in a phone number represents the area code). Our work extends those studies on chunking, but also incorporates other elements in that we apply number-chunking techniques to randomly generated PINs that are not associated with any meaningful information, evaluating the effects on both short-term and long-term memorability. To the best of our knowledge, we are the first to analyze the effects of chunking techniques specifically for PINs and to study so many different variations of chunking combinations (see Table 1) and PIN lengths. We are also the first to study the memorability of system-generated PINs, including the 4- and 6-digit PINs that many banks are currently using. Furthermore, previous studies have often been based on small-scale lab studies, with small numbers of participants; we have conducted a much larger-scale study using Mechanical Turk, recruiting a total of 9,114 participants (see Section 3.5).

3. METHODOLOGY

This section defines the key research questions and the hypotheses, provides an overview of the conducted user study, and explains the participant recruitment methodology.

3.1 Hypotheses

This work was motivated by research questions such as, how usable and memorable are system-generated 6-digit PINs compared to 4-digit PINs? Should banks also consider using 7- or 8-digit PINs? Can chunking techniques help improve the memorability of longer length system-generated PINs, and if so, how significant is the improvement?

Based on these research questions and our intuition, we defined the following three hypotheses.

1. The memorability of system-generated 6-digit PINs is worse than that of 4-digit PINs.
2. The memorability of system-generated 6-digit PINs is better than those of 7 and 8-digit PINs.
3. The memorability of longer (6-, 7- and 8-digit) system-generated PINs improves with chunking.

The user study and experiments were designed with the above hypotheses in mind. In Section 5, we discuss how the study results match up to these hypotheses.

3.2 PIN chunking policies

This section describes the 12 PIN chunking polices that we investigated (see Table 1), and explains why we chose these polices. Each chunking policy defines *the PIN length, how the numbers are chunked, and how the chunks are arranged.*

PIN lengths were defined first. Since banks (and other industrial entities) already use 4- and 6-digit PINs, we included them. Then, to test hypothesis 2 (see Section 3.1)

Table 1: PIN chunking policies. Each policy defines the PIN length, the number of digits each chunk contains, and the arrangement of chunks.

Policy	*Format*	*Example*
`4`	0000	8854
`6`	000000	480271
`6:2-4`	00–0000	48–0271
`6:4-2`	0000–00	4802–71
`7`	0000000	1685149
`7:3-4`	000–0000	168–5149
`7:4-3`	0000–000	1685–149
`8`	00000000	75357600
`8:4-4`	0000–0000	7535–7600
`8:2-2-4`	00–00–0000	75–35–7600
`8:2-4-2`	00–0000–00	75–3576–00
`8:4-2-2`	0000–00–00	7535–76–00

we included PIN lengths of 7 and 8 digits. The psychological literature on memorability suggests that most people can remember up to 4 digits without a problem over the short term [10]. Hence, we decided to include at least one chunk with 4 digits in all of the policies. We also decided that the smallest chunk in a policy should consist of at least two digits, because if we allow one-digit chunks, we could end up with policies with too many chunks in them. Based on those rules and intuition, PIN-chunking policies were designed. For instance, policy `8:2-2-4` says that a PIN would consist of 8 digits, where those 8 digits would be presented in three smaller chunks of 2, 2, and 4 digits, i.e., in the format of 00–00–0000.

3.3 User study design

Given the large number of PIN-chunking policies that we wanted to evaluate through empirical quantitative experiments, we chose to employ Amazon Mechanical Turk. To make the study as realistic as possible, we employed role-playing by simulating an online PIN setup page for a made-up bank, and informing each participant that he or she would use the PIN for card purchases (see Figure 1). Each participant was assigned a specific chunking policy (picked uniformly at random) and given a different system-generated PIN to remember. The given PIN was presented to the participant in the format defined by the chunking policy.

Our user study was designed following the dual memory model by Atkinson-Shiffrin [3] that postulates memories initially reside in a "short-term" memory for a limited time (20 to 30 seconds) while they are strengthening their association in the "long-term" memory. While later memory models (*e.g.*[4]) expand on the multi-store model proposed by Atkinson-Shiffrin they all agree that short-term memory (or working memory) has limited capacity and older items are wiped as new items enter. Further, rehearsing or recalling items while they are in the short-term memory causes the items to stay longer in the short-term memory while at the same time strengthening their association in the long-term memory.

Keeping the dual memory model in mind, our data collection involved two parts. The first part was meant to ensure that a PIN enters the long-term memory, and measure

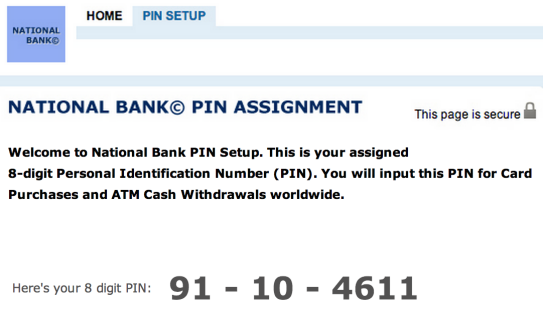

Figure 1: User study screenshot: Assigned PIN.

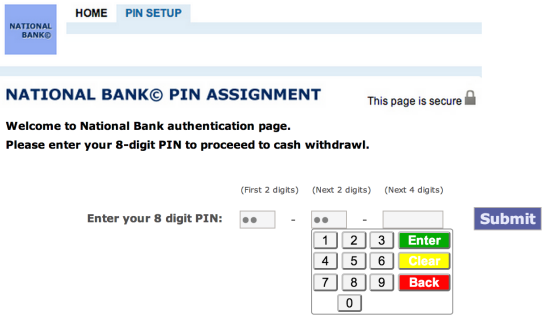

Figure 2: User study screenshot: PIN entry.

the short-term memorability[3] of that PIN. This part consisted of three steps. First, each participant was asked to complete three training tasks (rehearsing), to help them remember the assigned PIN (associate with long-term memory). Then, each participant was asked several questions related to cognition and memory strength, wiping out their short-term memory during the process. Third, each participant completed a short-term memorability test by entering the assigned PIN (see Figure 2). Two days (48 hours) after completing the first part, each participant that passed the short-term memorability test received an email inviting them to the second part of the study, in which we measured the long-term memorability. Two days was picked to study long-term memorability following the lead of other work in the community (e.g., [31, 19, 12, 28]). In the second part, each participant was simply asked to enter their PIN again.

Details of the different tasks and the order in which the participants were asked to complete them are as follows:

Study Part 1: Short-term Memorability

1. **PIN assignment:** Each participant was given a system-generated PIN, which was presented in the format defined by the randomly assigned chunking policy (see Figure 1). Participants were asked to remember their PINs.

2. **Remembrance training:** Each participant was asked to enter the correct PIN three times to help with memorization. If incorrect PINs were entered three times

[3]The terms short-term memory and long-term memory should not be confused with short-term memorability and long-term memorability. Both short-memorability and long-term memorability refer to recalling from long-term memory, but over different lengths of time.

consecutively, the correct PIN was revealed again so that the participant would have another chance to memorize it. The training session ended only when the participant entered the correct PIN three times. Using a universal keypad (seen in Figure 2), participants entered their PIN in the chunking format defined by the chunking policy picked at random for them.

3. **Cognition and memory strength questions:** Each participant was asked 18 cognition questions (e.g., "I prefer complex to simple problems") and 6 memory strength questions (see Table 2). Answering all of those questions was expected to take about 2 minutes. Those questions were asked to help us better understand the participants' characteristics, and to clear participants' short-term memory.

4. **Demographics and survey questions:** Each participant was asked five demographic questions and five survey questions (see Table 3) about the assigned PIN. The survey questions asked about their thoughts on the memorability and usability of the assigned PIN and the chunking policy, taking about a minute to complete.

5. **Enter PIN (short-term test):** Each participant was asked to enter the assigned PIN (see Figure 2) and was given three chances to enter it correctly. To simulate an existing banking scenario, we asked participants to enter the correct PIN to proceed with cash withdrawal.

Study Part 2: Long-term Memorability

6. **Enter PIN after two days (long-term test):** Two days (48 hours) after participating in the first part of the study, participants who passed the short-term memorability study received an email asking them to complete the second part of the study. The participants were asked to enter their PIN again and were given three chances to enter the correct one.

7. **PIN survey questions:** To understand how participants feel about long-term usability of the assigned PIN, the same five survey questions about the assigned PIN (Table 3) were asked again after completing the test.

Table 2: Memory strength questions

#	*Question*
MQ1	I have a difficult time remembering numerical information.
MQ2	I frequently get passwords and numbers mixed up in my head.
MQ3	I have a good memory for things I have done in the past week.
MQ4	I easily lose my train of thought.
MQ5	I have a good memory for phone numbers that I have dialed in the past.
MQ6	I frequently remember details of past events that other people have forgotten.

To minimize the chances that the participants would write down their PINs after the first part of the study, we did not disclose exactly what the participants would have to do in the second part and how they would be rewarded. We simply informed the participants that they might be invited to the second part of the study in two days. However, we informed those who returned to complete the second part that they could earn an extra bonus by getting the PIN right, providing incentives for them to try their best to recall the correct PIN.

Table 3: PIN survey questions

#	*Question*
SQ1	How difficult was it for you to remember the assigned PIN?
SQ2	Did you use an external storage (e.g., a sheet of paper or a text file) to write down the assigned PIN?
SQ3	Did you use any special technique (e.g., keypad patterns, assigning images to numbers, converting numbers to words) to help you remember the assigned PIN?
SQ4	If you answered "Yes" to Q3, what was the special technique that you used?
SQ5	Do you currently use a PIN that is equal to or longer than 6 digits?

3.4 User data collected

Throughout the 7 different stages of the user study (see above), we recorded the following information:

- **Assigned PIN and chunking policy.** We recorded the chunking policy and the PIN each participant was assigned.
- **Number of attempts made in entering PIN.** We recorded the number of attempts a participant made to enter the correct PIN in all of the training sessions and short-term and long-term memory tests.
- **Time taken to enter PIN.** Likewise, we measured the time it took each participant to enter a PIN for every attempt made, in all of the training sessions and short-term and long-term memory tests. Timing began when the participant first accessd the login screen and ended when the participant either entered the correct password or tried and failed all three attempts, capturing both successful and unsuccessful login attempts.
- **Memorability results.** We recorded the results of the memorability tests (i.e., whether a correct PIN was entered) for every attempt made, in all of the training sessions and short-term and long-term memory tests.
- **Survey answers.** We recorded participants' answers to the cognition questions, memory strength questions, and PIN survey questions (see Tables 2 and 3).

3.5 Mechanical Turk

Given the large number of PIN-chunking policies that we wanted to evaluate through empirical quantitative experiments, we chose to employ Amazon Mechanical Turk [1]. Every participant who completed the first part was rewarded with $0.50. Those who came back for the second part were rewarded with an additional $0.25 and another $0.25 if they

entered their PINs correctly. The intention of this extra bonus was to provide an incentive for participants to try their best to recall the correct PIN. As can be seen from the high short-term memorability scores in Table 5, participants did not need an extra monetary incentive to recall the PIN in the first part.

3.6 Statistical tests

We first performed the chi-square test on the proportion of successful logins and external storage usage to check whether proportions across all chunking policies are equal or not ($p < 0.05$). If chi-square test results indicated that not all proportions are equal, we performed Fisher's exact test (FET) to check whether a proportion in one chunking policy is significantly greater than that of another chunking policy ($p < 0.05$). All comparisons were corrected for multiple-testing using False Discovery Rate (FDR) estimation when appropriate.

As for authentication time, the Shapiro-Wilk's test was first used to show that the collected data is not normally distributed. To check whether all the policies have equal medians for authentication time, we performed the Kruskal-Wallis test ($p < 0.05$), showing that not all medians are equal. We then used the unpaired Mann-Whitney (MW) U test ($p < 0.05$) to measure the statistical confidence in the authentication time differences between chunking policies. All comparisons were corrected for multiple-testing using False Discovery Rate (FDR) estimation when appropriate.

Table 4: The demographics of the participants

Gender	
Male	5,180 (56.84%)
Female	3,855 (42.30%)
No answer	79 (0.86%)
Age group	
18–29	5,256 (57.67%)
30–39	2,285 (25.07%)
40–49	842 (9.24%)
50–59	488 (5.35%)
60 and over	168 (1.84%)
No answer	75 (0.83%)
Education	
Less than high school	63 (0.69%)
High school	2,825 (31%)
University	5,066 (55.58%)
Masters	768 (8.43%)
Doctoral	104 (1.14%)
Professional	177 (1.94%)
No answer	111 (1.22%)
Ethnicity	
African American	552 (6.06%)
Asian	769 (8.44%)
Caucasian	6,936 (76.10%)
Hispanic	503 (5.52%)
Other	198 (2.17%)
No answer	156 (1.71%)

4. RESULTS

This section presents the key results obtained from the user study, including the memorability results and the participants' responses regarding the difficulty in remembering their assigned PINs.

4.1 Demographics

As mentioned in Section 3.5, participants were recruited using Mechanical Turk. During the study period, a total of 9,114 participants completed the first part of the study, and of those 6,208 participants came back to complete the second part. A majority of the participants were Caucasian (76.10%), and more than half were in the age group of 18–29 (57.67%). 56.84% were male and 55.58% had a university degree. The details of the demographics are presented in Table 4.

4.2 Writing down PINs

In response to our survey question (see SQ2 in Table 3) participants reported using external storage (i.e., having their PIN written down) to store their PINs. The percentage of participants who reported using external storage to remember their PIN for short-term and long-term memorability tests is shown in column five of Tables 5 and 7 respectively.

The number of participants using some form of external storage during short-term memorability test steadily increased with the size of the PIN. In particular, in our study 5% of participants who were assigned a 4-digit PIN reported using external storage. For 6-, 7-, 8-digit PINs the number of users that reported using external storage ranged from $7-8\%$, $10-11\%$ and $11-14\%$ respectively. The chi-square test results showed that not all external storage usage proportions are equal ($\chi^2(11) = 65.34$, $p < 0.0001$). Hence, we used FET to identify differences across the policies that are statistically significant.

The observed increase in the percentage of users using external storage when compared to those with 4-digit PINs was found to be statistically significant for all policies with larger PINs ($p < 0.05$, pairwise corrected FET) except for `6` and `6:4-2`. The observed increase in the percentage of users using external storage when compared to those with 6-digit PINs (both chunked and non-chunked) was found to be statistically significant for all 8-digit PIN policies ($p < 0.05$, pairwise corrected FET) except for policy `6:2-4` vs. `8`, `6:4-2` vs. `8`, and `6:2-4` vs. `8:4-2-2`. Comparing 6-digit PINs with 7-digit PINs, the increase in external storage was found to be significant only for the case of policy `6` vs. `7:3-4` ($p < 0.05$, pairwise corrected FET).

Similar to what was observed during short-term memorability test, the number of participants who reported using external storage to remember the PIN increased with size of the PIN length in long-term memorability test, except for a slight dip when going from 4-digit to 6-digit (see column 5 in Table 7). Again, the chi-square test results showed that not all external storage usage proportions are equal ($\chi^2(11) = 33.18$, $p < 0.0005$).

The observed increase in the percentage of users using external storage when compared to those with 4-digit PINs was found to be statistically significant for all policies with 8-digit PINs (all $p < 0.05$, pairwise corrected FET) except for policy `8:4-2-2`. The observed increase in the percentage of users using external storage when compared to policy `6` was found to be statistically significant for all 7- and 8-digit PIN policies (all $p < 0.05$, pairwise corrected FET).

These observations are consistent with previous findings in literature that users tend to write down passwords when

Table 5: Short term memorability and average time taken to authenticate. Column '% correct PIN' represents the percentage of participants who entered the correct PIN in the short-term test *not counting* those who reported to have their PIN written down on paper or electronically to remember it (see 3.3). Column '% Ext. storage' represents the percentage of participants who reported using external storage. Column 'Time' is the median time taken to authenticate (considering both successful and unsuccessful results) and is measured in seconds. Column 'No. of attempts' is the average number of attempts for a successful login.

Policy	*# Participants*	*# Failed*	*% correct PIN*	*% Ext. storage*	*Time (s)*	*# attempt*	*σ*
4	722	5	99%	5%	9.1	1.0	0.2
6	714	19	97%	7%	11.1	1.1	0.3
6:2-4	717	12	98%	8%	12.9	1.1	0.3
6:4-2	709	14	98%	8%	12.8	1.1	0.3
7	678	25	96%	11%	13.3	1.1	0.4
7:3-4	658	11	98%	11%	13.8	1.1	0.3
7:4-3	669	19	97%	10%	14.1	1.1	0.4
8	682	27	96%	11%	14.7	1.2	0.5
8:4-4	670	16	98%	13%	15.5	1.1	0.4
8:2-2-4	644	16	98%	14%	16.6	1.1	0.4
8:2-4-2	647	14	98%	14%	17.4	1.1	0.4
8:4-2-2	667	12	98%	11%	16.4	1.1	0.4

they are required to remember what they perceive as complex passwords (e.g., [26, 25]. Since this paper focuses on the memorability (and not security) of system-generated PINs, we did *not* include participants who reported to have their PIN written down (on paper or electronically) in all of the following analyses.

4.3 Memorability of individual policies

We first present the short-term and long-term memorability results for individual chunking policies, comparing memorability of each policy against all other policies.

4.3.1 Short term

As shown in Table 5, in our study all of the PIN policies scored high in short-term memorability, ranging between 96% for non-chunked 8-digit PIN to 99% for 4-digit PIN. In our sample, as shown in Figure 3, chunked PINs had the same or slightly better memorability score (i.e., showed higher percentage) than non-chunked PINs of the same PIN length. The chi-square test results showed that not all short-term memorability scores are equal ($\chi^2(11) = 25.91$, $p < 0.01$). However, only the differences in short-term memorability between 4-digit and 7-digit (99% vs. 97%) and between 4-digit and 8-digit (99% vs. 96%) were found to be statistically significant (all $p < 0.005$, pairwise corrected FET). A summary of all of the *statistically significant* short-term memorability differences is presented in Table 6.

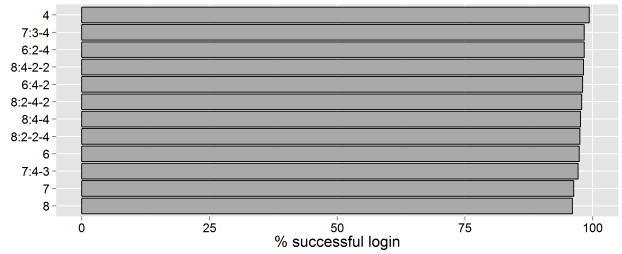

Figure 3: Sorted short-term memorability scores

Table 6: Statistically significant short-term memorability rate comparisons. 'p-value' is generated from pairwise corrected FET.

Superior policy/group	Inferior policy/group	p-value
Individual policy comparisons		
4 (99%)	7 (97%)	< 0.005
4 (99%)	8 (96%)	< 0.005
Policy group comparisons		
Chunk (98%)	No-Chunk (97%)	< 0.005
4 (99%)	7-Chunk (98%)	< 0.05
4 (99%)	8-Chunk (98%)	< 0.05
6-Chunk (98%)	7 (96%)	< 0.05
6-Chunk (98%)	8 (96%)	< 0.05
8-Chunk (98%)	8 (96%)	< 0.05
4-All (99%)	6-All (98%)	< 0.05
4-All (99%)	7-All (97%)	< 0.005
4-All (99%)	8-All (97%)	< 0.005

4.3.2 Long term

In contrast to the results for short-term memorability, we observed a significant decrease (up to 28%) in long-term memorability as we moved from 4-digit PINs to larger length PINs (see Table 7). Specifically, recall success rate for 4-digit PINs was at 74% while the rate for larger PIN lengths (including chunked PINs) varied from 46% to 57%. As expected, the chi-square test results showed that not all long-term memorability scores are equal across all of the policies ($\chi^2(11) = 79.31$, $p < 0.0001$).

The observed decrease in long-term memorability of larger length PINs when compared with that of 4-digit PINs was found to be statistically significant for all policies (chunked and non-chunked) with larger PINs ($p < 0.0001$, pairwise corrected FET). As expected, 4-digit PINs significantly outperformed larger length PINs in terms of memorability. Interestingly, however, the observed differences in memorability between non-chunked 6 (55%), 7(45%), and 8(50%) PINs were not found to be statistically significant. A summary of all of the *statistically significant* long-term memorability differences is presented in Table 8.

As seen in Figure 4, the recall success rates for all chunked

Table 7: Long term memorability and median time taken to authenticate.Column '% correct PIN' represents the percentage of participants who entered the correct PIN in the short-term test *not counting* those who reported to have their PIN written down on paper or electronically to remember it (see 3.3). Column '% Ext. storage' represents the percentage of participants who reported using external storage. Column 'Time' is the median time taken to authenticate (considering both successful and unsuccessful results) and is measured in seconds. Column 'No. of attempts' is the average number of attempts for a successful login.

Policy	*# Participants*	*# Failed*	*% correct PIN*	*% Ext. storage*	*Time (s)*	*# attempt*	*σ*
4	517	137	74%	6%	22.6	1.7	1.0
6	506	228	55%	5%	35.5	2.0	1.0
6:2-4	494	212	57%	8%	41.7	1.9	1.0
6:4-2	502	217	57%	8%	40.7	1.9	1.0
7	462	248	46%	10%	47.1	2.1	1.0
7:3-4	461	233	49%	10%	46.8	2.1	1.1
7:4-3	461	215	53%	10%	44.2	2.0	1.0
8	454	227	50%	11%	49.6	2.1	1.1
8:4-4	456	218	52%	10%	51.7	2.0	1.0
8:2-2-4	440	201	54%	12%	50.4	2.0	1.0
8:2-4-2	436	203	53%	12%	53.0	2.0	1.0
8:4-2-2	456	203	55%	10%	48.8	2.0	1.0

Table 8: Statistically significant long-term memorability rate comparisons.

Superior policy/group	Inferior policy/group	p-value
Individual policy comparisons		
4 (74%)	6 (55%)	< 0.0001
4 (74%)	6:2-4 (57%)	< 0.0001
4 (74%)	6:4-2 (57%)	< 0.0001
4 (74%)	7 (46%)	< 0.0001
4 (74%)	7:3-4 (49%)	< 0.0001
4 (74%)	7:4-3 (53%)	< 0.0001
4 (74%)	8 (50%)	< 0.0001
4 (74%)	8:4-4 (52%)	< 0.0001
4 (74%)	8:2-2-4 (54%)	< 0.0001
4 (74%)	8:2-4-2 (53%)	< 0.0001
4 (74%)	8:4-2-2 (55%)	< 0.0001
6:2-4 (57%)	7 (46%)	< 0.01
6:4-2 (57%)	7 (46%)	< 0.01
8:4-2-2 (55%)	7 (46%)	< 0.05
Policy group comparisons		
Chunk (54%)	No-Chunk (51%)	< 0.05
4 (74%)	6-Chunk (57%)	< 0.05
4 (74%)	7-Chunk (51%)	< 0.0001
4 (74%)	8-Chunk (54%)	< 0.0001
6-Chunk (57%)	7 (46%)	< 0.001
6-Chunk (57%)	7-Chunk (51%)	< 0.05
8-Chunk (54%)	7 (46%)	< 0.05
4-All (74%)	6-All (56%)	< 0.0001
4-All (74%)	7-All (50%)	< 0.0001
4-All (74%)	8-All (53%)	< 0.001
6-All (56%)	7-All (50%)	< 0.001

PINs were observed to be better than that of their corresponding non-chunked PINs. Surprisingly, none of the observed increases in memorability, when using chunked PINs, compared to their non-chunked peer policies, were found to be statistically significant. Interestingly, however, chunked 6-digit PIN policies (6:2-4 and 6:4-2) both outperformed (by 11 %) non-chunked 7-digit PINs when there was no statistically significant difference found in long-term memorability of non-chunked 6- and 7-digit PINs ($p < 0.01$, pairwise corrected FET). Further, policy 8:4-2-2 outperformed policy 7 that is shorter in length (55% vs. 46%) with statistical

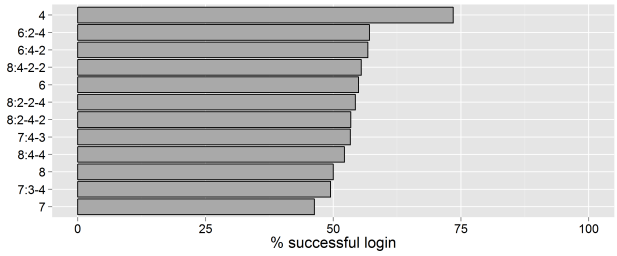

Figure 4: Sorted long-term memorability scores

significance ($p < 0.05$, pairwise corrected FET).

4.4 Memorability of chunking and non-chunking policy groups

Although we did not find any statistically significant improvement in memorability when chunking policies were compared against their non-chunked peers, Figures 3 and 4 do indicate that chunking policies might potentially outperform their non-chunked peers. To further probe effectiveness of chunking techniques, we bundled the chunking policies together and compared them as a group against the group of non-chunking policies. This section presents two such analyses.

4.4.1 Chunking policy group vs. non-chunking policy group

First, we divided the policies into a non-chunking group (No-Chunk) that consists of policies 6 and 7, and a chunking group (Chunk) that consists of policies 6:2-4, 6:4-2, 7:3-4, and 7:4-3. We excluded 4- and 8-digit PIN policies from this analysis because there is no 4-digit chunking policy, and there are twice as many 8-digit chunking policies as there are for 6-digit and 70digit which could skew the results.

As shown in Tables 9 and 10, the differences in the memorability scores between the two groups were about 1% in the short-term and about 4% in the long-term, which were statistically significant differences (all $p < 0.05$). Even though

Table 9: Short term memorability of the chunking policy group (Chunk) and the non-chunking policy group (No-chunk) for 6- and 7-digit PINs.

Policy	*# Participants*	*# Failed*	*% correct PIN*
No-chunk	2,074	71	97%
Chunk	5,381	114	98%

Table 10: Long term memorability of the chunking policy group (Chunk) and non-chunking policy group (No-chunk) for 6- and 7-digit PINs.

Policy	*# Participants*	*# Failed*	*% correct PIN*
No-Chunk	1,422	703	51%
Chunk	3,706	1,702	54%

none of the 6- and 7-digit chunking policies individually showed statistically meaningful improvement when compared to their non-chunked peer, when grouped together, they showed statistically significant difference against the non-chunked 6 and 7 policies.

4.4.2 Grouping chunking policies with the PIN length

Table 11: Short term memorability of chunking policies grouped by the PIN length.

Policy	*# Participants*	*# Failed*	*% correct PIN*
4	722	5	99%
6	714	19	97%
6-Chunk	1,426	26	98%
7	678	25	96%
7-Chunk	1,327	30	98%
8	682	27	96%
8-Chunk	2,628	58	98%

Table 12: Long term memorability of chunking policies grouped by the PIN length.

Policy	*# Participants*	*# Failed*	*% correct PIN*
4	517	137	74%
6	506	228	55%
6-Chunk	996	429	57%
7	462	248	46%
7-Chunk	922	448	51%
8	454	227	50%
8-Chunk	1,788	825	54%

In the second analysis, we grouped just the chunking policies together by their PIN length (i.e., three chunking groups of length 6 as 6-Chunk, 7 as 7-Chunk, and 8 as 8-Chunk) and compared them against their non-chunked peers as well as other chunking policy groups.

Table 11 shows the short-term memorability of those chunking policy groups. As shown in Table 6, only 8-Chunk showed statistically significant 2% improvement over its non-chunked peer policy 8 ($p < 0.05$, pairwise corrected FET). Group 6-Chunk showed statistically significant superiority over both 7 and 8 (all $p < 0.05$, pairwise corrected FET).

Long-term memorability of those chunking policy groups are shown in Table 12. In contrast to the short-term results, even policy 8-Chunk failed to show statistically significant improvement over 8. Policy 6-Chunk failed to show statistically significant superiority over 8 in the long-term, but still showed statistically significant improvement over 7 ($p < 0.001$, pairwise corrected FET). Group 6-Chunk also showed significant superiority over 7-Chunk in the long-term ($p < 0.05$, pairwise corrected FET). Similarly, policy 8-Chunk showed statistically significant superiority over 7 ($p < 0.05$, pairwise corrected FET). This can be explained by the best-performing individual 8-digit chunking policy 8:4-2-2 that outperformed policy 7 (see Table 8).

4.5 Memorability of PIN length groups

To further analyze the memorability differences between PINs of different lengths, we grouped all the policies of the same PIN length together, creating four groups of 4-All, 6-All, 7-All, and 8-All.

Table 13 shows the short-term results. As expected, group 4-All outperformed all other groups with statistical significance (all $p < 0.05$, pairwise corrected FET). Statistically significant differences are captured in Tables 6 and 8. Long-term results, presented in Table 14, were more interesting. Group 6-All at 56% showed statistically significant superiority over group 7-All (all $p < 0.001$, pairwise corrected FET), which showed the lowest memorability score of 50%. Groups 8-All (53%) and 7-All (50%), however, did not show statistically significant difference against each other. As expected, group 4-All outperformed all other groups again in the long-term with statistical significance (all $p < 0.001$, pairwise corrected FET).

Table 13: Short term memorability of four PIN length groups.

Policy	*# Participants*	*# Failed*	*% correct PIN*
4-All	761	5	99%
6-All	2,323	45	98%
7-All	2,245	55	97%
8-All	3,785	85	97%

Table 14: Long term memorability of four PIN length groups.

Policy	*# Participants*	*# Failed*	*% correct PIN*
4-All	548	137	74%
6-All	1,611	657	56%
7-All	1,534	696	50%
8-All	2,515	1052	53%

4.6 Time taken to authenticate

Tables 5 and 7 show median time taken to authenticate for the short-term test and the long-term test, respectively. Both successful and unsuccessful authentications were considered. Kruskal-Wallis test results showed that not all medians across the policies are equal, respectively, for short-term memorability ($\chi^2(11) = 1204.72$, $p < 0.0001$) and for long-term memorability ($\chi^2(11) = 360.79$, $p < 0.0001$). The Mann-Whitney U test was then used (since the time data was not normally distributed) to identify statistically significant differences in authentication times.

In the short-term memorability tests, policy 4 was the clear winner, outperforming all other policies with a median

of 9.1 seconds authentication time (all $p < 0.01$, pairwise corrected MW U test). Policy `6`, at 11.1 seconds, was the next best policy, and outperformed all other policies that required 6 or more digits to be entered (all $p < 0.01$, pairwise corrected MW U test). Similarly, policies `6:2-4` and `6:4-2` outperformed all 8-digit policies (all $p < 0.01$, pairwise corrected MW U test). Those Policies (`6:2-4` and `6:4-2`) also outperformed policies `7:3-4` and `7:4-3`, respectively (all $p < 0.01$, pairwise corrected MW U test). All 7-digit policies outperformed all chunked 8-digit policies with statistical significance ($p < 0.05$, pairwise corrected MW U test).

In the long-term memorability tests, policy `4` was the winner again with respect to authentication time (see Table 7). The observed median authentication time for policy `4` was 22.6 seconds, and its advantage over other policies was found to be statistically significant ($p < 0.01$, pairwise corrected MW U test). Interestingly, policy `6` did not fare as well as it did during short-term memorability tests. We recorded a median authentication time of 35.5 seconds, a significant increase relative to policy `4`. However, policy `6` was still lower than for all other policies we tested except for `6:4-2` (all $p < 0.05$, pairwise corrected MW U test).

4.7 Number of authentication attempts

Tables 5 and 7 also show the average numbers of authentication attempts made in the short-term and long-term tests, respectively. In the short-term test, the average number of attempts was around 1.1, with a standard deviation around 0.2-0.5. There was very little difference in the average values among all the policies, indicating that most participants entered the correct PIN on their first attempt.

In the long-term test, the average value rises to around 1.9 to 2.1, except for policy `4`, which averaged 1.7 attempts. This shows that the majority of participants made about two attempts in the long-term test. This explains the significant increase in average authentication time—we measured both successful and unsuccessful attempts—observed during long-term memorability test relative to the short-term memorability test.

4.8 User perception of recall difficulty

We compiled participants' responses to `SQ1` in Table 3, to gauge user perception of "recall difficulty" of PINs across different policies. The results for user perception of short-term and long-term recall difficulty are shown in Figures 5 and 6, respectively.

Not surprisingly, in the short-term test, the shorter the PIN length, the greater was the percentage of participants who felt that the PIN is easy to remember. The chi-square test results showed that not all recall difficulty proportions are equal ($\chi^2(44) = 604.60$, $p < 0.0001$). 90% of the participants felt 4-digit PINs are easy to remember, compared to only 61% who felt the same way about 8-digit PINs ($p < 0.01$, pairwise corrected FET). Policy `7:4-3` outperformed its non-chunked policy `7` with statistical significance ($p < 0.05$, pairwise corrected FET).

Similar trends in user perception were observed for the long-term test with one exception. In the long-term test, the chi-square test results also showed that not all recall difficulty proportions are equal ($\chi^2(44) = 257.53$, $p < 0.0001$). Specifically, the shorter the PIN length, the greater was the percentage of participants who felt that the PIN is easy to remember. However, all chunking policies with 8-digits except `8:4-4` did better than 7-digit PINs (although not found to be statistically significant) in the sample. In contrast to the short-term results, however, none of the chunked policies showed statistically significant improvement over their peer non-chunked policies.

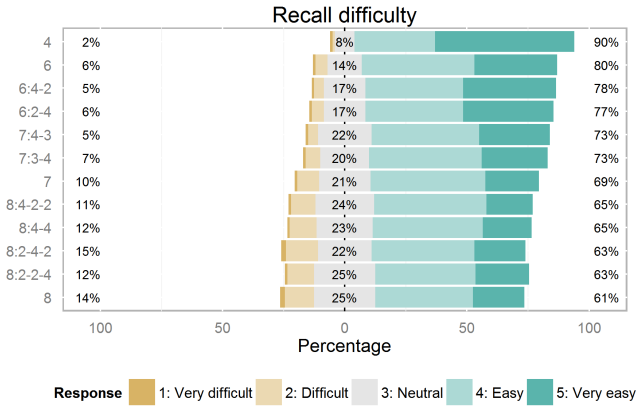

Figure 5: Results for short-term recall difficulty

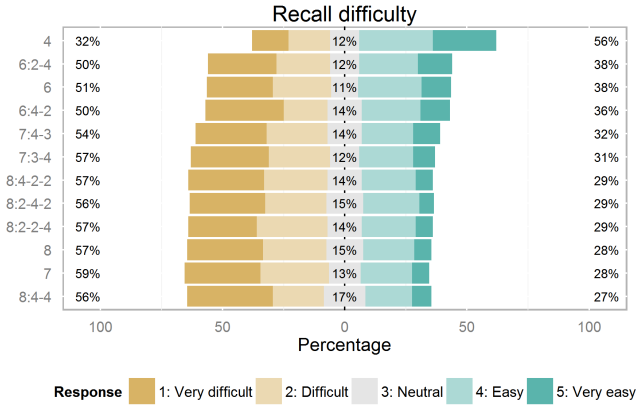

Figure 6: Results for long-term recall difficulty

4.9 Special remembrance techniques

Questions `SQ3` and `SQ4` from Table 3 asked the participants about any special techniques that they used to help them remember their PINs. Tables 15 and 16 show the relationship between the reported use of special techniques and the memorability scores for the short-term test and the long-term test, respectively. Special techniques offered no significant advantage or disadvantage in the short-term test, but in the long-term test, those who used special remembrance techniques clearly performed better than those who did not: the difference between the participants who correctly recalled their PINs was 36% ($p < 0.0001$, pairwise corrected FET). About 29% of the total number of participants reported using a special remembrance technique.

Among the 1,715 participants who used a special technique in the long-term test, 614 (36%) mentioned the use of 'keypad patterns', 166 (9.7%) mentioned the use of one of 'chunking', 'grouping', and 'splitting' technique, and 44 (2.6%) mentioned that they 'converted numbers to words' using letter associations on the keypad. Interestingly, from

the 166 participants who used some form of chunking technique, 90 (54%) were given non-chunked policies (6, 7, 8), indicating that those participants decided for themselves to use chunking.

Table 15: Use of special techniques: Short-term memorability.

Group	*# Participants*	*# Failed*	*% Successful login*
Used	2,391	41	98%
Not used	5,636	137	98%
No ans	150	12	—

Table 16: Use of special techniques: Long-term memorability.

Group	*# Participants*	*# Failed*	*% Successful login*
Used	1,715	321	81%
Not used	3,545	1,933	45%
No ans	385	288	—

4.10 Ownership of 6-digit or longer PINs

By asking SQ5 from Table 3, we found that 28% of the participants own 6-digit or longer PINs in real life. An interesting observation is that those who own 6-digit or longer PINs performed better in the long-term memorability tests than those who do not own one (see Tables 17 and 18). The long-term memorability score difference between the two groups was statistically significant: 60% versus 56% ($p < 0.01$, pairwise corrected FET). This result indicates that memorability could be improved over time with training.

Table 17: Ownership of 6-digit or longer PINs: Short-term memorability.

Group	*# Participants*	*# Failed*	*% Successful login*
Owns 6-digit PIN	2,269	47	98%
No 6-digit PIN	5,747	131	98%
No ans	161	12	—

Table 18: Ownership of 6-digit or longer PINs: Long-term memorability.

Group	*# Participants*	*# Failed*	*% Successful login*
Owns 6-digit PIN	1,479	592	60%
No 6-digit PIN	3,773	1,660	56%
No ans	393	290	—

5. DISCUSSION

Our discussion of results is organized into several topics, according to the hypotheses we set up in Section 3.1. We also offer recommendations for PIN policies in organizations.

5.1 6-digit versus 4-digit PINs

We hypothesized that "the memorability of system-generated 6-digit PINs is worse than 4-digit PINs." As apparent in Tables 5 and 7, while policy 4 clearly outperformed 6 in both short-term and long-term memorability; only the result for long-term memorability showed statistical significance. Since long-term memorability is the what is desired, our findings *accept* the first hypothesis. Further, the memorability score gap was considerable in the long-term test, in which 6-digit PINs scored 19 points lower than 4-digit PINs (almost 26% drop). 6-digit PINs also showed longer authentication times with statistical significance. Banks should consider all of those memorability and usability trade-offs when moving from 4- to 6-digit system-generated PINs.

5.2 Should banks consider using 7 and 8-digit PINs?

Our second hypothesis stated that "the memorability of system-generated 6-digit PINs is better than that of 7- and 8-digit PINs." Our results show that between policies 6, 7, and 8, there is no statistically significant difference in memorability, not providing enough evidence to accept the second hypothesis (see Tables 6 and 8).

As for authentication time, 6 outperformed both 7 and 8 in the short-term test, but only outperformed 7 in the long-term test. This indicates that 6 loses its shorter authentication time advantage over 8 over time. Looking at those results, there is no reason for banks to rule out 7- or 8-digit system-generated PINs if they are considering increasing the PIN length.

Our PIN length group analysis (see Section 4.5), which grouped all policies of the same PIN length together, showed that there is no statistically significant difference between groups 6-All and 8-All and between groups 7-All and 8-All, but showed statistically significant inferiority of 7-All against 6-All. Hence, if enhancing PIN security is a primary concern for a bank, length 8 should also be considered and carefully evaluated.

5.3 Can chunking techniques improve PIN memorability?

Our third hypothesis predicted that "the memorability of longer (6-, 7- and 8-digit) system-generated PINs improves with chunking." While we observed improvements in both short-term and long-term memorability when using chunked PINs (see Tables 5 and 7), our analysis did not show any statistically significant differences between the chunked and their peer non-chunked policies. Hence, we do not have sufficient evidence to accept the third hypothesis (see Tables 6 and 8).

Surprisingly, policies 8:4-2-2 showed statistically significant superiority of 9% in long-term memorability over 7. This was the only case where a policy of a longer PIN length outperformed a policy of a shorter PIN length with statistical significance. Similarly, When we grouped chunking policies of the same PIN length together (see Section 4.4.2) and compared them against other non-chunked and grouped chunking policies, group 8-Chunk (54%) showed statistically significant superiority of long-term memorability over 7 (46%). Further, while no statistically significant difference was found among long-term memorability of 6-, 7- and 8-digit policies, policies 6:2-4 and 6:4-2 did show statistically significant difference (57% vs. 46%) compared to

policy 7, and policy 6-Chunk outperformed both policy 7-Chunk and 7-digit PINs with statistical significance (57% vs 51%, and 57% vs. 46%; see Table 8).

Those mixed findings lead us to believe that, while chunking of system-generated random PINs may not be equally effective under all circumstances, they do show promise in certain cases and warrant a more focused study.

5.4 Policy recommendations

The findings of our study lead us to make the following recommendations for system-generated PINs.

- If a PIN length increase (from traditional 4-digit) is being considered, lengths 6 and 8 should all be considered.
- If 7- or 8- digit PIN lengths are being considered, chunking techniques such as 8:4-2-2 should be considered as chunking techniques seem to have some impact overall, and that policy in particular, can outperform shorter 7-digit PINs. However, the usability of the selected chunking policy should be studied more extensively (e.g., through a qualitative study) before deployment.

6. CONCLUSIONS AND FUTURE DIRECTIONS

We studied the memorability of system-generated PINs through a large-scale online user study, focusing on the effects of increasing the PIN length and applying number-chunking techniques that were traditionally applied to semantically meaningful chunks. Our results, not surprisingly, suggest that traditional 4-digit PINs have the best short-term and long-term memorability. While the memorability advantage of 4-digit PINs was small in the short-term, long-term memorability exhibited a significant drop when larger PIN lengths (6-, 7- and 8-digit) were used. What is interesting is that among 6-, 7-, and 8-digit PINs, we found no statistically significant difference in long-term memorability.

With regards to the effectiveness of chunking, we found that the number-chunking techniques used with larger PIN lengths did not provide a statistically significant improvement in memorability over their corresponding non-chunked PINs. However, chunked PINs did show significant improvements in some cases such as 8-digit chunking policy (0000-00-00) which exhibited statistically significant superiority in memorability against a non-chunked 7-digit policy (that is shorter in length). Further study is needed to understand this intriguing observation.

Our study used a 48 hour interval to study long-term memorability. It would be interesting to study long-term memorability of system-generated PINs using longer PIN recall intervals and compare findings as users may not necessarily use their PINS within 48 hours of assignment or use them every 48 hours. Similarly, it would be interesting to study how long-term memorability changes with multiple PIN recall sessions—especially where each recall session is also used as a remembrance training opportunity. Further, it would be interesting to study the impact on memorability when semantics are associated with chunks either through the use of mnemonics or training.

Our near term future work is to analyze the data collected to study correlations between PIN memorability and self-identified demographic and memory strength characteristics of participants.

7. ACKNOWLEDGEMENTS

This work was supported in part by the National Research Foundation of Korea (No. 2014R1A1A1003707), the ITRC (IITP-2015-H8501-15-1008), the NIPA (NIPA-2014-H0301-14-1010), the Information Trust Institute at University of Illinois, and the School of EECS at Oregon State University.

Authors would like to thank Andrew Patrick for shepherding the paper, and all the anonymous reviewers for their valuable feedback. Authors would also like to thank David Nicol of Information Trust Institute for supporting the initial study, and Ji Won Yoon for his help with the statistical analysis.

8. REFERENCES

[1] Amazon Mechanical Turk. https://www.mturk.com/mturk/welcome, 2014.

[2] hashcat advanced password recovery. http://hashcat.net/oclhashcat/, 2014.

[3] R. Atkinson and R. Shiffrin. Human memory: A proposed system and its control processes. volume 2 of *Psychology of Learning and Motivation*, pages 89 – 195. Academic Press, 1968.

[4] A. D. Baddeley and G. Hitch. Working memory. volume 8 of *The psychology of learning and motivation: Advances in research and theory*, pages 47–89. Academic Press, New York, NY, USA, 1974.

[5] M. Bishop. Password management. In *Compcon Spring '91. Digest of Papers*, pages 167–169, Feb 1991.

[6] J. Bonneau, C. Herley, P. van Oorschot, and F. Stajano. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In *Security and Privacy (SP), 2012 IEEE Symposium on*, pages 553–567, May 2012.

[7] A. S. Brown, E. Bracken, S. Zoccoli, and K. Douglas. Generating and remembering passwords. *Applied Cognitive Psychology*, 18(6):641–651, 2004.

[8] W. E. Burr, D. F. Dodson, and W. T. Polk. Electronic authentication guideline. Technical report, 2006.

[9] D. S. Carstens and L. C. Malone. Applying Chunking Theory in Organizational Password Guidelines. *Journal of Information, Information Technology, and Organizations*, 2006.

[10] N. Cowan. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1):87–114, 2001.

[11] T. J. Druzgal and M. D'esposito. Dissecting contributions of prefrontal cortex and fusiform face area to face working memory. *Journal of Cognitive Neuroscience*, 15(6):771–784, Aug. 2003.

[12] S. Fahl, M. Harbach, Y. Acar, and M. Smith. On the ecological validity of a password study. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, SOUPS '13, pages 13:1–13:13, New York, NY, USA, 2013. ACM.

[13] S. Gaw and E. W. Felten. Password management strategies for online accounts. In *Proceedings of the second symposium on Usable privacy and security*, SOUPS '06, pages 44–55, New York, NY, USA, 2006. ACM.

[14] F. Gobet, P. C. R. Lane, S. Croker, P. C. H. Cheng, G. Jones, I. Oliver, and J. M. Pine. Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, 5(6), June 2001.

[15] D. Goodin. Anatomy of a hack: How crackers ransack passwords like "qeadzcwrsfxv1331". `http://arstechnica.com/security/2013/05/how-crackers-make-minced-meat-out-of-your-passwords/`, May 2013.

[16] C. Herley and P. van Oorschot. A research agenda acknowledging the persistence of passwords. *Security Privacy, IEEE*, 10(1):28–36, Jan 2012.

[17] P. G. Inglesant and M. A. Sasse. The true cost of unusable password policies: password use in the wild. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 383–392, New York, NY, USA, 2010. ACM.

[18] S. P. Joseph Bonneau and R. Anderson. A birthday present every eleven wallets? The security of customer-chosen banking PINs. In *FC' 12: The 16 th International Conference on Financial Cryptography and Data Security*, 2012.

[19] P. Kelley, S. Komanduri, M. Mazurek, R. Shay, T. Vidas, L. Bauer, N. Christin, L. Cranor, and J. Lopez. Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms. In *Security and Privacy (SP), 2012 IEEE Symposium on*, pages 523–537, May 2012.

[20] H. Kim and J. H. Huh. PIN selection policies: Are they really effective? *Computers & Security*, 31(4):484–496, 2012.

[21] S. Komanduri, R. Shay, P. G. Kelley, M. L. Mazurek, L. Bauer, N. Christin, L. F. Cranor, and S. Egelman. Of passwords and people: measuring the effect of password-composition policies. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, pages 2595–2604, New York, NY, USA, 2011. ACM.

[22] C. Kuo, S. Romanosky, and L. F. Cranor. Human selection of mnemonic phrase-based passwords. In *Proceedings of the second symposium on Usable privacy and security*, SOUPS '06, pages 67–78, New York, NY, USA, 2006. ACM.

[23] M. A. Sasse, S. Brostoff, and D. Weirich. Transforming the 'weakest link' – a human/computer interaction approach to usable and effective security. *BT Technology Journal*, 19:122–131, July 2001.

[24] B. Schneier. Choosing Secure Passwords. `https://www.schneier.com/blog/archives/2014/03/choosing_secure_1.html`, March 2014.

[25] R. Shay and E. Bertino. A comprehensive simulation tool for the analysis of password policies. *International Journal of Information Security*, 8:275–289, August 2009.

[26] R. Shay, A. Bhargav-Spantzel, and E. Bertino. Password policy simulation and analysis. In *Proceedings of the 2007 ACM workshop on Digital identity management*, DIM '07, pages 1–10, New York, NY, USA, 2007. ACM.

[27] R. Shay, S. Komanduri, A. L. Durity, P. S. Huh, M. L. Mazurek, S. M. Segreti, B. Ur, L. Bauer, N. Christin, and L. F. Cranor. Can long passwords be secure and usable? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 2927–2936, 2014.

[28] R. Shay, S. Komanduri, A. L. Durity, P. S. Huh, M. L. Mazurek, S. M. Segreti, B. Ur, L. Bauer, N. Christin, and L. F. Cranor. Can long passwords be secure and usable? In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI '14, pages 2927–2936, New York, NY, USA, 2014. ACM.

[29] R. Shay, S. Komanduri, P. G. Kelley, P. G. Leon, M. L. Mazurek, L. Bauer, N. Christin, and L. F. Cranor. Encountering stronger password requirements: user attitudes and behaviors. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, SOUPS '10, pages 2:1–2:20, New York, NY, USA, 2010. ACM.

[30] M. A. Thornton and A. R. A. Conway. Working memory for social information: Chunking or domain-specific buffer? *NeuroImage*, 70:233–239, 2013.

[31] B. Ur, P. G. Kelley, S. Komanduri, J. Lee, M. Maass, M. L. Mazurek, T. Passaro, R. Shay, T. Vidas, L. Bauer, N. Christin, and L. F. Cranor. How Does Your Password Measure Up? The Effect of Strength Meters on Password Creation. In *Proceedings of the 21st USENIX Conference on Security Symposium*, Security '12, 2012.

[32] 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*, 65(8):744–757, 2007.

[33] J. Yan, A. Blackwell, R. Anderson, and A. Grant. Password memorability and security: empirical results. *Security Privacy, IEEE*, 2(5):25–31, 2004.