On the Security of Smartphone Unlock PINs

PHILIPP MARKERT and DANIEL V. BAILEY, Ruhr University Bochum, Germany
MAXIMILIAN GOLLA, Max Planck Institute for Security and Privacy, Germany
MARKUS DÜRMUTH, Ruhr University Bochum, Germany
ADAM J. AVIV, The George Washington University, USA

In this article, we provide the first comprehensive study of user-chosen four- and six-digit PINs (n = 1705) collected on smartphones with participants being explicitly primed for device unlocking. We find that against a throttled attacker (with 10, 30, or 100 guesses, matching the smartphone unlock setting), using six-digit PINs instead of four-digit PINs provides little to no increase in security and surprisingly may even decrease security. We also study the effects of blocklists, where a set of "easy to guess" PINs is disallowed during selection. Two such blocklists are in use today by iOS, for four digits (274 PINs) as well as six digits (2,910 PINs). We extracted both blocklists and compared them with six other blocklists, three for each PIN length. In each case, we had a small (four-digit: 27 PINs; six-digit: 29 PINs), a large (four-digit: 2,740 PINs; six-digit: 291,000 PINs), and a placebo blocklist that always excluded the first-choice PIN. For four-digit PINs, we find that the relatively small blocklist in use today by iOS offers little to no benefit against a throttled guessing attack. Security gains are only observed when the blocklist is much larger. In the six-digit case, we were able to reach a similar security level with a smaller blocklist. As the user frustration increases with the blocklists size, developers should employ a blocklist that is as small as possible while ensuring the desired security. Based on our analysis, we recommend that for four-digit PINs a blocklist should contain the 1,000 most popular PINs to provide the best balance between usability and security and for six-digit PINs the 2,000 most popular PINs should be blocked.

CCS Concepts: • **Security and privacy → Authentication**; **Usability in security and privacy;**

Additional Key Words and Phrases: Security, usability, authentication, PIN, blocklist, mobile, smartphone

ACM Reference format:
Philipp Markert, Daniel V. Bailey, Maximilian Golla, Markus Dürmuth, and Adam J. Aviv. 2021. On the Security of Smartphone Unlock PINs. *ACM Trans. Priv. Secur.* 24, 4, Article 30 (September 2021), 36 pages.
https://doi.org/10.1145/3473040

This research was supported by the research training group "Human Centered Systems Security" sponsored by the state of North Rhine-Westphalia, Germany, and funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2092 CASA – 390781972. This material is based upon work supported by the National Science Foundation under Grant No. 1845300. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We responsibly disclosed all our findings to Apple Inc.

Authors' addresses: P. Markert, D. V. Bailey, and M. Dürmuth, Ruhr University Bochum, Bochum, Germany; emails: philipp.markert@rub.de, danbailey@sth.rub.de, markus.duermuth@rub.de; M. Golla, Max Planck Institute for Security and Privacy, Bochum, Germany; email: maximilian.golla@csp.mpg.de; A. J. Aviv, The George Washington University, Washington, District of Columbia, USA; email: aaviv@gwu.edu.

2471-2566/2021/09-ART30 $15.00
https://doi.org/10.1145/3473040

1 INTRODUCTION

We provide the first study focused on the selection of **Personal Identification Numbers (PINs)** based on data collected from users specifically primed for the smartphone setting. While authentication on mobile devices has been studied in several contexts, including patterns [52] and passwords [37], little is known about PINs used for mobile authentication. Despite the rise of biometrics, such as fingerprint or facial recognition, devices still require PINs, e.g., after a restart or when the biometric fails. That is because the biometric does not replace knowledge-based authentication; access to a device is still possible with a PIN even when using a biometric. Moreover, the presence of a biometric may actually lead to a false sense of security when selecting knowledge-based authenticators [17].

Our study focuses on the PINs users choose to unlock their mobile devices. Previous work on PINs was primarily focused on the context of banking, e.g., as part of the *Chip-and-PIN* system [13], and also mainly relied on the analysis of digit sequences found in leaked text-based password datasets, since these data are more readily available [59]. Given the sparsity of information about PINs in the context of mobile authentication, we sought to fill this vital knowledge gap by conducting the first study ($n = 1705$) on the topic where participants either selected a four- or six-digit PIN, the two predominant PIN lengths used for device unlock. In addition to only allowing participants to complete the study on a smartphone, we also primed them specifically for the mobile unlock authentication setting, reminding participants that the selected "PIN protects [their] data and is used to unlock [their] smartphone." While our study cannot speak to the memorability of the selected PINs due to the short time duration, our qualitative feedback suggests that participants took this prompt seriously and selected relevant PINs.

PINs of four and six digits only provide security when paired with system controls like lockouts and delays that limit offline (or *unthrottled*) guessing. An unthrottled attacker who can bypass these controls can quickly guess all PIN combinations. We instead consider a *throttled* attacker model to empirically analyze the security of PINs when the system limits the guessing rate. This is usual in the smartphone-unlocking setting where pauses are enforced after a certain number of wrong guesses to slow attacks down. Guessing is then limited (or throttled) to, e.g., just 10, 30, or 100 attempts in a reasonable time window, such as a few hours. In such a model, it is essential to prioritize guessing resistance in the first few guesses. Our study found little benefit to longer six-digit PINs compared to four-digits. In fact, our participants tend to select more-easily guessed six-digit PINs when considering the first 40 guesses of an attacker.

As a mechanism for improving PIN selection, we also studied how PINs are affected by blocklisting. A blocklist is a set of "easy to guess" PINs, which triggers a warning to the user. Apple iOS devices show the warning *"This PIN Can Be Easily Guessed"* with a choice to *"Use Anyway"* or *"Change PIN."* Previous work in text-based passwords has shown that users choose stronger passwords due to a blocklist [31, 47], and recent guidance from NIST [25] concurs. To understand selection strategies in the presence of a blocklist, we conducted a between-subjects comparison of PIN selection using a number of different blocklists. This included two small (27 four-digit PINs and 29 six-digit PINs), two large (2,740 four-digit PINs and 291,000 six-digit PINs), and two blocklists (274 four-digit PINs and 2,910 six-digit PINs) in use today on iOS devices, which we extracted for this purpose. To determine if the experience of hitting a blocklist or the content of the blocklist itself drives the result, we included *placebo* blocklists that always excluded the participants' first choice. Finally, we included both enforcing and non-enforcing blocklists, where participants were able to "click through" and ignore the blocklist, the approach taken by iOS. Despite the popularity of blocklists and the positive impact on textual passwords, our results show that currently employed PIN blocklists are ineffective against a throttled attacker, in both the enforcing and non-enforcing

setting. This attacker performs nearly as well at guessing PINs as if there were no blocklist in use. To be effective, the blocklist would need to be much larger, leading to higher user frustration. Our results show that for four-digit PINs a blocklist of about 10% of the PIN space may be able to balance the security and usability needs, for six-digit PINs the same effect can be achieved by blocking 0.2% of the keyspace.

Finally, we collected both quantitative and qualitative feedback from our participants about their PIN selection strategies, perceptions of their PINs in the context of blocklists, and their thoughts about blocklisting generally. Overall, we find that despite having mostly negative sentiments about blocklist warnings, participants do perceive the PINs they select under a blocklist as more secure without impacting the memorability and convenience, except in situations of a very large blocklist. To summarize, we make the following contributions:

(1) We report on the security of four- and six-digit PINs as measured for smartphone unlocking, finding that in the throttled setting, the benefit of six-digit PINs is marginal and sometimes worse than that of four-digit PINs.
(2) Considering a realistic, throttled attacker model, we show how different blocklisting approaches influence PIN selection process for both security and usability, finding that blocklists in use today offer little to no added security.
(3) Through quantitative and qualitative feedback, we explore users' perception of security, memorability, and ease-of-use of PIN-based authentication, finding that participants perceive that blocklisting will improve their PINs without impacting usability, except for very large blocklists.
(4) We provide guidance for developers on choosing an appropriately-sized PIN blocklist that can influence the security in the throttled scenario, finding that a blocklist for four-digit PINs should consist of ∼1,000 PINs to have a noticeable impact while minimizing the negative effects. To achieve the same results in the six-digit case, ∼2,000 PINs should be blocked.

Note: We responsibly disclosed all our findings to Apple Inc.

2 RELATED WORK

Research on PIN authentication for mobile devices is related to the larger area of mobile authentication. User preferences for different unlock methods for Android devices were studied by Harbach et al. [28] in 2014. Since then, PINs have found new uses in encrypting mobile devices [4, 6, 41] and biometrics [17] that require a PIN as part of the keying material and for fallback authentication when biometrics fail. Today, PINs are also used in various situations in our everyday life, e.g., for gym lockers and safes, but also smarthomes and voicemail accounts [16, 32].

The work most closely related to this research is the analysis of PINs in the context of *Chip-and-PIN* systems done by Bonneau et al. [13], where they considered four-digit PIN creation strategies for banking customers for use with ATMs/credit cards. Bonneau et al. identified techniques used for selecting PINs, where choosing (birth) dates/years was the most popular—also true in our setting. As noted, an attacker can leverage the skewed distribution of PIN choices to improve the guessing strategy. As a countermeasure, Bonneau et al. proposed the use of a blocklist containing the 100 most popular PINs. From our analysis, it seems that their suggestion may have formed the basis for Apple iOS's four-digit blocklist.

Our work differs from Bonneau et al. in two significant ways. Foremost, Bonneau et al. were primarily concerned with payment cards, not smartphone unlock authentication. Second, Bonneau et al. did not collect new PINs but instead relied on digit sequences found in leaked passwords along with PINs collected without the benefit of a controlled experiment [3]. Our work aims for greater ecological validity by specifically priming users for this task. Our data suggests that using password leaks may be an imperfect approximation for how users choose PINs for unlock authentication.

Wang et al. [59] have also analyzed the security of PINs—in this case without any specific usage context. They report on comparing four- and six-digit PINs created by English and Chinese users. One counterintuitive finding is that six-digit PINs are less resistant to online attacks, despite the key space expansion from four- to six-digit PINs. Our results support the observation that in a rate limited guessing scenario there may actually be no benefit of using six-digit PINs at all, and in certain cases security even decreases. Wang et al. used PINs extracted from leaked, text-based password datasets, whereas we tend to increase the ecological validity of our results by collecting new PINs specifically primed for mobile authentication and the smartphone form-factor with its standard PIN layout.

Blocklists have been considered in the context of PINs by Kim et al. [33]. They tested blocklists for both four-digit as well as six-digit PINs and concluded that a reasonably sized blocklist could indeed increase the security. Kim et al. used *Shannon entropy* and *guessing entropy* as the strength metric and thus only consider an unthrottled, perfect knowledge attacker that will exhaustively guess the PIN space [12]. This is a questionable attacker model especially given the sparsity of their dataset. Kim et al. compared blocklists representing 2% and 32% of the possible PIN space and found the large blocklist led to lower Shannon-entropy and lower offline guessing-entropy PINs, perhaps due to the composition of Kim et al.'s large blocklist. In contrast, we show in our analysis of four-digit PINs that with a more realistic rate-limited, online attacker, a larger blocklist containing 27.4% of all possible PINs provides a benefit over a smaller one that blocklists only 2.7%, differing from the suggestion of Kim et al. regarding the effect of the size of the blocklist. We also make similar observations in our analysis of six-digit PINs.

Beyond PINs, another common knowledge-based mobile authentication mechanism are Android unlock patterns, whereby a user selects a pattern that connects points on a 3×3 grid. Uellenbeck et al. [52] showed that user selection of unlock patterns is highly biased, e.g., most patterns start in the upper left corner. These results have been confirmed by other works [7, 35, 39, 58]. Most relevant to our study, we compare the security of mobile unlock PINs to that of patterns and have obtained datasets from related work [7, 35, 52, 58].

While less common, according to Harbach et al. [28] and our own measurement (see Table 4), alphanumeric passwords are another option for users to unlock their mobile devices. For this reason, we also consider alphanumeric passwords in our comparisons with PINs, as available in leaked, text-based password datasets. Research has shown that the creation and use of passwords on mobile devices can be cumbersome and users may create weaker passwords than they would do on full-sized keyboards [26, 37, 46, 57, 63]. To counteract this, blocklists can be employed that is also the recommendation for password-based authentication in general [51].

2.1 Difference to Conference Version

Parts of this work have been presented at the 41st IEEE Symposium on Security and Privacy in May 2020 [36]. The results presented here substantially expand on this research by including data from 485 additional participants providing six-digit PINs, as well as five new analyses that were not part of the conference version. In detail, we aligned the Control-6-digit treatment to the size of Control-4-digit and added three completely new six-digit treatments: a non-enforcing iOS treatment (iOS-6-digit-nCt), as well as two data-driven ones (DD-6-digit-29, and DD-6-digit-291000). We provide extended analysis on the effects of using biometrics on the PIN strength (Section 5.1), and we further investigate the underlying PIN selection strategies to gather more insights into how users select their six-digit PINs (Section 5.2). Furthermore, we analyze the effects of blocklists on PIN creation and entry times (Section 6.1), investigate the bias in users' selection of certain

digits when composing their PIN (Section 6.4), and study the shifts in PIN selection strategies after encountering a blocklist warning (Section 6.5).

Summary of New Findings. Based on the new PIN data for six-digit PINs, we make new recommendations for the size of the six-digit blocklist; it should be ~2,000 PINs to ideally balance the positive and negative effects. The creation and entry times confirm this results as they are marginally affected compared to larger blocklist sizes. At the same time, we observe that users come up with more complex selection strategies and the composition of PINs at this blocklist size, leading to higher security. An attacker who wants to guess a six-digit PIN and is aware of a previously used four-digit PIN can abuse this knowledge.

3 BACKGROUND

In this section, we define our attacker model, describe the used datasets, and outline the extraction of the two iOS PIN blocklists that we evaluate in our user study.

3.1 Attacker Model

When studying guessing attackers, there are two primary threat models. An *unthrottled* attacker can guess *offline*, indefinitely, until all the secrets are correctly guessed, while a *throttled* attacker is limited in the number of guesses, sometimes called an *online* attack. Google's Android and Apple's iOS, the two most popular mobile operating systems, implement real-world rate limiting mechanisms to throttle attackers, because, otherwise, it would be possible to simply guess all PIN combinations. In our attacker model, we assume the rate limiting works as designed, and as such, it is appropriate to consider a throttled attacker when evaluating security as this best matches the reality of the attacks PINs must sustain for the mobile unlock setting.

The choice of the throttled attack model is further justified when considering mobile devices' **trusted execution environments (TEE)**, where the key for device encryption is stored in "tamper resistant" hardware and is "entangled" with the user's unlock secret [6]. This forces the attacker to perform decryption (unlock) attempts on the device itself in an online way. Moreover, the TEE is used to throttle the number of decryption attempts tremendously by enforcing rate limiting delays that also survive reboots. There is some research [27, 34, 48] and even tools [18, 29, 60] that exploit vulnerabilities in an attempt to escalate guessing to an unthrottled attacker. Moreover, there are companies that sell commercial solutions like Azimuth [40], Cellebrite [15], Elcomsoft [1], and GrayShift [14]. However, we consider such attacks out of scope. These exploits are usually bound to a specific OS or device version (e.g., iPhone 5) or can only be run within certain timeframes (e.g., 1 hour) after the last successful unlock [1, 61].

An overview of the currently enforced limits is given in Table 1. Apple's iOS is very restrictive and only allows up to 10 guesses [6] before the iPhone disables itself and requires a reset. Google's Android version 7 or newer are less restrictive with a first notable barrier at 30 guesses where the waiting time increases by 10 minutes. We define the upper bound for a reasonably invested throttled attacker at 100 guesses when the waiting starts to exceed a time span of 10 hours on Android [5], but we also report results for less determined attackers at 10 guesses (30 s) and 30 guesses (10.5 m) for Android. The iOS limit is 10 guesses (1.5 h) [6].

In our attacker model, we assume that the adversary has no background information about the owner of the device or access to other side-channels. In such a scenario, the best approach for an attacker is to guess the user's PIN in decreasing probability order. To derive this order, we rely on the best available PIN datasets, which are the Amitay-4-digit and RockYou-6-digit datasets as defined below. Again, we only consider an *un-targeted attacker* who does not have additional information about the victim. If the attacker is targeted, and is able to use other information and

Table 1. Rate Limiting on Mobile Operating Systems

To Make *n* Guesses	Accumulated Waiting Time	
	Android 7, 8, 9, 10, 11	iOS 9, 10, 11, 12, 13, 14
1–5 guesses	0 s	0 s
6 guesses	30 s	1 m 0 s
7 guesses	30 s	6 m 0 s
8 guesses	30 s	21 m 0 s
9 guesses	30 s	36 m 0 s
10 guesses	30 s	1 h 36 m 0 s
30 guesses	10 m 30 s	—
100 guesses	10 h 45 m 30 s	—
200 guesses	67 d 2 h 45 m 30 s	—

context about the victim, e.g., via shoulder-surfing attack [8, 10, 11, 46] or screen smudges [9], then the attacker would have significant advantages, particularly in guessing four- vs. six-digit PINs [10].

In other parts of this work, we make use of blocklists. In those cases, we consider an attacker that is aware and in possession of the blocklist. This is because the attacker can crawl the system's blocklist on a sample device, as we have done for this work. Hence, with knowledge of the blocklist, an informed attacker can improve the guessing strategy by *not* guessing known-blocked PINs and instead focusing on common PINs not on the blocklist.

3.2 Datasets

Perhaps the most realistic four-digit PIN data are from 2011 where Daniel Amitay developed the iOS application "Big Brother Camera Security" [3]. The app mimicked a lock screen allowing users to set a four-digit PIN. Amitay anonymously and surreptitiously collected 204,432 four-digit PINs and released them publicly [3]. While collected in an uncontrolled experiment, we apply the dataset (Amitay-4-digit) when guessing four-digit PINs, as well as to inform the selection of our "data-driven" blocklists. As there is no similar six-digit PIN data available to inform the attacker, we rely on six-digit PINs extracted from password leaks, similarly to Bonneau et al.'s [13] and Wang et al.'s [59] method. PINs are extracted from consecutive sequences of exactly *n*-digits in leaked password data. For example, if a password contains a sequence of digits of the desired length, then this sequence is considered as a PIN (e.g., PW: `ab3c123456d` → PIN: `123456`, but no six-digit PINs would be extracted from the sequence `ab3c1234567d`). By following this method, we extracted six-digit PINs from the *RockYou* password leak, which we refer to as RockYou-6-digit (2,758,490 PINs). We also considered six-digit PINs extracted from other password leaks, such as the *LinkedIn* [24] dataset, but found no marked differences between the datasets.

To provide more comparison points, we consider a number of other authentication datasets listed in Table 2. For example, we use a 3×3 Android unlock pattern dataset described by Golla et al. [22], combining four different datasets [7, 35, 52, 58]. It consists of 4, 637 patterns with 1, 635 of those being unique. In addition, we use a text-password dataset. Melicher et al. [37] found no difference in strength between passwords created on mobile and traditional devices considering a throttled guessing attacker. Thus, we use a random sample of 10,000 passwords from the LinkedIn [24] leak and use the *Pwned Passwords v7* [30] list to simulate a throttled guessing attacker to estimate the guessing resistance for the sampled LinkedIn passwords as a proxy for mobile text passwords.

Table 2. Datasets for Strength Estimations and Comparisons

Kind	Dataset	Samples
Four-digit PINs	Amitay-4-digit [3]	204,432
Four-digit PINs	RockYou-4-digit [59]	1,780,587
Six-digit PINs	RockYou-6-digit [59]	2,758,490
3 × 3 Patterns	"All" unlock patterns [22]	4,637
Passwords	LinkedIn [24]	10,000
Passwords	Pwned Passwords v7 [30]	*Top* 10,000

Fig. 1. The installation used to extract the iOS blocklists.

3.3 Extracting the iOS Blocklists

As part of our set of blocklists, we also consider a blocklist of "easily guessed" four-/six-digit PINs as used in the wild by Apple, which we obtained via brute-force extraction from an iPhone running iOS 12. We were able to verify that blocklisting of PINs is present on iOS 9 throughout the latest version iOS 14, and we also discovered that Apple updated their blocklist with the deployment of iOS 10 (e.g., the PIN 101471 is blocked on iOS 10.3.3 but is not on iOS 9.3.5). In theory, it is possible to extract the blocklist by reverse engineering iOS, yet we found a more direct way to determine the blocklist via brute force: During device setup, when a PIN is first chosen, there is no throttling. To test the membership of a PIN, one only needs to enter *all* the PINs, observe the presence of the blocklist warning, and then intentionally fail to re-enter the PIN to be able to start over. We constructed a device to automate this process using a Raspberry Pi Zero W equipped with a Pi Camera Module (8MP), as depicted in Figure 1. The Raspberry Pi emulates a USB keyboard,

which is connected to the iPhone. After entering a PIN, the camera of the Raspberry Pi takes a photo of the iPhone screen. The photo is sent to a remote server, where it is converted to grayscale and thresholded using *OpenCV*. Subsequently, the presence of the blocklist warning, as depicted in Figure 4, is detected by extracting the text in the photo using *Tesseract OCR*.

The extraction of all 10,000 four-digit PINs took ~9 hours. Testing all 1 million six-digit PINs took about 30 days using two setups. We repeated the process for four-digit PINs multiple times, tested lists of frequent six-digit PINs, and verified the patterns found in the PINs. Moreover, we validated all blocked PINs multiple times. We refer to these two lists as the iOS-4 and iOS-6 blocklists.[1] In total, the four-digit blocklist contains 274 PINs and includes common PINs as well as years from 1956 to 2015, but its composition is mostly driven by repetitions such as `aaaa`, `abab`, or `aabb`. The six-digit blocklist contains 2,910 PINs and includes common PINs as well as ascending and descending digits (e.g., `543210`), but its composition is, again, mostly driven by repetitions such as `aaaaaa`, `abcabc`, or `abccba`. The common PINs blocked by Apple overlap with a four-digit blocklist suggested by Bonneau et al. [13] in 2012 and the top six-digit PINs reported by Wang et al. [59] in 2017.

4 USER STUDY

In this section, we outline the treatment conditions, the user study, and the collected data. We also discuss limitations and our ethical considerations. Appendix A.1 outlines the entire questionnaire.

4.1 Study Protocol and Design

We conducted a user study of four- and six-digit PINs using Amazon **Mechanical Turk (MTurk)** with $n = 1{,}705$ participants. To mimic the PIN creation process in our browser-based study, participants were restricted to mobile devices by checking the user-agent string. We applied a 12-treatment, between-subjects study protocol for the PIN selection criteria, e.g., four- vs. six-digit with or without blocklisting. The specifics of the treatments are discussed in detail in Section 4.2. At the end of the study, we collected 851 and 854 PINs, four- and six-digits, respectively, for a total of 1,705 PINs as our core dataset. These PINs were all selected, confirmed, and recalled. We additionally recorded all intermediate PIN selections, such as what would happen if a selected PIN was *not* blocked and the participant did not have to select a different PIN. For more details of different kinds of PINs collected and analyzed, refer to Table 7. All participants were exposed to a set of questions and feedback prompts that gauged the security, memorability, and usability of their selected PINs, as well as their attitudes toward blocklisting events during PIN selection.

The survey itself consists of 10 parts. Within each part, to avoid ordering effects, we applied randomization to the order of the questions that may inform later ones; this information is also available in Appendix A.1. The parts of the survey are as follows:

(1) *Informed Consent*: All participants were informed of the procedures of the survey and had to provide consent. The informed consent notified participants that they would be required to select PINs in different treatments but did not inform them of any details about blocklisting that might be involved in that selection.
(2) *Agenda*: After being informed, participants were provided additional instructions and details in the form of an *agenda*. It stated the following: "You will be asked to complete a short survey that requires you to select a numeric PIN and then answer some questions about it afterwards. You contribute to research so please answer correctly and as detailed as possible."

[1] To foster future research on this topic, we share the described blocklists and the PIN datasets at https://this-pin-can-be-easily-guessed.github.io.

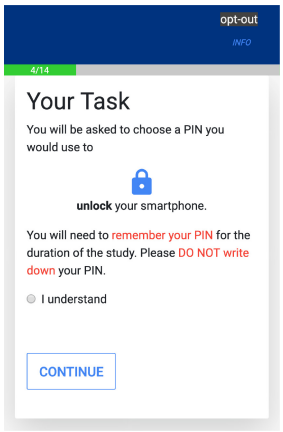


Fig. 2. Priming information provided before the participants were asked to create a PIN.

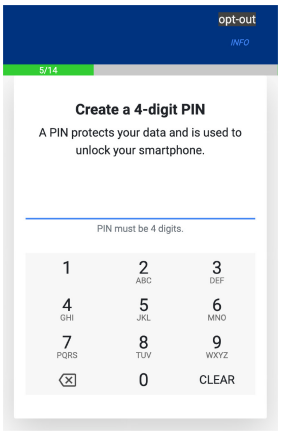


Fig. 3. The design of the page on which we asked the participants to create a PIN.

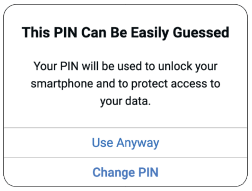


Fig. 4. Blocklist warning *with* the ability to "click through."

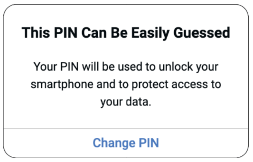


Fig. 5. Blocklist warning *without* the ability to "click through."

(3) *Practice*: Next, participants practiced with the PIN entry screen, which mimics typical PIN selection on mobile devices, including the "phoneword" alphabet on the virtual PIN pad. The purpose of the practice round was to ensure that participants were familiar with the interface prior to selecting a PIN. There was clear indication during the practice round that this was practice and that participants would begin the primary survey afterwards.
(4) *Priming*: After familiarization and before selection, participants were further primed about mobile unlock authentication and PINs using language similar to what iOS and Android use during PIN selection. A visual of the priming is in Figure 2. A lock icon was used to prime notions of security, and users were reminded that they will need to remember their PIN for the duration of the study without writing it down. Participants must click "I understand" to continue. The qualitative feedback shows that the priming was understood and followed with some participants even stating that they reused their actual PIN.
(5) *Creation*: The participants then performed the PIN creation on the page shown in Figure 3. The PIN was entered by touching the digits on the virtual PIN pad. As usual, users had to enter the PIN a second time to confirm it was entered correctly. Depending on the treatment (see Section 4.2), the users either selected a four- or six-digit PIN and did or did not experience a blocklist event. In Figure 4 and Figure 5 we depicted the two blocklist warnings that either allowed participants to "click through" the warning (or not). The feedback was copied to directly mimic the wording and layout of a blocklist warning used by Apple since iOS 12.
(6) *Blocklisting Followup*: After creation, we asked participants about their attitudes and strategies with blocklisting. If the participants experienced a blocklist event, then we referred back to that event in asking followup questions. Otherwise, we asked participants to "imagine" such an experience. These questions form the heart of our qualitative analysis (see Section 6.7).
(7) *PIN Selection Followup*: We asked a series of questions to gauge participants' attitudes toward the PIN they selected with respect to its security and usability, where usability was appraised based on ease of entry and memorability (see Section 6.6). As part of this questionnaire, we also asked an attention check question. We excluded the data of 19 participants, because we could not guarantee that they followed our instructions completely.

Table 3. Overview of Studied Treatments

	Treatment	Short Name	Blocklist	Size	Click-thr.
4 digits	Control-4-digit	Con-4	–	–	–
	Placebo-4-digit	Pla-4	First choice	1	✗
	iOS-4-digit-wCt	iOS-4-wC	iOS four-digit	274	✓
	iOS-4-digit-nCt	iOS-4-nC	iOS four-digit	274	✗
	DD-4-digit-27	DD-4-27	Top Amitay	27	✗
	DD-4-digit-2740	DD-4-2740	Top Amitay	2740	✗
6 digits	Control-6-digit	Con-6	–	–	–
	Placebo-6-digit	Pla-6	First choice	1	✗
	iOS-6-digit-wCt	iOS-6-wC	iOS six-digit	2910	✓
	iOS-6-digit-nCt	iOS-6-nC	iOS six-digit	2910	✗
	DD-6-digit-29	DD-6-29	Top RockYou	29	✗
	DD-6-digit-291000	DD-6-291000	Top RockYou	291000	✗

(8) *Recall:* On this page, participants were asked to recall their earlier selected PIN. Although the two prior parts formed distractor tasks we do not expect that the recall rates measured here speak broadly for the memorability of these PINs. As expected, nearly all participants could recall their selected PIN.
(9) *Demographics:* In line with best practice [43], we collected the demographics at the end, including participants' age, gender, IT background, and their current mobile unlock scheme.
(10) *Honesty/Submission:* Finally, we asked if the participants provided "honest" answers to the best of their ability. We informed them that they would be paid even if they indicated dishonesty. Using this information in combination with the attention check described above, we excluded the data of 19 participants to ensure the integrity of our data. After affirming honesty (or dishonesty), the survey concluded and was submitted.

4.2 Treatments

We used 12 different treatments: 6 treatments for four-digit PINs and 6 treatments for six-digit PINs. The details for each treatment can be found in Table 3.

4.2.1 Control Treatments. For each PIN length, we had a control treatment, *Control-4-digit* and *Control-6-digit*, that primed participants for mobile unlock authentication and asked them to select a PIN without any blocklist interaction. These PINs form the basis of our four- and six-digit mobile-authentication primed PIN dataset. In total, we have 231 control four-digit PINs and 236 control six-digit PINs. We also created two additional datasets, *First-Choice-4-digit* (851 PINs) and *First-Choice-6-digit* (854 PINs), by combining the control PINs with those chosen by participants from other treatments in their "first attempt" before having been subjected to any blocklist.

4.2.2 Blocklist Treatments. The remaining treatments considered PIN selection in the presence of a blocklist. There are two types of blocklist implementations: *enforcing* and *non-enforcing*. An enforcing blocklist does not allow to continue as long as the selected PIN is blocked; the user *must* select an unblocked PIN. A non-enforcing blocklist warns the user that the selection is blocked, but the user can choose to ignore the feedback and proceed anyway. We describe this treatment as providing the participant an option to *click through*. Otherwise, the treatment uses an enforcing blocklist. Visuals of the non-enforcing and enforcing feedback can be found in Figures 4 and 5.

Placebo Blocklist. As we wanted to determine if the experience of hitting a blocklist or the content of the blocklist itself drive the results, we included a *placebo* treatment for both four- and six-digit PINs (*Placebo-4-digit* and *Placebo-6-digit*). In this treatment, the user's first choice PIN was blocked, forcing a second choice. As long as the second choice differed from the first, it was accepted.

iOS Blocklist. For this treatment, we included the blocklists used on Apple's iOS 13. The four-digit iOS blocklist contains 274 PINs (2.74% of the available four-digit PINs), and the six-digit iOS blocklist contains 2,910 PINs (0.291% of the available six-digit PINs). These blocklists provide measurements of real scenarios for users selecting PINs on iOS devices. As iOS allows users to "click through" the blocklist warning and use their blocked PIN anyway, we implemented our blocklisting for the iOS condition in the same way (i.e., conditions *iOS-4-digit-wCt* and *iOS-6-digit-wCt*). To understand the effect of non-enforcing blocklists, we also tested enforcing versions of the iOS blocklists (*iOS-4-digit-nCt* and *iOS-6-digit-nCt*).

Data-Driven Blocklists. We considered two blocklists for each PIN length that are significantly smaller and larger than the iOS blocklist. The blocklists were constructed using the most frequently occurring PINs in the Amitay-4-digit and RockYou-6-digit dataset. We refer to the four-digit treatments as *DD-4-digit-27* and *DD-4-digit-2740*, because the blocklists contain 27 and 2,740 PINs respectively. Following this, we blocked the 29 most frequent PINs in the treatment *DD-6-digit-29* while 291,000 were blocked in *DD-6-digit-291000*.

When comparing the two data-driven four-digit blocklists and the one used in iOS, it can be seen that they are differently composed. While 22, i.e., 82%, of the PINs contained in DD-4-digit-27 are blocked in iOS, there are also 5 PINs that are not. Surprisingly, these PINs correspond to simple patterns like `0852`, which is a bottom-up pattern across the PIN pad or `1379`, the four corners of the pad chosen in a left-to-right manner. Now, when extending the comparison to the DD-4-digit-2740 blocklist, we see that 258 of the 274 PINs from the iOS blocklist, i.e., 92%, are also blocked by our large data-driven blocklist. The remaining 16 PINs all follow the same repetitive aabb scheme, e.g., `0033`, `4433`, or `9955`. Interestingly, only one of those PINs, `9933`, was selected in our study that shows that double repetitions are presumably not as common as Apple expects.

Similar observations can be made in the six-digit case when comparing the iOS blocklist with the two data-driven versions. There are 3 PINs (`159357`, `147852`, `246810`) in our DD-6-digit-29 blocklist with only 29 PINs that are not rejected by Apple's blocklist with 2,910 entries. Of those 3 PINs, at least `159357` and `147852` follow straightforward patterns that one may expect to be blocked. The intersection with the large data-driven blocklist covers 2,314 PINs, i.e., 80% of the iOS blocklist. The 596 PINs that are solely rejected by Apple follow three schemes: ababac (323 PINs), abccba (258 PINs), and abcabc (15 PINs). Again, those schemes are not very popular across our participants: only 7% of the PINs that were selected in our study follow them.

4.3 Recruitment and Demographics

Using Amazon's MTurk, we recruited a total of 1,944 participants. After excluding a portion due to invalid responses to attention tests or survey errors, we had 1,705 participants remaining. We required our participants to be 18 years or older, reside in the US (as checked by MTurk), and have at least an 85% approval rate on MTurk. The **Institutional Review Board (IRB)** approval required focusing on participants residing in the US, but there may be a secondary benefit to this: US residents often do not have *chip-and-PIN* credit cards (although, they do use four-digit ATM PINs), in contrast to residents in Europe or Asia, and thus may associate PIN selection more strongly with mobile device locking. In any case, participants were explicitly primed for the mobile

device unlock setting. Participants indicated they understood this instruction, and their qualitative responses confirm this.

We also reviewed all of the participants' responses for consistency, including answers to attention check questions, the honesty question, and speed of entry. We removed 19 who provided inconsistent data but did not "reject" any participants on Amazon Mechanical Turk. Participants were compensated with $1 (USD) for completion; the survey took on average 5 minutes for an hourly rate of $12.

Demographics and Background. As is typical on MTurk, our sample is relatively young and better educated than the general US population. Of the participants, 923 identified as male (54%) while 768 (45%) identified as female (1% identified as other or preferred not to say), and the plurality of our participants was between 25 and 34 years old (48%). Most participants had some college (21%) or a bachelor's degree (42%), and a few (11%) had a master's or doctoral degree. While 28% described having a technical background, 69% described not having one. We have the full details of the demographics responses in Appendix A.2 in Table 10.

Smartphone OS. We asked participants which operating system they use on their primary smartphone. Slightly more than half, 1,008 (59%), of the participants were Android users, while 676 (40%) were iOS users. We collected browser user-agent strings during the survey and confirmed similar breakdowns, suggesting most participants used their primary smartphone to take the survey. A detailed breakdown can be found in the Appendix A.3 in Table 11.

Unlock Schemes Usage. As we focus on mobile authentication, we were interested in learning about the kind of mobile authentication our participants use, recalling both biometric and knowledge-based authentication may be in use on a single device. We first asked if a biometric was used and then asked what authentication participants use instead or as a backup for the biometric, e.g., when it fails. While Table 4 shows a compressed description, a detailed breakdown can be found in the Appendix A.3 in Table 11. For knowledge-based authenticators, considered here, PINs are the most common: Forty-three percent described using a four-digit PIN, 22% using a six-digit PIN, and 3% using a PIN of other length. The second most common form of knowledge-based authentication are Android unlock patterns at 14%, and 57 participants (or 3%) reported using an alphanumeric password. In our study, 189 participants (11%) reported not using any locking method.

4.4 Ethical Considerations

All of the survey material and protocol was approved by our IRB. Beyond meeting the approval of our institution, we worked to uphold the ethical principles outlined in the Menlo Report [56].

In practicing *respect for persons* and *justice*, beyond informing and getting consent, we also sought to compensate participants fairly at least at the minimum wage of the municipality where the oversight was performed. Since some of our treatments may frustrate participants, e.g., where the blocklist was comparatively large (DD-4-digit-2740 & DD-6-digit-291000), we also compensated those who returned the survey and notified us of their frustration.

Additionally, as we are dealing with authentication information, we evaluated the ethics of collecting PINs and distributing blocklists in terms of *beneficence*. With respect to collecting PINs, there is risk in that participants may (and likely will) expose PINs used in actual authentication. However, there is limited to no risk in that exposure due to the fact that PINs are not linked to participants and thus cannot be used in a targeted attack. A targeted attack would need proximity and awareness of the victim, of which, neither is the case for this study. Meanwhile, the benefit of the research is high in that the goal of this research is to improve the security of mobile

Table 4. Usage of Mobile Unlock Authentication Schemes

Primary Scheme	No.	%	Secondary Scheme	No.	%
Fingerprint	779	46%	Four-digit PIN	387	50%
			Six-digit PIN	215	28%
			Pattern	109	14%
			Other	68	8%
Face	263	15%	Four-digit PIN	113	42%
			Six-digit PIN	104	40%
			Pattern	23	9%
			Other	23	9%
Other Biometric	33	2%	Four-digit PIN	10	30%
			Six-digit PIN	3	9%
			Pattern	16	49%
			Other	4	12%
Four-digit PIN	218	13%	*No secondary scheme used.*		
Six-digit PIN	59	4%			
Pattern	88	5%			
Other	76	4%			
None	189	11%			

authentication. Similarly, distributing blocklists increases social good and scientific understanding with minimal risk as a determined attacker likely already has access to this material.

Finally, we have described our procedures transparently and make our methods available when considering *respect for law and public interest*. We also do not access any information that is not already publicly available.

4.5 Limitations

There are a number of limitations in this study. Foremost among them is the fact that the participant sample is skewed toward mostly younger users residing in the US. However, as we described previously, there may be some benefit to studying PINs from US residents as they are less familiar with *chip-and-PIN* systems and may be more likely to associate PINs directly with mobile unlocking. We argue that our sample provides realizable and generalizable results regarding the larger ecosystem of PIN selection for mobile authentication. Further research would be needed to understand how certain populations, for example, more age-diverse ones select PINs [42]. For populations from different locations, there is some knowledge about the differences between English-speaking and Chinese users [59], but other populations have also not been studied yet.

Another limitation of the survey is that we are asking participants to select PINs while primed for mobile authentication and there is a risk that participants do not act the same way in the wild. We note that similar priming is used in the authentication literature for both text-based passwords for desktop [53, 54] and mobile settings [37], and these results generalize when compared to passwords from leaked password datasets [55]. We have similar results here. When compared to the most realistic dataset previously available, Amitay-4-digit, the most common four-digit PINs collected in our study are also present in similar distributions to Amitay [3]. Also, in analyzing the qualitative data, a number of participants noted that they used their real unlock PINs.

While this presents strong evidence of the effectiveness of mobile unlock priming, we, unfortunately, do not have any true comparison points, like what is available for text-based passwords. There is no obvious analog to the kinds of attacks that have exposed millions of text-based passwords that would similarly leak millions of mobile unlock PINs. Given the available evidence, we

argue that collecting PINs primed for mobile unlock authentication provides a reasonable approximation for how users choose PINs in the wild.

Due to the short, online nature of our study, we are limited in what we can conclude about the memorability of the PINs. The entirety of the study is only around 5 minutes, while mobile authentication PINs are used for indefinite periods, and likely carried from one device to the next. There are clear differences in these cases, and while we report on the recall rates within the context of the study, these results do not generalize.

Finally, we limited the warning messaging used when a blocklist event occurred. We made this choice based on evaluating the messaging as used by iOS, but there is a long line of research in appropriate security messaging [2, 19, 23, 50]. We do not wish to make claims about the quality of this messaging, and a limitation of this study (and an area of future work) is to understand how messaging affects changing strategies and click-through rates.

5 PIN SELECTION ON SMARTPHONES

In the following section, we discuss the security of both four- and six-digit PINs. Unless otherwise stated, our analyzed dataset consists of the PINs entered before any blocklist warning in Step (5) of the study. These "first choice" PINs (cf. Table 7) are unaffected by the blocklists.

5.1 Strength of Four- and Six-digit PINs

Entropy-based Strength Metrics. We analyzed PINs in terms of their mathematical metrics for guessing resistance based on entropy estimations. For this, we consider a *perfect knowledge* attacker who always guesses correctly (in perfect order) as described by Bonneau et al. [12]. The advantage of such an entropy estimation approach is that it always models a best-case attacker and does not introduce bias from a specific guessing approach. Our results are given in Table 5.

We report the β-success-rate, which measures the expected guessing success for a throttled adversary limited to β-guesses per account (e.g., $\lambda_3 = 3$ guesses). Moreover, we provide the Min-entropy H_∞ as a lower bound estimate that solely relies on the frequency of the most common PIN (1234, 123456). Finally, we present the partial guessing entropy (α-guesswork) G_α, which provides an estimate for an unthrottled attacker trying to guess a fraction α of all PINs. In three cases, the calculation of $\widetilde{G}_{0.2}$ is based on PINs occurring only once, due to the small size of the datasets. This constraint would result in inaccurate guessing-entropy values, which is why they are not reported.

For a fair comparison among the datasets that all differ in size, we downsampled all datasets to the size of the smallest dataset First-4 (851 PINs). We repeated this process 500 times, removed outliers using Tukey fences with $k = 1.5$. In Table 5 we report the median values. The low Min-entropy of the Rock-6 dataset is due to the fact that the PIN 123456 is over-represented. It is 21× more frequent than the second-most popular PIN. In contrast, the most common four-digit PIN occurs only 1.7× more often, leading to a higher H_∞ value. Overall, the PINs we collected, specifically primed for mobile authentication, have different (and *stronger*) strength estimations than PINs derived from leaked text-based password datasets. This is true for both the four- and six-digit PINs, which supports our motivation for conducting studies that collect PINs directly.

Guess Number-driven Strength Estimates. Next, we estimate the security of the PINs in regard to real-world guessing attacks. Our attacker guesses PINs in decreasing probability order based on the Amit-4, Rock-4, and Rock-6 datasets. When two or more PINs share the same frequency, i.e., it is not possible to directly determine a guessing order, we order those PINs using a Markov model [21]. We trained our model on the bi-grams (four-digit PINs) or tri-grams (six-digit PINs) of the respective attacking datasets that simulates the attacker with the highest success rate for each case without overfitting the problem.

Table 5. Guessing Difficulty for a Perfect Knowledge Attacker

Dataset	Size	Online Guessing (Success %) λ_3	λ_{10}	λ_{30}	Offline Guessing (bits) H_∞	$\widetilde{G}_{0.05}$	$\widetilde{G}_{0.1}$	$\widetilde{G}_{0.2}$
First-4	851	3.41%	6.23%	11.75%	5.65	7.07	7.81	—*
Amit-4†	204,432	9.28%	15.28%	22.91%	4.52	4.82	5.20	6.68
Rock-4†	1,780,587	8.23%	17.63%	30.67%	4.73	5.00	5.42	5.94
First-6†	854	5.05%	7.99%	13.04%	4.73	5.88	7.43	—*
Rock-6†	2,758,490	13.04%	15.51%	19.27%	3.10	3.10	3.10	7.41

†: For a fair comparison we downsampled the datasets to the size of First-4 (851 PINs).
★: We omit entries that are not sufficiently supported by the underlying data.

An overview of our guessing analysis can be found in Figure 6. In the throttled scenario, depicted in Figure 6(a), we find that guessing four-digit PINs with the Amitay-4-digit dataset (△) is the most effective attack. In contrast to the RockYou-4-digit dataset (▽) for which we extracted PINs from a password leak, the Amitay dataset consists of actual PINs (cf. Section 3.2). The fact that guessing PINs that the RockYou-4-digit dataset is less effective informs to use actual PIN data whenever possible and simulate our attacker by utilizing the Amitay dataset to estimate of four-digit PINs.

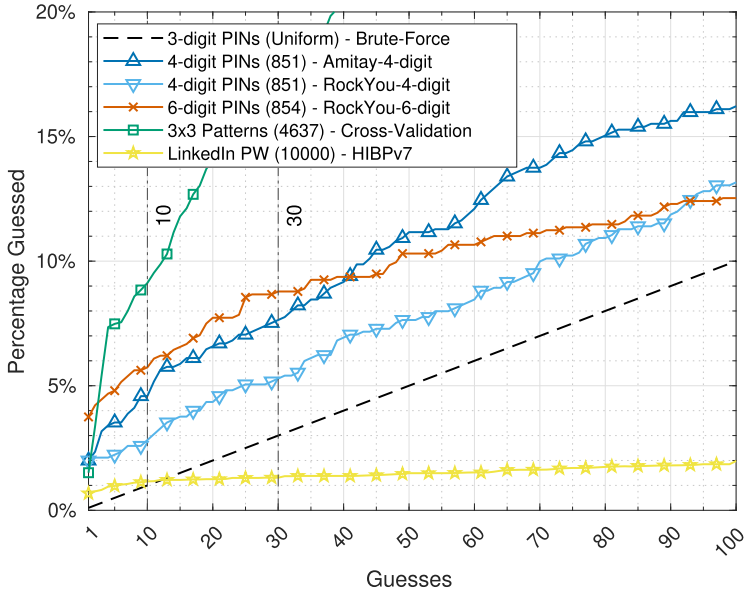
When comparing four- (△) and six-digit PINs (×), we see that guessing performance varies. For 10 guesses (the maximum allowed under iOS), we find 4.6% of the four-digit and 5.7% of the six-digit PINs are guessed. For 30 guesses (a less determined attacker on Android), 7.6% of the four-digit and 8.8% of the six-digit PINs are guessed, and for 100 guesses (a reasonable upper bound on Android), 16.2% of the four-digit and 12.5% of the six-digit PINs.

Somewhat counterintuitive is the weaker security for six-digit PINs for the first 40 guesses. Upon investigation, the most-common six-digit PINs are more narrowly distributed than their most-common four-digit counterparts. The most common six-digit PINs consist of simple PINs, such as 123456 as defined in Table 12 in Appendix A.4, and repeating digits. In contrast, the most common four-digit PINs consist of simple PINs, patterns, dates, and repeating digits. As a result, the most common six-digit PINs may actually be easier to guess and less diverse than the most common four-digit PINs.

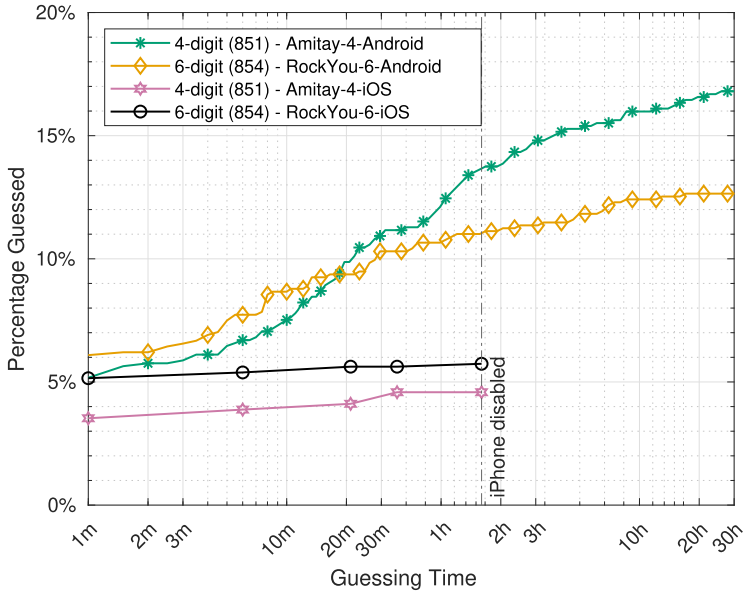
There could be many explanations for this counterintuitive finding. One explanation may be that users have more four-digit PIN sequences to draw on in choosing a PIN, such as dates, but have fewer natural six-digit analogs, and thus revert to less diverse, more easily guessed choices. We will present some evidence for this hypothesis in Section 5.2 where we analyze the selection strategies of six-digit PINs. Another explanation may be that users have a false sense of security that comes with six-digit PINs as they are "two digits more secure" than four-digit PINs. Thus, users do not feel that they need more complexity in their six-digit PIN choices. Either way, future research is needed to better understand this phenomenon, which has also been observed by Aviv et al. [7] in the context of increasing the size (3×3 vs. 4×4) of Android graphical unlock patterns.

Finally, we compare guessing resistance with other mobile authentication schemes including Android's graphical unlock patterns drawn on a 3×3 grid (□) and alphanumeric passwords (★), along with a uniform distribution of 3-digit PINs (–). In theory, a 3×3 grid allows 389,112 unique patterns, yet, the distribution of patterns is highly skewed [52]. When considering an attack throttled to 100 guesses, 35.5% of the patterns will be guessed. Against this attack, four- and six-digit PINs are twice as good. Password-based authentication, however, is the most secure scheme. After 100 guesses only 1.9% of the passwords are recovered.

Figure 6(b) shows the guessing time of an attacker due to rate limiting based on Table 1 for iOS and Android. iOS has stricter rate limiting with a maximum of 10 guesses that can be completed



(a) Guessing performance against mobile authentication systems based on the number of guesses.



(b) Guessing performance against four- and six digit PINs on Android and iOS based on the required time. For 4-digit PINs, we only show the success rate of an attack with Amit-4 as it outperforms Rock-4 (cf. Figure 6(a)).

Fig. 6. Guessing performance of a *throttled* attacker. The figure on the top is based on the number of guesses. The bottom figure is based on the required time and considers the rate limits of Android and iOS (cf. Table 1).

in 1 hour and 36 minutes, at which point an attacker compromises 4.6% of the four-digit PINs and 5.7% of the six-digit PINs. At the same time limit of roughly 1.5 h, an attacker on Android is able to compromise 13.6% of the four-digit PINs and 11.0% of the six-digit PINs because of less restrictive rate limiting.

Especially on iOS, rate limiting becomes more aggressive after the initial guesses. For example, the first 6 guesses on iOS can be done within a minute, while the first 8 guesses already take 21 minutes. An attacker with only one minute on iOS can compromise 3.5% of the four-digit PINs and 5.2% of the six-digit PINs. However, for 10 guesses that take 1h 36m on iOS, there are only marginal gains with 4.6% of the four-digit PINs and 5.7% of six-digit PINs compromised. Hence, after the first minute with 6 guesses on iOS, it does not greatly benefit the attacker to continue through the aggressive timeouts for 4 more guesses at 1h 36m. In contrast, an attacker on Android would benefit more from continuing to guess beyond the initial large increases in rate limiting. Note, in a targeted attack, there may be additional information or other motivations for the attacker not modeled here.

To summarize, we confirmed previous work from Wang et al. [59] that there is no evidence that six-digit PINs offer any security advantage over four-digit PINs considering a throttled guessing attacker with up to 40 guesses, which covers most mobile unlock authentication settings. Only when considering threat models where the attacker is allowed to guess more often, six-digit PINs start to exceed four-digit PINs in terms of their guessing resistance. To support this claim, we performed χ^2 tests ($\alpha = 0.05$) for both the four- and six-digit PINs guessed within 10 [4.6%, 5.7%], 30 [7.6%, 8.8%], and 100 guesses [16.2%, 12.5%]. The test for 10 ($p = 0.28$) and 30 guesses ($p = 0.39$) did not show a significant difference in PIN strength. For 100 guesses, however, we were able to observe that the six-digit PINs are significantly stronger than the four-digit ones ($p = 0.03$). This again highlights the importance of clearly defining threat model in terms of how many guesses the attacker is able to make when deciding on a certain PIN length.

Effect of Biometrics. Users who employ a biometric, cf. Table 4, do not need to provide their knowledge-based authenticator as often as users who solely rely on a PIN, pattern, or password. This may shift users toward more complex choices that are more cumbersome to type, but, owing to the biometric, only need to be provided on rare occasion like a device restart. Hence, the question arises: do users who authenticate with a biometric select more secure PINs?

To test this hypothesis, we split each the First-4 and First-6 dataset into two datasets, based on whether participants stated to use a biometric or not. As we primed our participants to select a PIN they would use to unlock their smartphone (cf. Figure 2), we have all the information required for this type of analysis. The security metrics for the "Biometric-used" and "No-biometric-used" datasets are shown in Table 7. The results do not support the hypothesis, but instead, participants who do not use a biometric tend to create more secure PINs. However, while the success rates of the attacker differ by up to 3% for 30 guesses when comparing Biometric-used-4 and No-biometric-used-4, we were not able to observe any significant differences using a χ^2 test ($\alpha = 0.05$).

5.2 Selection Strategies

In Step (6) of our study, we asked participants about their "strategy for choosing" their PIN. We analyzed the free-text responses to this question by building a codebook from a random sample of 314 PIN selection strategies using two coders. Inter-rater reliability between the coders measured by Cohen's kappa was $\kappa = 0.90$. Table 12 in Appendix A.4 shows the 10 most popular strategies.

While the selection strategies are diverse, most participants chose PINs that they perceive as memorable in general or based them on dates, especially birthdays and anniversaries. Other popular strategies are PIN pad patterns, choosing randomly, or selecting other kinds of meaningful numbers to the participants, like a zip code or a favorite number. While most of those strategies

Table 6. Overlap of the First-4 PINs with the Three Substring Lists Extracted from the First-6 PINs

Substring	Overlap No.	Overlap %	Top 5 PINs PIN	No.	%	Most Common Addition
Leftmost	196	23%	1234	17	9%	123456 (91%)
			2580	7	4%	*not distinct*
			6969	5	3%	696969 (80%)
			1212	4	2%	121212 (50%)
			1379	3	2%	*not distinct*
Middle	74	9%	0000	2	3%	000000 (100%)
			1111	2	3%	111111 (100%)
			2121	2	3%	121212 (100%)
			7777	2	3%	777777 (100%)
			9898	2	3%	898989 (100%)
Rightmost	116	14%	6969	5	4%	696969 (80%)
			4321	4	4%	654321 (100%)
			1212	3	3%	121212 (67%)
			4578	2	2%	124578 (100%)
			7777	2	2%	777777 (67%)

are common across both PIN lengths, 26 of the 33 participants (79%) who stated to choose digits randomly were asked to create a six-digit PIN. This again supports the intuition that users have less experience with six-digit PINs and start to run out of meaningful strategies earlier.

To further understand how users create six-digit PINs and to see if users take their four-digit selection strategy and just extend it to create a longer version, we now look at the four-digit substrings of the six-digit PINs. For this, we took the 855 First-Choice-6-digit PINs and created three lists extracting the 4 leftmost, the 4 middle, and the 4 rightmost digits of each PIN. For comparison, we overlapped those lists with the First-Choice-4-digit PINs. As can be seen in Table 6, the greatest overlap with 23%, occurs for the leftmost substring PINs, followed by the rightmost (14%). The substring PINs consisting of the 4 digits in the middle only overlap with the First-4 PINs by 9%. Moreover, all of the PINs we extracted for this list follow simple repetitions, strategies that are not specific for a certain PIN length. A similar conclusion can be drawn from the rightmost PINs, there is no indication that participants started with a four-digit PIN and added two digits on the left. Again, we see that the creation strategies can be used to create PINs of arbitrary length, mostly repetitions (e.g., 1212/121212), and sequences (e.g., 4321/654321). However, Table 6 also depicts two exceptions: 2580 and 1379. The former is a top-down walk, which allows for a simple four-digit PIN, yet, each of the seven participants who started a six-digit PIN this way ended up differently. A similar observation can be made for 1379, where each of the four corners is selected without an apparent addition for a six-digit PIN. Both cases suggest that there are participants who did not have an actual six-digit strategy but used one they had in mind for four-digits and added two digits. This also fits the overall impression that users are more familiar with four-digit PINs.

6 BLOCKLISTS AND PIN SELECTION

We now present results on our 10 blocklist treatments: five for each PIN length as shown in Table 7.

6.1 PIN Creation and Entry Times

The blocklist has an impact on the PIN creation time: An increase in the number of blocklist messages leads to increased creation time. The median creation time when receiving a blocklist message can be found in Table 7; a more detailed breakdown for each treatment can be seen in Figure 7.

Table 7. Security Metrics and Usage Times for PINs Considering Different Datasets and Treatments

	Name	Participants	Blocklist Hits	10 Guesses No.	10 Guesses %	30 Guesses No.	30 Guesses %	100 Guesses No.	100 Guesses %	Guess No. Median	Creation Time	Entry Time	Number of Attempts
Datasets	First-Choice-4-digit	851	—	39	5%	65	8%	138	16%	1 330	—	—	—
	Clicked-through-4	19	19	5	26%	6	32%	13	68%	50	—	—	—
	Biometric-used-4	533	—	28	5%	47	9%	91	17%	1347	—	—	—
	No-biometric-used-4	318	—	11	4%	18	6%	47	15%	1257	—	—	—
Treatments	Control-4-digit	231	—	11	5%	19	8%	39	17%	1 185	7.9 s	1.5 s	1.01
	Placebo-4-digit	122	122	5	4%	11	9%	19	16%	2 423	21.8 s	1.5 s	2.15
	iOS-4-digit-wCt	124	28	5	4%	8	6%	18	15%	1 405	10.4 s	1.4 s	1.17
	iOS-4-digit-nCt	126	21	4	3%	10	8%	14	11%	1 747	9.3 s	1.6 s	1.29
	DD-4-digit-27	121	5	4	3%	7	6%	18	15%	1 928	8.8 s	1.5 s	1.11
	DD-4-digit-2740	127	88	0	0%	0	0%	1	1%	2 871	25.4 s	1.6 s	2.98
Datasets	First-Choice-6-digit	854	—	49	5%	75	9%	107	13%	49 021	—	—	—
	Clicked-through-6	10	10	9	90%	9	90%	9	90%	1	—	—	—
	Biometric-used-6	542	—	33	6%	51	9%	68	13%	47 773	—	—	—
	No-biometric-used-6	312	—	16	5%	24	8%	39	13%	50 922	—	—	—
Treatments	Control-6-digit	236	—	15	6%	26	11%	35	15%	42 584	11.0 s	2.5 s	1.01
	Placebo-6-digit	117	117	3	3%	6	5%	10	9%	154 521	28.5 s	3.0 s	2.17
	iOS-6-digit-wCt	125	15	9	7%	9	7%	13	10%	40 972	11.9 s	2.6 s	1.06
	iOS-6-digit-nCt	125	16	2	2%	4	3%	6	5%	61 036	12.2 s	2.8 s	1.22
	DD-6-digit-29	126	12	1	1%	2	2%	7	6%	82 373	11.1 s	2.5 s	1.23
	DD-6-digit-291000	125	90	0	0%	0	0%	0	0%	324 621	45.2 s	3.5 s	3.94

In the four-digit case, there are obvious differences between the control treatments and the placebo and the large data-driven treatment DD-4-2740. The median creation time increases from 7.9 s for the Con-4 treatment to 21.8 s for Pla-4 and 25.4 s for DD-4-2740. Both differences are significant ($p < 0.001$) using a Kruskal-Wallis test followed by a Bonferroni-corrected pairwise Wilcoxon test. The differences for the remaining four-digit treatments iOS-4-wC, iOS-4-nC, and DD-4-27 are more subtle. The median creation time for the small data-driven treatment DD-4-27 only increases by 0.9 to 8.8 s, followed by the iOS-4-nC treatment (9.3 s), and iOS-4-wC (10.4 s). Moreover, we were able to observe significant differences for the latter comparison, i.e., Con-4 vs. iOS-4-wC ($p < 0.01$), whereas we were not for the comparisons of iOS-4-nC and DD-4-27 with the control treatment. We did not observe any significant differences between the four-digit treatments for entry time.

The situation is similar for six-digit PINs. As can be seen in both in Table 7 and Figure 7, the creation times for the Pla-6 and DD-6-291000 treatment increase compared to the control treatment, but both iOS treatments (iOS-6-wC and iOS-6-nC) and the small data-driven treatment DD-6-29 show minimal differences compared to control. We observed a significant differences for both Pla-6 and DD-6-291000 ($p < 0.001$) using a Kruskal–Wallis test followed by Bonferroni-corrected pairwise Wilcoxon tests. We did not find significant differences between iOS-6-wC, iOS-6-nC, nor DD-6-29.

The entry times are again not affected with one exception: the six-digit case. Participants required more time to enter the PIN they created in the large data-driven treatment DD-6-291000. The median here is 3.5 s compared to 2.5 s in the respective control treatment and this difference is also significant ($p < 0.001$) using the same statistical tests.

This suggests that blocklists, when properly sized, can lead to significant increases in the creation times that may in turn frustrated users, as we will explore in Section 6.6 and 6.7. However,

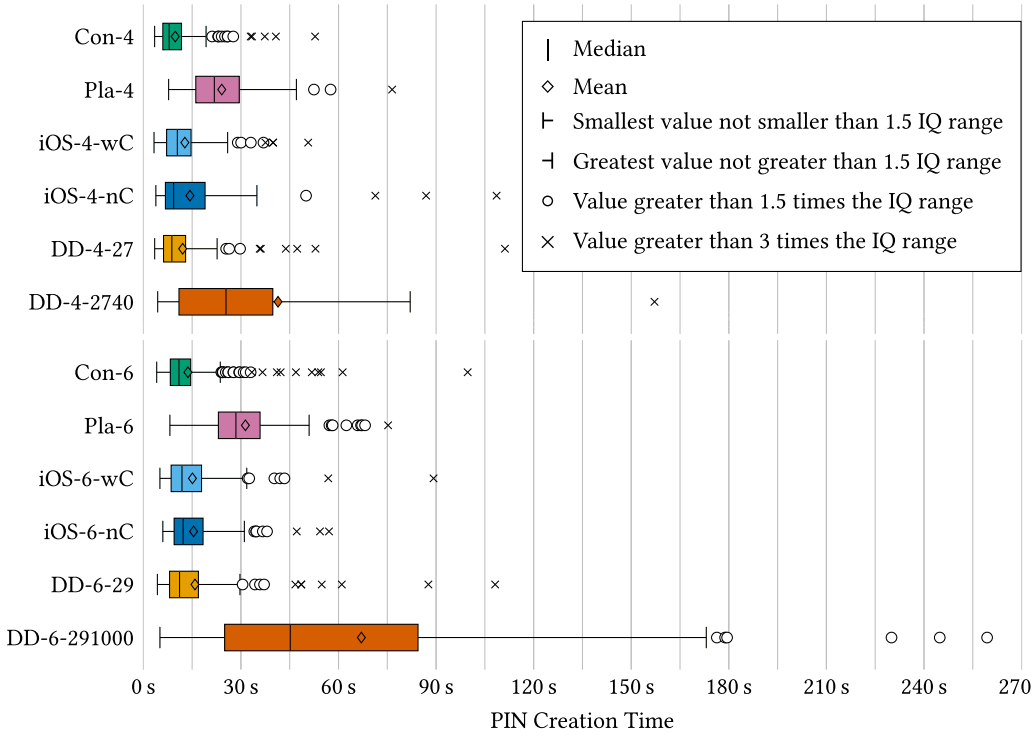


Fig. 7. PIN creation times for the different treatments. For the sake of clarity, we excluded two extrema from the plot: 1542.32 s from the DD-4-2740 treatment and 1105.13 s from DD-6-291000.

the subsequent usage of the PIN, as evidences by the entry time, is unaffected. Only in the case of a very large blocklist with six-digit PINs do we observe any meaningful increase in entry time.

6.2 Attacker's Knowledge of Blocklists

As described in Section 3.1, we assume the attacker knows which blocklisting strategy is used by the system and can optimize the guessing strategy by *not* guessing items on the blocklist. Here, we consider how much benefit this optimization provides. Table 8 shows the net gains and losses for guessing PINs when considering a blocklist-informed attacker.

Knowledge of the blocklist is unhelpful when considering the placebo (Pla-4 and Pla-6) and the click-through treatments (iOS-4-wC and iOS-6-wC). The blocklist is effectively of size one for the placebo as the first choice of a participant is dynamically blocked. Merely knowing that a PIN was blocked is of little help to the attacker. As there is no clear gain (or harm), we model a blocklist-knowledgeable attacker for the placebo treatments (see Table 7).

The case with a non-enforcing blocklist where users can click through the warning message is more subtle. If the attacker is explicitly choosing not to consider PINs on the blocklist, even though they may *actually* be selected due to non-enforcement, then the guessing strategy is harmed (negative in Table 8). None of the tested modifications of this strategy, e.g., by incorporating the observed click-through rate, lead to an improvement. As such, we consider an attacker that *does not* use the blocklist to change the guessing strategy for the click-through treatments (iOS-4-wC and iOS-6-wC). In the remaining treatments (iOS-4-nC, DD-4-27, DD-4-2740, iOS-6-nC, DD-6-29, and DD-6-291000), there are clear advantages when knowing the blocklist.

Table 8. Attacker's Gain from Blocklist Knowledge

Treatment	10 Guesses No.	10 Guesses %	30 Guesses No.	30 Guesses %	100 Guesses No.	100 Guesses %	Guess No. Median	Knowledge Beneficial
Pla-4	±0	±0%	±0	±0%	±0	±0%	±0	—
iOS-4-wC	−3	−2%	−4	−2%	−9	−8%	−303	✗
iOS-4-nC	**+3**	**+2%**	**+7**	**+6%**	**+3**	**+2%**	**+245**	✓
DD-4-27	**+4**	**+3%**	**+7**	**+6%**	**+5**	**+4%**	**+27**	✓
DD-4-2740	**±0**	**±0%**	**±0**	**±0%**	**+1**	**+1%**	**+2740**	✓
Pla-6	±0	±0%	±0	±0%	±0	±0%	±0	—
iOS-6-wC	−9	−7%	−5	−4%	−8	−6%	−7322	✗
iOS-6-nC	**+2**	**+2%**	**+2**	**+2%**	**+2**	**+2%**	**+1524**	✓
DD-6-29	**+1**	**+1%**	**+2**	**+2%**	**+2**	**+2%**	**+29**	✓
DD-6-291000	**±0**	**±0%**	**±0**	**±0%**	**±0**	**±0%**	**+291000**	✓

6.3 Blocklisting Impact on Security

We now consider how the different blocklists perform in terms of improving security. The primary results are in Table 7 where we report on the guessing performance against each treatment. As described in Section 3.1, there are certain rate limits implemented on Android and iOS, which is why we report on throttled attacks with 10, 30, and 100 guesses in terms of the number and percentage of correctly guessed PINs (No. and % columns). Furthermore, we provide the attacker's performance in an unthrottled setting based on the median guess number. The four-digit attacker is informed by the Amit-4 dataset, while the six-digit attacker employs the Rock-6 dataset. Both attackers guess in frequency order with knowledge of the blocklist where appropriate (see Section 6.2). Finally, Figure 8 shows the selection bias for the first digit of the PIN. Multiple treatments are visualized to show the effects of different blocklists (sampled down to 117 PINs for a fair comparison). Unfortunately, a heatmap does not provide a direct security metric and does not necessarily correlate with guessing difficulty, but this visualization suggest that PIN selection bias shifts when encountering a blocklist.

To analyze the security, we performed a multivariant χ^2 test comparison ($\alpha = 0.05$) for the PINs guessed within 10, 30, and 100 guesses across treatments. The test for 10 suggested some significant differences in the data ($p = 0.007$); however, we did not find any actual significant differences in the post hoc analysis (Bonferroni-corrected). For 30 guesses and 100 guesses the test also showed significant differences ($p < 0.001$); the results of the post hoc analyses are described below.

Smaller Blocklists. When looking at the four-digit treatments, there is little difference among Placebo-4-digit, iOS-4-digit-wCt, iOS-4-digit-nCt, and DD-4-digit-27, compared to Control-4-digit or First-Choice-4-digit. In our post hoc analyses (Bonferroni-corrected), we found no significant difference. The same holds when comparing the selection bias in the Control-4-digit treatment (Figure 8(a)) and the iOS-4-digit-wCt blocklist treatments (274 PINs, Figure 8(b)): there is a strong preference for selecting 1, but also 2 and 0 are common, as the first digit of the PIN. All other digits are selected similarly often. It remains unclear, whether this preference is owed to the digits or their position on the PIN pad (top, left). A common selection bias is also observed in graphical passwords [20, 22, 35, 52].

For our six-digit treatments, the situation is similar, yet, there is one exception: For 30 guesses, we observed a significant difference between the small data-driven blocklist and the control ($p < 0.01$). While this implies that it can make sense to employ a small blocklist in certain cases, we will show in Section 6.7 that blocklist warnings are associated with negative sentiments. Additionally, the selection bias in the control treatment and the small blocklists are comparable, as can be seen in Figure 8(e) and (f). Hence, it is hard to justify the combination of throttling and blocklists in general.

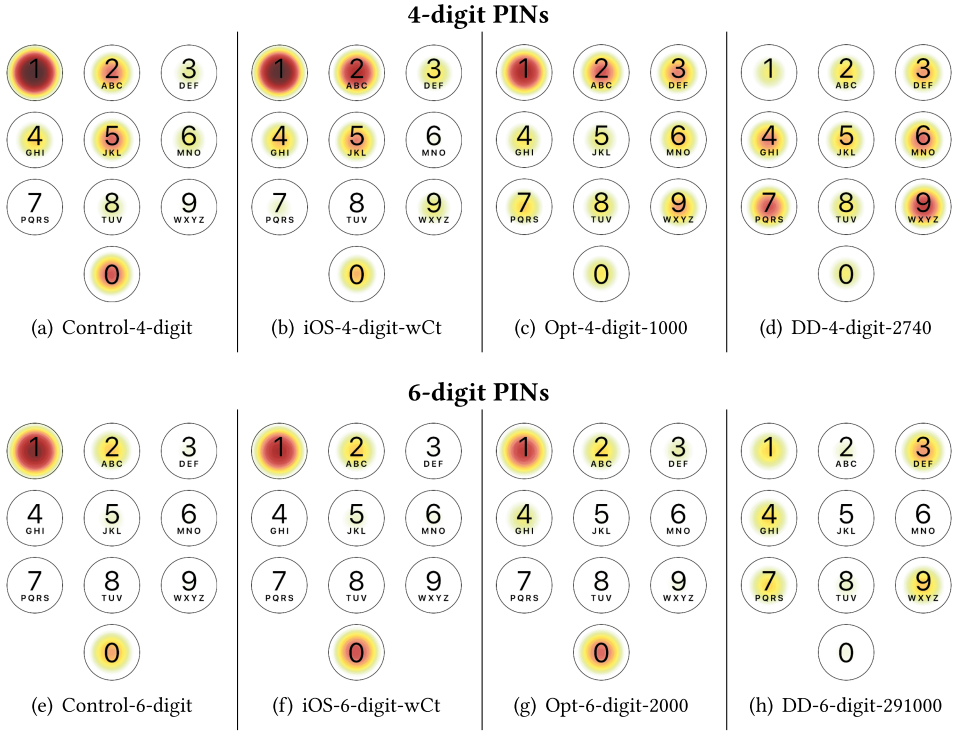


Fig. 8. Heatmaps showing the selection bias for the first digit when composing a PIN.

In the unthrottled setting, we see differences between the smaller and placebo blocklists. Notably, the smallest (DD-4-digit-27 and DD-6-digit-29) outperforms the larger iOS blocklists (iOS-4-digit-nCt and iOS-6-digit-nCt). We conjecture this may be due to iOS's inclusion of PINs based on repetitions that were chosen less often by our participants. As a result, in an unthrottled setting, blocklisting can offer real benefits. The median guess numbers for both four- and six-digit placebos suggest that just pushing users away from their first choice can improve security. Unfortunately, direct use of a placebo blocklist is unlikely to be effective and is problematic in practice as users will quickly figure out the deception.

Finally, these improvements to the unthrottled attack setting appear to be only of academic interest: Given the small key space, any attacker that is able to bypass the enforced rate limiting is able to exhaustively test all possible combinations [44]. For example, a tool from Elcomsoft is able to bypass the rate limiting on Apple's iPhone 5 and 5c. In this case, guessing all four-digit PINs takes about 12 minutes while enumerating all six-digit PINs takes 20.5 hours [1].

Large Blocklist. We also consider very large blocklists in the DD-4-digit-2740 and DD-6-digit-291000 treatment containing 2,740 PINs and 291,000 PINs respectively. These blocklists are bigger than their iOS counterparts, blocking 27.4% in the four- and 29.1% of the key space in the six-digit case. At this scale, we do see noticeable effects on the security in the throttled setting. In the four-digit case, the attacker finds only 1% of four-digit PINs after 100 guesses. Our χ^2 tests support this, for 100 guesses we found a significant difference ($p < 0.001$). For post hoc analyses (Bonferroni-corrected) we found significant differences between the large DD-4-2740 blocklist and Con-6 ($p < 0.01$) as well as the treatments: Con-4 ($p < 0.001$), Pla-4 ($p < 0.01$), iOS-4-wC

($p < 0.05$), and DD-4-27 ($p < 0.05$). As expected, the selection bias of the first digit also becomes less pronounced (cf. Figure 8(d)).

In the six-digit case, we make similar observations for the guessing routine although we already start to see significant differences for 30 guesses when comparing the DD-6-291000 and the control treatment ($p < 0.01$). For 100 guesses the guessing success of the attacker in the DD-6-291000 treatment is significantly lower than for all four-digit treatments: Con-4 ($p < 0.001$), Pla-4 ($p < 0.01$), iOS-4-wC ($p < 0.05$), DD-4-27 ($p < 0.01$), as well as the six-digit control treatment ($p < 0.001$). Again, participants also choose their first digit more equally distributed as can be seen in Figure 8(h). All of this suggests that a larger blocklist can improve security in a throttled setting.

While similar positive security results are present for the unthrottled setting, we show in Section 6.6 that the larger blocklist also leads to a perceived lower usability, and thus it is important to balance the user experience with security gains.

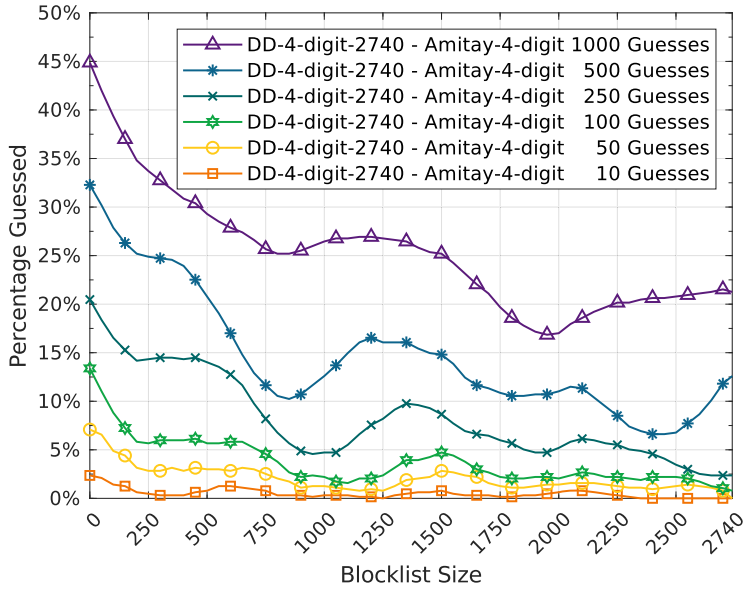
Correctly Sizing a Blocklist. While there is a clear benefit to having a large blocklist, it is important to consider the right size of a blocklist to counteract negative usability and user experience issues. Data from the large data-driven treatments enable us to simulate how users would have responded to shorter blocklists. In our user study, we collected not only the final PIN accepted by the system, but also all $n - 1$ intermediate (first-choice, second-choice, and so on) PINs rejected due to the blocklist. Consider a smaller blocklist that would have permitted choice $n - 1$ to be the final PIN, rather than n. To simulate that smaller blocklist size, we use choice $n - 1$.

The results of the simulation are shown in Figure 9. We observe that there are several troughs and peaks in the curves in both figures. We speculate that these relate to changes in user choices as they move from their first choice PIN to their second choice PIN, and so on due to the expanding blocklist restrictions. For example, entering the first trough, the attacker is most disadvantaged when it is no longer possible to rely on guessing only first choice PINs and second choice PINs need to be considered. Eventually, the blocklist has restricted all first choice PINs, whereby the attacker can now take advantage of guessing popular second choices that results in a peak. These cycles continue until the blocklist gets so large that few acceptable PINs remain, and the attacker's advantage grows steadily by guessing the remaining PINs not on the blocklist.

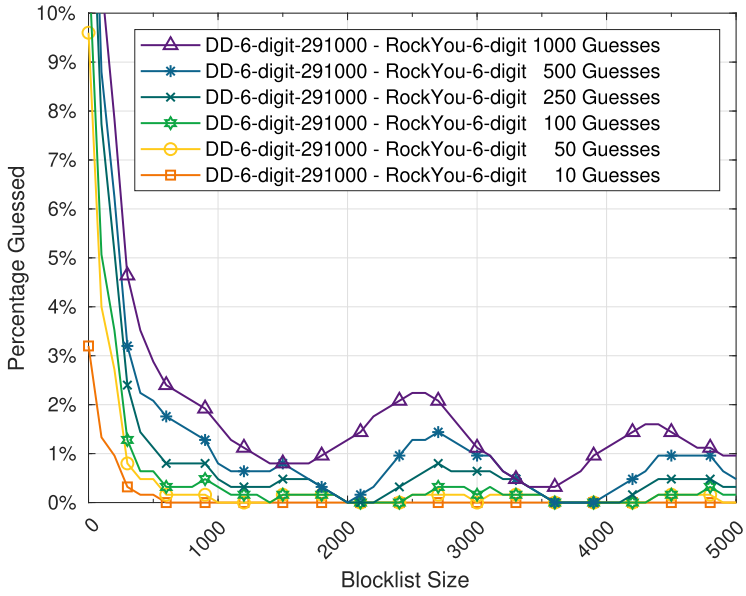
Based on these cycles, we conclude that an appropriately-sized blocklist should be based on one of the troughs where an attacker is most disadvantaged to maximize the security gained in the throttled setting. As we are also concerned about minimizing user discomfort and frustration (e.g, PIN creation time, see Section 6.1), the first trough appears the most ideal. As can be seen in Figure 9(a), for four-digit PINs the first trough occurs at about 1000 PINs (10% of the four-digit PIN space) throttled at 100 guesses. A similar suggestion can be drawn from the simulation for six-digit PINs in Figure 9(b), however, due to the overall larger key space, a blocklist with 2000 PINs only corresponds to 0.2% of all possible selections. In Figure 8, we refer to these two recommended blocklists as Opt-four-digit-1000 and Opt-6-digit-2000, respectively. For four-digit PINs, the distribution shown in Figure 8(c) is more equally distributed across all digits. When comparing the selection bias for the optimal six-digit blocklist to the control treatment in Figure 8(g), the bias toward starting a PIN with 1 is less pronounced. Other digits, like 3 and 4, become more popular. We do not observe equally aligned distribution, but the attackers success rate is sufficiently low when blocking only 0.2% of the keyspace. In contrast, the ideal four-digit blocklist rejects 10% of all possible PINs.

6.4 Enforcing the Blocklist

To test the effect of a click-through option, we compared the enforcing treatment for each length (iOS-4-nC/iOS-6-nC) with its non-enforcing counterpart (iOS-4-wC/iOS-6-wC). In neither of the



(a) For throttled attackers, limited to 100 guesses, a blocklist of ~10 % of the key space (~1000 PINs) is ideal.



(b) For throttled attackers, limited to 100 guesses, a blocklist of ~0.2 % of the key space (~2000 PINs) is ideal.

Fig. 9. Blocklist size recommendations for four- (upper figure) and six-digit PINs (lower figure).

Table 9. Changes of Participants' PIN Selection Strategies across Treatments

Treatment	Hits	Selection vs. Changing Strategy				Edit Distance	
		Sample	Same	Minor	New	Mean	SD
Pla-4	122	29	35%	24%	41%	3.20	0.90
iOS-4-wC	9★	9	0%	44%	56%	3.11	0.87
iOS-4-nC	21	21	19%	29%	52%	3.24	0.92
DD-4-27	5	5	40%	40%	20%	3.20	0.75
DD-4-2740	88	29	14%	24%	62%	3.39	0.76
Pla-6	117	28	28%	18%	54%	4.59	1.41
iOS-6-wC	5★	5	0%	40%	60%	4.40	1.20
iOS-6-nC	16	16	6%	50%	44%	4.00	1.54
DD-6-29	12	12	33%	33%	33%	5.25	0.72
DD-6-291000	90	29	14%	21%	65%	4.82	1.13

★: Hit blocklist, and did not click-through.

two comparisons, we observed significant differences. This suggests that using a click-through option does not reduce security in the throttled attacker setting despite the fact that clicked-through PINs are extremely weak (see row Clicked-through-4 in Table 7). These results seem to be driven by the fact that it is uncertain whether the user clicked through (see Table 8). In an enforcing setting, the attacker can leverage the blocklist but is equally challenged in guessing the remaining PINs.

We also investigated why participants chose to ignore and click through the warning. From 28 participants who saw a blocklist warning in the iOS-4-wC treatment, we observed a click-through-rate of 68% (19 participants). In the respective six-digit treatment iOS-6-wC, 10 of 15, i.e., 67%, ignored the warning. This is twice the rate at which TLS warnings are ignored (∼30%) [49]. Furthermore, we asked the 29 participants who pressed "*Use Anyway*" about their motivations. The 3 most observed answers are *Memorability Issues:* "Because this is the number I can remember," *Incomplete Threat Models:* "Many people don't tend to try the obvious PIN as they think it's too obvious so people won't use it," and *Indifference*: "I don't give [sic] about the warning. Security is overrated." These findings are similar to prior work where users do not follow external guidance for a number of reasons [45, 62]. In older versions of iOS, the blocklist warning message was "*Are You Sure You Want to Use This PIN? This PIN is commonly used and can be easily guessed.*" with the safe option "*Choose New PIN*" in bold and the unsafe click-through option saying "*Use PIN*." We observed that Apple changed this wording with iOS 11 to what is depicted in Figure 4. Considering that TLS warning design research started with similarly high click-through-rates of around 70% [2], we hope that new designs can also improve blocklist warning CTRs [49].

6.5 PIN Changing Strategies

In our study, we asked 485 participants who faced a blocklist how their creation strategy changed in response to the warning. We sampled 183 responses (∼10% of our total number of participants) and grouped them into three categories: participants who continued using the "Same" strategy, participants who made "Minor" changes to the strategy, and participants who came up with a completely "New" strategy. Examples for those cases can be found in the Appendix A.4 in Table 13. Two coders independently coded the data. Inter-rater reliability between the coders measured by Cohen's kappa was $\kappa = 0.92$. The detailed results for each treatments are shown in Table 9.

About 50% of the participants choose a new strategy when confronted with a blocklist warning. Only participants of the DD-4-27 and DD-6-29 treatment with a very small blocklist, tended to keep their pre-warning strategy. The edit distances vary slightly across the treatments and

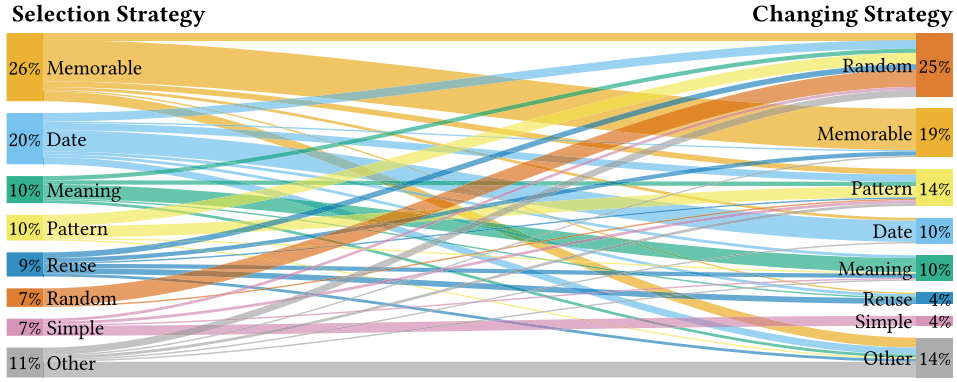


Fig. 10. Participants' PIN selection (first choice) and changing strategies (final choice) for $n = 183$.

support this self-reported behavior: Participants in the four-digit scenario changed on average three digits with the standard deviation showing that some participants changed their PIN completely while some participants only changed two digits. The same conclusion can be drawn from the edit distances in the six-digit case with one difference: participants in the DD-6-29 treatment changed more digits on average. This is particularly interesting, because the blocklist is by far the smallest, which suggests that users may be more willing to change their PIN if the warning does not appear to be arbitrary.

To analyze how participants changed their PIN selection, we mapped the initial selection strategies to the final ones. The result is shown in Figure 10. First, 25% of the participants stated to have changed their PIN into something random (cf. Table 12 in Appendix A.4). While there are 7% of the participants who already had this strategy, we observe a shift from all of the other selection strategies to a random PIN that shows the effectiveness of the blocklist warnings. Moreover, we see that participants usually do not change their PIN to be "memorable," a "date," or "simple." Furthermore, we also see that a certain number of participants stick to their strategy. While we already described that this decision is influenced by the treatment (cf. Table 9), we are now able to see that the selection strategy also influences this decision. For example, nearly all participants who initially selected a random PIN, held on to this approach. This is less distinct across other strategies, yet, participants who stuck to their selection strategy are always the largest group. The only two exceptions are participants who reused a PIN or selected it based on a pattern, they tended to change their strategy after seeing a blocklist warning.

6.6 User Perception

We analyzed participants' perceptions regarding PIN selections with respect to security and usability. Participants were asked to complete the phrase "*I feel the PIN I chose is*" with three different adjectives: "*secure*, *memorable*, and *convenient*." The phrases were displayed randomly and participants responded using a Likert scale. The results are shown in Figure 11. To compare these results, we converted the Likert responses into weighted averages on a scale of −2 to +2. As the weighted averages are not normally distributed, tested using the Shapiro–Wilk test ($p < 0.001$), we tested for initial differences using a Mann–Whitney U test, followed with post hoc, pairwise tests using Dunn's-test comparisons of independent samples with a Bonferroni correction.

We found that there are significant differences across treatments when considering Likert responses for *security*. For the four-digit PINs, post hoc analysis did not indicate any significant differences. One explanation for this overall high confidence in the security of the PIN choice,

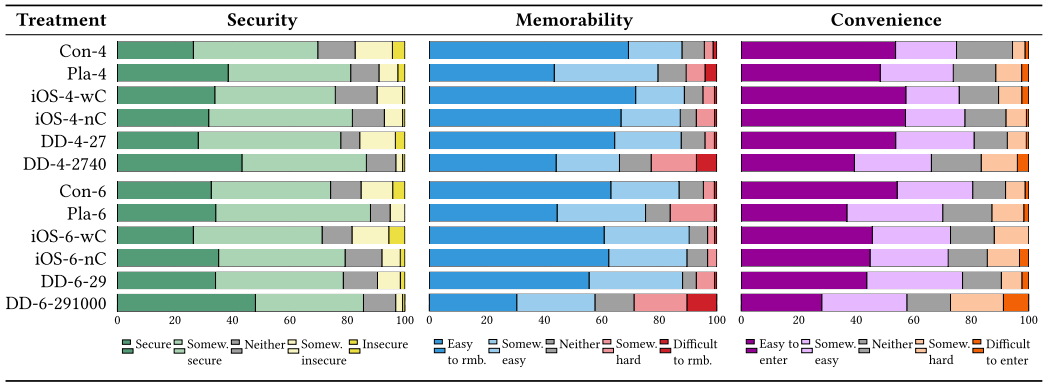


Fig. 11. Participants' perception of their PIN's security (*Secure – Insecure*), memorability (*Easy to remember–Difficult to remember*), and convenience (*Easy to enter – Difficult to enter*).

may be the familiarity with four-digit PINs. In contrast to this, participants in the DD-6-291000 treatment perceive their PINs as more secure compared to participants of the six-digit control ($p < 0.05$), and iOS-6-wC treatment ($p < 0.01$). Here, the large portion (72%) of participants who encountered the blocklist may have lead to a change in the overall perception.

For *memorability* we also found significant differences among the treatments. In post hoc analysis we found that increased interaction with the blocklist led to lower perceived memorability of PINs, as evidenced by the Pla-4 ($p < 0.001$), DD-4-2740 ($p < 0.05$), Pla-6 ($p < 0.001$), and DD-6-291000 ($p < 0.001$) treatments compared to their respective control treatments. The DD-4-2740 and DD-6-291000 showed the most significant differences with other treatments. Again, this is likely due to the fact that many participants encountered a blocklist warning sometimes even for multiple PIN choices and were thus relying on not just second-choice PINs, but also third- and fourth-choice, and so on. PINs that are perceived to be less memorable.

The responses to perceived *convenience* also show significant differences, however, post hoc analysis revealed limited effects when considering pairwise comparisons. In general, participants perceived their four-digit or six-digit PINs at the same convenience level across treatments. However, there is one exception: PINs created in the DD-6-291000 treatment are perceived as significantly more difficult to enter than PINs in the six-digit control treatment ($p < 0.01$), iOS-6-wC ($p < 0.05$), DD-6-29 ($p < 0.05$), and all four-digit treatments ($p < 0.001$). As for the memorability, this suggests that while users may be comfortable with their first-choice six-digit PIN, there is much higher perceived *inconvenience* when having to conform with a large blocklist.

6.7 User Sentiment

To gain insight into participants' sentiments regarding blocklisting, we asked "*Please describe three general feelings or reactions that you had after you received this warning message*" or "*would have had"* if the participant did not encounter a blocklist. Accompanying the prompt are three free-form, short text fields. A codebook was constructed by two individual coders summarized in Appendix A.5 in Table 14. For each of the four categories (blocklist hit experienced vs. imagined, four- vs. six-digit PINs, non-enforcing vs. enforcing, different blocklist sizes), 21 individuals' responses were randomly selected. Again, two individual raters were tasked with coding the responses. The inter-rater reliability, computed using Cohen's kappa, was $\kappa = 0.98$.

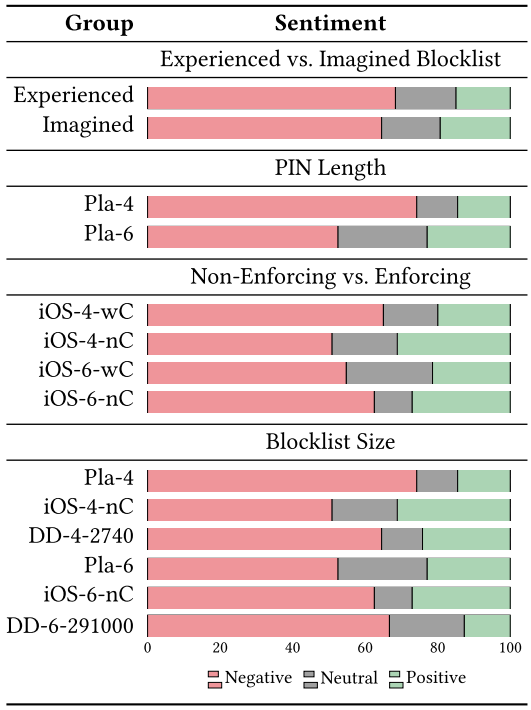


Fig. 12. Participants' sentiment: We split the participants into four categories and classified their feelings in terms of sentiment using EmoLex [38].

Using the NRC Word-Emotion Association Lexicon [38], we classified assigned codes in terms of sentiment (positive, negative, or neutral) for Figure 12. EmoLex maps individual English words (in this case, codes assigned by our coders) to exactly one sentiment. For example, "indifference," is labeled with the "negative" sentiment. As expected, participants generally had a negative reaction to the blocklist warning message.

While overall, participants expressed negative sentiments toward blocklist messages, which may be expected as warning messages are not often well received by users [2], we only observed significant differences in a single comparison. Using a χ^2 test, we found that there was significant difference ($p < 0.05$) in the proportion of negative sentiment when considering PIN length for the two placebo treatments. As both groups always experienced a blocklist event, a higher negative sentiment exists for the placebo blocklist with four digits. This might be because users were confused and angered by the warning as the blocklist event was arbitrary. However, in the six-digit PIN case, less familiarity with six-digit PINs may have led to less negative reactions.

Interestingly, participants in general consider displaying warnings about weak PIN choices to be appropriate although they cannot imagine that their own choice might be considered insecure. Moreover, sentiments are similar for those who hit the blocklist and those who imagined having done so. This suggests that future research on blocklist warning design may benefit from simply asking participants to imagine such events.

7 CONCLUSION AND RECOMMENDATIONS

This article presents the first comprehensive study of PIN security as primed for the smartphone unlock setting. In the smartphone unlock setting, developers have adopted notable

countermeasures—throttling, blocklisting, PIN length—which we consider as part of our analysis. Using a throttled attacker model, we find that six-digit PINs offer little to no advantage, and sometimes make matters worse. Also, we find that blocklists in use on today's mobile operating systems are not designed reasonably. In some cases, they need to be larger to affect security at all, while they are oversized in other cases, needlessly impairing the user experience.

Given this information, we offer a number of recommendations to mobile developers.

- In a throttled scenario, simply increasing the PIN length is of little benefit. In our results, were only able to observe a significant difference between four- and six-digit PINs for an attacker that performs at least 100 guesses. As this exceeds most attacking scenarios for mobile authentication, developers should carefully articulate an alternative threat model to justify the adoption of longer PINs. Observe that without throttling, an attacker could quickly try all four- and six-digit PINs.
- On iOS, with only 10 possible guesses, we could not observe any security benefits when a blocklist is deployed, either for four- or six-digit PINs. On Android, where 100 guesses are feasible, we find that a blocklist would be beneficial. Such a blocklist would need to contain the 1000 most popular PINs in the four-digit case or the 2,000 most popular for six-digit PINs, to increase the security of the chosen PINs while minimizing user frustration.
- We observe that the increase in terms of the perceived security is only significant when users are forced to conform with a large six-digit blocklist as compared to selecting a PIN in presence of a large four-digit blocklist (as was the case in the data-driven treatments). This may suggest users are less familiar with selecting six-digit PINs, an observation our analysis of the selection strategies supports. Yet, a detailed exploration of the reasons for this are left to future investigation.
- While we observed advantages for using a placebo blocklist in the unthrottled settings, we do not recommend implementing a placebo blocklist, as users will simply game it once the deception is known.

A APPENDIX

A.1 Survey Instrument

Questions for participants who **hit** the blocklist.

We noticed that you received the following warning while choosing your PIN:

[A screenshot of the same warning message that the participant saw during the study.]

People use different strategies for choosing their PINs. Below, we will ask about your strategy.

(1) Prior to seeing the warning above, what was your strategy for choosing your PIN?
Answer: _____

(2) After receiving the warning message, please describe how or if your strategy changed when choosing your PIN.
Answer: _____

The "Extra" question was only asked if the participant had the option to ignore the warning and did so by clicking "Use Anyway."

(Extra) You selected "Use Anyway" when choosing your final PIN. Please describe why you did not change your final PIN after seeing this warning message.
Answer: _____

(3) Please describe three general feelings or reactions that you had after you received this warning message.
Feeling 1: _____ Feeling 2: _____ Feeling 3: _____

Please select the answer choice that most closely matches how you feel about the following statements:

(4) My initial PIN creation strategy caused the display of this warning.
○ Strongly agree ○ Agree ○ Neutral ○ Disagree ○ Strongly Disagree

Questions for participants who did **not hit** the blocklist.

People use different strategies for choosing their PINs. Below, we will ask about your strategy.

(1) What was your strategy for choosing your PIN?
Answer: _____

Imagine you received the following warning message after choosing your PIN:

[A screenshot of the warning message as in Figure 4 or Figure 5.]

(2) Please describe how or if your strategy would change as a result of the message.
Answer: _____

(3) Please describe three general feelings or reactions that you would have had after you received this warning message.
Feeling 1: _____ Feeling 2: _____ Feeling 3: _____

Please select the answer choice that most closely matches how you feel about the following statements:

(4) My PIN creation strategy would cause this warning message to appear.
○ Strongly agree ○ Agree ○ Neutral ○ Disagree ○ Strongly Disagree

From now on all participants saw the same questions.

(5) It is appropriate for smartphones to display warning messages about PIN security.
○ Strongly agree ○ Agree ○ Neutral ○ Disagree ○ Strongly Disagree

Please select the answer choice that most closely matches how you feel about the following statements referring to the final PIN you chose:

The order of questions 6, 7, and 9 was chosen randomly for each participant. The attention check question was always the 8th question.

(6) I feel the PIN I chose is:
○ Secure ○ Somewhat secure ○ Neither secure nor insecure ○ Somewhat insecure ○ Insecure

(7) I feel the PIN I chose is:
○ Easy to remember ○ Somewhat easy to remember ○ Neither easy nor hard to remember ○ Somewhat hard to remember ○ Difficult to remember

(8) What is the shape of a red ball?
○ Red ○ Blue ○ Square ○ Round

(9) I feel the PIN I chose is:
○ Easy to enter ○ Somewhat easy to enter ○ Neither easy nor hard to enter ○ Somewhat hard to enter ○ Difficult to enter

(10) What is your age range?
○ 18-24 ○ 25-34 ○ 35-44 ○ 45-54 ○ 55-64 ○ 65-74 ○ 75 or older ○ Prefer not to say

(11) With what gender do you identify?
○ Male ○ Female ○ Non-Binary ○ Other ○ Prefer not to say

(12) What is the highest degree or level of school you have completed?
○ Some high school ○ High school ○ Some college ○ Trade, technical, or vocational training ○ Associate's Degree ○ Bachelor's Degree ○ Master's Degree ○ Professional Degree ○ Doctorate ○ Prefer not to say

(13) Do you use any of the following biometrics to unlock your primary smartphone? (Select all that apply)
□ Fingerprint □ Face □ Iris □ Other biometric □ I do not use a biometric □ I do not use a smartphone □ Prefer not to say

If the participant stated they use a biometric in question 13:

14A) How do you unlock your smartphone, if your biometric fails or when you reboot your primary smartphone?
○ None ○ Pattern ○ 4-digit PIN ○ 6-digit PIN ○ PIN of other length ○ Alphanumeric password ○ I use an unlock method not listed here ○ I do not use a smartphone ○ Prefer not to say

If the participant stated they do not use a biometric in question 13:

14B) What screen lock do you use to unlock your primary smartphone?
○ None ○ Pattern ○ 4-digit PIN ○ 6-digit PIN ○ PIN of other length ○ Alphanumeric password ○ I use an unlock method not listed here ○ I do not use a smartphone ○ Prefer not to say

(15) What is the operating system of your primary smartphone?
○ Android ○ iOS (iPhone) ○ Other ○ I do not use a smartphone ○ Prefer not to say

(16) Which of the following best describes your educational background or job field?
○ I have an education in, or work in, the field of computer science, computer engineering or IT.
○ I do not have an education in, nor do I work in, the field of computer science, computer engineering or IT.
○ Prefer not to say to say

(17) Please indicate if you have honestly participated in this survey and followed instructions completely. You will not be penalized/rejected for indicating 'No' but your data may not be included in the analysis:
○ Yes ○ No

(18) Please feel free to provide any final feedback you may have in the field below.
Answer: _____

A.2 Demographics

Table 10. Overall Demographics of the Participants

		Male		Female		Other		Total	
		No.	%	No.	%	No.	%	No.	%
What is your age range?		923	54%	768	45%	14	1%	1705	100%
	18–24	125	7%	87	5%	5	0%	217	13%
	25–34	461	27%	350	21%	5	0%	816	48%
	35–44	231	14%	195	11%	2	0%	428	25%
	45–54	72	4%	84	5%	0	0%	156	9%
	55–64	24	1%	47	3%	0	0%	71	4%
	65–74	10	1%	5	0%	0	0%	15	1%
	Prefer not to say	0	0%	0	0%	2	0%	2	0%
What is the highest degree or level of school you have completed?		923	54%	768	45%	14	1%	1705	100%
	Some High School	3	0%	4	0%	0	0%	7	0%
	High School	95	6%	66	4%	3	0%	164	10%
	Some College	208	12%	142	9%	5	0%	355	21%
	Training	33	2%	28	2%	0	0%	61	4%
	Associates	85	5%	104	6%	2	0%	191	11%
	Bachelor's	389	23%	321	19%	2	0%	712	42%
	Master's	82	5%	86	5%	0	0%	168	10%
	Professional	13	1%	8	0%	0	0%	21	1%
	Doctorate	14	1%	9	0%	0	0%	23	1%
	Prefer not to say	1	0%	0	0%	2	0%	3	0%
Which of the following best describes your educational background or job field?		923	54%	768	45%	14	1%	1705	100%
	Tech	360	21%	109	7%	3	0%	472	28%
	No Tech	534	31%	638	37%	8	0%	1180	69%
	Prefer not to say	29	2%	21	1%	3	0%	53	3%

For the sake of clarity, we grouped answers for *Non-Binary*, *Other*, and *Prefer not to say* under *Other*.

A.3 Device Usage

Table 11. Device Usage of the Participants

		Male		Female		Other		Total	
		No.	%	No.	%	No.	%	No.	%
Do you use any of the following biometrics to unlock your primary smartphone?		923	54%	768	45%	14	1%	1705	100%
	Fingerprint	504	30%	395	23%	7	0%	906	53%
	Face	161	9%	102	6%	0	0%	263	15%
	Iris	41	3%	17	1%	0	0%	58	4%
	Other Biometric	19	1%	26	2%	0	0%	45	3%
	No Biometric	299	18%	266	16%	5	0%	570	34%
	No Smartphone	2	0%	0	0%	0	0%	2	0%
	Prefer not to say	28	2%	28	2%	2	0%	58	4%
How do you unlock your smartphone, if your biometric fails or when you reboot your		594	55%	474	44%	7	1%	1075	100%
primary smartphone?	None	2	0%	5	0%	0	0%	7	1%
	Pattern	93	9%	55	5%	0	0%	148	14%
	4-digit PIN	262	24%	245	23%	3	0%	510	47%
	6-digit PIN	177	16%	141	14%	4	0%	322	30%
	PIN of other length	20	2%	12	1%	0	0%	32	3%
	Alphanumeric	30	3%	12	1%	0	0%	42	4%
	Other method	6	1%	2	0%	0	0%	8	1%
	No smartphone	1	0%	0	0%	0	0%	1	0%
	Prefer not to say	3	0%	2	0%	0	0%	5	0%
What screen lock do you use to unlock your primary smartphone?		329	52%	294	47%	7	1%	630	100%
	None	85	13%	104	17%	0	0%	189	30%
	Pattern	54	8%	32	5%	2	0%	88	13%
	4-digit PIN	115	18%	101	16%	2	0%	218	36%
	6-digit PIN	32	4%	27	4%	0	0%	59	8%
	PIN of other length	8	1%	3	0%	0	0%	11	2%
	Alphanumeric	8	1%	7	1%	0	0%	15	3%
	Other method	10	2%	4	1%	0	0%	14	2%
	No smartphone	0	0%	1	0%	0	0%	1	0%
	Prefer not to say	17	3%	15	2%	3	0%	35	6%
What is the operating system of your primary smartphone?		923	54%	768	45%	14	1%	1705	100%
	Android	592	35%	408	24%	8	0%	1008	59%
	iOS	323	19%	349	21%	4	0%	676	40%
	Other	2	0%	4	0%	0	0%	6	0%
	No smartphone	0	0%	0	0%	0	0%	0	0%
	Prefer not to say	6	0%	7	0%	2	0%	15	1%

Note, for the biometrics question, participants selected all that apply. For the sake of clarity, we grouped answers for *Non-Binary*, *Other*, and *Prefer not to say* under *Other*.

A.4 PIN Selection and Changing Strategies

Table 12. PIN Selection Strategies

Code Name	Frequency	Description	Example PIN	Sample from the Study
Memorable	77	Memorability was the main concern	`2827` / `777888`	"A number easy to remember."
Date	65	Special date like anniversary, birthday, graduation day	`1987` / `112518`	"A date I won't forget."
Pattern	37	Visualized a pattern on the PIN pad	`2580` / `137955`	"The numbers on how they appeared on the PIN pad."
Random	33	Randomly chosen digits	`4619` / `568421`	"Random numbers that do not repeat."
Meaning	27	Personal meaning; Familiar or significant number	`6767` / `769339`	"I chose my favorite numbers and used them repeatedly."
Reuse	18	Reused PIN from a different device/service	`0596` / `260771`	"The one I normally use."
Simple	16	Simplistic, comfortable, easy	`0000` / `123987`	"To just chose an easy PIN."
Word	12	Textonyms; Converted a word to a number	`2539` / `567326`	"Dog name."
System	10	User's established systematic strategy	`0433` / `041512`	"I used the numbers from the current time 04:33 PM."
Phone	7	(Partial) phone number	`1601` / `407437`	"I used the first four digits of a friend's phone number."

Above, we list the top 10 selection strategies. Two coders independently coded the a sample of 314 answers. The level of agreement among the coders, measured by Cohen's kappa, was $\kappa = 0.90$. Question: "*People use different strategies for choosing their PINs. Below, we will ask about your strategy. What was your strategy for choosing your PIN?*"

Table 13. PIN Changing Strategies

Code Name	Frequency	Description	Use Case	Strategy	Sample from the Study
Same	37	Same strategy for both	Selection Change	Date Date	"Birthday of relative." "Chose another birthday."
Minor	51	Slight modification of strategy	Selection Change	Meaning Meaning++	"It's one I remember, a number with personal significance." "I changed one number in the sequence to get the app to accept it."
New	95	New strategy that is different	Selection Change	Date Phone	"I used my girlfriend's birthday." "I changed my strategy to a memorable phone number's last 4 digits."

Above we list and explain our codes for the changing strategies of participants that encountered a blocklist and in response changed their PIN. Two coders independently coded a sample of 183 answers. The level of agreement among the coders, measured by Cohen's kappa was $\kappa = 0.92$. Question: "*After receiving the warning message, please describe how or if your strategy changed when choosing your PIN.*"

A.5 Feelings and Sentiments

Table 14. Participants' Feelings about the Blocklist Warning

Code Name	Frequency	Sample from the Study	Sentiment
Annoyance	125	"Annoyed by this message."	Negative
Worried	81	"I am worried about my PIN's security."	Negative
Frustrated	56	"This message frustrates me."	Negative
Surprised	53	"Surprised to see this message."	Neutral
Indifference	48	"Don't care about this message."	Negative
Thinking	47	"Thinking about my PIN's security."	Neutral
Acceptance	46	"I agree with this message."	Positive
Fear	43	"Afraid of attackers."	Negative
Compelling	41	"Motivated to change my PIN."	Positive
Doubt	39	"I distrust the veracity of this message."	Negative
Confusion	35	"This message is confusing."	Negative
Angry	32	"Angry this message appeared."	Negative
Cautious	30	"Cautious about my PIN."	Positive
Happy	24	"Happy my PIN will be stronger."	Positive
Curiosity	19	"I wonder why this message appeared."	Positive
Shame	19	"Ashamed my PIN wasn't strong."	Negative
Remember	17	"I might forget my PIN."	Neutral
Alert	15	"I'm now more aware."	Neutral
Disappointed	14	"Disappointed seeing this warning."	Negative
Safe	13	"Confident this PIN will be safe."	Positive

We coded and analyzed these feelings from a sample of 182 participants that encountered a blocklist. We also included 21 participants that only imagined hitting a blocklist. Above, we list the top 20 reported feelings. Two coders independently coded the data and the level of agreement between the coders, measured by Cohen's kappa was $\kappa = 0.98$. Question: "*Please describe three general feelings or reactions that you had after you received this warning message.*" or "*Please describe three general feelings or reactions that you would have had after you received this warning message.*"

ACKNOWLEDGMENTS

We thank Flynn Wolf, Timothy J. Forman, Leah Flynn, and Joseph Bonneau for their assistance.

REFERENCES

[1] Oleg Afonin. 2020. iPhone 5 and 5c Passcode Unlock with iOS Forensic Toolkit. Retrieved May 14, 2021 from https://blog.elcomsoft.com/2020/08/iphone-5-and-5c-passcode-unlock-with-ios-forensic-toolkit/.

[2] Devdatta Akhawe and Adrienne Porter Felt. 2013. Alice in warningland: A large-scale field study of browser security warning effectiveness. In *Proceedings of the USENIX Security Symposium*. USENIX, 257–272.

[3] Daniel Amitay. 2011. Most Common iPhone Passcodes. Retrieved May 14, 2021 from http://danielamitay.com/blog/2011/6/13/most-common-iphone-passcodes.

[4] Android Open Source Project. 2018. Full-Disk Encryption—Storing the Encrypted Key. Retrieved May 14, 2021 from https://source.android.com/security/encryption/full-disk#storing_the_encrypted_key.

[5] Android Open Source Project. 2020. Android 11: GateKeeper. Retrieved May 14, 2021 from https://android.googlesource.com/platform/system/gatekeeper/+/refs/heads/android11-release/gatekeeper.cpp#268.

[6] Apple, Inc. 2021. Apple Platform Security. Retrieved May 14, 2021 from https://manuals.info.apple.com/MANUALS/1000/MA1902/en_US/apple-platform-security-guide.pdf.
[7] Adam J. Aviv, Devon Budzitowski, and Ravi Kuber. 2015. Is bigger better? Comparing user-generated passwords on 3x3 vs. 4x4 grid sizes for android's pattern unlock. In *Proceedings of the Annual Computer Security Applications Conference*. ACM, 301–310.
[8] Adam J. Aviv, John T. Davin, Flynn Wolf, and Ravi Kuber. 2017. Towards baselines for shoulder surfing on mobile authentication. In *Proceedings of the Annual Conference on Computer Security Applications*. ACM, 486–498.
[9] Adam J. Aviv, Katherine Gibson, Evan Mossop, Matt Blaze, and Jonathan M. Smith. 2010. Smudge attacks on smartphone touch screens. In *Proceedings of the USENIX Workshop on Offensive Technologies*. USENIX, 1–7.
[10] Adam J. Aviv, Flynn Wolf, and Ravi Kuber. 2018. Comparing video based shoulder surfing with live simulation and towards baselines for shoulder surfing on mobile authentication. In *Proceedings of the Annual Conference on Computer Security Applications*. ACM, 453–466.
[11] Farid Binbeshr, Miss Laiha Mat Kiah, Lip Yee Por, and A. A. Zaidan. 2021. A systematic review of pin-entry methods resistant to shoulder-surfing attacks. *Comput. Secur.* 101 (Feb. 2021).
[12] Joseph Bonneau. 2012. The science of guessing: analyzing an anonymized corpus of 70 million passwords. In *Proceedings of the IEEE Symposium on Security and Privacy*. IEEE, 538–552.
[13] Joseph Bonneau, Sören Preibusch, and Ross Anderson. 2012. A birthday present every eleven wallets? the security of customer-chosen banking PINs. In *Financial Cryptography and Data Security*. Springer, 25–40.
[14] Thomas Brewster. 2018. Mysterious $15,000 "GrayKey" Promises To Unlock iPhone X For The Feds. Retrieved May 14, 2021 from https://www.forbes.com/sites/thomasbrewster/2018/03/05/apple-iphone-x-graykey-hack/.
[15] Thomas Brewster. 2018. The Feds Can Now (Probably) Unlock Every iPhone Model In Existence. Retrieved May 14, 2021 from https://www.forbes.com/sites/thomasbrewster/2018/02/26/government-can-access-any-apple-iphone-cellebrite/.
[16] Maria Casimiro, Joe Segel, Lewei Li, Yigeng Wang, and Lorrie Faith Cranor. 2020. A quest for inspiration: How users create and reuse PINs. In *Who Are You?! Adventures in Authentication Workshop*. 1–7.
[17] Ivan Cherapau, Ildar Muslukhov, Nalin Asanka, and Konstantin Beznosov. 2015. On the impact of touch ID on iPhone passcodes. In *Proceedings of the Symposium on Usable Privacy and Security*. USENIX, 257–276.
[18] Justin Engler and Paul Vines. 2013. Electromechanical PIN Cracking with Robotic Reconfigurable Button Basher (and C3BO). Retrieved May 14, 2021 from https://doi.org/10.5446/38941.
[19] Adrienne Porter Felt, Alex Ainslie, Robert W. Reeder, Sunny Consolvo, Somas Thyagaraja, Alan Bettes, Helen Harris, and Jeff Grimes. 2015. Improving SSL warnings: Comprehension and adherence. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, 2893–2902.
[20] Maximilian Golla, Dennis Detering, and Markus Dürmuth. 2017. EmojiAuth: Quantifying the security of emoji-based authentication. In *Proceedings of the Workshop on Usable Security*. ISOC.
[21] Maximilian Golla and Markus Dürmuth. 2018. On the accuracy of password strength meters. In *Proceedings of the ACM Conference on Computer and Communications Security*. ACM, 1567–1582.
[22] Maximilian Golla, Jan Rimkus, Adam J. Aviv, and Markus Dürmuth. 2019. Work in progress: on the in-accuracy and influence of android pattern strength meters. In *Proceedings of the Workshop on Usable Security and Privacy*. ISOC.
[23] Maximilian Golla, Miranda Wei, Juliette Hainline, Lydia Filipe, Markus Dürmuth, Elissa Redmiles, and Blase Ur. 2018. "What was that site doing with my facebook password?" Designing password-reuse notification. In *Proceedings of the ACM Conference on Computer and Communications Security*. ACM, 1549–1566.
[24] Jeremi M. Gosney ("epixoip"). 2016. How LinkedIn's Password Sloppiness Hurts Us All. Retrieved May 14, 2021 from https://arstechnica.com/?post_type=post&p=892339.
[25] Paul A. Grassi, James L. Fenton, and William E. Burr. 2017. Digital Identity Guidelines—Authentication and Lifecycle Management: NIST Special Publication 800-63B.
[26] Kristen K. Greene, Melissa A. Gallagher, Brian C. Stanton, and Paul Y. Lee. 2014. I can't type that! P@$$w0rd entry on mobile devices. In *Human Aspects of Information Security, Privacy, and Trust*. Springer, 160–171.
[27] Gregor Haas, Seetal Potluri, and Aydin Aysu. 2021. iTimed: Cache attacks on the apple a10 fusion SoC. *Cryptology ePrint Archive* Report 2021/464 (April 2021), 1–16.
[28] Marian Harbach, Emanuel von Zezschwitz, Andreas Fichtner, Alexander De Luca, and Matthew Smith. 2014. It's a hard lock life: A field study of smartphone (Un)Locking behavior and risk perception. In *Proceedings of the Symposium on Usable Privacy and Security*. USENIX, 213–230.
[29] Andrew Horton ("urbanadventurer") and Community. 2020. Android-PIN-Bruteforce – Bruteforcing the Lockscreen PIN. Retrieved May 14, 2021 from https://github.com/urbanadventurer/Android-PIN-Bruteforce.
[30] Troy Hunt. 2020. Pwned Passwords. Retrieved May 14, 2021 https://haveibeenpwned.com/Passwords.

[31] Patrick Kelley, Saranga Kom, Michelle L. Mazurek, et al. 2012. Guess again (and again and again): Measuring password strength by simulating password-cracking algorithms. In *Proceedings of the IEEE Symposium on Security and Privacy*. IEEE, 523–537.

[32] Hassan Khan, Jason Ceci, Jonah Stegman, Adam J. Aviv, Rozita Dara, and Ravi Kuber. 2020. Widely reused and shared, infrequently updated, and sometimes inherited: A holistic view of PIN authentication in digital lives and beyond. In *Proceedings of the Annual Computer Security Applications Conference*. ACM, 249–262.

[33] Hyoungshick Kim and Jun Ho Huh. 2012. PIN selection policies: Are they really effective? *Comput. Secur.* 31, 4 (Jun. 2012), 484–496.

[34] Oleksiy Lisovets, David Knichel, Thorben Moos, and Amir Moradi. 2021. Let's take it offline: Boosting brute-force attacks on iPhone's user authentication through SCA. *IACR Trans. Cryptogr. Hardw. Embed. Syst.* 2021, 3 (Jun. 2021), 1–24.

[35] Marte Løge, Markus Dürmuth, and Lillian Røstad. 2016. On user choice for android unlock patterns. In *Proceedings of the European Workshop on Usable Security*. ISOC.

[36] Philipp Markert, Daniel V. Bailey, Maximilian Golla, Markus Dürmuth, and Adam J. Aviv. 2020. This PIN can be easily guessed: Analyzing the security of smartphone unlock PINs. In *Proceedings of the IEEE Symposium on Security and Privacy*. IEEE, 286–303.

[37] William Melicher, Darya Kurilova, Sean M. Segreti, Pranshu Kalvani, Richard Shay, Blase Ur, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Michelle L. Mazurek. 2016. Usability and security of text passwords on mobile devices. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, 527–539.

[38] Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Comput. Intell.* 29, 3 (2013), 436–465.

[39] Collins W. Munyendo, Miles Grant, Philipp Markert, Timothy J. Forman, and Adam J. Aviv. 2021. Using a blocklist to improve the security of user selection of android patterns. In *Proceedings of the Symposium on Usable Privacy and Security*. USENIX, 1–19.

[40] Ellen Nakashima and Reed Albergotti. 2021. Australian Firm Azimuth Unlocked the San Bernardino Shooter's iPhone for the FBI. Retrieved May 14, 2021 from https://www.washingtonpost.com/technology/2021/04/14/azimuth-san-bernardino-apple-iphone-fbi/.

[41] Lily Hay Newman. 2019. Google's Making it Easier to Encrypt Even Cheap Android Phones. Retrieved May 14, 2021 from https://www.wired.com/story/android-encryption-cheap-smartphones/.

[42] Lina Qiu, Alexander De Luca, Ildar Muslukhov, and Konstantin Beznosov. 2019. Towards understanding the link between age and smartphone authentication. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, 163:1–163:10.

[43] Elissa M. Redmiles, Yasemin Acar, Sascha Fahl, and Michelle L. Mazurek. 2017. *A Summary of Survey Methodology Best Practices for Security and Privacy Researchers*. Technical Report CS-TR-5055. UM Computer Science Department.

[44] Thomas Reed. 2018. GrayKey iPhone Unlocker Poses Serious Security Concerns. Retrieved May 2021 from https://blog.malwarebytes.com/?p=22342.

[45] Karen Renaud and Melanie Volkamer. 2015. Exploring mental models underlying PIN management strategies. In *Proceedings of the World Congress on Internet Security*. IEEE, 19–21.

[46] Florian Schaub, Ruben Deyhle, and Michael Weber. 2012. Password entry usability and shoulder surfing susceptibility on different smartphone platforms. In *Proceedings of the International Conference on Mobile and Ubiquitous Multimedia*. ACM, 13:1–13:10.

[47] Richard Shay, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Alain Forget, Saranga Komanduri, Michelle L. Mazurek, William Melicher, Sean M. Segreti, and Blase Ur. 2015. A spoonful of sugar?: The impact of guidance and feedback on password-creation behavior. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, 2903–2912.

[48] Sergei Skorobogatov. 2017. The bumpy road towards iphone 5c NAND mirroring. In *Proceedings of the Hardware Security Conference & Training (HardwearIO'17)*. 1–55.

[49] Emily Stark. 2019. The URLephant. In *Proceedings of the USENIX Enigma Conference*. USENIX.

[50] Joshua Sunshine, Serge Egelman, Hazim Almuhimedi, Neha Atri, and Lorrie Faith Cranor. 2009. Crying wolf: An empirical study of SSL warning effectiveness. In *Proceedings of the USENIX Security Symposium*. USENIX, 399–416.

[51] Joshua Tan, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2020. Practical recommendations for stronger, more usable passwords combining minimum-strength, minimum-length, and blocklist requirements. In *Proceedings of the ACM Conference on Computer and Communications Security*. ACM, 1407–1426.

[52] Sebastian Uellenbeck, Markus Dürmuth, Christopher Wolf, and Thorsten Holz. 2016. Quantifying the security of graphical passwords: The case of android unlock patterns. In *Proceedings of the ACM Conference on Computer and Communications Security*. ACM, 161–172.

[53] Blase Ur, Felicia Alfieri, Maung Aung, Lujo Bauer, Nicolas Christin, Jessica Colnago, Lorrie Faith Cranor, Henry Dixon, Pardis Emami Naeini, Hana Habib, Noah Johnson, and William Melicher. 2017. Design and evaluation of a data-driven password meter. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. ACM, 3775–3786.

[54] Blase Ur, Fumiko Noma, Jonathan Bees, Sean M. Segreti, Richard Shay, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2015. "I added '!' at the end to make it secure": Observing password creation in the lab. In *Proceedings of the Symposium on Usable Privacy and Security*. USENIX, 123–140.

[55] Blase Ur, Sean M. Segreti, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Saranga Komanduri, Darya Kurilova, Michelle L. Mazurek, William Melicher, and Richard Shay. 2015. Measuring real-world accuracies and biases in modeling password guessability. In *Proceedings of the USENIX Security Symposium*. USENIX, 463–481.

[56] U.S. Department of Homeland Security. 2012. The Menlo Report. Retrieved May 14, 2021 from https://www.caida.org/publications/papers/2012/menlo_report_actual_formatted/.

[57] Emanuel von Zezschwitz, Alexander De Luca, and Heinrich Hussmann. 2014. Honey, i shrunk the keys: Influences of mobile devices on password composition and authentication performance. In *Proceedings of the Nordic Conference on Human-Computer Interaction*. ACM, 461–470.

[58] Emanuel von Zezschwitz, Malin Eiband, Daniel Buschek, Sascha Oberhuber, Alexander De Luca, Florian Alt, and Heinrich Hussmann. 2016. On quantifying the effective passsword space of grid-based unlock gestures. In *Proceedings of the Conference on Mobile and Ubiquitous Multimedia*. ACM, 201–212.

[59] Ding Wang, Qianchen Gu, Xinyi Huang, and Ping Wang. 2017. Understanding human-chosen PINs: Characteristics, distribution and security. In *Proceedings of the ACM Asia Conference on Computer and Communications Security*. ACM, 372–385.

[60] Gareth Watts ("gwatts") and Community. 2015. Pinfinder—iOS Screen Time & Restrictions Passcode Finder. Retrieved May 14, 2021 from https://github.com/gwatts/pinfinder.

[61] Chris Welch. 2018. Apple Releases iOS 11.4.1 and Blocks Passcode Cracking Tools Used by Police. Retrieved May 14, 2021 from https://www.theverge.com/2018/7/9/17549538/.

[62] Sonia Secher Wichmann. 2011. Self-determination theory: The importance of autonomy to well-being across cultures. *J. Humanist. Counsel.* 50, 1 (Mar. 2011), 16–26.

[63] Yulong Yang, Janne Lindqvist, and Antti Oulasvirta. 2014. Text entry method affects password security. In *Learning from Authoritative Security Experiment Results*. USENIX, 11–20.

Received January 2021; revised May 2021; accepted June 2021