2020 IEEE Symposium on Security and Privacy

# This PIN Can Be Easily Guessed: Analyzing the Security of Smartphone Unlock PINs

Philipp Markert* Daniel V. Bailey* Maximilian Golla† Markus Dürmuth* Adam J. Aviv‡
philipp.markert@rub.de danbailey@sth.rub.de maximilian.golla@csp.mpg.de markus.duermuth@rub.de aaviv@gwu.edu
* Ruhr University Bochum, † Max Planck Institute for Security and Privacy, ‡ The George Washington University

***Abstract*—In this paper, we provide the first comprehensive study of user-chosen 4- and 6-digit PINs ($n = 1220$) 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 6-digit PINs instead of 4-digit PINs provides little to no increase in security, and surprisingly may even decrease security. We also study the effects of blacklists, where a set of "easy to guess" PINs is disallowed during selection. Two such blacklists are in use today by iOS, for 4-digits (274 PINs) as well as 6-digits (2910 PINs). We extracted both blacklists compared them with four other blacklists, including a small 4-digit (27 PINs), a large 4-digit (2740 PINs), and two placebo blacklists for 4- and 6-digit PINs that always excluded the first-choice PIN. We find that relatively small blacklists in use today by iOS offer little or no benefit against a throttled guessing attack. Security gains are only observed when the blacklists are much larger, which in turn comes at the cost of increased user frustration. Our analysis suggests that a blacklist at about 10 % of the PIN space may provide the best balance between usability and security.**

## I. 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 [39] and passwords [28], 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 biometric authentication 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 [14].

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 [11] and also mainly relied on the analysis of digit sequences found in leaked text-based password datasets since this data is more readily available [46].

Given the sparsity of explicit information about PINs in the context of mobile unlock authentication, we sought to fill this vital knowledge gap by conducting the first study ($n = 1220$) on the topic where participants either selected a 4- or 6-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 our participants specifically for the mobile unlock authentication setting, reminding them that the selected "PIN protects [their] data and is used to unlock [their] smartphone." While our study cannot speak to memorability of selected PINs due to the short time duration, our qualitative feedback suggests that participants took this prompt seriously and selected relevant PINs.

PINs of 4 and 6 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 in order 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 6-digit PINs as compared to 4-digit PINs. In fact, our participants tend to select more-easily guessed 6-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 blacklisting. A blacklist 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 blacklist [24], [36], and recent guidance from NIST [20] concurs.

To understand selection strategies in the presence of a blacklist, we conducted a between-subjects comparison of PIN selection using a number of different blacklists. This included one small (27 4-digit PINs), one large (2740 4-digit PINs), and two blacklists (274 4-digit PINs and 2910 6-digit PINs) in use today on iOS devices, which we extracted for this purpose. To determine if the experience of hitting a blacklist or the content of the blacklist itself drives the result, we included a *placebo* blacklist that always excluded the participants' first choice. Finally, we included both enforcing and non-enforcing blacklists, where participants were able to "click through" and ignore the blacklist, the approach taken by iOS.

Despite the popularity of blacklists and the positive impact on textual passwords, our results show that currently employed PIN blacklists are ineffective against a throttled attacker, in both the enforcing and non-enforcing setting. This attacker

DOI 10.1109/SP40000.2020.00100

performs nearly as well at guessing 4-digit PINs as if there were no blacklist in use. To be effective, the blacklist would need to be much larger, leading to higher user frustration. Our results show that a blacklist of about 10 % of the PIN space may be able to balance the security and usability needs.

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

To summarize, we make the following contributions:

1) We report on the security of 4- and 6-digit PINs as measured for smartphone unlocking, finding that in the throttled setting, the benefit of 6-digit PINs is marginal and sometimes worse than that of 4-digit PINs.
2) Considering a realistic, throttled attacker model, we show how different blacklisting approaches influence PIN selection process for both security and usability, finding that blacklists 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 blacklisting will improve their PINs without impacting usability, except for very large blacklists.
4) We provide guidance for developers on choosing an appropriately-sized PIN blacklist that can influence the security in the throttled scenario, finding that a 4-digit PIN blacklist needs to be about 10 % of the key space to have a noticeable impact.

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

## II. 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. [22] in 2014. Since then, PINs have found new uses in encrypting mobile devices [3], [5], [30] and biometrics [14] which require a PIN as part of the keying material and for fallback authentication when biometrics fail.

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. [11], where they considered 4-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 blacklist containing the 100 most popular PINs. From our analysis, it seems that their suggestion may have formed the basis for Apple iOS's 4-digit blacklist.

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 [2]. Our research aims for greater ecological validity by specifically priming users for this task. Our data further suggests that using password leaks may be an imperfect approximation for how users choose PINs for unlock authentication.

Wang et al. [46] have also analyzed the security of PINs – in this case without any specific usage context. They report on comparing 4- and 6-digit PINs created by English and Chinese users. One counter-intuitive finding is that 6-digit PINs are less resistant to online attacks, despite the key space expansion from 4- to 6-digit PINs. Our results support the observation that in a rate limited guessing scenario there may actually be no benefit of using 6-digit PINs at all and in certain cases security even decreases. Yet, 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.

Blacklists have been considered in the context of PINs by Kim et al. [25]. They tested blacklists for both 4-digit as well as 6-digit PINs, and concluded that a reasonably-sized blacklist 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 [10]. This is a questionable attacker model especially given the sparsity of their dataset. Kim et al. compared blacklists representing 2 % and 32 % of the possible PIN space and found the large blacklist led to lower Shannon-entropy and lower offline guessing-entropy PINs, perhaps due to the composition of Kim et al.'s large blacklist. In contrast, we show that with a more realistic rate-limited, online attacker, a larger blacklist containing 27.4 % of all possible PINs provides a benefit over a smaller one that blacklists only 2.7 %, differing from the suggestion of Kim et al. regarding the effect of the size of the blacklist.

Beyond PINs, another common knowledge-based mobile authentication mechanism are Android unlock patterns, whereby a user selects a pattern that connects points on a 3x3 grid. Uellenbeck et al. [39] 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 [6], [27], [45]. Most relevant to our study, we compare the security of mobile unlock PINs to that of patterns and have obtained datasets from related work [6], [27], [39], [45].

While less common, according to Harbach et al. [22] and our own measurement (see Table IV), 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 [21], [28], [35], [44], [49].

TABLE I
DATASETS FOR STRENGTH ESTIMATIONS AND COMPARISONS.

| Kind | Dataset | Samples | Use |
|---|---|---|---|
| 4-digit PINs | Amitay-4-digit [2] | 204 432 | Strength |
| 6-digit PINs | RockYou-6-digit [46] | 2 758 490 | Strength |
| 4-digit PINs | RockYou-4-digit [46] | 1 780 587 | Comparison |
| Unlock patterns | "All" – 3x3 patterns [17] | 4 637 | Comparison |
| Passwords | LinkedIn [19] | 10 000 | Comparison |
| Passwords | Pwned Passwords v4 [23] | *Top* 10 000 | Comparison |

TABLE II
RATE LIMITING ON MOBILE OPERATING SYSTEMS.

| To Make *n* Guesses | Accumulated Waiting Time<br>Android 7, 8, 9, 10 | <br>iOS 9, 10, 11, 12, 13 |
|---|---|---|
| 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 | - |

## III. BACKGROUND

### A. 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 [5]. 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 which also survive reboots.[1]

An overview of the currently enforced limits is given in Table II. Apple's iOS is very restrictive and only allows up to 10 guesses [5] 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 [4], 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) [5].

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 being attacked. If the attacker is targeted, and is able to use other information and context about the victim, e.g., via shoulder-surfing attack [35], [7], [9] or screen smudges [8], the attacker would have significant advantages, particularly in guessing 4- vs. 6-digit PINs [9].

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

### B. Datasets

Perhaps the most realistic 4-digit PIN data is from 2011 where Daniel Amitay developed the iOS application "Big Brother Camera Security" [2]. The app mimicked a lock screen allowing users to set a 4-digit PIN. Amitay anonymously and surreptitiously collected 204 432 4-digit PINs and released them publicly [2]. While collected in an uncontrolled experiment, we apply the dataset (Amitay-4-digit) when guessing 4-digit PINs, as well as to inform the selection of our "data-driven" blacklists.

As there is no similar 6-digit PIN data available to inform the attacker, we rely on 6-digit PINs extracted from password leaks, similar to Bonneau et al.'s [11] and Wang et al.'s [46] 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, this sequence is considered as a PIN (e.g., PW: `ab3c123456d` → PIN: `123456`, but no 6-digit PINs would be extracted from the sequence `ab3c1234567d`).

By following this method, we extracted 6-digit PINs from the *RockYou* password leak, which we refer to as RockYou-6-digit (2 758 490 PINs). We also considered 6-digit PINs extracted from other password leaks, such as the *LinkedIn* [19] 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 I. For example,

[1] While there are tools by Cellebrite [13] and GrayShift [12] that exploit vulnerabilities in an attempt to escalate guessing to an unthrottled attacker, we consider such attacks out of scope. These exploits are usually bound to a specific device or OS version or can only be run within certain timeframes (e.g., 1 hour) after the last successful unlock [47].

we use a 3x3 Android unlock pattern dataset described by Golla et al. [17], combining four different datasets [6], [27], [39], [45]. It consists of 4637 patterns with 1635 of those being unique. In addition, we use a text-password dataset. Melicher et al. [28] 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 [19] leak and use the *Pwned Passwords v4* [23] list to simulate a throttled guessing attacker to estimate the guessing resistance for the sampled LinkedIn passwords as a proxy for mobile text passwords.

### C. Extracting the iOS Blacklists

As part of our set of blacklists, we also consider a blacklist of "easily guessed" 4/6-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 blacklisting of PINs is present on iOS 9 throughout the latest version iOS 13, and we also discovered that Apple updated their blacklist with the deployment of iOS 10 (e.g., the PIN `101471` is blacklisted on iOS 10.3.3, but is not on iOS 9.3.5).

In theory, it is possible to extract the blacklist by reverse engineering iOS, yet, we found a more direct way to determine the blacklist 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 and observe the presence of the blacklist 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 blacklist warning, as depicted in Figure 4, is detected by extracting the text in the photo using *Tesseract OCR*.

The extraction of all 10 000 4-digit PINs took $\sim$ 9 hours. Testing all 1 million 6-digit PINs took about 30 days using two setups in parallel. To ensure accuracy, we repeated the process for 4-digit PINs multiple times, tested lists of frequent 6-digit PINs, and verified the patterns found in the PINs. Moreover, we validated all blacklisted PINs multiple times. We refer to these two lists as the iOS-4 and iOS-6 blacklists. [2]

In total, the 4-digit blacklist 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 6-digit blacklist contains 2910 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 blacklisted by Apple overlap with a 4-digit blacklist suggested by Bonneau et al. [11] in 2012 and the top 6-digit PINs reported by Wang et al. [46] in 2017.

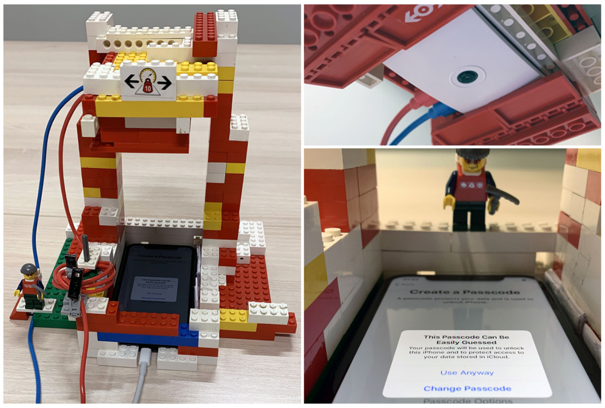

Create a Passcode
A passcode protects your data and is used to unlock iPhone.
This Passcode Can Be Easily Guessed
Your passcode will be used to unlock this iPhone and to protect access to your data stored in iCloud.
Use Anyway
Change Passcode
Passcode Options

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

## IV. User Study

In this section, we outline the specifics of the treatment conditions, the user study protocol, and the collected data. We will also discuss any limitations of the study as well as ethical considerations. Please refer to Appendix A for the specific wording and layouts of the questions.

### A. Study Protocol and Design

We conducted a user study of 4- and 6-digit PINs using Amazon Mechanical Turk (MTurk) with $n = 1220$ participants over a period of three weeks. 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 9-treatment, between-subjects study protocol for the PIN selection criteria, e.g., 4- vs. 6-digit with or without blacklisting. The specifics of the treatments are discussed in detail in Section IV-B. At the end of the study, we collected 851 and 369 PINs, 4- and 6-digits respectively, for a total of 1220 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* blacklisted 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 VI.

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 towards blacklisting 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. The parts of the survey are:

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 blacklisting 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

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

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."

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 IV-B), the users either selected a 4- or 6-digit PIN and did or did not experience a blacklist event. In Figure 4 and Figure 5 we depicted the two blacklist warnings which either allowed participants to "click through" the warning (or not). The feedback was copied to directly mimic the wording and layout of a blacklist warning used by Apple since iOS 12.
6) *Blacklisting Followup*: After creation, we asked participants about their attitudes and strategies with blacklisting. If the participants experienced a blacklist event, 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 VI-F).
7) *PIN Selection Followup*: We asked a series of questions to gauge participants' attitudes towards the PIN they selected with respect to its security and usability, where usability was appraised based on ease of entry and memorability (see Section VI-E). As part of this questionnaire, we also asked an attention check question. We excluded the data of 12 participants because we could not guarantee that they followed our instructions completely.
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 [32], we collected the demographics of the participants at the very end, including age, gender, IT background, and their current mobile unlock authentication.
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 12 participants to ensure the integrity of our data. After affirming honesty (or dishonesty), the survey concluded and was submitted.

### B. Treatments

We used 9 different treatments: 6 treatments for 4-digit PINs and 3 treatments for 6-digit PINs. The naming and description of each treatment can be found in Table III, as well as the number of participants (non-overlapping, between-subjects) exposed to each treatment.

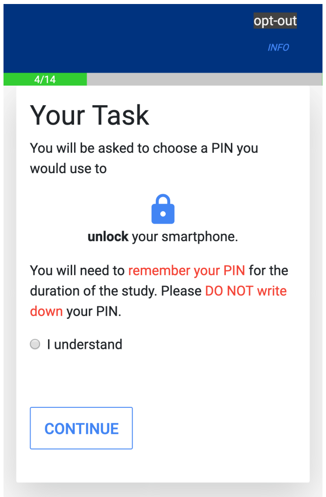
*1) Control Treatments:* For each PIN length, we had a control treatment, **Control-4-digit** and **Control-6-digit**, that simply primed participants for mobile unlock authentication and asked them to select a PIN without any blacklist interaction. These PINs form the basis of our 4- and 6-digit mobile-authentication primed PIN dataset. In total, we have 231 control 4-digit PINs and 127 control 6-digit PINs. We decided to have a larger sample of 4-digit PINs to better validate our methodology compared to other datasets.

We sometimes refer to two datasets, **First-Choice-4-digit** and **First-Choice-6-digit**. These combine the control PINs with those chosen by participants from other treatments in their "first attempt" before having been subjected to any blacklist. The First-Choice-4-digit dataset contains 851 4-digit PINs while First-Choice-6-digit consists of 369 6-digit PINs.

*2) Blacklist Treatments:* The remaining treatments considered PIN selection in the presence of a blacklist. There are two types of blacklist implementations: *enforcing* and *non-enforcing*. An enforcing blacklist does not allow to continue as long as the selected PIN is blacklisted; the user *must* select a non-blacklisted PIN. A non-enforcing blacklist warns the user that the selection is blacklisted, 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 blacklist. Visuals of the non-enforcing and enforcing feedback can be found in Figure 4 and 5, respectively.

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

*b) iOS Blacklist:* For this treatment, we included the blacklists used on Apple's iOS 12. The 4-digit iOS blacklist contains 274 PINs (2.74 % of the available 4-digit PINs), and



Your Task

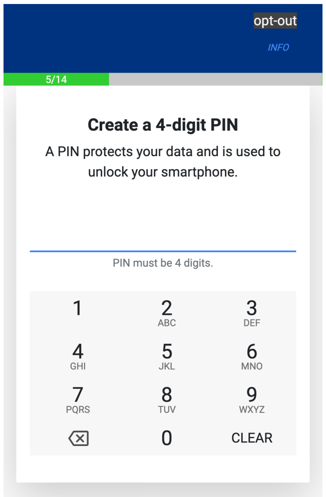
You will be asked to choose a PIN you would use to

**unlock** your smartphone.

You will need to remember your PIN for the duration of the study. Please DO NOT write down your PIN.

○ I understand

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

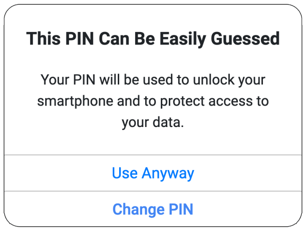


**Create a 4-digit PIN**

A PIN protects your data and is used to unlock your smartphone.

PIN must be 4 digits.

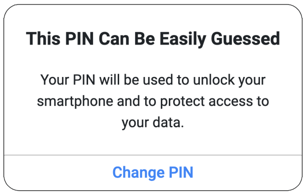
Fig. 3. The design of the page on which we asked the participants to create a PIN.



**This PIN Can Be Easily Guessed**

Your PIN will be used to unlock your smartphone and to protect access to your data.

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



**This PIN Can Be Easily Guessed**

Your PIN will be used to unlock your smartphone and to protect access to your data.

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

the 6-digit iOS blacklist contains 2910 PINs (0.291 % of the available 6-digit PINs). These blacklists provide measurements of real scenarios for users selecting PINs on iOS devices. As iOS allows users to "click through" the blacklist warning and use their blacklisted PIN anyway, we implemented our blacklisting 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 blacklists, we also had an enforcing version of the iOS blacklist for 4-digits (**iOS-4-digit-nCt**).

*c) Data-Driven Blacklists:* We considered two 4-digit blacklists that are significantly (10x) smaller (27 PINs) and (10x) larger (2740 PINs) than the iOS blacklist. The blacklists were constructed using the 27 and 2740 most frequently occurring PINs in the Amitay-4-digit dataset, and we refer to them as **DD-4-digit-27** and **DD-4-digit-2740**. When comparing these two data-driven blacklists and the one used in iOS, it can be seen that they have different compositions. While 22, i.e., 82 % of the PINs contained in DD-4-digit-27 are blacklisted in iOS, there are also 5 PINs which 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. Similar observations can be made when comparing the iOS and the large DD-4-2740 blacklist. Of 274 PINs which are rejected by iOS, 258, i.e., 92 %, are also blocked by our large data-driven blacklist. 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 which shows that double repetitions are presumably not as common as Apple expects.

### C. Recruitment and Demographics

Using Amazon's Mechanical Turk (MTurk), we recruited a total of 1452 participants. After excluding a portion due to invalid responses to attention tests or survey errors, we had 1220 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 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 4-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 that this was indeed the case.

TABLE III
OVERVIEW OF STUDIED TREATMENTS.

| | Treatment | Short | Blacklist | 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 4-digit | 274 | ✓ |
| | iOS-4-digit-nCt | iOS-4-nC | iOS 4-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 6-digit | 2910 | ✓ |

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 12 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.

*a) Demographics and Background:* As typical on MTurk, our sample is relatively young and better educated than the general US population. Of the participants, 619 identified as male (51%) while 590 (48%) identified as female (1 % identified as other or preferred not to say), and the plurality of our participants were between 25 and 34 years old (47 %). Most participants had some college (23 %) or a bachelor's

TABLE IV
USAGE OF MOBILE UNLOCK AUTHENTICATION SCHEMES.

| Primary Scheme | No. | % | Secondary Scheme | No. | % |
|---|---|---|---|---|---|
| Fingerprint | 573 | 47 % | 4-digit PIN | 285 | 50 % |
| | | | 6-digit PIN | 148 | 26 % |
| | | | Pattern | 84 | 14 % |
| | | | Other | 56 | 10 % |
| Face | 162 | 13 % | 4-digit PIN | 82 | 51 % |
| | | | 6-digit PIN | 50 | 31 % |
| | | | Pattern | 15 | 9 % |
| | | | Other | 15 | 9 % |
| Other Biometric | 24 | 2 % | 4-digit PIN | 5 | 21 % |
| | | | 6-digit PIN | 3 | 13 % |
| | | | Pattern | 14 | 58 % |
| | | | Other | 2 | 8 % |
| 4-digit PIN | 165 | 14 % | *No secondary scheme used.* | | |
| 6-digit PIN | 37 | 3 % | | | |
| Pattern | 60 | 5 % | | | |
| Other | 59 | 5 % | | | |
| None | 140 | 11 % | | | |

degree (39 %), and few (12 %) had a master's or doctoral degree. While 26 % described having a technical background, 71 % described not having one. We have the full details of the demographics responses in Appendix B in Table IX.

*b) Smartphone OS:* We asked participants which operating system they use on their primary smartphone. Slightly more than half, 698 (57 %), of the participants were Android users, while 506 (42 %) 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 C in Table X.

*c) 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 IV shows a compressed description, a detailed breakdown can be found in the Appendix C in Table X. For knowledge-based authenticators, considered here, PINs are the most common: 44 % described using a 4-digit PIN, 20 % using a 6-digit PIN, and 3 % using a PIN longer than 6 digits. The second most common form of knowledge-based authentication are Android unlock patterns at 14 %, and 44 participants (or 4 %) reported using an alphanumeric password. In our study, 140 participants reported not using any locking method.

### D. Ethical Considerations

All of the survey material and protocol was approved by our Institutional Review Board (IRB). Beyond meeting the approval of our institution, we worked to uphold the ethical principles outlined in the Menlo Report [43].

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 blacklist was quite large (DD-4-digit-2740), 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 blacklists 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 authentication. Similarly, distributing blacklists 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.

### E. Limitations

There are a number of limitations in this study. Foremost among them is the fact that the participant sample is skewed towards 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 more age-diverse [31] and location-diverse populations select PINs.

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 [41], [40] and mobile settings [28], and these results generalize when compared to passwords from leaked password datasets [42]. We have similar results here. When compared to the most realistic dataset previously available, Amitay-4-digit, the most common 4-digit PINs collected in our study are also present in similar distributions to Amitay [2]. 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 blacklist 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 [38], [1], [15], [18]. 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.

## V. PIN Selection on Smartphones

In the following section, we discuss the security of both 4- and 6-digit PINs. Unless otherwise stated, our analyzed dataset consists of the PINs entered before any blacklist warning in Step (5) of the study. These so-called "first choice" PINs (cf. Table VI) are unaffected by the blacklists.

### A. Strength of 4- and 6-digit PINs

*a) 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. [10]. 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 V.

We report the $\beta$-success-rate, which measures the expected guessing success for a throttled adversary limited to $\beta$-guesses per account (e.g., $\lambda_3$ = 3 guesses). Moreover, we provide the Min-entropy $H_\infty$ 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 ($\alpha$-guesswork) $G_\alpha$, which provides an estimate for an unthrottled attacker trying to guess a fraction $\alpha$ of all PINs in the dataset. 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 which all differ in size, we downsampled First-4, Amit-4, Rock-4, and Rock-6 to the size of the smallest dataset First-6 (369 PINs) in our calculations. We repeated this process 500 times, removed outliers using Tukey fences with $k = 1.5$. In Table V 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 4-digit PIN occurs only 1.7× more often, leading to a lower $H_\infty$ value.

TABLE V
Guessing difficulty for a perfect-knowledge attacker.

| | Online Guessing (Success %) | | | Offline Guessing (bits) | | | |
|---|---|---|---|---|---|---|---|
| **Dataset** | $\lambda_3$ | $\lambda_{10}$ | $\lambda_{30}$ | $H_\infty$ | $\widetilde{G}_{0.05}$ | $\widetilde{G}_{0.1}$ | $\widetilde{G}_{0.2}$ |
| First-4† | 3.79 % | 7.86 % | 16.80 % | 5.72 | 6.60 | 7.11 | -⋆ |
| Amit-4† | 9.49 % | 16.26 % | 26.29 % | 4.53 | 4.74 | 5.16 | 6.33 |
| Rock-4† | 8.67 % | 18.70 % | 32.79 % | 4.72 | 4.94 | 5.23 | 5.81 |
| First-6 | 6.23 % | 10.30 % | 15.72 % | 4.53 | 5.19 | 6.57 | -⋆ |
| Rock-6† | 13.28 % | 16.53 % | 21.95 % | 3.10 | 3.10 | 3.07 | -⋆ |

†: For a fair comparison we downsampled the datasets to the size of First-6 (369 PINs).

⋆: We omit entries which are not sufficiently supported by the underlying data.

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 studied in the previous work. This is true for both the 4- and 6-digit PINs, which supports our motivation for conducting studies that collect PINs directly.

*b) Guess Number-Driven Strength Estimates:* Next, we estimate the security of the PINs in regard to real-world guessing attacks. For this, we consider an attacker as described in Section III-A. 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, Golla et al. [16] suggests ordering those PINs using a Markov model. We trained our model on the bi-grams (4-digit PINs) or tri-grams (6-digit PINs) of the respective attacking datasets which simulates the attacker with the highest success rate for each case without overfitting the problem.

An overview of our guessing analysis can be found in Figure 6. In the throttled scenario, depicted in Figure 6(a), we find attacking 4-digit PINs with the Amitay-4-digit dataset (△) is more effective than using RockYou-4-digit (▽). We simulate the stronger attacker by utilizing the Amitay dataset in subsequent strength estimations of 4-digit PINs.

When comparing 4- (△) and 6-digit PINs (×), we see that guessing performance varies. For 10 guesses (the maximum allowed under iOS), we find 4.6 % of the 4-digit and 6.5 % of the 6-digit PINs are guessed. For 30 guesses (a less determined attacker on Android), 7.6 % of the 4-digit and 8.9 % of the 6-digit PINs are guessed and for 100 guesses (a reasonable upper bound on Android), 16.2 % of the 4-digit and 13.3 % of the 6-digit PINs.

Somewhat counter-intuitive is the weaker security for 6-digit PINs for the first 40 guesses. Upon investigation, the most-common 6-digit PINs are more narrowly distributed than their most-common 4-digit counterparts. The most common 6-digit PINs consist of simple PINs, such as `123456` as defined in Table XII in Appendix E, and repeating digits. In contrast, the most common 4-digit PINs consist of simple PINs, patterns, dates, and repeating digits. As a result, the most common 6-digit PINs may actually be easier to guess and less diverse than the most common 4-digit PINs.

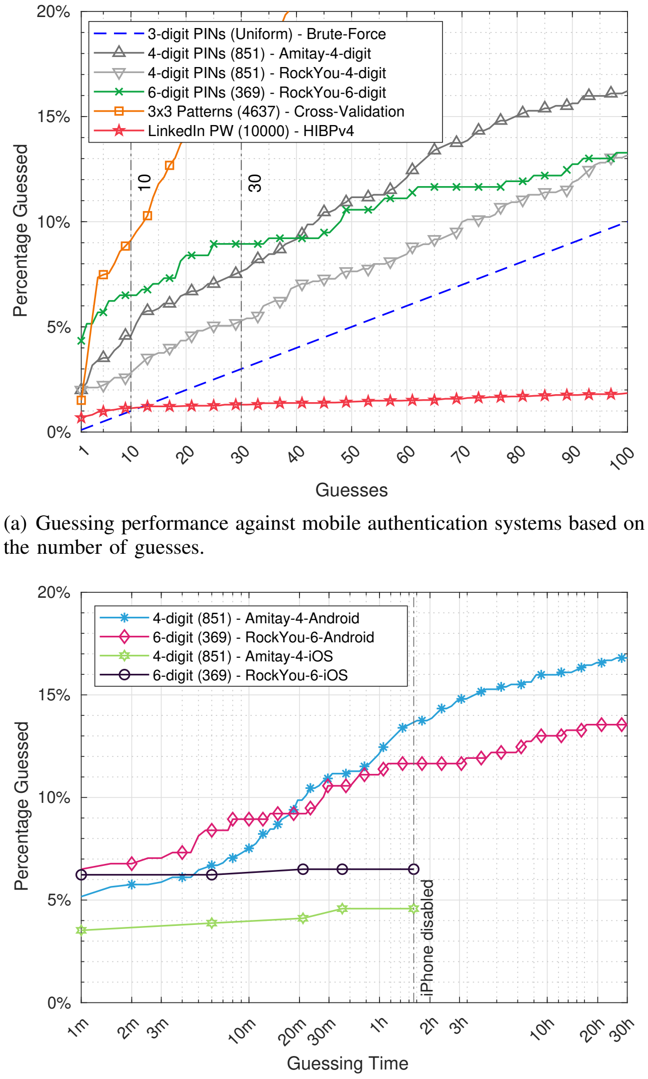
There could be many explanations for this counter-intuitive finding. One explanation may be that users have more 4-digit PIN sequences to draw on in choosing a PIN, such as dates, but have fewer natural 6-digit analogs, and thus revert to less diverse, more easily guessed choices. Another explanation may be that users have a false sense of security that comes with 6-digit PINs as they are "two digits more secure" than 4-digit PINs. Thus, users do not feel that they need more complexity in their 6-digit PIN choices. Either way, future research is needed to better understand this phenomenon, which has also been observed by Aviv et al. [6] in the context of increasing the size (4x4 vs. 3x3) of Android graphical unlock patterns.

Finally, we compare guessing resistance with other mobile authentication schemes including Android's graphical unlock patterns drawn on a 3x3 grid (□) and alphanumeric passwords (★), along with a uniform distribution of 3-digit PINs (–). In theory, a 3x3 grid allows 389 112 unique patterns, yet, the distribution of patterns is highly skewed [39]. When considering an attack throttled to 100 guesses, 35.5 % of the patterns will be guessed. Against this attack, 4- and 6-digit PINs are twice as good. Password-based authentication, on the other hand, 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 II for iOS and Android. iOS has stricter rate limiting with a maximum of 10 guesses that can be completed in 1h 36m, at which point an attacker compromises 4.6 % of the 4-digit PINs and 6.5 % of the 6-digit PINs. At the same time limit of roughly 1.5 h, an attacker on Android is able to compromise 13.6 % of the 4-digit PINs and 11.7 % of the 6-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 is able to compromise 3.5 % of the 4-digit PINs and 6.2 % of the 6-digit PINs. But there are only marginal gains for 10 guesses which take 1h 36m on iOS with 4.6 % of the 4-digit PINs and 6.5 % of 6-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. Of course, in a targeted attack setting, there may be additional information or other motivations for the attacker not modeled here.

To summarize, in line with previous work from Wang et al. [46], we found no evidence that 6-digit PINs offer any security advantage over 4-digit PINs considering a throttled guessing attacker, which is the relevant threat model for mobile unlock authentication. To support this claim, we performed $\chi^2$ tests ($\alpha = 0.05$) for both the 4- and 6-digit PINs guessed within 10 [4.6 %, 6.5 %], 30 [7.6 %, 8.9 %], and 100 guesses [16.2 %, 13.3 %]. Neither the test for 10 guesses showed a significant difference ($p = 0.16$) in PIN strength, nor the tests for 30 ($p = 0.44$) or 100 guesses ($p = 0.19$).



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

(b) Guessing performance against 4- and 6-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 different rate limits of Android and iOS (cf. Table II).

### B. 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 200 PIN selection strategies using two coders. Inter-rater reliability between the coders measured by Cohen's kappa was $\kappa = 0.92$. The 10 most popular strategies are shown in Appendix E in Table XII. We found no difference in the top 5 selection strategies between 4- and 6-digit PINs.

While the set of selection strategies is diverse, we found that many of the participants chose their PINs based on dates, especially birthdays and anniversaries. Followed by that are PINs that were perceived memorable by participants who have selected something "easy to remember." Also popular are patterns on the PIN pad and PINs that have some meaning to the participants like a partial ZIP code or a favorite number.

TABLE VI
SECURITY METRICS AND CREATION TIMES FOR PINS CONSIDERING DIFFERENT DATASETS AND TREATMENTS.

| | Name | Participants | Blacklist 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 | - | - | - |
| Treatments | Control-4-digit | 231 | - | 11 | 5 % | 19 | 8 % | 39 | 17 % | 1 185 | 7.9 s | 1.48 s | 1.01 |
| | Placebo-4-digit | 122 | 122 | 5 | 4 % | 11 | 9 % | 19 | 16 % | 2 423 | 21.8 s | 1.52 s | 2.15 |
| | iOS-4-digit-wCt | 124 | 28 | 5 | 4 % | 8 | 6 % | 18 | 15 % | 1 405 | 10.4 s | 1.36 s | 1.17 |
| | iOS-4-digit-nCt | 126 | 21 | 4 | 3 % | 10 | 8 % | 14 | 11 % | 1 747 | 9.3 s | 1.58 s | 1.29 |
| | DD-4-digit-27 | 121 | 5 | 4 | 3 % | 7 | 6 % | 18 | 15 % | 1 928 | 8.8 s | 1.47 s | 1.11 |
| | DD-4-digit-2740 | 127 | 88 | 0 | 0 % | 0 | 0 % | 1 | 1 % | 2 871 | 25.4 s | 1.55 s | 2.98 |
| Datasets | First-Choice-6-digit | 369 | - | 24 | 7 % | 33 | 9 % | 49 | 13 % | 39 389 | - | - | - |
| | Clicked-through-6 | 10 | 10 | 9 | 90 % | 9 | 90 % | 9 | 90 % | 1 | - | - | - |
| Treatments | Control-6-digit | 127 | - | 7 | 6 % | 12 | 9 % | 18 | 14 % | 36 822 | 11.5 s | 2.52 s | 1.01 |
| | Placebo-6-digit | 117 | 117 | 3 | 3 % | 6 | 5 % | 10 | 9 % | 154 521 | 28.5 s | 2.98 s | 2.17 |
| | iOS-6-digit-wCt | 125 | 15 | 9 | 7 % | 9 | 7 % | 13 | 10 % | 40 972 | 11.9 s | 2.56 s | 1.06 |

## VI. BLACKLISTS AND PIN SELECTION

We now present results on our 7 blacklist treatments: 5 treatments for 4-digit PINs and 2 treatments for 6-digit PINs as shown in Table VI.

### A. Attacker's Knowledge of Blacklists

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

Knowledge of the blacklist is unhelpful when considering the placebo (Pla-4 and Pla-6) and the click-through treatments (iOS-4-wC and iOS-6-wC). The blacklist is effectively of size one for the placebo as the first choice of a participant is dynamically blacklisted. 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 blacklist-knowledgeable attacker for the placebo treatments (see Table VI).

The case with a non-enforcing blacklist where users can click through the warning message is more subtle. If the attacker is explicitly choosing not to consider PINs on the blacklist, even though they may *actually* be selected due to non-enforcement, the guessing strategy is harmed (negative in Table VII). 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 blacklist to change their 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), there are clear advantages when knowing the blacklist.

### B. Blacklisting Impact on Security

We now consider how the different blacklists perform in terms of improving security. The primary results are in Table VI where we report on the guessing performance against each treatment. As described in Section III-A, 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). In addition, we provide the attacker's performance in an unthrottled setting based on the median guess number. The 4-digit attacker is informed by the Amit-4 dataset, while the 6-digit attacker employs the Rock-6 dataset. Both attackers guess in frequency order with knowledge of the blacklist where appropriate (see Section VI-A).

TABLE VII
ATTACKER'S GAIN FROM BLACKLIST 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 | ✗ |

We performed a multivariant $\chi^2$ test comparison ($\alpha = 0.05$) for the PINs guessed within 10, 30, and 100 guesses across treatments. The test for 10 and 30 guesses did not show any significant difference ($p = 0.21$ and $p = 0.10$); the test for 100 guesses did ($p < 0.01$), as described below.

*a) Smaller Blacklists:* In the throttled setting with 100 guesses, there is little difference among iOS-4-digit-wCt (15 %), iOS-4-digit-nCt (11 %), DD-4-digit-27 (15 %), Placebo-4-digit (16 %), compared to Control-4-digit (17 %) and First-Choice-4-digit (16 %). Our post-hoc analyses (Bonferroni-corrected for multiple testing) results support this, as we found no significant difference between the smaller blacklists. It is therefore hard to justify the combination of throttling and small blacklists, especially as blacklist warnings are associated with negative sentiments (see Section VI-F).

In the unthrottled setting, though, we see some differences between the smaller and placebo blacklist cases. Notably, the smallest blacklist (DD-4-digit-27) outperforms the 10× larger iOS blacklist (iOS-4-digit-nCt). We conjecture this may be due to iOS' inclusion of PINs based on repetitions which were chosen less often by our participants. As a result, in an unthrottled setting, blacklisting can offer real benefits. The median guess numbers for both 4- and 6-digit placebos suggest that just pushing users away from their first choice can improve security. Unfortunately, direct use of a placebo blacklist is unlikely to be effective and is problematic in practice as users will quickly figure out the deception.

Finally, we reiterate that these improvements to the unthrottled attack setting appear to be only of academic interest: given the small key space, it is reasonable to assume that all possible combinations can be exhaustively tested within minutes [33].

*b) Large Blacklist:* We also consider a very large blacklist in the DD-4-digit-2740 treatment containing 2740 PINs, 10× bigger than the 4-digit iOS blacklist and blocking 27.4 % of the key space. At this scale, we do see noticeable effects on security in the throttled setting. Even after 100 guesses, the attacker finds only 1 % of 4-digit PINs. Our $\chi^2$ tests support this, for 100 guesses we found a significant difference ($p < 0.01$). For post-hoc analyses (Bonferroni-corrected) we found a significant difference between the large DD-4-2740 blacklist and Con-6 ($p < 0.01$) as well as all other 4-digit treatments: Con-4 ($p < 0.001$), Pla-4 ($p < 0.01$), iOS-4-wC ($p < 0.01$), iOS-4-nC ($p < 0.05$), and DD-4-27 ($p < 0.01$). This suggests that a larger blacklist can improve security in a throttled setting.

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

*c) Correctly Sizing a Blacklist:* While there is a clear benefit to having a large blacklist, it is important to consider the right size of a blacklist to counteract negative usability and user experience issues. This leads to the question: *Can a smaller blacklist provide similar benefits in the throttled setting and if so, what is an appropriately sized blacklist?*

Data from the DD-4-digit-2740 treatment enables us to simulate how users would have responded to shorter blacklists. 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 blacklist. Consider a smaller blacklist that would have permitted choice $n-1$ to be the final PIN, rather than $n$. To simulate that smaller blacklist size, we use choice $n-1$.

The results of the simulation are shown in Figure 7. We observe that there are several troughs and peaks in the curves. 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 blacklist 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 blacklist has restricted all first choice PINs, whereby the attacker can now take advantage of guessing popular second choices which results in a peak. These cycles continue until the blacklist gets so large that few acceptable PINs remain, and the attacker's advantage grows steadily by guessing the remaining PINs not on the blacklist.

Based on these cycles, we conclude that an appropriately-sized blacklist 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 Table VI), the first trough appears the most ideal, which occurs at about 10 % of the 4-digit PIN space throttled at 100 guesses. We expect similar results for 6-digit blacklists as well, but we do not have sufficient examples of selected 6-digit PINs to perform the same analysis. Future research would be needed to confirm this. However, as 4-digit PINs are still exceedingly common, this result provides important insight to developers to improve user selection of 4-digit PINs.

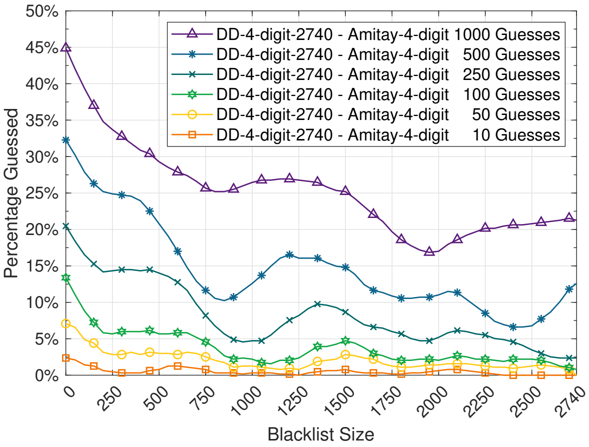


Fig. 7. Blacklist size recommendation: For throttled attackers, limited to 100 guesses, a blacklist of ∼10 % of the key space (∼1150 PINs) is ideal.

## C. Enforcing the Blacklist

In the 4-digit case, we compared the enforcing (iOS-4-digit-nCt) with the non-enforcing (iOS-4-digit-wCt) blacklist and found that enabling a click-through option does not show significant security 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 VI).

These results seem to be driven by the fact that it is uncertain whether the user clicked through (see Table VII). In an enforcing setting, the attacker can leverage the blacklist but is equally challenged in guessing the remaining PINs.

We also investigated why participants chose to ignore and click through the warning messages. From the 28 participants who saw a blacklist warning in the iOS-4-wC treatment, we observed a click-through-rate (CTR) of 68 % (19 participants). In the respective 6-digit treatment iOS-6-wC, 10 out of 15, i.e., 67 %, ignored the warning. This is twice the rate at which TLS warnings are ignored (∼ 30 %) [37].

We also 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 [26], [48], [34].

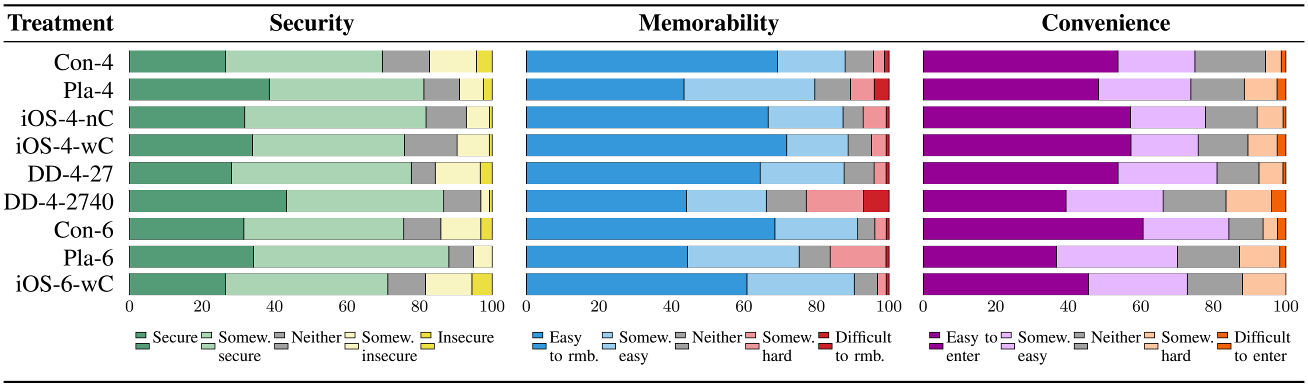


Fig. 8. Participants' perception of their PIN's security (*Secure – Insecure*), memorability (*Easy to remember – Difficult to remember*), and convenience (*Easy to enter – Difficult to enter*).

TABLE VIII
PARTICIPANTS' PIN CHANGING.

| | | Selection vs. Changing Strategy | | | | Edit Distance | |
|---|---|---|---|---|---|---|---|
| Treatment | Hits | Sample | Same | Minor | New | Mean | SD |
| Pla-4 | 122 | 29 | 10 | 7 | 12 | 3.20 | 0.90 |
| iOS-4-wC | 9⋆ | 9 | 0 | 4 | 5 | 3.11 | 0.87 |
| iOS-4-nC | 21 | 21 | 4 | 6 | 11 | 3.24 | 0.92 |
| DD-4-27 | 5 | 5 | 2 | 2 | 1 | 3.20 | 0.75 |
| DD-4-2740 | 88 | 29 | 4 | 7 | 18 | 3.39 | 0.76 |
| Pla-6 | 117 | 28 | 8 | 5 | 15 | 4.59 | 1.41 |
| iOS-6-wC | 5⋆ | 5 | 0 | 2 | 3 | 4.40 | 1.20 |

⋆: Hit blacklist, and did not click-through.

In older versions of iOS, the blacklist 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*" (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 CTRs of around 70 % [1], we hope that new designs can also improve blacklist warning CTRs [37].

## D. PIN Changing Strategies

In our study, we asked 367 participants who faced a blacklist how their creation strategy changed in response to the warning. We sampled 126 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 E in Table XIII. Two coders independently coded the data. Inter-rater reliability between the coders measured by Cohen's kappa was $\kappa = 0.96$. The results are shown in Table VIII.

About 50 % of the participants choose a new strategy when confronted with a blacklist warning. Only participants of the DD-4-27 treatment with a very small blacklist, tended to keep their pre-warning strategy. The edit distances vary slightly across the treatments and support this self-reported behavior: participants in the 4-digit scenario changed on average 3 digits with the standard deviation showing that some participants changed their PIN completely while some participants only changed 2 digits. The same conclusion can be drawn from the edit distances in the 6-digit case.

## E. 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 8. 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, pair-wise 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*. Post-hoc analysis indicates that the presence of a blacklist for 4-digit PINs increases the security perception of the final PIN selected. This is supported by considering the 4-digit placebo treatment (Pla-4) compared to the 4-digit control (Con-4). In the placebo treatment, every participant interacted with a blacklist, and there is a significant increase in security perceptions ($p < 0.01$). We see similar differences for the large blacklist treatment DD-4-2740 ($p < 0.001$), where again, a large portion (70 %) of participants encountered the blacklist. We did not see significant differences for 6-digit PIN users after encountering the blacklist. This may be because there is a pre-existing notion that 6-digit PINs are secure.

For *memorability* we also found significant differences among the treatments. In post-hoc analysis we found that increased interaction with the blacklist led to lower perceived memorability of PINs, as evidenced by the Pla-4 ($p < 0.001$), DD-4-2740 ($p < 0.001$), and the Pla-6 ($p < 0.01$) treatments compared to their respective control treatments. The DD-4-2740 showed the most significant differences with other treatments, likely due to the fact that many participants encountered the blacklist for multiple PIN choices and thus were relying on not just second-choice PINs, but also third- and fourth-choice, etc. PINs that are perceived to be less memorable.

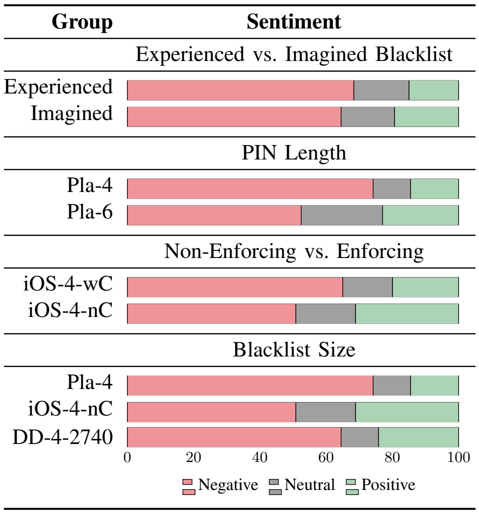


Fig. 9. Participants' sentiment: We split the participants into four categories and classified their feelings in terms of sentiment using EmoLex [29].

The responses to perceived *convenience* also show significant differences, however, post-hoc analysis revealed limited effects when considering pair-wise comparisons. In general, participants perceived their 4-digit PINs at the same convenience level across treatments. While we did not see a significant difference between convenience levels between 4- and 6-digit PINs, the perceived convenience of 6-digit PINs may be precarious because we observed a significant difference ($p < 0.01$) between the 6-digit placebo and control treatments. This suggests that while a user may be comfortable with their first-choice 6-digit PIN, there is much higher perceived *inconvenience* for their second-choice 6-digit PIN.

### *F. User Sentiment*

To gain insight into participants' sentiments regarding blacklisting, 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 blacklist. Accompanying the prompt are three free-form, short text fields. A codebook was constructed by two individual coders summarized in Appendix D in Table XI. For each of the four categories (blacklist hit experienced vs. imagined, 4- vs. 6-digit PINs, non-enforcing vs. enforcing, different blacklist 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 $\kappa = 0.91$.

Using the NRC Word-Emotion Association Lexicon [29], we classified assigned codes in terms of sentiment (positive, negative, or neutral) for Figure 9. 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 blacklist warning message.

While overall, participants expressed negative sentiments towards blacklist messages, which may be expected as warning messages are not often well received by users [1], we only observed significant differences in a single comparison.

Using a $\chi^2$ 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 blacklist event, a higher negative sentiment exists for the placebo blacklist with 4-digits. This might be because users were confused and angered by the warning as the blacklist event was arbitrary. However, in the 6-digit PIN case, less familiarity with 6-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 blacklist and those who imagined having done so. This suggests that future research on blacklist warning design may benefit from simply asking participants to imagine such events.

## VII. Conclusion and Recommendations

This paper 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, blacklisting, PIN length—which we consider as part of our analysis. Using a throttled attacker model, we find that 6-digit PINs offer little to no advantage, and sometimes make matters worse. Also, blacklists would have to be far larger than those in use on today's mobile operating systems to affect security.

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, there was no significant difference between 4- and 6-digit PINs within the first 100 guesses. To justify the adoption of longer PINs, developers should carefully articulate an alternative threat model. Observe that without throttling, an attacker could quickly try all 4- and 6-digit PINs.
- On iOS, with only 10 possible guesses, we could not observe any security benefits when a blacklist is deployed, either for 4- or 6-digit PINs. On Android, where 100 guesses are feasible, we find that a blacklist would be beneficial for 4-digit PINs. However, such a blacklist would need to contain roughly 10% of the PIN space which is much more than currently deployed blacklists. More research is needed to test the effectiveness of blacklists for 6-digit PINs.
- We observe that the perceived convenience is lower when users are forced to select a second 6-digit PIN as compared to selecting a second 4-digit PIN (as was the case in the placebo treatments). This may suggest users are less familiar with selecting 6-digit PINs, but the reasons for this are left to future investigation.
- While we observed advantages for using a placebo blacklist in the unthrottled settings, we do not recommend implementing a placebo blacklist, as users will simply game it once the deception is known.

## Acknowledgments

This research was supported by the research training group "Human Centered Systems Security" sponsored by the state of North Rhine-Westphalia, Germany, and the German Research Foundation (DFG) within the framework of the Excellence Strategy of the Federal Government and the States – 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 also wish to thank Flynn Wolf, Timothy J. Forman, and Leah Flynne for their assistance, and we thank Joseph Bonneau for his revision guidance as well as the feedback from the anonymous reviewers.

## References

[1] Devdatta Akhawe and Adrienne Porter Felt. Alice in Warningland: A Large-Scale Field Study of Browser Security Warning Effectiveness. In *USENIX Security Symposium*, SSYM '13, pages 257–272, Washington, District of Columbia, USA, July 2013. USENIX.

[2] Daniel Amitay. Most Common iPhone Passcodes, June 2011. http://danielamitay.com/blog/2011/6/13/most-common-iphone-passcodes, as of April 1, 2020.

[3] Android Open Source Project. Full-Disk Encryption – Storing the Encrypted Key, August 2018. https://source.android.com/security/encryption/full-disk#storing_the_encrypted_key, as of April 1, 2020.

[4] Android Open Source Project. Android 10: GateKeeper – ComputeRetryTimeout Function, September 2019. https://android.googlesource.com/platform/system/gatekeeper/+/refs/heads/android10-release/gatekeeper.cpp#261, as of April 1, 2020.

[5] Apple, Inc. Apple Platform Security, December 2019. https://manuals.info.apple.com/MANUALS/1000/MA1902/en_US/apple-platform-security-guide.pdf, as of April 1, 2020.

[6] Adam J. Aviv, Devon Budzitowski, and Ravi Kuber. Is Bigger Better? Comparing User-Generated Passwords on 3x3 vs. 4x4 Grid Sizes for Android's Pattern Unlock. In *Annual Computer Security Applications Conference*, ACSAC '15, pages 301–310, Los Angeles, California, USA, December 2015. ACM.

[7] Adam J. Aviv, John T. Davin, Flynn Wolf, and Ravi Kuber. Towards Baselines for Shoulder Surfing on Mobile Authentication. In *Annual Conference on Computer Security Applications*, ACSAC '17, pages 486–498, Orlando, Florida, USA, December 2017. ACM.

[8] Adam J. Aviv, Katherine Gibson, Evan Mossop, Matt Blaze, and Jonathan M. Smith. Smudge Attacks on Smartphone Touch Screens. In *USENIX Workshop on Offensive Technologies*, WOOT '10, pages 1–7, Washington, District of Columbia, USA, August 2010. USENIX.

[9] Adam J. Aviv, Flynn Wolf, and Ravi Kuber. Comparing Video Based Shoulder Surfing with Live Simulation and Towards Baselines for Shoulder Surfing on Mobile Authentication. In *Annual Conference on Computer Security Applications*, ACSAC '18, pages 486–498, Puerto Rico, USA, December 2018. ACM.

[10] Joseph Bonneau. The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords. In *IEEE Symposium on Security and Privacy*, SP '12, pages 538–552, San Jose, California, USA, May 2012. IEEE.

[11] Joseph Bonneau, Sören Preibusch, and Ross Anderson. A Birthday Present Every Eleven Wallets? The Security of Customer-Chosen Banking PINs. In *Financial Cryptography and Data Security*, FC '12, pages 25–40, Kralendijk, Bonaire, February 2012. Springer.

[12] Thomas Brewster. Mysterious $15,000 "GrayKey" Promises To Unlock iPhone X For The Feds, March 2018. https://www.forbes.com/sites/thomasbrewster/2018/03/05/apple-iphone-x-graykey-hack/, as of April 1, 2020.

[13] Thomas Brewster. The Feds Can Now (Probably) Unlock Every iPhone Model In Existence, February 2018. https://www.forbes.com/sites/thomasbrewster/2018/02/26/government-can-access-any-apple-iphone-cellebrite/, as of April 1, 2020.

[14] Ivan Cherapau, Ildar Muslukhov, Nalin Asanka, and Konstantin Beznosov. On the Impact of Touch ID on iPhone Passcodes. In *Symposium on Usable Privacy and Security*, SOUPS '15, pages 257–276, Ottawa, Canada, July 2015. USENIX.

[15] Adrienne Porter Felt, Alex Ainslie, Robert W. Reeder, Sunny Consolvo, Somas Thyagaraja, Alan Bettes, Helen Harris, and Jeff Grimes. Improving SSL Warnings: Comprehension and Adherence. In *ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 2893–2902, Seoul, Republic of Korea, April 2015. ACM.

[16] Maximilian Golla and Markus Dürmuth. On the Accuracy of Password Strength Meters. In *ACM Conference on Computer and Communications Security*, CCS '18, pages 1567–1582, Toronto, Ontario, Canada, October 2018. ACM.

[17] Maximilian Golla, Jan Rimkus, Adam J. Aviv, and Markus Dürmuth. Work in Progress: On the In-Accuracy and Influence of Android Pattern Strength Meters. In *Workshop on Usable Security and Privacy*, USEC '19, San Diego, California, USA, February 2019. ISOC.

[18] Maximilian Golla, Miranda Wei, Juliette Hainline, Lydia Filipe, Markus Dürmuth, Elissa Redmiles, and Blase Ur. "What was that site doing with my Facebook password?" Designing Password-Reuse Notification. In *ACM Conference on Computer and Communications Security*, CCS '18, pages 1549–1566, Toronto, Canada, November 2018. ACM.

[19] Jeremi M. Gosney ("epixoip"). How LinkedIn's Password Sloppiness Hurts Us All, June 2016. https://arstechnica.com/?post_type=post&p=892339, as of April 1, 2020.

[20] Paul A. Grassi, James L. Fenton, and William E. Burr. Digital Identity Guidelines – Authentication and Lifecycle Management: NIST Special Publication 800-63B, June 2017.

[21] Kristen K. Greene, Melissa A. Gallagher, Brian C. Stanton, and Paul Y. Lee. I Can't Type That! P@$$w0rd Entry on Mobile Devices. In *Human Aspects of Information Security, Privacy, and Trust*, HAS '14, pages 160–171, Heraklion, Greece, June 2014. Springer.

[22] Marian Harbach, Emanuel von Zezschwitz, Andreas Fichtner, Alexander De Luca, and Matthew Smith. It's a Hard Lock Life: A Field Study of Smartphone (Un)Locking Behavior and Risk Perception. In *Symposium on Usable Privacy and Security*, SOUPS '14, pages 213–230, Menlo Park, California, USA, July 2014. USENIX.

[23] Troy Hunt. I've Just Launched "Pwned Passwords" V2 With Half a Billion Passwords for Download, February 2018. https://www.troyhunt.com/ive-just-launched-pwned-passwords-version-2/, as of April 1, 2020.

[24] Patrick Kelley, Saranga Kom, Michelle L. Mazurek, et al. Guess Again (and Again and Again): Measuring Password Strength by Simulating Password-Cracking Algorithms. In *IEEE Symposium on Security and Privacy*, SP '12, pages 523–537, San Jose, California, USA, May 2012. IEEE.

[25] Hyoungshick Kim and Jun Ho Huh. PIN Selection Policies: Are They Really Effective? *Computers & Security*, 31(4):484–496, June 2012.

[26] Fiona Lee. When the Going Gets Tough, Do the Tough Ask for Help? Help Seeking and Power Motivation in Organizations. *Organizational Behavior and Human Decision Processes*, 72(3):336–363, December 1997.

[27] Marte Løge, Markus Dürmuth, and Lillian Røstad. On User Choice for Android Unlock Patterns. In *European Workshop on Usable Security*, EuroUSEC '16, Darmstadt, Germany, July 2016. ISOC.

[28] William Melicher, Darya Kurilova, Sean M. Segreti, Pranshu Kalvani, Richard Shay, Blase Ur, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Michelle L. Mazurek. Usability and Security of Text Passwords on Mobile Devices. In *ACM Conference on Human Factors in Computing Systems*, CHI '16, pages 527–539, San Jose, CA, USA, May 2016. ACM.

[29] Saif M. Mohammad and Peter D. Turney. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

[30] Lily Hay Newman. Google's Making it Easier to Encrypt Even Cheap Android Phones, February 2019. https://www.wired.com/story/android-encryption-cheap-smartphones/, as of April 1, 2020.

[31] Lina Qiu, Alexander De Luca, Ildar Muslukhov, and Konstantin Beznosov. Towards Understanding the Link Between Age and Smartphone Authentication. In *ACM Conference on Human Factors in Computing Systems*, CHI '19, pages 163:1–163:10, Glasgow, Scotland, United Kingdom, May 2019. ACM.

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

[33] Thomas Reed. GrayKey iPhone Unlocker Poses Serious Security Concerns, March 2018. https://blog.malwarebytes.com/?p=22342, as of April 1, 2020.

[34] Karen Renaud and Melanie Volkamer. Exploring Mental Models Underlying PIN Management Strategies. In *World Congress on Internet Security*, WorldCIS '15, pages 19–21, Dublin, United Kingdom, October 2015. IEEE.

[35] Florian Schaub, Ruben Deyhle, and Michael Weber. Password Entry Usability and Shoulder Surfing Susceptibility on Different Smartphone Platforms. In *International Conference on Mobile and Ubiquitous Multimedia*, MUM '12, pages 13:1–13:10, Ulm, Germany, December 2012. ACM.

[36] Richard Shay, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Alain Forget, Saranga Komanduri, Michelle L. Mazurek, William Melicher, Sean M. Segreti, and Blase Ur. A Spoonful of Sugar?: The Impact of Guidance and Feedback on Password-Creation Behavior. In *ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 2903–2912, Seoul, Republic of Korea, April 2015. ACM.

[37] Emily Stark. The URLephant. In *USENIX Enigma Conference*, Enigma '19, Burlingame, California, USA, January 2019. USENIX.

[38] Joshua Sunshine, Serge Egelman, Hazim Almuhimedi, Neha Atri, and Lorrie Faith Cranor. Crying Wolf: An Empirical Study of SSL Warning Effectiveness. In *USENIX Security Symposium*, SSYM '09, pages 399–416, San Diego, California, USA, June 2009. USENIX.

[39] Sebastian Uellenbeck, Markus Dürmuth, Christopher Wolf, and Thorsten Holz. Quantifying the Security of Graphical Passwords: The Case of Android Unlock Patterns. In *ACM Conference on Computer and Communications Security*, CCS '13, pages 161–172, Berlin, Germany, November 2016. ACM.

[40] 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. Design and Evaluation of a Data-Driven Password Meter. In *ACM Conference on Human Factors in Computing Systems*, CHI '17, pages 3775–3786, Denver, Colorado, USA, May 2017. ACM.

[41] Blase Ur, Fumiko Noma, Jonathan Bees, Sean M. Segreti, Richard Shay, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. "I Added '!' at the End to Make It Secure": Observing Password Creation in the Lab. In *Symposium on Usable Privacy and Security*, SOUPS '15, pages 123–140, Ottawa, Ontario, Canada, July 2015. USENIX.

[42] Blase Ur, Sean M. Segreti, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Saranga Komanduri, Darya Kurilova, Michelle L. Mazurek, William Melicher, and Richard Shay. Measuring Real-World Accuracies and Biases in Modeling Password Guessability. In *USENIX Security Symposium*, SSYM '15, pages 463–481, Washington, District of Columbia, USA, August 2015. USENIX.

[43] U.S. Department of Homeland Security. The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research, August 2012. https://www.caida.org/publications/papers/2012/menlo_report_actual_formatted/, as of April 1, 2020.

[44] Emanuel von Zezschwitz, Alexander De Luca, and Heinrich Hussmann. Honey, I Shrunk the Keys: Influences of Mobile Devices on Password Composition and Authentication Performance. In *Nordic Conference on Human-Computer Interaction*, NordiCHI '14, pages 461–470, Helsinki, Finland, October 2014. ACM.

[45] Emanuel von Zezschwitz, Malin Eiband, Daniel Buschek, Sascha Oberhuber, Alexander De Luca, Florian Alt, and Heinrich Hussmann. On Quantifying the Effective Passsword Space of Grid-Based Unlock Gestures. In *Conference on Mobile and Ubiquitous Multimedia*, MUM '16, pages 201–212, Rovaniemi, Finland, December 2016. ACM.

[46] Ding Wang, Qianchen Gu, Xinyi Huang, and Ping Wang. Understanding Human-Chosen PINs: Characteristics, Distribution and Security. In *ACM Asia Conference on Computer and Communications Security*, ASIA CCS '17, pages 372–385, Abu Dhabi, United Arab Emirates, April 2017. ACM.

[47] Chris Welch. Apple Releases iOS 11.4.1 and Blocks Passcode Cracking Tools Used by Police, July 2018. https://www.theverge.com/2018/7/9/17549538/, as of April 1, 2020.

[48] Sonia Secher Wichmann. Self-Determination Theory: The Importance of Autonomy to Well-Being Across Cultures. *Journal of Humanistic Counseling*, 50(1):16–26, March 2011.

[49] Yulong Yang, Janne Lindqvist, and Antti Oulasvirta. Text Entry Method Affects Password Security. In *Learning from Authoritative Security Experiment Results*, LASER '14, pages 11–20, Arlington, Virginia, USA, October 2014. USENIX.

# Appendix

## A. Survey Instrument

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

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 blacklist.

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 easy 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: ________________________

## B. Demographics

TABLE IX

OVERALL DEMOGRAPHICS OF THE PARTICIPANTS. FOR THE SAKE OF CLARITY, WE GROUPED ANSWERS FOR *Non-Binary*, *Other*, AND *Prefer not to say* UNDER *Other*.

| | **Male** | | **Female** | | **Other** | | **Total** | |
|---|---|---|---|---|---|---|---|---|
| | **No.** | **%** | **No.** | **%** | **No.** | **%** | **No.** | **%** |
| **What is your age range?** | 619 | 51 % | 590 | 48 % | 11 | 1 % | 1220 | 100 % |
| 18–24 | 85 | 7 % | 76 | 6 % | 4 | 0 % | 165 | 13 % |
| 25–34 | 309 | 25 % | 267 | 22 % | 4 | 0 % | 580 | 47 % |
| 35–44 | 147 | 12 % | 145 | 12 % | 2 | 0 % | 294 | 24 % |
| 45–54 | 56 | 5 % | 63 | 5 % | 0 | 0 % | 119 | 10 % |
| 55–64 | 16 | 1 % | 35 | 3 % | 0 | 0 % | 51 | 4 % |
| 65–74 | 6 | 1 % | 4 | 0 % | 0 | 0 % | 10 | 1 % |
| Prefer not to say | 0 | 0 % | 0 | 0 % | 1 | 0 % | 1 | 0 % |
| **What is the highest degree or level of school you have completed?** | 619 | 51 % | 590 | 48 % | 11 | 1 % | 1220 | 100 % |
| Some High School | 2 | 0 % | 3 | 0 % | 0 | 0 % | 5 | 0 % |
| High School | 63 | 5 % | 52 | 4 % | 2 | 0 % | 117 | 10 % |
| Some College | 154 | 13 % | 116 | 10 % | 5 | 0 % | 275 | 23 % |
| Training | 23 | 2 % | 23 | 2 % | 0 | 0 % | 46 | 4 % |
| Associates | 63 | 5 % | 82 | 7 % | 1 | 0 % | 146 | 12 % |
| Bachelor’s | 236 | 19 % | 235 | 19 % | 2 | 0 % | 473 | 39 % |
| Master’s | 54 | 5 % | 66 | 5 % | 0 | 0 % | 120 | 9 % |
| Professional | 11 | 1 % | 4 | 0 % | 0 | 0 % | 15 | 1 % |
| Doctorate | 12 | 1 % | 9 | 1 % | 0 | 0 % | 21 | 2 % |
| Prefer not to say | 1 | 0 % | 0 | 0 % | 1 | 0 % | 2 | 0 % |
| **Which of the following best describes your educational background or job field?** | 619 | 51 % | 590 | 48 % | 11 | 1 % | 1220 | 100 % |
| Tech | 231 | 30 % | 83 | 7 % | 3 | 0 % | 317 | 26 % |
| No Tech | 368 | 19 % | 491 | 40 % | 7 | 1 % | 866 | 71 % |
| Prefer not to say | 20 | 2 % | 16 | 1 % | 1 | 0 % | 37 | 3 % |

## C. Device Usage

TABLE X

ANSWERS OF PARTICIPANTS REGARDING THEIR DEVICE USAGE. 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*.

| | **Male** | | **Female** | | **Other** | | **Total** | |
|---|---|---|---|---|---|---|---|---|
| | **No.** | **%** | **No.** | **%** | **No.** | **%** | **No.** | **%** |
| **Do you use any of the following biometrics to unlock your primary smartphone?** | 619 | 51 % | 590 | 48 % | 11 | 1 % | 1220 | 100 % |
| Fingerprint | 329 | 27 % | 310 | 25 % | 7 | 1 % | 646 | 53 % |
| Face | 95 | 8 % | 67 | 5 % | 0 | 0 % | 162 | 13 % |
| Iris | 24 | 2 % | 11 | 1 % | 0 | 0 % | 35 | 3 % |
| Other Biometric | 13 | 1 % | 15 | 1 % | 0 | 0 % | 28 | 2 % |
| No Biometric | 209 | 17 % | 206 | 17 % | 3 | 0 % | 418 | 34 % |
| No Smartphone | 1 | 0 % | 0 | 0 % | 0 | 0 % | 1 | 0 % |
| Prefer not to say | 19 | 2 % | 22 | 2 % | 1 | 0 % | 42 | 3 % |
| **How do you unlock your smartphone, if your biometric fails or when you reboot your primary smartphone?** | 390 | 51 % | 362 | 48 % | 7 | 1 % | 759 | 100 % |
| None | 2 | 0 % | 5 | 1 % | 0 | 0 % | 7 | 1 % |
| Pattern | 65 | 8 % | 48 | 6 % | 0 | 0 % | 113 | 15 % |
| 4-digit PIN | 183 | 24 % | 186 | 25 % | 3 | 0 % | 372 | 49 % |
| 6-digit PIN | 98 | 13 % | 99 | 13 % | 4 | 1 % | 201 | 26 % |
| PIN of other length | 12 | 2 % | 10 | 1 % | 0 | 0 % | 22 | 3 % |
| Alphanumeric | 21 | 3 % | 11 | 2 % | 0 | 0 % | 32 | 4 % |
| Other method | 6 | 1 % | 2 | 0 % | 0 | 0 % | 8 | 1 % |
| No smartphone | 0 | 0 % | 0 | 0 % | 0 | 0 % | 0 | 0 % |
| Prefer not to say | 3 | 0 % | 1 | 0 % | 0 | 0 % | 4 | 1 % |
| **What screen lock do you use to unlock your primary smartphone?** | 229 | 50 % | 228 | 50 % | 4 | 0 % | 461 | 100 % |
| None | 58 | 13 % | 82 | 18 % | 0 | 0 % | 140 | 30 % |
| Pattern | 36 | 8 % | 23 | 5 % | 1 | 0 % | 60 | 13 % |
| 4-digit PIN | 83 | 18 % | 81 | 18 % | 1 | 0 % | 165 | 36 % |
| 6-digit PIN | 20 | 4 % | 17 | 4 % | 0 | 0 % | 37 | 8 % |
| PIN of other length | 6 | 1 % | 2 | 1 % | 0 | 0 % | 8 | 2 % |
| Alphanumeric | 6 | 1 % | 6 | 1 % | 0 | 0 % | 12 | 3 % |
| Other method | 7 | 2 % | 4 | 1 % | 0 | 0 % | 11 | 2 % |
| No smartphone | 0 | 0 % | 1 | 0 % | 0 | 0 % | 1 | 0 % |
| Prefer not to say | 13 | 3 % | 12 | 2 % | 2 | 1 % | 27 | 6 % |
| **What is the operating system of your primary smartphone?** | 619 | 51 % | 590 | 48 % | 11 | 1 % | 1220 | 100 % |
| Android | 382 | 31 % | 310 | 25 % | 6 | 1 % | 698 | 57 % |
| iOS | 232 | 19 % | 270 | 22 % | 4 | 1 % | 506 | 42 % |
| Other | 1 | 0 % | 4 | 0 % | 0 | 0 % | 5 | 0 % |
| No smartphone | 0 | 0 % | 0 | 0 % | 0 | 0 % | 0 | 0 % |
| Prefer not to say | 4 | 1 % | 6 | 1 % | 1 | 0 % | 11 | 1 % |

## D. Feelings and Sentiments

TABLE XI

As part of our questionnaire, we asked participants for 3 feelings about the blacklist warning. We coded and analyzed these feelings from a sample of 130 participants that encountered a blacklist. We also included 21 participants that only imagined hitting a blacklist. Below, 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 $\kappa = 0.91$. 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.*"

| **Code Name** | **Frequency** | **Sample from the Study** | **Sentiment** |
|---:|:---:|:---|:---|
| Annoyance | 92 | "Annoyed by this message." | Negative |
| Frustrated | 45 | "This message frustrates me." | Negative |
| Worried | 41 | "I am worried about my PIN's security." | Negative |
| Indifference | 34 | "Don't care about this message." | Negative |
| Surprised | 32 | "Surprised to see this message." | Neutral |
| Fear | 32 | "Afraid of attackers." | Negative |
| Doubt | 32 | "I distrust the veracity of this message." | Negative |
| Thinking | 31 | "Thinking about my PIN's security." | Neutral |
| Acceptance | 27 | "I agree with this message." | Positive |
| Compelling | 26 | "Motivated to change my PIN." | Positive |
| Cautious | 26 | "Cautious about my PIN." | Positive |
| Confusion | 22 | "This message is confusing." | Negative |
| Happy | 19 | "Happy my PIN will be stronger." | Positive |
| Shame | 18 | "Ashamed my PIN wasn't strong." | Negative |
| Remember | 13 | "I might forget my PIN." | Neutral |
| Angry | 13 | "Angry this message appeared." | Negative |
| Curiosity | 12 | "I wonder why this message appeared." | Positive |
| Alert | 10 | "I'm now more aware." | Neutral |
| Safe | 8 | "Confident this PIN will be safe." | Positive |
| Sadness | 7 | "Sad this message appeared." | Negative |

## E. PIN Selection and Changing Strategies

TABLE XII

We coded and analyzed a sample of 200 PIN selection strategies. Below, we list the top 10 selection strategies. Two coders independently coded the data. The level of agreement among the coders, measured by Cohen's kappa, was $\kappa = 0.92$. Question: "*People use different strategies for choosing their PINs. Below, we will ask about your strategy. What was your strategy for choosing your PIN?*"

| **Code Name** | **Frequency** | **Description** | **Example PIN** | **Sample from the Study** |
|---:|:---:|:---|:---|:---|
| Date | 59 | Special date like anniversary, birthday, graduation day | 1987 / 112518 | "A date I won't forget." |
| Memorable | 37 | Memorability was the main concern | 2827 / 777888 | "A number easy to remember." |
| Pattern | 24 | Visualized a pattern on the PIN pad | 2580 / 137955 | "The numbers on how they appeared on the PIN pad." |
| Meaning | 20 | Personal meaning; Familiar or significant number | 6767 / 769339 | "I chose my favorite numbers and used them repeatedly." |
| Random | 14 | Randomly chosen digits | 4619 / 568421 | "Random numbers that do not repeat." |
| Reuse | 12 | Reused PIN from a different device/service | 0596 / 260771 | "The one I normally use." |
| Word | 9 | Textonyms; Converted a word to a number | 2539 / 567326 | "Dog name." |
| Simple | 9 | Simplistic, comfortable, easy | 0000 / 123987 | "To just chose an easy PIN." |
| System | 8 | User's established systematic strategy | 0433 / 041512 | "I used the numbers from the current time 04:33 PM." |
| Phone | 3 | (Partial) phone number | 1601 / 407437 | "I used the first four digits of a friend's phone number." |

TABLE XIII

We coded and analyzed a sample of 126 PIN changing strategies of participants that encountered a blacklist and in response changed their PIN. Below we list and explain our codes. Two coders independently coded the data. The level of agreement among the coders, measured by Cohen's kappa was $\kappa = 0.96$. Question: "*After receiving the warning message, please describe how or if your strategy changed when choosing your PIN.*"

| **Code Name** | **Frequency** | **Description** | **Use Case** | **Strategy** | **Sample from the Study** |
|:---:|:---:|:---|:---|:---|:---|
| Same | 28 | Same strategy for both | Selection<br>Change | Date<br>Date | "Birthday of relative."<br>"Chose another birthday." |
| Minor | 33 | Slight modification of strategy | Selection<br>Change | Meaning<br>Meaning++ | "It's one I remember, a number with personal significance."<br>"I changed one number in the sequence to get the app to accept it." |
| New | 65 | New strategy that is different | Selection<br>Change | Date<br>Phone | "I used my girlfriend's birthday."<br>"I changed my strategy to a memorable phone number's last 4 digits." |