The Personal Identification Chord: A Four Button Authentication System for Smartwatches

Ian Oakley[1], Jun Ho Huh[2], Junsung Cho[3], Geumhwan Cho[3], Rasel Islam[1] and Hyoungshick Kim[3]
[1]Department of Human Factors Engineering, UNIST, Republic of Korea
[2]Samsung Research, Samsung Electronics, Republic of Korea
[3]Department of Computer Science and Engineering, Sungkyunkwan University, Republic of Korea
ian.r.oakley@gmail.com,junho.huh@samsung.com,js.cho@skku.edu
geumhwan@skku.edu,rasel@unist.ac.kr,hyoung@skku.edu

ABSTRACT

Smartwatches support access to a wide range of private information but little is known about the security and usability of existing smartwatch screen lock mechanisms. Prior studies suggest that smartwatch authentication via standard techniques such as 4-digit PINs is challenging and error-prone. We conducted interviews to shed light on current practices, revealing that smartwatch users consider the ten-key keypad required for PIN entry to be hard to use due to its small button sizes. To address this issue, we propose the Personal Identification Chord (PIC), an authentication system based on a four-button chorded keypad that enables users to enter ten different inputs via taps to one or two larger buttons. Two studies assessing usability and security of our technique indicate PICs lead to increases in setup and (modestly) recall time, but can be entered accurately while maintaining high recall rates and may improve guessing entropy compared to PINs.

CCS CONCEPTS

• **Security and privacy** → **Authentication**; *Usability in security and privacy*;

KEYWORDS

Smartwatch screen lock, user authentication, personal identification number (PIN), personal identification chord (PIC)

ACM Reference Format:
Ian Oakley[1], Jun Ho Huh[2], Junsung Cho[3], Geumhwan Cho[3], Rasel Islam[1] and Hyoungshick Kim[3]. 2018. The Personal Identification Chord: A Four Button Authentication System for Smartwatches. In *ASIA CCS '18: 2018 ACM Asia Conference on Computer and Communications Security, June 4–8, 2018, Incheon, Republic of Korea.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3196494.3196555

1 INTRODUCTION

Smartwatches are rapidly developing into powerful standalone computing devices integrating technologies such as voice assistants, electronic SIM cards, phone connections, the capability to make financial transactions [18] and advanced fitness monitoring. They store, present and, through technologies such as Google's proximity-based Smart Lock [11], mediate access to a broad spectrum of personal information. However, recent literature suggests that the diminutive input spaces on smartwatches may dissuade users from securing their devices: usability studies of PIN entry report optimal error rates of between 7.5% [25] and 11% [35], which we argue is sufficient to represent a barrier to adoption. In addition, smartwatch PIN entry interfaces suffer from further disincentives: they typically need to be summoned with an explicit command and then occupy the entire surface of a smartwatch – first further adding to entry times, requiring users to pay careful attention to the smartwatch face immediately after donning their smartwatches, and then obstructing core device functions, such as viewing the time. In this way, authentication techniques such as LG's Knock Code [30] that are designed to be used without a Graphical User Interface (GUI), may be a better fit to the smartwatch form factor.

This paper argues that current smartwatch authentication techniques, such as PIN, are a relatively poor fit for the combination of

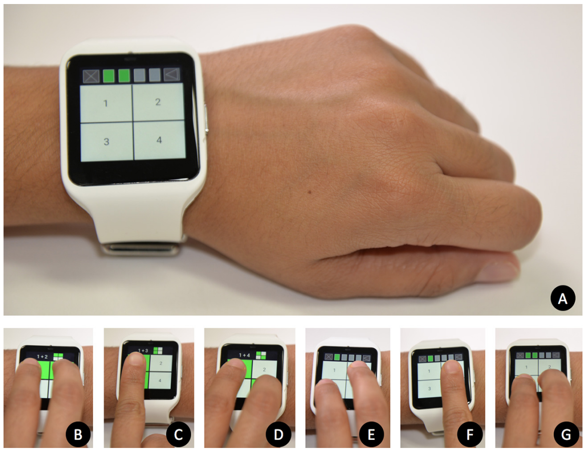


Figure 1: PIC Unlock Screen. Top image (A) shows a PIC unlock screen on a smartwatch running the interface from study 3. PICs are a chorded input system composed of four single-tap inputs achieved by selecting the keys labeled 1 to 4, and six dual-taps inputs achieved by selecting any pair of these keys as shown in figures B through G. Dual-taps can be entered with either two fingers (B, D, E, G) or a finger placed flat over two keys (C, E).

ASIA CCS '18, June 4–8, 2018, Incheon, Republic of Korea
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5576-6/18/06…$15.00
https://doi.org/10.1145/3196494.3196555

users' fat fingers [29] and the form factor's inherently small screens. PIN buttons are small and hard to hit, issues that may affect both performance in entry tasks [34], and impact the adoption rate of authentication systems [27]. This paper explores this issue in four ways. Firstly, to better understand those usability issues, we conducted an interview study (N = 10) with current watch owners. Secondly, to mitigate the concerns they raised, we designed the Personal Information Chord (PIC), a novel authentication interface for smartwatches based on four large buttons and chorded input. Thirdly, to understand input performance on smartwatches, we characterized and compared performance with standard PIN against PIC input in a simple single session keypress-level usability study (N = 21). Fourthly, to evaluate the security and usability of PIC and PIN as authentication mechanisms, we conducted a comprehensive two-day study in which participants were required to create and recall their own PINs/PICs (N = 120).

The results from the interview study indicate that smartwatch users are generally concerned with entering four digit PINs on small watch screens using small buttons, and prefer a GUI free interface for unlocking their watches. Data from the usability study indicate these concerns may be overstated: during standard input, PIN entry is rapid (0.75s) and accurate (1.34% errors). During a more challenging input condition with the GUI obscured, PIN input error rates increase to 9.05% suggesting that authentication in non-optimal input scenarios (e.g. while mobile, distracted) may be more problematic. Performance with PIC, designed to mitigate such problems, is modestly slower than PIN in standard input (by 0.13s), while achieving lower error rates compared to PIN in the more challenging GUI free input condition (errors are stable at between 5.0% and 5.5% in both GUI and GUI free conditions). Results from the final study flesh out these assertions. In terms of usability, PIC setup requires longer than PIN setup (by 12-34s) and, when using a standard UI, PIC entry is marginally slower than PIN entry (0.4-1.4s). Recall rates (97-100%) are high for both schemes. In terms of security, PICs provide a modest increase in partial guessing entropy over PINs. Furthermore, 50% of PIN participants used personal information (e.g., birthdays) to select memorable PINs, a behavior that the non-numeric structure of PIC precluded. Based on these outcomes, we suggest that while the costs of using PIC are increases in setup and, more modestly, recall time, the benefits it brings to non-optimal input settings (reduced error rates) and potential increases in resistance to guessing attacks makes it an interesting and viable alternative to PIN for smartwatches.

In sum, this paper makes the following contributions: 1) an interview study (N = 10) capturing usability issues with existing smartwatch screen lock GUIs; 2) PICs, a novel authentication input technique explicitly designed to include large targets that better fit the smartwatch form factor; 3) a keypress level study (N = 21) of PINs and PICs; 4) a multi-session PIN/PIC recall study (N = 120) capturing metrics such as setup time, recall time, recall rate and reporting an in-depth security analysis of the generated PINs/PICs with respect to guessing entropy.

2 FIRST STUDY: IDENTIFYING DESIGN REQUIREMENTS

We conducted an interview study with existing smartwatch users focusing on screen lock experiences, in order to derive design requirements in this space.

2.1 Methodology

The semi-structured interview was conducted at a large IT company with ten current smartwatch users. It involved an exercise: if screen lock was not enabled, participants enabled it. They were then asked to remove their smartwatches, wait a few minutes, then put them back on, and authenticate – smartwatches typically only require authentication when donned. This activity ensured some baseline experience with smartwatch lock screens. We also asked them to note down their daily routines, as relating to taking off and putting on their smartwatches. The remainder of the interview covered smartwatch locking behaviors and perceptions, and participants' opinions about the screen lock mechanisms on their smartwatches in terms of usability and how they could be improved. Participants were compensated with retail vouchers worth about ten USD.

2.2 Results

Participants were aged between 29 and 37, with a mean of 33, all Asian and right handed and from diverse backgrounds: arts, engineering, computers and architecture. All wore their watch on their left wrists. Six participants used Samsung Gear S2, one participant Samsung Gear S3, and three participants the Apple Watch Series 1 (two were using the smaller Apple Watch). Only two participants (both with Apple Watch) were using 4-digit PINs to lock their screens. The remaining Apple Watch participant had previously used it (for about a year) but deactivated the feature because it locked the screen too frequently. PIN was the only screen lock option available on all smartwatches in the study. Participants also reported donning their smartwatches between one and four times a day. One participant reported that he or she puts on the watch once in the morning around 7 before going to work, and takes it off just once at night around 9 when he or she returns home. This routine of putting on the watch just once per day was the longest (approximately 24 hour) interval between donning routines reported in the study.

Two researchers used *open coding* to analyze interview responses separately then discussed and reviewed all codes until they reached consensus. There were a number of very clear usability responses: eight out of ten participants raised issues with "small button" sizes, and three participants echoed this with concerns about "small screen" sizes. The prominence of these codes leads to our first design requirement:

> **Requirement 1:** Smartwatch screen lock user interfaces should be designed with buttons larger than those of existing PIN keypads.

In terms of usability enhancements, bio-metric approaches were frequently suggested: fingerprint (four participants) or vein (two). Two participants mentioned that they would like to unlock without looking at their watch screens (eyes-free [7]), and two suggested that unlocking should not require a GUI (GUI-free). Motivations for these recommendations stemmed from a wish for greater convenience and ease during PIN entry: a desire to unlock while walking or otherwise mobile, and a sense that existing numerical keypads are a poor fit for small watch screens. We derived our second design requirement from these comments:

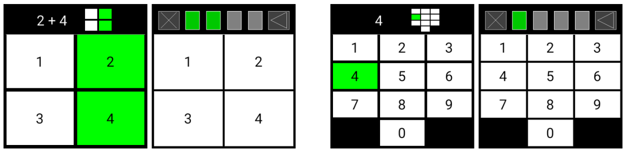
> **Requirement 2:** Smartwatch screen locks should be designed to work under eyes-free and GUI-free conditions.

3 PERSONAL INFORMATION CHORD SCHEME DESIGN

To fulfill these requirements, we propose the Personal Identification Chord (or PIC), a novel input scheme for authentication on smartwatches. It is designed to achieve three objectives. Firstly, to reduce the number of on-screen targets presented to users. In this way, the targets can be enlarged (Requirement 1), facilitating selection. Secondly, to be used without a dedicated GUI to remove the need for a UI event to summon an authentication input screen that then obscures core content such as a watch-face (Requirement 2). This approach mirrors techniques such as LG's Knock Code for smartphones [30]. This is a specialized input technique designed to support rapid unlocking when the device deactivated or in sleep mode. Thirdly, to maintain the password space and resistance to brute force attack of a standard PIN by enabling multiple simultaneous selections, as in a chorded keyboard. PIC realizes these objectives by presenting four equally sized targets labelled one, two, three and four in the space reserved for the 10 buttons in a numeric keypad. In addition to *single-tap* selections of the four individual keys, PIC enables all six possible *dual-tap* selections of pairs of keys for a total of ten separate input symbols. The set of dual-taps is illustrated on a smartwatch in Figure 1. As with a standard PIN, a PIC is composed of a sequence of four (single or dual) taps, yielding 10,000 possible options. The ultimate design goal for PICs is to encourage the use of authentication on small wearables such as smartwatches by making systems more approachable, reliable, and efficient while maintaining security.

PICs were implemented on a Sony Smartwatch 3 using the Processing programming language. This device features a 30mm square multi-touch capable capacitive touch screen, natively enabling detection of two simultaneous taps. However, some dual-taps, specifically those requiring touches to a pair of vertically aligned keys (i.e., 1+3 and 2+4) are challenging to achieve with a pair of touches – for example, to align the index and middle finger vertically on the screen requires an awkward rotation of the wrist. To facilitate these inputs, we also incorporated the ability to select a pair of targets with a single touch that covers them both. For vertically aligned pairs of keys, this takes the form of a finger laid flat on the left or right side of the watch, as illustrated in Figure 1. Our PIC implementation also devotes the top 20% (6mm) of the screen to feedback and/or instructional interface elements, and facilitates accurate selection with small 0.6mm spaces between keys. The ultimate size of the four PIC keys is 157 pixels wide (14.7mm) by 125 pixels high (11.7mm).

To implement the touch sensing functionality required for PICs, we modified the Android kernel to capture and process the raw sensor data from the touch screen, an approach followed by numerous authors seeking to design novel touch interfaces [17, 33]. We then manually generated all touch events in the system. Specifically, we adapted the open source implementation introduced in Gil et al. [10]. The data provided by this approach is an intensity map of touch activations over the whole sensor area: a seven by seven grid for the Sony Smartwatch 3. Through continuous testing during development we determined an appropriate process to clean up the raw data and create a clear and accurate touch image. We used a threshold to exclude data of less than 30% of the maximum sensor reading and applied a log gamma correction to the remainder. We then processed each touch through a combination of flood-filling to identify separate touches followed by calculating image moments to identify the key properties of centroid, orientation and major/minor axis length (as in [33]). We excluded touches where we were unable to derive lengths for both major and minor axes and considered the temporally central point of each touch in order to generate selections. We classified touches as selecting the vertically aligned keys of 1+3 and 2+4 when they satisfied three conditions. Firstly, a centroid in the left (1+3) or right (1+4) portion of the screen. Secondly, a major axis length exceeding 53.6% of the touch screen, equivalent to covering 3.75 of the seven sensors or 16.2mm of real space. Finally, the angle of the touch had to be 20 degrees from vertical. These thresholds were derived through subjective experimentation. We processed all other touches based solely on their centroid. The system ran at approximately 80 samples per second, rapid enough to support fluid touch input and our empirical objectives.



(a) PIC (b) PIN

Figure 2: PIC Study UIs. Figures show conditions in the usability and practice (left) and authentication (right) studies. Usability study interfaces include textual and graphical instructions at the top of the smartwatch, while authentication study interfaces include a PIN/PIC entry feedback bar as well as back and clear buttons.

4 SECOND STUDY: INDIVIDUAL TAP USABILITY

We conducted an initial study to assess usability of the PIC taps on a smartwatch against the baseline of a standard numeric PIN and in an *GUI-free* setting in which graphical representations of the on-screen targets were not shown – the targets were present, just blacked out. This condition represents the GUI-free, and to a limited extent the eyes-free, input scenarios highlighted by our interview participants.

4.1 Methodology

The studied featured a repeated measures design with four conditions derived from the two binary variables: input-mode (PIN/PIC)

Figure 3: Time and error data from the second study. Bars show standard error.

Figure 4: Subjective data from the second study. Bars show standard error. BORG-CR10 scores range from 0 to 10; TLX from 0 to 20

and GUI (shown/hidden). The PIN condition was a standard 4 3 arrangement of a 0 9 numerical grid and was implemented in the same screen area, and featured the same inter-target spacing, as the PIC system: each target was 103 pixels wide (9.65mm) by 60 pixels high (5.63mm). The input-mode variable was fully balanced among participants, while the hidden condition for each input-mode was always presented immediately after the shown condition. This ensured participants practiced with each GUI before completing tasks without it. Each condition featured four blocks of 50 trials, with each block including five repetitions of the ten possible PIN items or PIC taps. Items in each block were presented in a random order and the first block of trials was treated as practice and discarded. In the shown conditions, participants were required to complete all trials correctly any erroneous trials were returned to the pool of incomplete trials and displayed again later in the block. In contrast, in the hidden conditions, participants were not required to complete error trials, ensuring that (regardless of the difficulty of the hidden input task) the study would be completed in a reasonable period of time.

Each trial followed a simple structure: participants tapped the screen to start, a fixation spot was displayed for 500ms, followed by the experimental instructions and (in the shown conditions) the graphical targets. Participants then made the requested selection by touching and releasing the screen. Figure 2 shows the experimental screen in both shown conditions it includes identical instructions in both textual and graphical forms at the top of the smartwatch in the area typically reserved for displaying PIN entry progress. Hidden conditions were identical to shown conditions, except that the screen area showing the PIN or PIC keys was completely blank a black rectangle. For each trial, we logged trial correctness and two measures of task time: preparation time referring to the period after the fixation spot disappeared until the first touch to the screen and; touch time referring to the duration the finger was in contact with the screen. Directly after each of the four conditions participants completed the NASA TLX [15] measure of subjective workload and the Borg CR10 [6] measure of perceived exertion in order to capture more qualitative aspects of performance.

4.1.1 Demographics. In total, 22 participants were recruited, but one failed to complete the study. Of the remaining, nine were female and they were aged between 20 and 30, with a mean of 22. All were recruited from the local student body and self-rated their experience with computers, touchscreens, and smartphones as high (4.71/5, 4.2/5, 5/5) while mean performance with smartwatches, smart-glasses, and other wearables was reported to be low (1.3/5). One participant owned a smartwatch and reported a high level of experience (5/5). Participants were compensated with approximately USD 10. In total, the study included 12,600 trials (21 participants 4 conditions 3 blocks 5 repetitions 10 items/taps).

4.2 Results

Figure 3 shows the preparation time, touch time and error data for each condition while the subjective data is shown in Figure 4. All data passed normality checks and were analyzed with two-way repeated measures ANOVA on the binary factors of input-mode and GUI. In the interests of brevity, Table 1 reports only significant results at $p < 0.05$. The two measures of time show significant but numerically modest (~50 100ms) variations in performance with small to medium effect sizes. The standard UI combination instantiated in the PIN/shown condition performs optimally. This conclusion is replicated in the error results, but with an interesting wrinkle the relatively strong interaction, supported by an inspection of the raw data, suggests that while the PIN condition resulted in a sharp rise in errors between the shown and hidden conditions, this effect was largely absent in the PIC condition participants were able to enter PIC taps in hidden conditions without impacting error rates. The subjective data showed similar patterns. Given strong general trend evident in the raw TLX scores, we opted to analyze only overall workload in addition to the Borg CR10 data. While the results again favor the PIN/shown combination, we note the absolute values suggest that participants did not struggle with any of the tasks in the study: TLX hovers around the midpoint of the scale, while Borg CR10 data is best characterized as light, indicating that participants could continue the task without difficulty. In this data, we once again observed close similarity between the shown and hidden PIC data. The stability of ratings between these conditions reinforces the idea that the participants entered PIC taps as easily without as with a GUI to cue them. We argue these results support the continued study of the PIC technique: performance is fast, accurate, resilient to challenging input conditions, and does not subjectively burden users.

5 THIRD STUDY: PIC AND PIN USABILITY

This section presents a user study comparing authentication performance of PINs and PICs. The user study was designed to specifically measure recall-rate, input entry, and input accuracy performances in a more realistic screen unlocking scenario. We considered two research questions while designing the user study. Firstly, how secure and readily recalled are PICs compared to PINs? Secondly, can we improve security of PICs by mandating the use of dual-taps? Answering these questions will provide a balanced assessment of the potential of the PIC technique.

5.1 Methodology

5.1.1 User study design. To investigate these questions, we evaluated the effectiveness of a standard PIN against three PIC policies (detailed in the following section), in a two session study conducted at two universities. Sessions were spread over two days and participants were screened for availability on both days. They were compensated with retail vouchers worth approximately USD 10 for each session. The study followed a between-groups design in which each participant completed study tasks for a single policy (experimental condition) in order to avoid practice and/or order effects. Before running the real user study, we conducted pilot studies with 35 participants to fix bugs, finalize the protocol, and address unclear instructions and descriptions. All of the three studies presented in this paper were IRB approved.

5.1.2 PIC policies. The four policies in the study are as follows:

PIN-original : This condition used the PIN input-mode from the first study, and enforced a password policy based that used on the Apple Watch, a popular wearable device. PINs were required to be four digits in length and a warning was presented if users attempted to create PINs with repeated digits or pairs, ascending numerals or common dates (e.g., birth years), lexical or geometrical codes.

PIC-free : This condition used the PIC input-mode from the first study and enforced parts of the PIN password policy. Years and lexical or geometric content were omitted due to a lack of digits or similarly configured keys.

PIC-dual : This condition used the PIC input-mode and PIC-free policy. In addition, users were required to compose their password using at least one dual-tap. If they attempted to enter a PIC without using a dual-tap, they were told it was invalid and required to choose another PIC. This policy resembles a commonly used password complexity policy that mandates the use of at least one special character.

PIC-dual-rand : This final condition implemented a policy similar to that of PIC-free , except that participants were required to use a specific dual-tap provided to them as part of the study instructions. We borrow this policy idea from SysPal [9] their findings show that mandating one random point can significantly improve pattern security with minimal compromise in recall-rate. The required dual-taps were equally sampled across all participants in this policy (i.e., an equal number of participants were assigned each dual-tap).

Table 1: Significant ANOVA results from the second study.

Measure	Comparison	Outcome		
Prep Time	Interaction	F(1,20) = 10.15	p = 0.005	$\eta^2_p = 0.34$
	GUI	F(1,20) = 22.58	p < 0.001	$\eta^2_p = 0.53$
Touch Time	Interaction	F(1,20) = 17.97	p < 0.001	$\eta^2_p = 0.47$
	GUI	F(1,20) = 17.01	p < 0.001	$\eta^2_p = 0.46$
Errors	Interaction	F(1,20) = 32.02	p < 0.001	$\eta^2_p = 0.60$
	GUI	F(1,20) = 16.23	p = 0.001	$\eta^2_p = 0.44$
Workload	Interaction	F(1,20) = 22.31	p < 0.001	$\eta^2_p = 0.53$
	Input-Mode	F(1,20) = 13.97	p = 0.001	$\eta^2_p = 0.41$
	GUI	F(1,20) = 19.48	p < 0.001	$\eta^2_p = 0.49$
Borg CR10	Interaction	F(1,20) = 22.83	p < 0.001	$\eta^2_p = 0.53$
	GUI	F(1,20) = 9.26	p = 0.006	$\eta^2_p = 0.32$

5.1.3 System. To evaluate the four policies in a realistic setting, we developed the application used in the second study into a system that resembles real-world smartwatch PIN screen lock setup and unlock GUIs. Using this application, we collected the participants' behavioral data to examine how they choose a PIN or PIC and used it to unlock the Sony smartwatch we provided. This application adjusted thresholds in classifier based on the usability study results, most importantly using a 30 threshold for angle and enabling single long horizontal touches to select dual-taps 1+2 and 3+4. It also included a standard PIN entry grid with four feedback boxes that were greyed out when no PIN/PIC items had been entered and turned successively green from left to right as PIN/PIC items were produced. These boxes could also be used to present textual PIN/PIC reminders. The top bar of the app also included a back button that deleted a single PIN/PIC item and a clear button that cleared all items. When four items were entered, the system automatically committed the PIN or PIC. The updated GUI is shown in Figure 2.

5.1.4 Procedure. This section provides details of the data collection procedures in the order participants completed them.

1. Practice: Participants first completed a five to ten minute training session based on the system described in study two and using the input-mode appropriate for their assigned policy condition. They first completed a shown session composed of five blocks of each possible PIN/PIC input with a standard GUI and then a similar GUI-hidden condition. In total, this session involved entering 100 PIN/PIC items.

2. PIN/PIC setup: Each participant was randomly assigned to one of the four policies and given appropriate instructions to create a PIN or PIC. For PIC-dual this entailed instructions to use at least one dual-tap. For PIC-dual-rand participants were given paper showing the specific dual-tap they were required to use. Borrowing from Uellenbeck et al.'s [31] virtual sweet method, the participants were also told there was a further voucher worth approximately USD 5 in the watch. In order to get the voucher, they would have to remember their PIN or PIC in a follow up session one day later, and also generate a PIN or PIC that prevented other participants from accessing their watch. The intention was to encourage the participants to generate PINs or PICs that are both easy to recall and secure. Participants then entered a PIN or PIC to set it up, and dealt with any policy warnings or failures by either clicking past them (for warnings) or restarting the setup process. Finally, they re-entered an identical PIN or PIC to confirm. Confirmation failures led to starting afresh. The dual-tap that participants in the

Table 2: Survey questions asked after each recall test. For Q2 and Q3, the seven-level Likert item format was very difficult, difficult, somewhat difficult, neutral, somewhat easy, easy, and very easy.

#	Question	Answers
Q1	Did you use an external storage (e.g., a sheet of paper) to write down your PIN?	yes/no
Q2	How difficult was it for you to enter your PIN?	Likert scale
Q3	How difficult was it for you to remember your PIN?	Likert scale
Q4	Did you use any special technique (e.g., use of birth dates or tap rhythms) to help you create and remember your PIN?	yes/no
Q5	If you answered Yes to Q4, what was the special technique that you used?	Open ended

PIC-dual-rand policy were required to use did not change at any point during this process.

3. PIN/PIC memorization: Each participant was asked to enter the correct PIN or PIC three times to help with memorization. If incorrect PINs or PICs were entered five times consecutively, the correct PIN or PIC was revealed again so that the participant would have another chance to memorize it.

4. Puzzle: Each participant was asked to complete a moderately challenging lexical and graphical puzzle, which takes about 3 minutes to complete.

5. Demographics questions: Each participant was asked demographic questions such as ethnicity, age, gender and level of education. We also asked questions about participants' previous experiences with smartwatches.

6. Recall: Each participant was asked to enter his or her chosen PIN or PIC within five attempts. A failure to do so terminated the study. To prevent the participants from cheating, we ensured they did not look at any external storage during this or any other recall stage.

7. Recall survey: Participants were asked to answer the survey questions listed in Table 2. Only those who correctly recalled their PIN or PIC were invited to the next stage.

8. Recall-hidden: Participants were asked to enter their PINs or PICs within five attempts in an hidden input condition identical to that used in the hidden condition in the second study.

9. Recall-hidden survey: Participants answered the survey questions listed in Table 2. Only those who correctly entered their PIN/PIC in stage 8 were invited to the next stage.

10. Day2-Recall: Steps 6 through 9 were repeated 24 hours later in the second recall test. If participants successfully completed this recall test, they were given the additional USD 5 sweet voucher.

The 24 hour break period in this study (between steps 9 and 10) was selected to reflect the real-world smartwatch unlocking frequencies captured in our initial interviews: the longest interval between two consecutive watch donning (unlocking) routines was 24 hours. The study structure also reflects the Atkinson-Shiffrin dual memory model [1]. This model postulates that human memories initially reside in a short-term memory for a limited time (20 to 30 seconds). Short-term memory has limited capacity and older items are wiped as new items enter. Further, rehearsing or recalling items while they are in the short-term memory causes the items to stay longer in the short-term memory. Based on Atkinson-Shiffrin memory model, the memorization tasks in stage 3 will help participants remember their selected PINs or PICs. The puzzle in stage 4 is intended to wipe out the short-term memory of selected PIN or PIC information. Subsequently, the participants were asked to complete two sessions of recall tests to check whether they can remember their PINs or PICs.

Table 3: Mean time (sec) taken to complete a single task, and error rate (%) in Study 3 practice sessions (: mean, : standard deviation).

Policy	Shown				Hidden			
	Task-Time		Error Rate		Task-Time		Error Rate	
PIN	0.78	0.10	1.79	2.14	0.91	0.12	11.57	8.10
PIC	0.94	0.14	4.49	4.44	0.90	0.11	3.93	4.06

5.1.5 User data collected. Throughout the study, we recorded the following information:

Selected PIN/PIC policy: For each participant, we recorded the selected PIN or PIC and the assigned policy.

Setup time: We measured the time it took participants to set up their PINs or PICs, starting from when they first saw the input screen and ending when they successfully met all policy requirements and confirmed their PINs or PICs.

Unlock attempts: For recall and recall-hidden tests on both days, we recorded the number of attempts each participant made in entering the selected PIN or PIC.

Unlock time: For recall and recall-hidden tests on both days, we measured the time it took each participant to complete an authentication attempt. We divided this into preparation time and entry time. Preparation time started when the unlock screen was displayed and finished on the first tap to the screen. Entry time started from the participant's first touch of the screen and ended when the participant either entered the correct PIN or PIC, or failed to enter the PIN or PIC within five attempts.

Recall rate: For all of the recall and recall-hidden tests, we recorded whether a correct PIN or PIC was entered for each attempt made.

Survey answers: We recorded the participants' responses to the survey questions in Table 2.

5.2 Results

5.2.1 Demographics. A total of 120 participants completed the first day sessions tests; all returned for the second day. 30 were assigned to each policy. System errors caused setup times to be lost for four participants (all from the PIC-free policy) and data for one participant's day two recall session was corrupt (in the PIN-original policy). All of the participants were Asian, in the 18 29 age group, with a mean of 23, and 52.5% were male. 83.3% had a high school diploma, 15% had a university degree, and 1.7% had a Master or Doctoral degree. We recruited individuals regardless of their ownership of or previous experiences with smartwatches. Six

Table 4: Mean time (sec) taken to set up a PIN or PIC, mean number of mismatched PIN/PIC confirmations and policy failures that occurred while setting up a PIN or PIC (: mean, : standard deviation). Only a single participant received a policy warning (PIN policy).

Policy	Setup time		Mismatched		Policy Fail	
PIN-original	11.10	9.15	0.03	0.18	N/A	N/A
PIC-free	23.14	14.65	0.06	0.25	N/A	N/A
PIC-dual	45.00	34.10	0.26	0.58	0.00	0.00
PIC-dual-rand	44.78	40.17	0.20	0.55	0.16	0.74

participants owned a smartwatch but none of them reported using a screen lock.

5.2.2 Practice. We opted to analyze all trials in the practice condition to better characterize the usability of PICs when users are first exposed to them. To minimize the number of tests, we first calculated task-time as the sum of preparation and touch time, then removed a single outlier participant (more than three SDs from the mean). Error data were analyzed as in the second study. All data from this session are shown in Table 3. A two-way mixed methods ANOVA on task-time indicated all effects were significant: the interaction (F(1,1) = 71.4, p < 0.001, ^2_p = 0.38), input-mode (F(1,117) = 11.1, p < 0.001, ^2_p = 0.38) and GUI (F(1,117) = 26.76, p < 0.001, ^2_p = 0.09). The strong effect in the interaction indicates that while PICs were slower to enter in the *shown* condition, this difference was absent in the *hidden* condition. Error data failed normality tests, but showed a significant main effect of GUI using CI corrected Kruskal-Wallis (^{2}(1) = 6.2, p = 0.01). As PIC error data are flat across the GUI condition, we attribute this effect to the steep increase in errors in the PIN condition. These findings mirror those from the second study and serve to reinforce the idea that participants were capable users of the PIC technique from the outset. Despite receiving less training, the participants performed well (error rates are slightly lower than in the second study) including in challenging input conditions, such as the *hidden* GUI condition, where traditional input techniques, such as PIN, suffered a sharp uptick in error rate.

5.2.3 Setup time. As shown in Table 4, the mean time taken to set up a PIN or PIC varied considerably: from a mean of 11.1 seconds with PIN-original to 45 seconds with PIC-dual. The main effect of these differences was significant with a Kruskal-Wallis test (^{2}(3) = 42.23, p < 0.001). Corrected *post-hoc* tests showed that the PIN-original policy led to significant lower setup times than all other policies: PIC-free (^{2}(1) = 18.95, p < 0.006), PIC-dual (^{2}(1) = 27.39, p < 0.006), and PIC-dual-rand (^{2}(1) = 26.47, p < 0.006).

5.2.4 Recall rate, time, and attempts. We calculated and analyzed the proportion of participants who successfully recalled their PINs or PICs in the two recall tests and two recall-hidden tests to compare the recall effects of the four policies. A single participant in the PIC-dual-rand policy failed the recall test on day 2. There was no statistically significant difference in any of the polices (all p = 1:0, corrected FET).

Table 5: Mean preparation and entry time (sec) taken to complete authentication across the four policies (: mean, : standard deviation).

Policy	First Test				Second Test			
	Shown		Hidden		Shown		Hidden	
	Prep.	Entry	Prep.	Entry	Prep.	Entry	Prep.	Entry
PIN-original	1.15	1.34	1.34	5.58	3.14	1.89	1.18	2.42
	1.07	0.52	0.87	13.11	8.21	2.02	0.29	1.30
PIC-free	1.09	1.75	1.25	2.32	1.6	2.24	1.07	2.50
	0.58	0.77	0.47	2.04	1.08	1.53	0.42	1.68
PIC-dual	1.2	2.81	1.51	2.08	1.65	2.63	1.16	2.2
	0.79	2.22	0.91	1.03	1.69	1.66	0.47	1.39
PIC-dual-rand	1.02	2.45	1.46	2.41	2.08	2.82	1.26	2.87
	0.58	2.42	0.99	3.12	2.08	2.01	0.6	2.88

Recall times for all sessions are presented in Table 5 while recall attempts are in Table 6. Recall preparation times were not significant while recall entry times for the shown condition in both first and second tests showed significant main effects with a Kruskal-Wallis test (^{2}(3) = 18.44, p < 0.001 and ^{2}(3) = 19.99, p < 0.001, respectively). Corrected *post-hoc* tests indicated that the PIN-original policy led to reduced recall entry times compared to all other policies on both days (all p < 0.05 or lower). A similar analysis of recall attempts showed significant main effects in first recall-hidden (^{2}(3) = 12.66, p = 0.005) and second recall (^{2}(3) = 9.37, p = 0.025) tests. Corrected *post-hoc* tests showed a single significant difference: in the first recall-hidden test, the PIC-dual-rand policy required significantly fewer attempts to authenticate than the PIN-original policy (^{2}(1) = 7.78, p = 0.03).

5.2.5 External storage usage. After completing each recall test, we asked the participants about the use of an external storage (see Q1 in Table 2). Just three participants reported using an external storage, noting down their PINs or PICs after the first session. We ensured that no one cheated during all recall tests though.

5.2.6 Input difficulty. Based on the participants' responses to Q2 in Table 2, we estimated PIN/PIC input difficulty across four different policies. Responses collected after day 2 recall-hidden session indicate that PIC-dual and PIC-free may be easier to enter than PINs when GUIs are hidden - see Figure 5.

Table 6: Mean number of authentication attempts made across the four policies (: mean, : standard deviation).

Policy	First Test		Second Test	
	Shown	hidden	Shown	hidden
PIN-original	1.00	1.33	1.00	1.10
	0.00	0.80	0.00	0.31
PIC-free	1.03	1.07	1.00	1.07
	0.18	0.25	0.00	0.25
PIC-dual	1.07	1.03	1.00	1.03
	0.25	0.18	0.00	0.18
PIC-dual-rand	1.07	1.00	1.10	1.10
	0.36	0.00	0.31	0.31

Figure 5: Second recall-hidden test input difficulty.

Table 7: Usage frequency of the four tap groups used across all PIC policies. For the single group, the counted frequencies were divided by half for normalization purposes.

Category	PIC-free		PIC-dual		PIC-dual-rand	
Single	24	(25.00%)	28.5	(31.15%)	23	(23.71%)
Vertical	17	(17.71%)	15	(16.39%)	21	(21.65%)
Horizontal	26	(27.08%)	14	(15.03%)	21	(21.65%)
Diagonal	29	(30.21%)	34	(37.16%)	32	(32.99%)

6 PIC AND PIN SECURITY

This section presents security results for PINs and PICs, including guessing entropy and per tap or item usage frequencies.

6.1 PIC taps and PIN items used

Biases in PIC tap and PIN item selection could weaken their security. To study their security implications, we analyzed the usage frequencies of each of the ten PIC taps and PIN items. Figure 6 shows the usage ratios of all ten taps and items across the four policies arranged in descending (usage frequency) order. Overall, the usage frequencies of the taps used in PIC-free were more evenly distributed than those in other policies. We also analyzed the usage frequencies of the start/end taps and items (see Figure 8 and 9 in Appendix A). Appendix A shows the most frequently used start taps in PIC-dual and PIC-dual-rand were `2+3' (26.67%) while `3' was the most frequently used in PIC-free (23.33%). In PIN-original , `0' was the most frequently used item and `5' and `6' have never been used as a start tap. The usage frequencies of end taps and items seem evenly distributed except for PIC-dual. The most frequently used end tap in PIC-dual is `4' (26.67%). Interestingly, the most frequently used end item for PIN-original was `0' which is same for the case of the start PIN item. To better understand the characteristics of tap selection in PICs, we also categorized the PIC taps into four groups and normalized the proportions. They were: Single = {1, 2, 3, 4}, vertical = {1+3, 2+4}, horizontal = {1+2, 3+4}, and diagonal = {1+4, 2+3}. Usages frequencies are shown in Table 7. Unexpectedly, the diagonal taps were most popularly used in all PIC policies (30.21 37.16%). Figure 7 shows the frequency ratio of each tap group being used in each of the four PIC positions (indexes). These graphs indicate that diagonal taps are used more often in the first and second positions for all PIC policies.

6.2 Repeated use of PIN items and PIC taps

One security concern with PICs is that users could repeatedly use one or two taps to create chords that are easy to remember and enter. For example, a PIC consisting of 2+3, 2+3, 1, 2 re-uses 2+3. If such selection behaviors are common, an attacker could try to create rules based on repetitive tap selection patterns, and perform smarter guessing attacks on PICs. To measure the severity of such attacks, we counted the number of PICs that consist of a tap used twice or more (see Table 8). Unexpectedly, all PIC policies, PIC-free (26.67%), PIC-dual (36.67%), and PIC-dual-rand (30%), contained smaller percentages of PICs consisting of a repeating tap compared to PINs (46.67%). We surmise that PICs may be more robust against rule-based attacks that guess using repeating taps.

To further analyze the patterns of repeated taps in PIC policies, we classified taps into the categories of single (S) and dual (D) tap. Table 9 shows the results. There were 16 PIC-free passwords (53.33%) including at least 3 consecutively repeating taps (i.e., DDDD, SSSS, DDDS, DSSS, SSSD and SDDD patterns), which is substantially greater than the 6 PIC-dual passwords (20%) and the 12 PIC-dual-rand passwords (40%). This implies that PIC-free users more frequently used the same finger(s) to select taps in their

Table 8: The number of PINs and PICs that contain a tap used twice or more across the four policies. X represents a tap or item that is used more than twice in a given PIC/PIN, and ? represents any tap or item. Any represents the total number PINs or PICs that contain an X without recounting PINs/PICs that belong to multiple X patterns.

Pattern	PIN-original		PIC-free		PIC-dual		PIC-dual-rand	
XX??	3	(10.00%)	1	(3.33%)	0	(0.00%)	2	(6.67%)
X?X?	2	(6.67%)	4	(13.33%)	3	(10.00%)	0	(0.00%)
X??X	3	(10.00%)	5	(16.67%)	3	(13.33%)	4	(13.33%)
?XX?	3	(10.00%)	1	(3.33%)	3	(13.33%)	1	(3.33%)
?X?X	3	(10.00%)	1	(3.33%)	2	(6.67%)	0	(0.00%)
??XX	2	(6.67%)	2	(6.67%)	1	(3.33%)	2	(6.67%)
Any	14	(46.67%)	8	(26.67%)	11	(36.67%)	9	(30.00%)

Table 9: The repeated usage patterns of PIC taps with the single (S) and dual (D) tap categories.

Pattern	PIC-free		PIC-dual		PIC-dual-rand	
DDDD	6	(20.00%)	1	(3.33%)	3	(10.00%)
SSSS	1	(3.33%)	0	(0.00%)	0	(0.00%)
DDDS	3	(10.00%)	2	(6.67%)	3	(10.00%)
DSSS	2	(6.67%)	2	(6.67%)	0	(0.00%)
SSSD	1	(3.33%)	1	(3.33%)	2	(6.67%)
SDDD	3	(10.00%)	0	(0.00%)	4	(13.33%)
DDSD	0	(0.00%)	1	(3.33%)	3	(10.00%)
DSDD	0	(0.00%)	2	(6.67%)	2	(6.67%)
SSDS	1	(3.33%)	0	(0.00%)	1	(3.33%)
SDSS	0	(0.00%)	1	(3.33%)	1	(3.33%)
DDSS	2	(6.67%)	5	(16.67%)	3	(10.00%)
SSDD	5	(16.67%)	1	(3.33%)	2	(6.67%)
DSSD	3	(10.00%)	1	(3.33%)	2	(6.67%)
SDDS	2	(6.67%)	4	(13.33%)	0	(0.00%)
DSDS	1	(3.33%)	5	(16.67%)	2	(6.67%)
SDSD	0	(0.00%)	4	(13.33%)	2	(6.67%)

(a) PIN-original (b) PIC-free (c) PIC-dual (d) PIC-dual-rand

Figure 6: Ratio of each PIC tap and PIN item used, sorted in a descending order of usage ratio.

(a) 1st (b) 2nd (c) 3rd (d) 4th

Figure 7: Ratio of each tap group (single, vertical, horizontal, and diagonal) used in four PIC positions (indexes).

Table 10: Comparison of bits of information with across all policies.

Policy				
	0.1	0.4	0.7	1.0
PIN-original	10.24	11.27	11.86	12.21
PIC-free	10.45	11.45	12.00	12.33
PIC-dual	10.03	11.16	11.80	12.19
PIC-dual-rand	10.01	11.19	11.84	12.21

PICs (because switching between `S' and `D' types may require switching between one- and two-finger inputs).

6.3 Guessing entropy

To compare the robustness of the four policies against guessing attacks, we calculated partial guessing entropy estimates [3] because some attackers may only be interested in stealing just a fraction of an entire password set. This is a popular technique for estimating the average number of trials needed to successfully guess a fraction ( ) of an entire password set. Because our collected samples of PICs and PINs only represent a small portion of the theoretically possible password space, we employed the 2-gram Markov model [23] to estimate the occurrence likelihood of every possible PIC or PIN. To cover rare n-gram cases, we used the Laplace smoothing: the frequency of each n-gram is incremented by one.

For more intuitive comparison of entropy estimates, entropy estimates can be represented in bits of information. The converted results are shown in Table 10. Overall, PIC-free showed the highest partial guessing entropy estimates in bits of information. Contrary to our expectations, mandating the use of dual-taps (PIC-dual and PIC-dual-rand) did not increase guessing entropy estimates, achieving lower entropy estimates compared to PIC-free . There is a impactful difference between PIC-free and other three policies. As increases, the differences between PIN-original and PIC-free decreases slightly but remains considerable. These results indicate that PIC-free PICs are more robust against guessing attacks compared to PINs even when is large.

6.4 Remembrance techniques

After PIN or PIC setup, the participants were asked about the use of special techniques to create their PINs (Q3 of Table 2). Remembrance techniques were widely used, but varied considerably in type PINs tended to be numerical and PICs spatial or rhythmic. We grouped use of special dates (e.g., birthday), student IDs, phone numbers, and SAT scores into a category of personal information, data that can be exploited to perform more effective guessing attacks (e.g., [5]).

We analyzed the use of techniques in this category across all policies revealing that they were significantly more common with PIN than any PIC policy: 15 out of 30 PIN participants used personal information to create PINs compared to just 0, 1, and 2 in policies PIC-free , PIC-dual , and PIC-dual-rand , respectively. These differences were all significant ($p < 0.05$, FET with CI adjustment). We argue these differences are due to the compound dual-tap labels (e.g., 1+2 or 2+3) making it difficult for PIC participants to integrate semantically meaningful numerical content into their PICs. Inherently, PIC's four buttons prevent use of personal info (e.g. dates, IDs) in most cases, and we argue this enhances security against informed guessing attacks using personal information [5].

However, PICs are rich enough to support a range of novel remembrance techniques: shapes (e.g., hourglass, symmetrical shape), tapping rhythms, and ease of transition between touch poses. 9 participants of PIC-free , 10 participants of PIC-dual , and 10 participants of PIC-dual-rand used those techniques in the PIC policies, respectively.

6.5 2-gram tapping sequence frequencies

To investigate how the PIC remembrance techniques mentioned above may be exploited to perform guessing attacks on PICs, we analyzed the frequencies of all possible 2-gram tapping sequences (see Appendix B). In general, the distribution of 2-gram tapping sequence probabilities were similar between all four policies. From PIC-free , the top four frequently used 2-gram sequences were 1, 2, 3, 2, 1+2, 1+4, and 1+3, 2+4 (all with probability of 0.24). This indicates that the use of a single-tap was often followed by another single-tap, and the use of a dual-tap was often followed by another dual-tap. Such characteristics may be exploited to perform an informed guessing attack but we would need to collect more PIC data in the future to firmly establish this. In contrast, with PIC-dual and PIC-dual-rand , we noticed higher probabilities of 2-gram sequences consisting of both single- and dual-tap. As for PIN-original , 1, 0, and 8, 1 were the top two frequently used 2-gram sequences. In general, digit 1 was popularly used as the second digit.

7 DISCUSSION

7.1 PIN performance

PIN performed well in all studies. In the prolonged input tests in the second study, and in the practice sessions of the third study, performance in the shown conditions was rapid and accurate: task times and error rates of approximately 800ms and 2%. This highlights a contrast between users' doubts about their ability to acquire small targets on watches (as noted in the initial interviews) and their performance in this task. The low error rates are also striking when compared with prior studies. Hara [13], for example, reports error rates of 12.96% with 7mm targets that drop to 2.22% errors with 10mm targets. With the 9.65×5.63mm targets in our PIN system, we expected figures between these values. Possible explanations for the improved performance in our system are that wide, rectangular targets may be easier to select than expected, or the customized touch input system we used increased accuracy. The strong performance of shown PIN input was maintained in the third study: it led to rapid, accurate authentication sessions. Again this data is in contrast to the lower performance reported in recent work. In more purely usability focused work, Nyugen and Menon [25] report error rates in watch PIN authentication to be 7.5%, while Zhao et al. [35] indicate this may be as high as 11%. These differences highlight the importance of conducting studies of authentication that closely match real life use. Both these prior studies use multiple policies per participant in studies that involve just one session—performance in these atypical situations may not match performance in more realistic settings like the single policy multi-day method used in this paper.

PIN performance in the hidden conditions was split. Both usability tests highlight poor accuracy (errors of 8–11%), but these differences were not maintained in authentication sessions in the third study. This speaks to the importance of using multiple methods for evaluating authentication input performance—the prolonged input sessions and reliable performance data captured by usability studies can accurately highlight challenging input conditions, but the impact these variations have on real authentication performance (i.e., in short, sporadic input sessions) is not clear. However, we argue that the usability costs of PIN in the hidden conditions, as observed by the usability studies, will ultimately (over time) impact its viability in these settings.

7.2 PIC performance

PIC performance in the usability sessions was also good: quick (overall mean: 910ms) and reasonably accurate (4.8%). Although prior authors have examined both multiple simultaneous taps [21] and area-based input [26] on smartwatches, this study is the first to combine these techniques with a set of inputs requiring a range of different poses: single finger tap, double finger tap, and finger flat. The data from this study indicates that participants were able to switch between those poses with a very limited cost to input efficiency. PIC performance was also maintained between shown and hidden conditions: in contrast to the sharp increase in error rate with PIN, data are almost flat in both studies. High accuracy with PICs was also maintained in the authentication study: we note no significant differences in recall rates across the four recall tests. This data supports the claim that PIC is a reliable technique for authenticating on smartwatches, and that it may be a particularly useful approach in GUI-free situations, such as authenticating on a screen in power saving mode (as in LG's Knock Code [30]) or on a watch-face. The participants' responses to the question about input difficulty in recall tests also suggest that PICs may be easier to enter when GUIs are hidden (see Figure 5). We argue this data indicates that PICs satisfy the second requirement from the initial interview study: they work well in GUI-free conditions. While this conclusion may also apply to eyes-free conditions, we acknowledge that further studies would be required to formally establish this.

PIC led to weaker results in terms of the key usability metric of setup time—the three policy conditions in the third study logged 23-45 seconds vs 11 seconds for PIN. While this is likely due to a range of factors (e.g., novelty effects), this difference may be sufficient to dissuade users from adopting PIC in practice. Additionally, we note that PIC entry times were lower than PIN entry times (by up to 1.47 seconds) in the majority of recall tests, suggesting that PIC entry may modestly more difficult (e.g., laborious, challenging, or unfamiliar) than PIN entry. Future work should target reducing PIC setup and recall times.

7.3 Comparing PIC and PIN security and recall rate

The third study sought to establish the recall-rate and security of PICs compared to PINs. The results are positive. Our results failed to show statistically significant difference in recall rates between the two schemes. Furthermore, the PIC-free policy outperformed PINs in terms of guessing entropy, showing that PICs may be more robust to guessing attacks. This can be explained by the less skewed distribution of individual taps used in PIC-free (see Figure 6). In total, 46% of PINs contained repeating items, which can facilitate guessing attacks that exploit repetitive item patterns. In contrast, only 27% of PIC-free PICs contained repeating tap patterns (see Table 8). This observation is strengthened by additional evidence suggesting PINs may be easier to compromise: a significantly higher number of participants reported using personal information such as birth dates or ID numbers to compose their PINs. In contrast, PIC participants reported using rhythms and shapes as mnemonics. This evidence suggests that PICs may be stronger against rule-based guessing attacks that exploit personal information [5].

7.4 Implications of mandating dual-taps

The second research question in the third study sought to explore if PIC security can be improved with policies that mandate dual-taps. The answer is no; guessing entropy was highest for PIC-free . This was unexpected and highlights an incorrect assumption that dual-taps would be, proportionately, less frequently used than single taps. The ease with which participants were able to enter many of the dual-taps made them a frequent component of users' PICs. Building on the data from the third study, a more appropriate policy might be based on encouraging users to include the most infrequently used taps: 1+2, 2+4, 4 and 1+3. We note two of these taps are horizontal dual-taps, suggesting participants may have been reluctant to use this category in particular (see Table 7).

7.5 Limitations

A number of limitations impact this work. In the recall study, the sample size is sufficient only to make predictions about guessing entropy; larger samples would be required to fully establish it. Furthermore, the virtual sweet voucher may have distorted the participants' PIN and PIC selection behaviors. PINs and PICs collected in the study may be stronger than those used to protect real watches. However, since the same incentive applied to all policies, both PIN and PIC selections would have been affected in a similar way.

8 RELATED WORK

Textual passwords are still the most popularly implemented authentication method as they are easy to deploy and people are already familiar with them [4, 16]. Indeed, even if alternative schemes such as biometrics are used to secure a device, traditional passwords typically remain in place as a parallel authentication system and to maintain and manage biometric credentials, access rights and data (e.g, Apple's Touch ID). However, entering textual passwords on small virtual keyboards available on mobile devices can be difficult [12, 19, 24, 28, 32]. Zezschwitz et al. [32] found that passwords used on mobile devices are shorter and contain fewer symbols and uppercase letters. Melicher et al. [24] also found that creating passwords on mobile devices takes significantly longer, and is more error prone.

Only a few studies concentrated on the analysis of screen lock usability for smartwatches. Nguyen and Memon [25] evaluated the usability of popular locking mechanisms for smartwatches, and found that conventional PIN and patterns are more usable than the newer draw PIN and voice PIN methods. Zhao et al. [35] conducted a similar study, focusing on how watch sizes and layouts (square and circular) affect the usability of PINs and patterns. The most preferred method was patterns even though 75% of participants were concerned with their finger movements on small watch screens. They used mobile phones (and not smartwatches) and simulators to conduct that study. As far as we know, we are the first to fully implement a smartwatch-focused password entry system, and study its usability and security in a multi-day study.

Schaub et al. [28] evaluated the usability of textual password entry on mobile devices. Jakobsson and Akavipat used multi-word passphrases with auto-correction to improve the speed of password entry [19]. Numerous graphical patterns [9, 31], and biometric authentication schemes [2, 22] have been proposed for mobile devices. Despite many screen lock schemes becoming available, PIN is still popular: about 33.6% of mobile users use PINs [14]. Moreover, before activating biometric authentication, most mobile devices require users to first set up their dominant locks in the form of PINs, passwords, or patterns [8]. Hence, the overall security is often determined by the robustness of user chosen PINs. But we also know that most users choose easy-to-remember PINs that are vulnerable to dictionary attacks. Bonneau et al. [5] showed that many users use memorable dates (e.g., birth dates) as PINs, and an effective guessing attack would involve brute-forcing PINs with dates. The results presented in the Remembrance Techniques section confirm this, showing that half of the PIN participants used their personal information to create easily recallable PINs.

PIN complexity policies can help users choose stronger PINs. Kim et al. [20] studied the effectiveness of numerous PIN complexity policies, showing that enforcing a blacklist of popularly used PINs can help users choose more secure PINs that are also memorable. As mentioned in the PIC policies section, Apple uses a warning policy to help users avoid using weak PINs like 0000, 1111, or 1234. In contrast to existing literature on PIN complexity policies, however, our analyses of the PIC-dual and PIC-dual-rand policies showed that not all policies are effective in improving PIC security, and such policies must be designed carefully to address the PIC selection biases discovered in this paper.

9 CONCLUSIONS

The paper explores authentication on smartwatches. It first captures current opinions and behaviors via an interview study with current watch owners. Based on their concerns about the difficulty and inconvenience of authentication, we propose PICs, a novel authentication input based on chorded input on four large, easily targeted buttons. Two studies then assess the value of the PIC design. A keypress level usability study suggests that PIC performance is modestly slower and more error prone than PIN during standard input, but leads to fewer errors in a challenging, GUI-free input condition. A recall study shows both PICs and PINs achieve high recall rates and input accuracy, with PIC requiring additional time for setup and, more modestly, for recall. Security analyses based on an objective assessment of partial guessing entropy, and a subjective assessment of PIN/PIC selection strategies indicate that PICs may offer improved resistance to brute force attacks and to attacks based on knowledge about a watch owner. Future work will seek to establish the real-world impact of these variations by assessing input performance of PIN and PIC during real smartwatch use and also include comparisons with other common authentication techniques such as pattern lock. We will also investigate guessing attacks on PICs with a larger sample, capture additional data on PIC generation strategies and explore the susceptibility of PIC entry to other common forms of attack such as shoulder surfing.

ACKNOWLEDGMENT

This work was supported in part by Samsung Electronics Samsung Research and the ITRC (Information Technology Research Center) support program (IITP-2017-2015-0-00403). The authors would like to thank all the anonymous reviewers and Blase Ur for their valuable feedback.

REFERENCES

[1] Richard C Atkinson and Richard M Shiffrin. 1968. Human memory: A proposed system and its control processes. The psychology of learning and motivation 2 (1968), 89 195.

[2] Rasekhar Bhagavatula, Blase Ur, Kevin Iacovino, Su Mon Kywe, Lorrie Faith Cranor, and Marios Savvides. 2015. Biometric Authentication on iPhone and Android: Usability, Perceptions, and Influences on Adoption. In Proceedings of Network and Distributed Systems Symposium Workshop on Usable Security.

[3] Joseph Bonneau. 2012. The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords. In Proceedings of the 33rd IEEE Symposium on Security and Privacy. 538 552. https://doi.org/10.1109/SP.2012.49

[4] Joseph Bonneau, Cormac Herley, Paul C. van Oorschot, and Stajano Frank. 2012. The Quest to Replace Passwords: A Framework for Comparative Evaluation of Web Authentication Schemes. In Proceedings of the 33rd IEEE Symposium on Security and Privacy. 553 567. https://doi.org/10.1109/SP.2012.44

[5] Joseph Bonneau, Sören Preibusch, and Ross J. Anderson. 2012. A Birthday Present Every Eleven Wallets? The Security of Customer-Chosen Banking PINs. In Proceedings of the 16th International Conference on Financial Cryptography and Data Security. 25 40. https://doi.org/10.1007/978-3-642-32946-3_3

[6] Gunnar A Borg. 1982. Psychophysical bases of perceived exertion. Med. sci sports exerc. 14, 5 (1982), 377 381.

[7] Stephen Brewster, Joanna Lumsden, Marek Bell, Malcolm Hall, and Stuart Tasker. 2003. Multimodal `Eyes-free' Interaction Techniques for Wearable Devices. In Proceedings of the 21st Annual ACM Conference on Human Factors in Computing Systems (CHI '03). 473 480. https://doi.org/10.1145/642611.642694

[8] Ivan Cherapau, Ildar Muslukhov, Nalin Asanka, and Konstantin Beznosov. 2015. On the Impact of Touch ID on iPhone Passcodes. In Proceedings of the 11st Symposium On Usable Privacy and Security (SOUPS '15). 257 276. https://www.usenix.org/conference/soups2015/proceedings/presentation/cherapau

[9] Geumhwan Cho, Jun Ho Huh, Junsung Cho, Seongyeol Oh, Youngbae Song, and Hyoungshick Kim. 2017. SysPal: System-Guided Pattern Locks for Android. In Proceedings of the 38th IEEE Symposium on Security and Privacy. 338 356. https://doi.org/10.1109/SP.2017.61

[10] Hyunjae Gil, DoYoung Lee, Seunggyu Im, and Ian Oakley. 2017. TriTap: Identifying Finger Touches on Smartwatches. In Proceedings of the 35th Annual ACM Conference on Human Factors in Computing Systems (CHI '17). 3879 3890. https://doi.org/10.1145/3025453.3025561

[11] Google. 2017. Google Smart Unlock. https://get.google.com/smartlock/. (2017). [Online; accessed 19-Sept-2017].

[12] Kristen K. Greene, Melissa A. Gallagher, Brian C. Stanton, and Paul Y. Lee. 2014. I Can't Type That! P@$$w0rd Entry on Mobile Devices. In Proceedings of the 2nd International Conference on Human Aspects of Information Security, Privacy, and Trust. 160 171. https://doi.org/10.1007/978-3-319-07620-1_15

[13] Kiyotaka Hara, Takeshi Umezawa, and Noritaka Osawa. 2015. Effect of Button Size and Location When Pointing with Index Finger on Smartwatch. Springer International Publishing, Cham, 165 174. https://doi.org/10.1007/978-3-319-20916-6_16

[14] Marian Harbach, Emanuel von Zezschwitz, Andreas Fichtner, Alexander De Luca, and Matthew Smith. 2014. It's a Hard Lock Life: A Field Study of Smartphone (Un)Locking Behavior and Risk Perception. In Proceedings of the 10th Symposium On Usable Privacy and Security (SOUPS 2014). 213 230. https://www.usenix.org/conference/soups2014/proceedings/presentation/harbach

[15] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In Human Mental Workload, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139 183. https://doi.org/10.1016/S0166-4115(08)62386-9

[16] Cormac Herley and Paul C. van Oorschot. 2012. A Research Agenda Acknowledging the Persistence of Passwords. IEEE Security & Privacy 10, 1 (2012), 28 36. https://doi.org/10.1109/MSP.2011.150

[17] Christian Holz, Senaka Buthpitiya, and Marius Knaust. 2015. Bodyprint: Biometric User Identification on Mobile Devices Using the Capacitive Touchscreen to Scan Body Parts. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15). 3011 3014. https://doi.org/10.1145/2702123.2702518

[18] Gabriel Jakobson and Steven Rueben. 2013. Commercial transactions via a wearable computer with a display. (Nov. 18 2013). US Patent App. 13/998,623.

[19] Markus Jakobsson and Ruj Akavipat. 2011. Rethinking passwords to adapt to constrained keyboards. (2011). http://www.markus-jakobsson.com/fastwords.pdf

[20] Hyoungshick Kim and Jun Ho Huh. 2012. PIN selection policies: Are they really effective? Computers & Security 31, 4 (2012), 484 496. https://doi.org/10.1016/j.cose.2012.02.003

[21] Benjamin Lafreniere, Carl Gutwin, Andy Cockburn, and Tovi Grossman. 2016. Faster Command Selection on Touchscreen Watches. In Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems (CHI '16). 4663 4674. https://doi.org/10.1145/2858036.2858166

[22] Alexander De Luca, Alina Hang, Emanuel von Zezschwitz, and Heinrich Hussmann. 2015. I Feel Like I'm Taking Selfies All Day!: Towards Understanding Biometric Authentication on Smartphones. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15). 1411 1414. https://doi.org/10.1145/2702123.2702141

[23] Jerry Ma, Weining Yang, Min Luo, and Ninghui Li. 2014. A Study of Probabilistic Password Models. In Proceedings of the 35th IEEE Symposium on Security and Privacy. 689 704. https://doi.org/10.1109/SP.2014.50

[24] William Melicher, Darya Kurilova, Sean M. Segreti, Pranshu Kalvani, Richard Shay, Blase Ur, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Michelle L. Mazurek. 2016. Usability and Security of Text Passwords on Mobile Devices. In Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems (CHI '16). 527 539. https://doi.org/10.1145/2858036.2858384

[25] Toan Nguyen and Nasir Memon. 2017. Smartwatches Locking Methods: A Comparative Study. In Proceedings of the 13rd Symposium On Usable Privacy and Security (SOUPS '17). Santa Clara, CA. https://www.usenix.org/conference/soups2017/workshop-program/way2017/nguyen

[26] Ian Oakley, Carina Lindahl, Khanh Le, DoYoung Lee, and MD. Rasel Islam. 2016. The Flat Finger: Exploring Area Touches on Smartwatches. In Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems (CHI '16). 4238 4249. https://doi.org/10.1145/2858036.2858179

[27] M. A. Sasse, S. Brostoff, and D. Weirich. 2001. Transforming the `Weakest Link' a Human/Computer Interaction Approach to Usable and Effective Security. BT Technology Journal 19 (July 2001), 122 131. Issue 3. https://doi.org/10.1023/A:1011902718709

[28] Florian Schaub, Ruben Deyhle, and Michael Weber. 2012. Password entry usability and shoulder surfing susceptibility on different smartphone platforms. In Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia (MUM '12). 13. https://doi.org/10.1145/2406367.2406384

[29] Katie A. Siek, Yvonne Rogers, and Kay H. Connelly. 2005. Fat Finger Worries: How Older and Younger Users Physically Interact with PDAs. In Proceedings of the IFIP TC13 International Conference on Human-Computer Interaction (INTERACT '05). 267 280. https://doi.org/10.1007/11555261_24

[30] Ben Sin. 2017. The Galaxy S8 And Pixel Should Copy LG's Knock Code. https://www.forbes.com/sites/bensin/2017/03/02/the-galaxy-s8-and-pixel-should-copy-lgs-knock-code. (2017). [Online; accessed 19-Sept-2017].

[31] Sebastian Uellenbeck, Markus Dürmuth, Christopher Wolf, and Thorsten Holz. 2013. Quantifying the security of graphical passwords: the case of android unlock patterns. In Proceedings of the 20th ACM Conference on Computer and Communications Security. 161 172. https://doi.org/10.1145/2508859.2516700

[32] Emanuel von Zezschwitz, Alexander De Luca, and Heinrich Hussmann. 2014. Honey, I shrunk the keys: influences of mobile devices on password composition and authentication performance. In Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational. 461 470. https://doi.org/10.1145/2639189.2639218

[33] Robert Xiao, Julia Schwarz, and Chris Harrison. 2015. Estimating 3D Finger Angle on Commodity Touchscreens. In Proceedings of the 2015 International Conference on Interactive Tabletops & Surfaces (ITS '15). 47 50. https://doi.org/10.1145/2817721.2817737

[34] Chun Yu, Hongyi Wen, Wei Xiong, Xiaojun Bi, and Yuanchun Shi. 2016. Investigating Effects of Post-Selection Feedback for Acquiring Ultra-Small Targets on Touchscreen. In Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems (CHI '16). 4699 4710. https://doi.org/10.1145/2858036.2858593

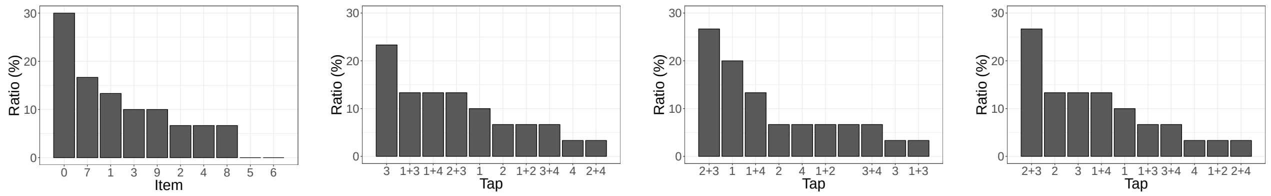
[35] Yue Zhao, Zhongtian Qiu, Yiqing Yang, Weiwei Li, and Mingming Fan. 2017. An Empirical Study of Touch-based Authentication Methods on Smartwatches. In Proceedings of the ACM International Symposium on Wearable Computers (ISWC '17). 122 125. https://doi.org/10.1145/3123021.3123049

A USAGE FREQUENCIES OF THE START/END PIC TAPS AND PIN ITEMS

Figures 8 and 9 show the usage frequencies of start/end taps/items.

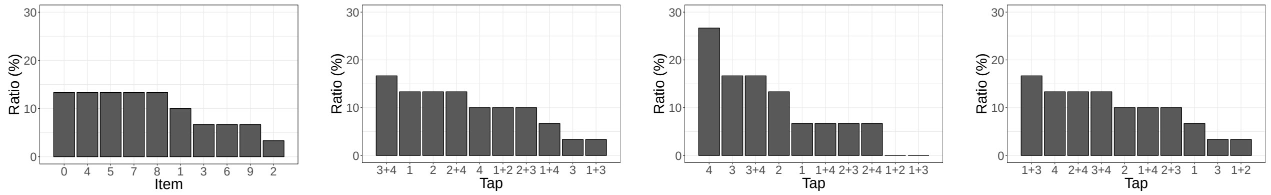
B 2-GRAM TAPPING SEQUENCES

Figure 10 shows the frequencies of all 2-gram tapping sequences.



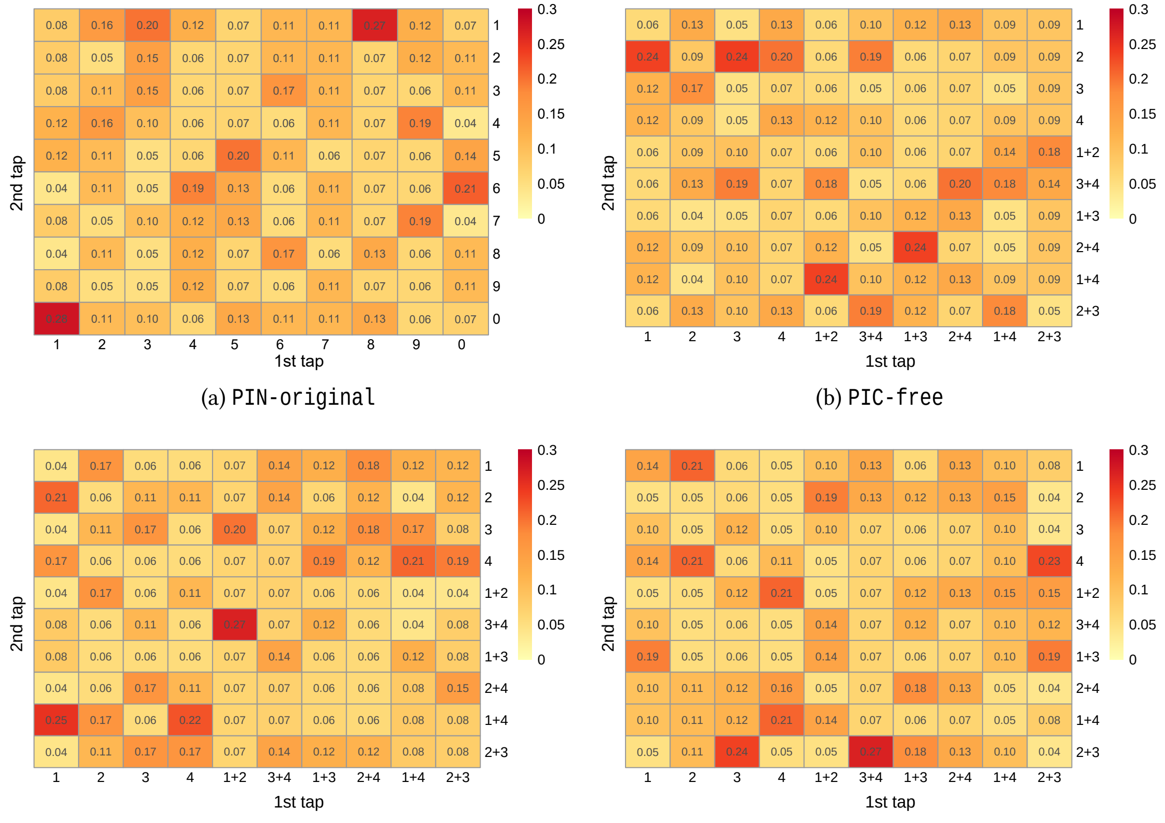
(a) PIN-original (b) PIC-free (c) PIC-dual (d) PIC-dual-rand

Figure 8: Ratio of the start PIC tap and PIN item used for each policy, sorted in a descending order of usage ratio.



(a) PIN-original (b) PIC-free (c) PIC-dual (d) PIC-dual-rand

Figure 9: Ratio of the end PIC tap and PIN item used for each policy, sorted in a descending order of usage ratio.



(a) PIN-original (b) PIC-free

(c) PIC-dual (d) PIC-dual-rand

Figure 10: Probability distribution of all possible 2-gram tapping sequences. The x-axis refers to the first tap, and the y-axis refers to the second tap in a given 2-gram tapping sequence.