Smartwatches Locking Methods: A Comparative Study

Toan Nguyen
New York University, School of Engineering
2 Metrotech Center, Brooklyn, New York, USA
toan.v.nguyen@nyu.edu

Nasir Memon
New York University, School of Engineering
2 Metrotech Center, Brooklyn, New York, USA
memon@nyu.edu

ABSTRACT

Smartwatches are rapidly emerging to be the next generation of personal devices from the smartphone era due to their novel form factor and broad applications. However, their emergence also poses new challenges to securing user information. An important challenge is preventing unauthorized access to private information stored on the watch, for which a locking method is typically used. Due to smartwatches' limited display, the performance of locking methods offered on smartwatches may suffer from the fat-finger problem and is currently unknown. In this paper, we present the first study to evaluate different locking methods for smartwatches. We contribute to the ongoing research trend in authentication for smartwatches with a reference benchmark and interesting insights for future work.

Keywords

User Authentication; Lock Screen; Smartwatches; Wearables; Security; Usability

1. INTRODUCTION

Smartwatches are on the rise; their market has shown an 18% increase per year in global sales since 2013 [10]. Their emergence is mostly because of their wide range of applications including communication functions, fitness tracking, and financial conducting capabilities [8, 14]. Currently, applications that utilize a smartwatch as a vault to sensitive information (i.e., passwords, credit card information) [11], or a token to conveniently unlock other devices [9, 4] and vehicles [18], have been introduced. As a result, a plethora of private information is stored on smartwatches that needs to be protected from unauthorized access. Typically, this protection is deployed in the form of a locking method, which will be activated in some scenarios: a) the smartwatch is taken off from the user's wrist; b) after some timeout; c) the watch is disconnected from its paired phone. Recently, research have shown that most smartwatches only offer a regular PIN lock or a Pattern Lock to the users [7, 12]. These are two matured methods on the phones, but their performance on smartwatches may suffer from the fat-finger problem [16] due to smartwatches' small display. In addition, being wearable gives smartwatches many chances of being used when the user is in the motion. Thus, the performance of the locking methods in different conditions needs to be explored.

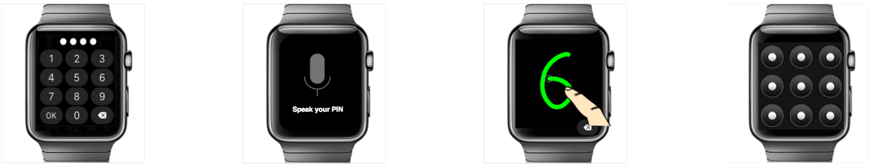
In this work, we conducted a lab study to evaluate actual performance and users' perceptions of different locking methods on smartwatches in different conditions. Specifically, we explore following four locking methods (illustrated in Figure 1).

- **Regular PIN**: This method requires users to type a short (4- or 6-digit) code to unlock their watch.
- **Voice PIN**: This method allows users to speak their PIN to the watch, which then recognizes and validates the PIN using speech-to-text techniques. Note that this method does not verify the speaker.
- **Draw PIN**: This two-factor authentication method, proposed in [17, 13], lets users draw their PIN on the touchscreen instead of typing it. The verification includes verifying the correctness of the PIN and the behavioral characteristics of the users.
- **Pattern Lock**: This method asks users to draw a secret pattern on a grid of 3x3 dots to unlock the watch.

Regular PIN and Pattern Lock are selected because they are state-of-the-art methods that are currently available on smartwatches shipped to customers. Their performance has been studied on smartphones [5], but not on the smartwatches. As we observe that PINs are still popular, we choose Voice PIN and Draw PIN which allow different ways to input PINs. Voice PIN utilizes voice, which is one of the main input channels on smartwatches. Draw PIN lets users draw on the whole display each digit in the PIN, one by one, to log in. The advantage of Draw PIN is that drawing is natural to most users. Also, drawing characteristics are unique to each person, which can be used as a second authentication factor [17, 13].

The preliminary results showed that the performances of Regular PIN and Pattern Lock are somewhat similar to those of smartphones. Interestingly, although Draw PIN is significantly more secure than other methods due to its two-factor nature, it is not suitable for unlocking daily as users frequently made more authentication errors. In contrast,

Symposium on Usable Privacy and Security (SOUPS) 2017, July 12–14, 2017, Santa Clara, California.



(a) Regular PIN (b) Voice PIN (c) Draw PIN (d) Pattern Lock

Figure 1: Four (un)locking methods in this study

Voice PIN is the least secure method, but thanks to its convenience, it was preferred in contexts in which users are moving or at their comfortable places like at home. Users were aware of the trade-off between usability and security. However, they favored the method with shorter input time. They also showed their interest in having multiple locking methods to use in different contexts. Our contribution is twofold. First, to our knowledge, we are the first to conduct a comparative evaluation of locking methods for smartwatches. We shed light on the performance of current methods which can be used as a benchmark for future work. Second, we introduce open problems and discuss interesting avenues for future research on smartwatch authentication.

2. METHOD

In this section, we describe details of our study. Our goal is to answer following questions.

- What is the performance of different locking methods on smartwatches? How does the performance change in different conditions, namely, sitting and walking?
- What are users' perceptions of each method? Do users' perceptions match with the actual performance of each method?
- Do users prefer multiple locking methods to choose from depending on the context they are in when authenticating? What are some use cases of these methods?

2.1 Study Design

Our study, which was approved by the IRB from our institution, was conducted using a repeated measures factorial design. The independent variables were *Locking Method* and *Condition*. The *Locking Method* includes four levels: *Regular PIN, Voice PIN, Draw PIN, and Pattern Lock*. The *Condition* comprises two levels: *Sitting* and *Walking*. This resulted in 8 experiment settings, which were presented to the participants using an 8x8 Latin square for counterbalancing.

The dependent variables were *performance* of locking methods including *error rate* and *input time*, and *users' perceptions* of each method. Error rate indicates how often an unsuccessful login occurs when using a method. The error rate is defined as the ratio between the unsuccessful login trials of the user and the total number of trials. In the case of Voice PIN and Draw PIN, the number of unsuccessful trials also includes the number of mistakes that are not caused by the user but rather from the implementation of the method. For example, Voice PIN may fail to recognize a spoken PIN because the user speaks too fast. These errors contribute to the usability level of each method and thus, need to be considered in the evaluation. *Input time* is the amount of time taken by a user to input her secret (PIN, pattern). It primarily determines how fast a locking method is since the time needed to make a login decision after the user input is very small. *Users' perceptions* of each method are separated into *perceived usability* and *perceived security*. *Perceived usability* was collected using the popular System Usability Scale (SUS) questionnaire [2]. *Perceived security* was collected by asking users to score security level of each method from 1 to 5 with 5 is the highest security.

2.2 Apparatus

We implemented four locking methods on a Samsung Gear Live smartwatch. All methods were implemented to run offline on the watch. This is to reflect the scenarios when the watch is disconnected from the phone and when the user authenticating in an offline mode, i.e., in a subway without an internet connection. Implementing Regular PIN and Pattern Lock was trivial. We implemented Voice PIN using the popular open source speech recognition toolkit CMUSphinx [3]. CMUSphinx has a lightweight speech recognition engine called PocketSphinx, which is specifically tuned for handheld and mobile devices. The PocketSphinx APIs enable us to capture user voice inputs, translate them to text, and recognize them. For the purpose of this study, we only recognize digits which are spoken in English. We trained our app to recognize a PIN even if it was spoken in different ways. For example, PIN "1234" can be spoken by a user as "twelve thirty-four" or "one thousand two hundred and thirty-four," and it can still be recognized correctly by the app. However, from our experiment, we observed that most users speak each digit in the PIN separately and sequentially (i.e., "one two three four,") which makes the recognition task much easier and more accurate. For this study, we implemented Voice PIN so that it will stop recording once it recognizes four digits inputted or after 300 ms timeout.

Draw PIN was implemented following technical details presented in [13]. Draw PIN has two components. The first component is a PIN Content Analyzer, which verifies the correctness of an entered PIN. If the PIN is invalid, the access to the watch is rejected immediately. Otherwise, it will be passed to the second component, a Drawing Behavior Analyzer, which verifies the drawing behavior of the user and grants access to the watch based on the verification result. When Draw PIN is invoked for the first time, the user needs to choose a PIN and draw it several times to train the classifier (enrollment phase). Following the Draw PIN paper, we required five samples from a user to train her model [13].

2.3 Participants

30 participants (average age: 27.5, range: 18-35) were re-

cruited for our study through a mailing list and a Facebook posting. The participant pool comprised undergraduates, graduates and faculty members from our institution. 12 participants were female, and 7 participants had owned or used smartwatches before this study. All participants have been using smartphones for more than five years. All participants expressed that they would be concerned if someone gains access to their smartphone or smartwatch and believe that it is important to have a locking method to keep others away. However, one participant did not use a locking method on his phone in exchange for convenience.

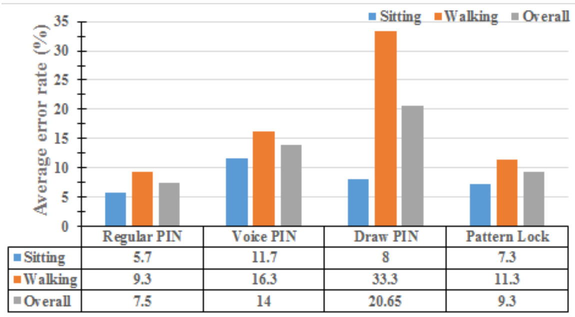
2.4 Study procedure

The study was conducted in a quiet office with the same surrounding setting for all participants. They performed authentication on the same smartwatch and were asked to wear it on the wrist that they normally wear or would wear a watch. The study included three sessions.

Introductory session: An experimenter explained to each participant the purpose of our study as well as what she would do during the study. The participant was asked to sign a consent form if she decided to participate. The participant was then asked to fill out a demographic survey. After that, the participant was allowed to operate the smartwatch and instructed to get familiar with four locking methods. For methods using PIN (Regular PIN, Voice PIN, Draw PIN), the participant was asked to choose a PIN which was used across the three methods. She was asked to draw the PIN five times on the watch display to train her Draw PIN model. For the Pattern Lock method, the participant was asked to choose a pattern with a length >= 4 (default minimum length as required by Android OS). This was the enrollment phase. In the verification phase, she was asked to practice with each method until she felt confident with the method and was able to successfully log in at least five times when sitting and five times when walking. After this practice, all apps were reset. To minimize the learning effect, the participant enrolled again for all methods with the same PIN and pattern to prepare for the authentication phase.

Authentication session: The participant performed authentication in the setting chosen for her from the Latin square, which was an ordered combination of method and condition. She unlocked the watch using each method ten times when sitting and ten times when walking around the room.

Survey session: At the end of the study, the participant was asked to fill out four System Usability Scale (SUS) questionnaires [2] for the four locking methods and a final exit survey. SUS has ten questions, with each having five response options from strongly agree to strongly disagree. It has become an industry standard and has been used widely to gather user's usability perception for various products and services. We also added a question at the end of each SUS questionnaire to ask participants to score security level of each method on a scale of 1–5 with 5 being the highest security level. In the exit survey, we first asked about the users' attitude toward having multiple locking methods on their watch from which they can choose based on their surrounding. In the second question, we asked about users' preference of locking methods in different contexts or use cases of locking methods presented in this study. The two questions are as follows.



	Regular PIN	Voice PIN	Draw PIN	Pattern Lock
■ Sitting	5.7	11.7	8	7.3
■ Walking	9.3	16.3	33.3	11.3
■ Overall	7.5	14	20.65	9.3

Figure 2: Average users' error rate of four methods in two conditions and overall

1. Would you like to have a multiple locking methods to choose from in different scenarios? (From 1–No to 5–I would really like to)
2. In following contexts, which method would you use to unlock your watch (you can choose multiple choices)?
   - A quiet place like office or classroom
   - A public place like a busy subway or a party
   - At home with family or friends
   - At home alone
   - Jogging or biking
   - You feel that somebody may look at your watch when you are authenticating

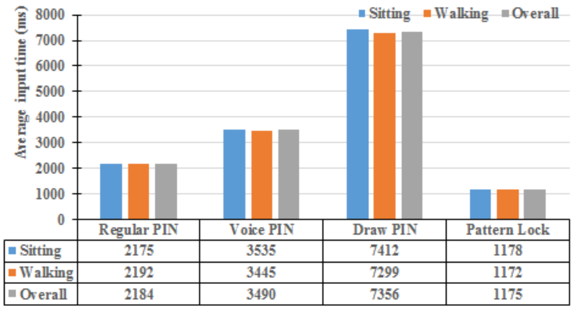
3. RESULT

Our data include (10 unlock trials per condition per user per method × 2 conditions × 30 users × 4 methods) = 2400 unlock trials. In this section, we present our analysis on this data to answer our research questions.

3.1 Error rate

For this analysis, the error rate of each participant in each condition in each method was calculated and considered as an independent data item. Statistical tests were done on the set of these per-user data. Figure 2 presents the average error rates of each method in two conditions and overall. Our tests showed that the error rate was not normally distributed. Thus we conducted a Friedman ANOVA and post hoc analysis with Wilcoxon signed-rank. There was a statistically significant difference in the *error rate* depending on which *locking method* was used ($\chi^2(3) = 21.77, p < .001$). Using a Wilcoxon signed-rank test with a Bonferroni correction resulted in a significance level of $\alpha < .00083$. Post hoc analysis using Wilcoxon signed-rank tests showed significantly higher errors using Draw PIN (mean=20.7%, sd=19.21%) vs Regular PIN (mean=7.5%, sd=9.85%) ($Z = -4.54, p < .0001$) and Draw PIN vs Pattern Lock (mean=9.3%, sd=10.23%) ($Z = -3.71, p < .0001$). However, we found no significant differences between other pairs.

Furthermore, *Condition* was a significant factor for the *error rate*. The Wilcoxon signed-rank test showed a significantly higher error rate when walking compared to sitting ($Z = -5.41, p < .0001$). This is reasonable because, in the



	Regular PIN	Voice PIN	Draw PIN	Pattern Lock
Sitting	2175	3535	7412	1178
Walking	2192	3445	7299	1172
Overall	2184	3490	7356	1175

Figure 3: Average input time (milliseconds) of four methods in two conditions and overall

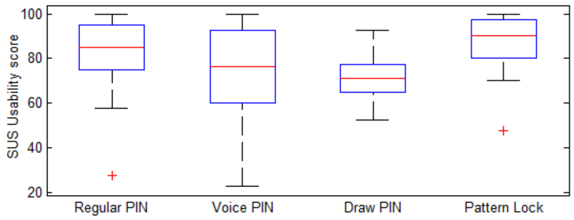


Figure 4: Usability score distribution of each locking method as rated by participants.

moving state, participants' hands and the watch were shaking. Thus, it was harder for them to input PIN or pattern on the display, which resulted in more input mistakes. This is especially true for Draw PIN because the drawing behavior would change dramatically when participants were walking. We can see that its average error rate increased four-fold from 8% in sitting condition to 33.3%.

3.2 Input time

Figure 3 presents average input time of each method in two conditions and overall. Since the input time distribution was skewed, we again used a Friedman ANOVA and post hoc analysis with Wilcoxon signed-rank. We found statistically significant difference in the *input time* according to *locking method* used ($\chi^2(3) = 170.78, p < .0001$). Using a Wilcoxon signed-rank test with a Bonferroni correction resulted in a significance level of $\alpha < .00083$. Post hoc analysis using Wilcoxon signed-rank tests showed that Draw PIN (mean=7355 ms, sd=879 ms) was significant slower than Voice PIN (mean=3490 ms, sd=482 ms) ($Z = -6.74, p < .0001$), Voice PIN was significant slower than Regular PIN (mean = 2183 ms, sd=566 ms) ($Z = -6.65, p < .0001$), and Regular PIN was significant slower than Pattern Lock (mean=1175 ms, sd=451 ms) ($Z = -6.21, p < .0001$).

However, we found that *Condition* was not a significant factor to *input time* (the Wilcoxon signed-rank test showed $Z = -5.41, p < .0001$). To our surprise, walking did not increase input time. This might be because, in the lab condition, the participants were not distracted by surrounding conditions as compared to real-world scenarios where they have to pay attention to obstacles on their way.

3.3 Users' perceived usability

Response to SUS questionnaire from each participant was converted to a usability score on a scale from 0 to 100. An average score of 68 indicates that a system is usable [15]. Overall, all methods were perceived as usable as their average scores were all above 68 (usability distribution of each method was presented in Figure 4). Running a Friedman ANOVA on usability scores of participants, we found statistically significant difference in the *perceived usability* according to *locking method* used ($\chi^2(3) = 29.23, p < .0001$). Using a Wilcoxon signed-rank test with a Bonferroni correction resulted in a significance level of $\alpha < .00083$. Post hoc analysis using Wilcoxon signed-rank tests showed that Draw PIN (mean=71, sd=2.07) was perceived as significantly less usable than Regular PIN (mean=83, sd=2.75) ($Z = -3.63, p < .0001$) and also significantly less usable than Pattern Lock (mean=87, sd=2.23) ($Z = -3.95, p < .0001$). However, there were no significant differences in perceived usability between other pairs.

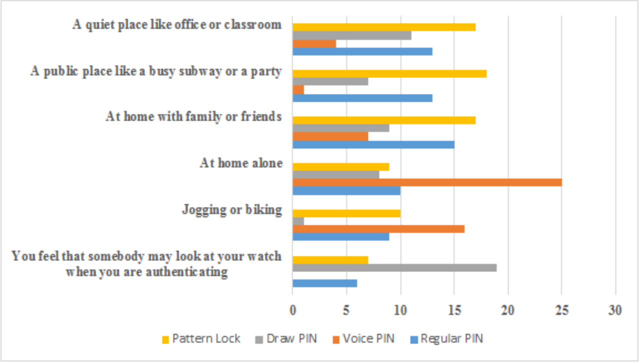


Figure 5: Preferences of locking methods in different contexts as rated by participants.

We ran a multiple regression to predict *users' perceived usability* (SUS score) based on *error rate* and *input time*. A significant regression equation was found ($F(2, 117) = 8.30, p < .0001$) with an $R^2 = .124$. Participants' predicted SUS score is equal to $87.66 - 9.925 \times a - .002 \times b$, where a is coded or measured as *error rate* and b is coded as *input time*. However, only *input time* was a significant predictor of SUS score ($p = .001$) while *error rate* was not ($p = .42$). The results indicate that participants perceived a method as less usable if they need more time to input the password (PIN, pattern) using that method.

3.4 Users' perceived security

A Friedman ANOVA and post hoc analysis with Wilcoxon signed-rank test was conducted for the users' security rating of each method. There was a statistically significant difference in the perceived security depending on the *locking method* used ($\chi^2(3) = 23.58, p < .0001$). Using Wilcoxon signed-rank test with a Bonferroni correction resulted in a significance level of $\alpha < .00083$. Post hoc analysis using Wilcoxon signed-rank tests showed the participants perceived that Draw PIN (mean=3.87, sd=.17) is significantly more secure than Voice PIN (mean=1.93, sd=.23) ($Z = 4.20, p < .0001$). However, there were no significant differences found between other pairs.

3.5 Qualitative data

As mentioned before, in the exit survey we first asked if participants would like to have multiple locking methods to choose from based on their surrounding. Results showed that participants like the idea (average=3.37, sd=1.12). Only one participant chose not to have this option. As it turned out, this participant also did not have any locking method on her phone for her convenience. We then asked participants to choose which method they would use in different contexts. Results were depicted in Figure 5. As we can see, most of the time, participants still preferred Regular PIN and Pattern Lock over other methods. However, when there was a security risk of an observation attack, they would opt to use a more secure method (Draw PIN). None of them chose Voice PIN in this context. In a scenario where they were on the move, majority of participants chose Voice PIN for convenience thanks to its hands-free and eyes-free feature. Voice PIN also was chosen when participants were alone at home. This indicates that even though Voice PIN is not secure and potentially awkward in social places, it is still useful in certain contexts. Draw PIN, which offers two-factor verification, should be used in scenarios or applications where extra security level is needed. However, it is not a good option for unlocking the watch in normal contexts because of its high error rate and low speed.

4. DISCUSSION AND CONCLUSION

Although the error rate of Regular PIN on smartwatches was more than twice as high as that of smartphones (average: 7.5% compared to 3.1% [5]), they were both less than 10%. On the other hand, the error rate of Pattern Lock on smartwatches was slightly better (average: 9.3% compared to 12.1% on smartphones [5]). Nevertheless, using PINs resulted in fewer errors than using Pattern. Our finding is consistent with the previous result on smartphones [5]. In terms of input time, PIN (average: 2184 ms) and Pattern (1175 ms) were slightly slower than that of PIN (average: 1963 ms) and Pattern (average: 910 ms) on smartphones [5]. These are interesting observations worthy of further investigation. An important difference is that our study was conducted in a lab environment whereas Harbach et al. in [5] conducted a field study. We plan to extend our study to a longitudinal field study in future work. Nonetheless, from insights provided by the results, it seems like the fat-finger problem is actually not a significant problem for authentication on smartwatches.

Our results also suggest that users tend to favor a locking method that is fast to input. This should be taken into consideration of future research on authentication for smartwatches. Approaches that require significantly more time to input than the PIN lock and Pattern lock will unlikely to be widely accepted by users.

Having multiple locking methods to use in different contexts was appreciated by the users. Especially, having multiple ways to input a secret (i.e., PINs can be entered by typing, drawing or speaking) is an intriguing idea. Combining with context-aware techniques [1, 6], this approach can potentially enhance user experience and improve the security of authentication tasks.

We are aware that our study was limited as a lab pilot study and the results may not totally reflect real-world performances. Nevertheless, this controlled environment allows us to give the same condition to each participant and each method. Therefore, the comparison results of different locking methods will likely be the same in a field study. In addition, our results are limited in error rate, input time, and users' perception. There are other aspects of authentication, like those used by Harbach et. al [5], i.e., the time before unlock, recovery time, and types of errors. We hope to take these aspects into consideration in a more rigorous field study in future work.

In conclusion, we have presented the first study that evaluates different locking methods on smartwatches, including built-in PIN and Pattern Lock. Insights from our results will benefit and spark interesting questions for future work on smartwatch authentications.

5. REFERENCES

[1] J. E. Bardram, R. E. Kjær, and M. Ø. Pedersen. Context-aware user authentication–supporting proximity-based login in pervasive computing. In *International Conference on Ubiquitous Computing*, pages 107–123. Springer, 2003.

[2] J. Brooke. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

[3] CMU. Cmusphinx, open source speech recognition toolkit, 2016. Retrieved May 25, 2017 from http://cmusphinx.sourceforge.net/.

[4] Google. Google smart lock, 2016. Retrieved May 25, 2017 from https://get.google.com/smartlock/.

[5] M. Harbach, A. De Luca, and S. Egelman. The anatomy of smartphone unlocking: A field study of android lock screens. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 4806–4817. ACM, 2016.

[6] E. Hayashi, S. Das, S. Amini, J. Hong, and I. Oakley. Casa: context-aware scalable authentication. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, page 3. ACM, 2013.

[7] HP. Internet of things security study – smartwatches, 07 2015. Retrieved May 25, 2017 from https://www.ftc.gov/system/files/documents/public_comments/2015/10/00050-98093.pdf.

[8] A. Inc. Apple pay with your apple watch, 2016. Retrieved May 25, 2017 from https://support.apple.com/en-us/HT204506.

[9] A. Inc. Automatically unlock your mac with your apple watch, 2016. Retrieved May 25, 2017 from https://support.apple.com/en-us/HT206995.

[10] B. Insider. The smartwatch report: Forecasts, adoption trends, and why the market isn't living up to the hype, 09 2016. Retrieved May 25, 2017 from http://www.businessinsider.com/smartwatch-and-wearables-research-forecasts-trends-market-use-cases-2016-9.

[11] LastPass. Lastpass for apple watch is here!, 2015. Retrieved May 25, 2017 from https://blog.lastpass.com/2015/04/lastpass-for-apple-watch-is-here.html/.

[12] MobileIron. Mobileiron analysis of smartwatch security risks to enterprise data, 2015. Retrieved May 25, 2017 from https://www.mobileiron.com/sites/default/files/whitepapers/files/smartwatch-security-1.2-EN.pdf.

[13] T. V. Nguyen, N. Sae-Bae, and N. Memon. DRAW-A-PIN: Authentication using finger-drawn pin on touch devices. *Computers & Security*, Volume 66:115 – 128, 2017.

[14] Samsung. Samsung pay on gear s2 beta, 2016. Retrieved May 25, 2017 from http://www.samsung.com/us/support/owners/app/samsung-pay-gear.

[15] J. Sauro. Measuring usability with the system usability scale (sus). http://www.measuringu.com/sus.php. Accessed: 2016-08-12.

[16] K. A. Siek, Y. Rogers, and K. H. Connelly. Fat finger worries: how older and younger users physically interact with pdas. In *IFIP Conference on Human-Computer Interaction*, pages 267–280. Springer, 2005.

[17] T. Van Nguyen, N. Sae-Bae, and N. Memon. Finger-drawn pin authentication on touch devices. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 5002–5006. IEEE, 2014.

[18] T. Verge. You can remotely start your hyundai with your apple watch now, 07 2015. Retrieved May 25, 2017 from http://www.theverge.com/2015/7/1/8877445/hyundai-blue-link-apple-watch.