Usability and Security Perceptions of Implicit Authentication: Convenient, Secure, Sometimes Annoying

Hassan Khan, Urs Hengartner, and Daniel Vogel
Cheriton School of Computer Science
University of Waterloo
Waterloo, ON Canada
{h37khan,urs.hengartner,dvogel}@uwaterloo.ca

ABSTRACT

Implicit authentication (IA) uses behavioural biometrics to provide continuous authentication on smartphones. IA has been advocated as more usable when compared to traditional explicit authentication schemes, albeit with some security limitations. Consequently researchers have proposed that IA provides a middle-ground for people who do not use traditional authentication due to its usability limitations or as a second line of defence for users who already use authentication. However, there is a lack of empirical evidence that establishes the usability superiority of IA and its security perceptions. We report on the first extensive two-part study (n = 37) consisting of a controlled lab experiment and a field study to gain insights into usability and security perceptions of IA. Our findings indicate that 91% of participants found IA to be convenient (26% more than the explicit authentication schemes tested) and 81% perceived the provided level of protection to be satisfactory. While this is encouraging, false rejects with IA were a source of annoyance for 35% of the participants and false accepts and detection delay were prime security concerns for 27% and 22% of the participants, respectively. We point out these and other barriers to the adoption of IA and suggest directions to overcome them.

1. INTRODUCTION

Recent studies on locking behaviour of smartphone users have shown that 40% or more users did not use any authentication mechanism on their devices [8, 15, 20]. Furthermore, participants of these studies most commonly cited "inconvenience" as the reason for not using any authentication mechanism on their devices [8, 15]. The traditional explicit authentication (EA) mechanisms (such as PIN, pattern lock, facial and fingerprint recognition) provide all-or-nothing access control. However, smartphone sessions are short and frequent, and PIN entry for every short session is inconvenient for users and unnecessary for some situations [16]. In a field study, Harbach et al. [15] demonstrated that smartphone users considered unlock screens unnecessary in 24% of the situations and they spent up to 9% of the time they use their smartphones to deal with the unlock screens.

To address these usability issues with EA, researchers have proposed implicit authentication[1] (IA). IA employs behavioural biometrics to continuously and transparently recognize and validate the identity of smartphone users. Some of the recent touch behaviour-based IA schemes provide high accuracy rates (≥95%) and have gained traction in technology media news with claims like: *"Identifying someone by the way they tap and swipe on a touchscreen might be the more natural, unobtrusive future of smartphone biometrics"* [22]. Researchers have proposed to use IA as a middle ground for those smartphone users who do not configure any EA on their devices due to EA's usability issues [18, 31] or as a second line of defense in case the EA mechanism on the device is compromised [12, 19, 31].

The focus of the majority of IA research is on improving the accuracy of IA schemes with existing behavioural biometrics and by leveraging new sensors like orientation [3]. The usability evaluation of IA has largely been ignored. Existing IA literature has assumed without empirical evidence that since IA authenticates without requiring explicit input, it is more usable. For instance, when Shi et al. introduced the term IA with a behaviour-based scheme, they postulated that *"this is a meaningful approach, whether used to increase usability or increase security"* [25].

Given that IA does not require explicit input to authenticate the device user for every session, intuitively it seems that IA should reduce the amount of time spent on authentication. However, despite the reasonably high detection accuracy of some IA schemes, these schemes are still subject to false rejects, false accepts and detection delays (these terms are explained in § 2.1), which could introduce new usability issues and affect users' security perceptions. If the IA detection model is unsure about the user's identity, it naturally resorts to an interrupt-authenticate approach in which the current task is pushed to the background and the user has to explicitly authenticate to establish their identity [9, 19]. This interrupt-authenticate approach for false rejects is quite different from consistent EA authentication. It remains unclear how it affects usability in terms of annoyance and task performance in terms of time and error. Similarly, it is not obvious whether the usability-security trade-off offered by IA overcomes the perception of security given the risks of false accepts and delay in detection of an intruder.

[1]The terms active authentication, continuous authentication, implicit authentication and transparent authentication have been used in the literature interchangeably.

Symposium on Usable Privacy and Security (SOUPS) 2015, July 22–24, 2015, Ottawa, Canada.

To find answers to these questions, we conducted an ambitious user study (n=37) to empirically evaluate usability claims and security perceptions of IA. Our study uses a two-part design combining a controlled lab experiment with a field study. In a controlled lab environment, we compare IA with the preferred EA scheme of each participant. Then, a subsequent field study captured real-life experiences with IA false rejects over three days. We use a pseudo-IA scheme for both parts to simulate representative false reject rates in a controlled fashion and avoid potential confounds from idiosyncratic behaviour caused by specific IA schemes. During both parts of the study, we logged measurement data and collected feedback through surveys for a quantitative analysis and we conducted interviews for a qualitative analysis and to gather insights into subjective perceptions of IA.

In terms of usability, we show that IA decreased the overall task completion time without affecting the task error rate. While there was no statistically significant difference in the system usability scale score [4] between IA and EA, IA was favoured by 91% of the participants in terms of convenience (26% more compared to EA). However, annoyance due to false rejects is a potential issue with IA: 35% of the participants found this to be somewhat or very annoying.

In terms of security perceptions, we found that although detection delay and false accepts were serious concerns, 82% of participants were satisfied with the level of protection provided by IA. Consequently, 33% of participants were interested in using it as a primary authentication mechanism, 30% were interested in using it as a secondary authentication mechanism and another 30% wanted to try and test IA more to see if it satisfied their security needs.

Feedback gathered during the interviews further supports these findings and provides insights for how to mitigate the effect of false rejects and how to address deployment issues such as opacity of IA and operating threshold customization. We believe our findings provide important evidence and guidelines for future IA research and deployments.

2. BACKGROUND AND RELATED WORK

In this section we first provide a brief background on IA and its terminology. We then survey the related literature on usability evaluations and security perceptions of IA.

2.1 Background on Implicit Authentication

IA employs behavioural biometrics to continuously and transparently recognize and validate the identity of smartphone users. To provide IA on smartphones, many IA schemes have been proposed by the research community that rely on a diverse set of behavioural biometrics including a user's device usage patterns [25, 26], touchscreen input behaviour [12, 19, 24, 31], keystroke patterns [7, 10, 13], and gait patterns [11, 21]. IA schemes create a normal behavioural profile of the device owners using their behavioural data and it then monitors the real-time device usage to detect anomalous behaviour. The anomaly detection based origins of IA result in the same inherent limitations including: *false rejects (FR):* when an access attempt by a legitimate user is rejected due to variations in the user behaviour or imperfections in the behavioural profile; *false accepts (FA):* when an access attempt by an adversary is granted due to behavioural similarities between the adversary and the legitimate user; *delayed detection:* due to the unavailability of behavioural data, an adversary may be able to use the device for some duration before she fails authentication; and *training delay:* time spent to collect usage data in order to create the behavioural profile of the user.

Due to the security limitations of IA, researchers have proposed to use IA only as a middle-ground for the smartphone owners who do not configure any EA on their devices due to EA's usability issues [18, 31] or as a second line of defense in case the EA on the device is compromised [12, 19, 31]. For both of these deployment scenarios, researchers have suggested to interrupt and then explicitly authenticate the user (using an "administrator password") in case the behavioural profile of the current user does not match that of the device owner [9, 18, 19]. We use the term *interrupt-authenticate* to describe this phenomenon. Some related terms that we use throughout this paper are also described: a *true accept (TA)* is when an access attempt by a legitimate user is granted. A *true reject (TR)* is when an access attempt by an adversary is rejected. Finally, *operating threshold* is used to define the desired values for the negatively correlated FA and FR entries (by increasing the operating threshold, FRs can be decreased at the cost of increased FAs and vice versa).

2.2 Related Work

Usability evaluation of EA on smartphones is a well researched area [2, 15, 28, 29]. However, it is only partly related to our work since some of the evaluation metrics for EA (such as time-to-authenticate and memorability) are not applicable to IA. Another avenue partially related (due to the overlapping limitations) to our work is the usability evaluation of anomaly detection systems. However, to the best of our knowledge, the literature on the usability evaluation of anomaly detection systems only discusses challenges in determining an appropriate operating threshold and does not evaluate the security and usability perceptions due to FAs and FRs [23, 30]. Therefore, in this section, we only discuss the related work in the field of IA.

While there are dozens of papers on IA, the focus of contemporary IA research is on using novel behavioural biometrics for authentication and on improving the accuracy of IA. The usability issues surrounding IA have been ignored except for the work by Clarke et al. [5] and Crawford and Renaud [6]. Clarke et al. developed a prototype on a personal computer for an IA scheme that employed a combination of face, voice and keystroke biometrics to continuously authenticate users. They evaluated their prototype using 27 participants and found that 92% of the participants considered it more secure in comparison to the traditional forms of authentication. The participants were also asked to rate the convenience on a 5-point Likert scale and although the responses were mixed, a slight skew towards the system being convenient existed. While Clarke et al. are the only authors who provide a usability evaluation of the IA scheme that they proposed, their evaluation was limited because: (i) it was not a strictly behavioural biometric-based scheme since they used a combination of physiological (facial recognition) and behavioural biometrics (voice and keystroke data); and (ii) participants evaluated the prototype on a personal computer instead of a mobile device.

More related to our work is the recent work by Crawford and Renaud [6] in which they determine the security perceptions of IA by conducting an in-lab study with 30 participants. They provided a smartphone with a pseudo-IA scheme and asked the participants to perform tasks that

required different levels of security. The participants were divided into three groups: (G1) the participants started with a low device confidence level. If a participant wanted to perform a task of medium/high security level, she may increase the device confidence by providing a matching keystroke or voice biometric or by explicitly authenticating; (G2) participants were always successfully implicitly authenticated (0% FR rate); and (G3) participants always failed implicit authentication (100% FR rate). G2 and G3 were used to get the perceptions of distrustful and frustrated participants, respectively. Crawford and Renaud found that 73% of participants felt IA was more secure than EA and 90% indicated that they would consider adopting it. While Crawford and Renaud provide the only in-depth study on the security perceptions of IA, it has some limitations including: (i) no usability evaluation is performed; (ii) annoyance due to FRs is not quantified; and (iii) security perceptions due to FAs and detection delays are not evaluated.

3. GOALS

We divide our goals to investigate IA usability and perceived security into seven questions. Later, we organize our study results around these seven questions.

Our main goal regarding IA usability was to test established usability metrics and commonly accepted usability assumptions, as captured by the following research questions:

U1 Does IA decrease the overall task completion time and the authentication overhead when compared to EA?

U2 Do the interrupt-authenticates in IA increase the error rate of the primary task?

U3 Are fewer but less predictable authentication interrupts of IA less annoying or tolerable as compared to EA or no authentication at all?

U4 Does IA score higher on the system usability scale [4] as compared to EA?

U1 and U2 address standard usability metrics for time and error as they may be affected by the interrupt-authenticate model of IA. These metrics have never been evaluated directly with IA, but they have been used to measure performance impact of similar task interruptions with personal computers [1] and they have been implied as benefits of IA in previous work [25]. By answering U3, we will test levels of annoyance caused by FRs and through U4, we test claims of higher perceived usability for IA compared to the primary authentication baselines of EA [5, 6] and no authentication.

Our main goal for the security perceptions of IA was to explore the following research questions:

S1 Are the security properties of current IA schemes (such as the FA rate) acceptable to users?

S2 Is the perception of IA security better than common current authentication schemes?

S3 Are smartphone users interested in adopting IA?

S1 has never been explored in the IA literature. S2 has been explored in previous studies [5, 6] and we attempt to validate these prior findings. In addition to evaluating the overall perceived level of security, we elicit the perceived level of security against different types of adversaries, different device states and different types of tasks. Finally, answering S3 provides an indication of IA deployment potential from a human-centric perspective since it essentially combines security perceptions and usability.

4. STUDY

We use a two-part study for our evaluations. The first part is a lab-based experiment where each participant performs simulated tasks with IA and with their current authentication scheme. This provides highly controlled, quantitative results. Measuring annoyance and other subjective feedback caused by IA interruptions is more ecologically valid when evaluated with real tasks over a longer time period, so the second part is a three-day field study where participants used IA on their own smartphone. For experimental control, both parts use a pseudo-IA scheme (described below). Our methodology was reviewed and approved by the IRB of our university.

4.1 Participants

The in-lab study was completed by 37 participants and 34 of those same participants completed the field study. Three participants dropped due to technical issues (two participants had device encryption enabled and one participant reported a broken device). We recruited these participants using multiple sources including: (i) an advertisement on Craigslist and Kijiji in November of 2014, under the "other jobs" section; and (ii) on the university-wide mailing list. The title of the advertisement was "Participate in a research study on the efficacy of a novel authentication scheme on smartphones" and it stated that the study was about the evaluation of a novel authentication scheme and adults who owned and used an Android-based smartphone for over six months could participate. Those interested were requested to fill out an online screening survey (provided in Appendix B), which collected information about their age, gender, profession, security preferences, smartphone make and model, amount of time they have used a smartphone, and email address. Participants were paid $35 ($10 for each of two in-lab sessions and $15 for the field study).

Participant demographics, current authentication schemes, and authentication preferences are summarized in Table 1. Current authentication scheme by age group is provided in Figure 1. Overall, our participant pool has good diversity by profession, age, and current authentication scheme. For our research questions, this kind of diversity is important. Similar to the prior studies [8, 15], the top reason our participants gave for not using any authentication scheme was inconvenience. Furthermore, about half of the participants who used some authentication scheme agreed that it was inconvenient or annoying at times. The annoyance was split by current authentication scheme: PIN users were significantly more likely (53% more) to find their authentication scheme inconvenient as compared to the pattern lock users (Fisher's Exact Test, $p = 0.028$).

In terms of current authentication, 14 participants used no authentication, eleven used Android's Pattern Lock, nine used a four-digit PIN and three used other schemes (two participants used a password and one participant used a longer PIN conforming to his company policy). We use the participants' current authentication scheme as an independent between-subject variable, which we refer to as *Use*. Where relevant, we summarize results using groups and subgroups formed by this variable. Specifically, *DontUseAuth* is the

n = 37	
Gender:	56% Female
	43% Male
Occupation:	32% Employed
	30% Grad student
	24% Undergrad student
	13% Unemployed/retired
IT experience:	22% Studied/worked in IT
Current authentication scheme:	38% None
	24% PIN
	30% Pattern lock
	8% Other
Sharing habits:	51% Never
	41% Rarely (once a month)
	5% Occasionally (once a week)
	3% Daily
Top reasons for not using any authentication:	8/14 It's a hassle/takes time
	5/14 Nothing to hide
	3/14 Never thought about it
Top reasons for using authentication:	19/23 Protected if lost/stolen
	18/23 Protected when unattended
	12/23 Someone casually picking it
	10/23 Unwanted disclosures
Protecting against:	18/23 Strangers
	12/23 Coworkers
	8/23 Friends/roommates
	7/23 Spouse/own children
Thoughts on authentication:	13/23 It is inconvenient sometimes
	10/23 It is easy
	3/23 It takes time

Table 1: Demographics and security preferences of the study participants

group of participants who reported that they do not use any authentication and *UseAuth* refers to the group of participants who reported that they use some EA scheme. We further separate *UseAuth* into two common EA schemes: *UsePIN* for the subgroup of *UseAuth* participants who reported using a PIN and *UsePAT* for the subgroup of *UseAuth* participants using a pattern lock.

4.2 Apparatus

We developed two Android apps, Explicit Authentication and Implicit Authentication, that executed on the devices of the participants during both parts of the study. For the lab-based experiment, the apps presented a series of tasks to the participants. The apps contained authentication screens to authenticate the participants using a PIN, Android's Pattern Lock or a six character password. We used the Android Open Source Project's UI and implementation[2] for the pattern lock and used a UI identical to that of Android for PIN and password screens. We simulated authentication on the participants' devices using our apps (explained in S 4.3.3) to accurately measure the time spent on authentication. For the field study, the Implicit Authentication app executed as a background service to simulate FRs. Although IA was simulated, participants were told that all biometric data remained solely on their device.

[2] http://code.google.com/p/android-lockpattern/

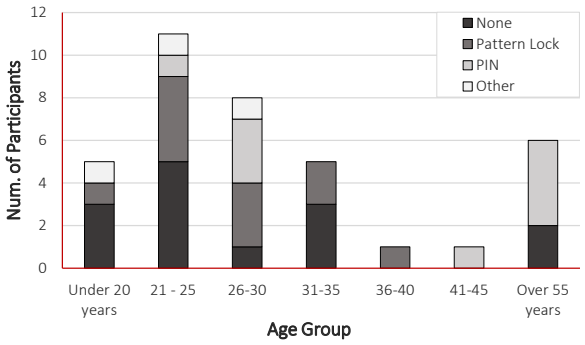


Figure 1: The "Age Group" – "Current Authentication Scheme" distribution for study participants

Deception: In both parts of the study we used a pseudo-IA scheme that ostensibly employed the touch behaviour biometric. The pseudo-IA scheme interrupt-authenticated users by triggering authentication screens during their interaction with the device to simulate FRs. Our pseudo-IA scheme was configurable to simulate different FR rates and detection delays. We used a pseudo-IA scheme because: it was not possible to intercept touch input events for IA without rooting the device due to the constraints imposed by Android [18]; and a pseudo-IA scheme enabled us to strictly control the frequency and the detection delay of FRs. Participants were told that the IA scheme on their device was fully operational and that one of the in-lab sessions served as training for the IA algorithm (details provided in § 4.3). Furthermore, to reduce the chance of participants discovering this deception by testing it with other users, we asked them not to share their devices during the field study arguing that sharing would interfere with the brief re-training phases required by our early-stage IA scheme. There are limitations of using a pseudo-IA and we discuss these limitations in § 7.

Configuration parameters selection: A challenging aspect was the selection of IA configuration parameters. By choosing different parameter values, we can change the behaviour of IA schemes. For instance, there is a negative correlation between FAs and FRs, and increasing detection delay reduces the number of FRs. While there is no recommended operating threshold, we inspected different operating thresholds because: (i) high accuracy input behaviour-based IA schemes have deployment constraints as compared to low accuracy schemes that can be readily deployed [17]; (ii) studies report a high degree of variability for FR rate between users [7]; (iii) variability in FR rate for an individual user may be caused by the type of activity that the user is performing [9]. We chose a representative and realistic FR rate range based on previous work [17]. We evaluated three FR rates: 5%, 10% and 20% that have a corresponding FA rate range between 3%-18% and 5-30 seconds of detection delay for different IA schemes. We discuss operating threshold configurations specific to each experiment in § 4.3 and § 4.4.

4.3 Part 1: Controlled Lab Experiment

The purpose of the controlled lab experiment is to: (i) introduce the participants to IA; (ii) perform an A/B testing of IA with non-IA (*UseAuth* or *DontUseAuth*); (iii) demonstrate an ostensible TR; (iv) elicit initial feedback on usability and security perceptions of IA; (v) collect data to eval-

uate the performance metrics of U1 and U2; and (vi) test the pseudo-IA scheme on the participants' devices without any EA scheme configured (this was not possible in the field study due to data security and privacy threats without EA).

4.3.1 Task

Our two apps presented a series of experiment tasks on the participant's own smartphone. In the experiment, each *task* represented a device usage session. The participant waited for the device to ring (with vibrate) to indicate it was time to perform a task. For the Explicit Authentication app, the participant turned the screen on, performed authentication if required, completed an activity, and turned the screen off. For the Implicit Authentication app, instead of the authentication at the beginning, the participant was interrupted and authenticated in the middle of an activity (frequency and timing of interrupt-authenticates are discussed in § 4.3.2). There was no time limit to complete a task. To simulate longer breaks between real device usage sessions, the participant waited a random time between 8-15 seconds before performing the next task.

We chose a subset of activities from the primary activities proposed by Bailey and Konstan [1]. We chose those activities that were abstract representations of common mobile activities and were diverse in terms of difficulty and cognitive load, enabling us to inspect error rate and interrupt-authenticate overhead. A description of these activities by increasing level of difficulty due to higher mental loads on working memory (based on the rankings of [1]) are provided below (see also screen captures in Appendix A):

- *Input Activity:* entering a sequence of characters. Our activity required participants to enter a nine digit number displayed on the screen into an input field. For the input number, we carefully chose permutations that did not overlap with the local area codes and did not have any consecutive or repeating integers. This activity is representative of common smartphone activities like entering search queries, composing texts, and entering emails.
- *Selection Activity:* selecting multiple items from a list of items. We used a list of words with selection checkboxes arranged in a 12-row x 3-column table. Thirty-six words were randomly chosen from a base set of six words. Participants had to select each word in the table that matched a given target word (taken from the base set). This activity is representative of common smartphone activities such as choosing a number to dial, choosing an app to launch or scanning the results of a search.
- *Reading Activity:* reading and comprehending information. Participants read a 7-10 sentence narrative passage from Wikipedia and then answered two multiple choice questions regarding its content. This activity is representative of common smartphone reading and comprehension activities such as reading emails or web browsing.

4.3.2 Design

Participants completed the lab-based experiment in two sessions held on different days. Each session lasted between 45-60 minutes including introduction, pre-survey, experiment tasks, post-survey, and interview. In each session, experimental tasks were completed under one of two within-subject conditions: the *IA* condition when they used IA and *non-IA* when they did not use IA. The order of IA and non-IA sessions was counterbalanced across participants.

Each session had 30 task trials with each task showing one activity. There were ten instances of each of the three activity types. Since the activity types were of varying difficulty level, we did not use simple random sampling to select their order since some orderings could introduce a confounding effect (e.g. the first 10 tasks are all difficult activities). Instead, we constrained the random presentation order by creating five blocks of six tasks where the tasks have two variations of each activity. We then permuted the order of tasks within each block using the first 5 rows of a 6x6 Latin square. This counterbalanced the varying difficulty levels of activity types. The same order of blocks and tasks was used across IA and non-IA sessions across all participants to create an unbiased comparison and to make sessions directly comparable.

For the non-IA session, participants were assigned the same authentication scheme that they used currently on their device (which could be an EA scheme or no authentication). For the interrupt-authenticate caused by a simulated FR in the IA session, *UseAuth* participants used their current authentication scheme while *DontUseAuth* participants were assigned a scheme that they preferred to use. Although we could have assigned a random authentication scheme to *DontUseAuth* participants, this could have introduced negative bias from dislike or inexperience with the assigned scheme. In both sessions, we were not interested in the memorability of the secret. Participants could write down their secret or reset it in between tasks if they wished.

While the input activity naturally generated tapping data, we rendered the reading activity and the selection activity in such a way that participants had to swipe to scroll to see their content. This led them to believe that their interactions were used as a biometric. We also used deception in terms of training by telling the participants who tested IA in the their first session that the data from the first few tasks was used for training. The participants who tested IA in the second session were told that the data from the first session was used for training.

We used a 20% FR rate (six interrupt-authenticates in total, twice for each type of activity) and a detection delay between 5-10 seconds. A lower detection delay (2-4 seconds) was used for the shorter input activity.

4.3.3 Procedure and Data Collection

The shortlisted participants were asked to bring their devices to the lab. We started the first session by showing a two minutes video introducing the apps and activities[3]. Participants were introduced to IA using a three minutes video before the IA session, which explained the operations of touch behaviour-based IA and the associated FAs, FRs and operating threshold[4]. A researcher was available during these video demonstrations to answer any questions. After the briefing, participants downloaded and installed the app for the session through Google Play Store. They were then asked to set the current authentication scheme to 'None' and turn on 'airplane mode' on their devices. This eliminated notifications or interruptions during the experimental tasks and enabled our app to control all authentications.

[3] http://youtu.be/qDQm_0ad6Pw
[4] http://youtu.be/HUR2-bxBtI8

After device setup, participants completed the STAI survey [27] to provide us with their current state of anxiety. The participants were then asked to launch the app to configure an authentication secret and then complete the main experiment tasks. After completing all tasks, they provided another measurement on anxiety by completing the STAI survey again. The participants were asked to complete a post-survey (provided in Appendix C) for the non-IA and the IA sessions. This survey consisted of 12 questions regarding usability and security perceptions of the authentication schemes that they tested. The participants used the survey to rate their perceived level of security (overall and for different adversaries, device states, and different tasks) and usability (in terms of convenience, annoyance, time consuming and tiring). Participants who tested any authentication scheme during a session were asked to complete the system usability scale (SUS) survey [4] after the session. The SUS survey was modified (provided in Appendix D) to explicitly inform the participants that it was evaluating the authentication scheme, not the apps. We also changed the word 'system' to 'method' and we dropped the question 'I found the various functions in this system were well integrated' because it was not applicable to our evaluations. Finally, a semi-structured interview (provided in Appendix E.1) of 10-15 minutes gained insight into survey answers. The interviews were recorded and later transcribed.

4.4 Part 2: Field Study

The field study was conducted after the lab study with the same participants. The main purpose of the field study was to gather realistic data on potential annoyance due to FRs. In addition, we wanted to subject participants to different operating thresholds to determine a tolerable one in terms of the frequency of FRs.

4.4.1 Task

The task for the field study was for participants to use their device as usual and handle simulated IA FRs (experienced as interrupt-authenticate screens) as they occurred. After each interruption the participant also provided brief feedback through an in-situ pop-up.

Each FR interrupted the current smartphone app with an authentication screen requiring an explicit authentication. These were the same simulated authentication screens used in the lab experiment. The background service in the Implicit Authentication app monitored two events: the `ACTION_SCREEN_ON` event to keep track of when the users turned on their screens and the `ACTION_USER_PRESENT` event to know when the users were present on their devices after dismissing the lock screen. An interrupt-authenticate was triggered after k `ACTION_USER_PRESENT` events (k was controlled during the study, details are in the Design section).

We were also interested in measuring the annoyance of each FR. After an interrupt-authenticate, we performed experience sampling with a simple in-situ feedback screen (Figure 14d). It asked the participants about their current annoyance on a 5-point Likert scale ("Very annoying" - "Not annoying at all"). The feedback screen also displayed the current operating threshold and the associated security strength of that threshold (in terms of the proportion of strangers that the IA scheme would likely protect against).

4.4.2 Design

We conducted the field study for three days to measure annoyance for different operating thresholds. FRs were simulated after every k `ACTION_USER_PRESENT` events and we randomly chose a value of k for each day to reflect high, medium and low accuracy corresponding to 20%, 10% and 5% FR rates, respectively. Since we were unable to determine when a participant interacted with the touch screen after an `ACTION_USER_PRESENT` event, we simulated a FR by choosing a random delay between 15-30 seconds. If the participants turned off the screen before the delay timeout, they were authenticated in the next session with a reduced delay. The delay value is decreased by five seconds each time down to a minimum delay value of ten seconds to ensure that the participants with short sessions also experienced FRs. We did not simulate a FR for the sessions when the call state of the device was ringing or off-hook.

We allowed participants to adjust the operating threshold if they wished. To ensure that the participants did not set the operating threshold to zero, the participants' adjusted value was only effective for 30 minutes and after that it was reset. This mitigated the possibility of participants killing the background service if interrupts became too irritating (participants felt they had some control) and gathered data to study the potential need and utility for users to control the trade-off between usability and security. The briefing video explained the trade-off when setting different threshold values. We report frequency of adjustment and discuss the need for such a control in § 6.3. We told participants that the IA scheme automatically adjusted the threshold value after brief re-training phases but they could change it depending on their desired level of protection. Participants were informed about this behaviour and that they could dismiss the interrupt-authenticate by pressing the home button but we asked them to avoid doing so except in extreme cases.

4.4.3 Procedure and Data Collection

After the second in-lab study session, participants were briefed about the background service executing on their devices and the interface of the in-situ feedback pop-up. The researcher then performed an ostensible demonstration of a true reject on their device to lead them to believing that the IA scheme was behaving as expected. After the completion of the three day usage period of the pseudo-IA scheme, the participants were instructed to contact us through emails to arrange for an in-person semi-structured interview (provided in Appendix E.2) of 10-15 minutes and to collect the remuneration. The participants were also debriefed about the deception at the end of this interview.

During the field study, we logged the `ACTION_SCREEN_ON` and `ACTION_USER_PRESENT` events. From the in-situ feedback, we logged the level of annoyance of IA and the adjusted value of operating threshold.

5. RESULTS

The quantitative and qualitative results of the controlled lab experiment and field study are presented together organized by the research questions raised in § 3. A discussion is provided after the results for each research question.

For the in-lab study, participants completed all tasks in 25 minutes on average (*median = 23, sd = 4.2*). Dur-

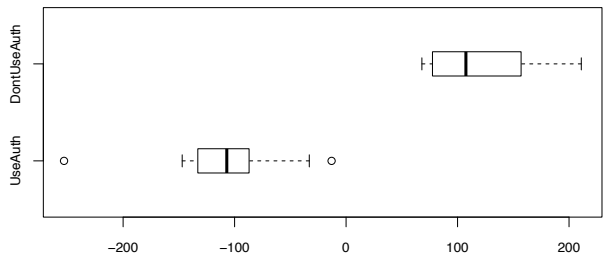


Figure 2: Change in the overall task completion time for the non-IA session as compared to the IA session

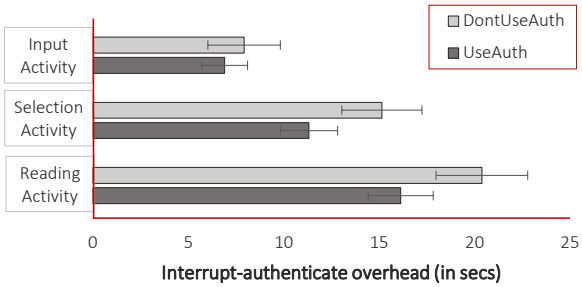


Figure 3: Interrupt-authenticate overhead for different activities (error bars represent ±95% confidence interval)

ing the IA session, participants witnessed 222 FRs in total. For the field study, on average our app logged 104 `ACTION_SCREEN_ON` events (*median = 57, sd = 65*) and 63 `ACTION_USER_PRESENT` events (*median = 42, sd = 28*) per participant per day. In total 10,608 `ACTION_SCREEN_ON` events and 6,420 `ACTION_USER_PRESENT` events for 34 participants were logged across three days. During the field study, participants also provided feedback against 693 FRs (98, 214, and 381 for low, medium and high operating threshold, respectively) and dismissed 42 authentications and feedbacks.

For qualitative analysis of the semi-structured interviews, the researchers coded all participant responses using the grounded theory approach [14] with meetings to achieve consensus. For test statistics, we use a t-test when comparing continuous data between subjects (such as between *UseAuth* and *DontUseAuth* or between *UsePIN* and *UsePAT*). We use a paired t-test when comparing continuous data for the within subjects condition (IA and *Non-IA*). We use a chi-square test for participants' responses to categorical Likert scale questions.

5.1 Usability Evaluation of IA

5.1.1 U1: Effect of IA on overall task completion time and authentication overhead

Overall task completion time is the total time to complete all 30 tasks including EA authentications or IA interrupt-authenticates if present. We calculate the increase or decrease in this time for each individual participant for their IA session compared to their non-IA session. This relative measure compensates for inter-participant differences due to reading level, motor skills, etc. The time differences are aggregated by *UseAuth* and *DontUseAuth* participants (Figure 2). Intuitively, the IA session should take less time as compared to the non-IA session for *UseAuth* participants due to fewer authentications, and more time when compared to the non-IA session for *DontUseAuth* participants. For *UseAuth* participants, the overall task completion time on average decreased by 100 seconds (*median = -103; mean = -101; sd = 59*). A paired t-test between the completion times of the IA and non-IA session for the *UseAuth* participants indicates that they are significantly different ($t = -3.6$, $p = 0.01$). The overall task completion time for *DontUseAuth* participants increased by 120 seconds on average (*median = 107; mean = 122; sd = 52*) for the IA session. A paired t-test between the completion times of the IA and non-IA session for the *DontUseAuth* participants indicates that they are significantly different ($t = 5.2$, $p = 0.014$).

We also evaluate the interrupt-authenticate overhead, defined as additional time taken for IA interrupted tasks compared to non-interrupted tasks. For our tasks, interrupt-authenticate overhead is the difference between the average completion times of an activity, with each activity type analysed separately. Figure 3 shows that on average, *DontUseAuth* participants had an interrupt-authenticate overhead of 8, 15, and 20 seconds for the interrupted input, selection, and reading activities, respectively. Similarly, on average, the *UseAuth* participants had an interrupt-authenticate overhead of 7, 11, and 16 seconds for the input, selection, and reading activities, respectively. A t-test for interrupt-authenticate for each activity reveals that the difference is not significant for the input activity between IA interrupted tasks and non-interrupted tasks ($t = 1.6$, $p = 0.11$), but the difference is significant for the selection ($t = 7.8$, $p = 0.002$) and the reading activity ($t = 10.5$, $p = 0.002$).

Discussion: While these results indicate that IA imposes an interrupt-authenticate overhead for the individual interrupted tasks, the total completion time decreased by 7.1% for *UseAuth* participants because they did not authenticate for every task. For *DontUseAuth* participants, we observe 8.8% increase in the total completion time due to interrupt-authenticates. It should be noted that the performance gains (or losses) will be more pronounced when the number of device usage sessions increases. The interrupt-authenticate overhead was primarily due to the unpredicted or sudden "lock-out" and the context switch as pointed out by the participants:

> *"It pops-out very suddenly... in between the tasks at times. I got tensed because I was worried about completing the task without making the pop-up appear... getting pop-up in the middle of task was quite distracting"* (P10)

> *"I generally lost my train of thought when it popped up that authentication"* (P37)

5.1.2 U2: Effect of IA on the task error rate

We classified errors using simple correctness checks built into the apps. An error occurred when: entered numbers mismatched in the input activity; incorrect answers were provided to a question in the reading activity; or a target word was missed or a non-target word was selected in the selection activity. We calculated the error rate separately for the 222 interrupted tasks from the IA session and for the 222 uninterrupted tasks located at the same task index from the EA session (Figure 4).

A t-test indicates that the differences in error rates across uninterrupted and interrupted tasks are not statistically significant for input ($t = -1.0$, $p = 0.69$), selection ($t = 1.5$, p

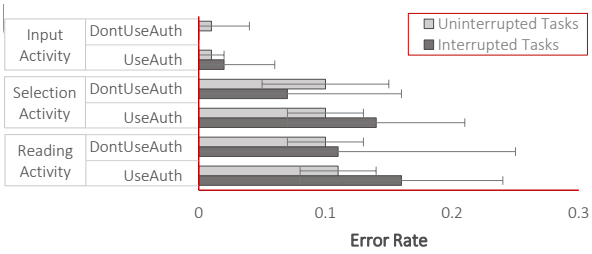


Figure 4: Error rate between interrupted tasks from the IA session and corresponding uninterrupted tasks from the non-IA session (error bars represent ±95% confidence interval)

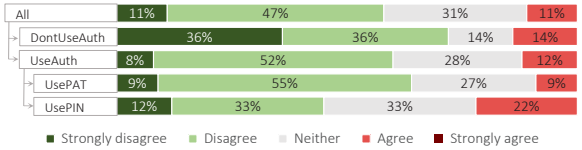


Figure 5: Responses for "Do you agree with the statement 'I think this method is annoying'?"

= *0.32*) and reading activities (*t = -1.8, p = 0.30*). There is no evidence that the interrupt-authenticate model increases the error rate.

Discussion: Our results agree with Bailey and Konstan's [1] findings for personal computers where interruptions did not increase the error rate of interrupted tasks. However, they also found that the expectancy of interruptions caused more errors overall (due to a higher load on the cognitive resources). While our participants complained about the unpredictability of interrupt-authenticates, a paired t-test reveals that there is no significant difference between the error rates of the IA and non-IA session (*t = 0.84, p = 0.4*).

5.1.3 U3: Effect of fewer but less predictable authentication interrupts on annoyance

After the in-lab sessions, participants answered survey questions regarding annoyance. The first question asked if they thought the overall experience of IA was annoying (see Figure 5). Overall, 58% did not say IA was annoying, 11% considered IA as annoying while the rest were neutral. Furthermore, significantly fewer *UseAuth* participants (12% less) thought IA was not annoying compared to *DontUseAuth* participants ($\chi^2(1) = 5.1, p = 0.02$). There is also a significant difference in annoyance for participants based on the type of EA currently used: significantly more *UsePIN* participants (22% more) found IA to be annoying compared to *UsePAT* participants ($\chi^2(1) = 4.09, p = 0.04$). We suspect that IA's lower perceived level of protection by *UsePIN* participants (discussed in § 5.2.1) and consequent low utility is responsible for this.

The second question asked participants how annoying the IA *interrupt-authenticates* were (see Figure 6). Overall, 35% of the participants found them to be annoying (32% somewhat annoying and 3% very annoying), 44% found them to be tolerable and 21% found them to be not annoying. Furthermore, significantly more *UseAuth* participants (28% more) found interruptions to be annoying as compared to *DontUseAuth* participants ($\chi^2(1) = 9.4, p = 0.002$). We did not find evidence for increased anxiety (which could be

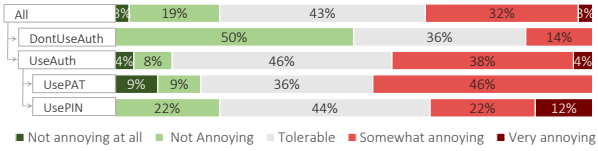


Figure 6: Responses for "How annoying were the interruptions for authentication?"

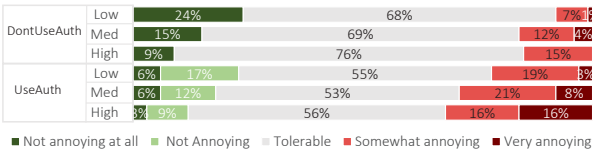


Figure 7: Annoyance against three operating thresholds from the in-situ feedback survey of the field study

linked to annoyance) during the lab sessions: a pair-wise t-test does not indicate any changes in anxiety across the participants for the IA and the non-IA session (*t = 10.6, p = 0.82*).

During the field study, we subjected participants to three different operating thresholds corresponding to 5%, 10% and 20% FR rate with a goal to determine an *acceptable threshold* (operating thresholds with "tolerable" or better annoyance ratings). The feedback of participants for annoyance across these FR rates is provided in Figure 7. *DontUseAuth* participants found interrupt-authenticates to be more acceptable for different thresholds as compared to *UseAuth* participants. More specifically, for low, medium and high FR rates, interrupt-authenticates for *DontUseAuth* participants were significantly more likely (14%, 13% and 17% more) to be acceptable as compared to *UseAuth* participants ($\chi^2(1) = 11.4, p < 0.001$), ($\chi^2(1) = 8.2, p = 0.004$) and ($\chi^2(1) = 13.2, p < 0.001$), respectively. Figure 7 also illustrates responses in terms of the proportion of interrupt-authenticates that are annoying between low-medium and medium-high thresholds, while differences between medium-high thresholds are negligible. More specifically when *DontUseAuth* participants were subjected to the low threshold, interrupt-authenticates were significantly more acceptable (8% more) as compared to the medium threshold ($\chi^2(1) = 4.7, p = 0.03$). On the other hand, for *DontUseAuth* participants the difference in terms of proportion of acceptable interrupt-authenticates between medium-high threshold was insignificant — 84% vs. 85% ($\chi^2(1) = 0.07, p = 0.78$), respectively. These observations for inter-threshold level correlations across *DontUseAuth* participants were also true for *UseAuth* participants.

Discussion: Although the majority of participants were not annoyed with IA, it is clear that interrupt-authenticates can cause moderate levels of annoyance, more so for users who currently use EA. During the qualitative interviews, we asked the participants for the cause of this annoyance and 10/37 participants indicated the unpredictability of the interrupt-authenticate as the reason:

> *"I think it is a little bit annoying because there is a little stress [when] you don't know what will happen"* (P15)

> *"I am ready to enter a password before I start doing anything whereas for implicit authentication it catches*

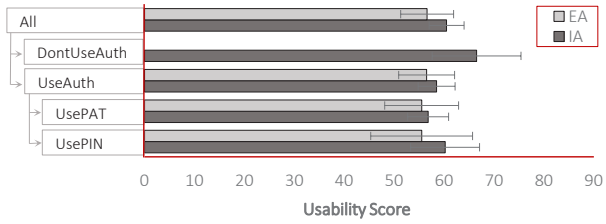


Figure 8: Average system usability scale (SUS) score for IA and non-IA sessions (±95% confidence interval)

me off-guard " (P14)

4/37 participants were more annoyed due to a perceived error at the part of IA:

"It is the unpredictability of it... I know that I have to enter my PIN every time and this becomes annoying... it would be frustrating because you don't know what was wrong that you did, with PIN you know because it is something wrong that you entered " (P14)

"It sure was annoying. I use my phone a lot when I am watching TV and at times my device turns off due to inactivity. That's my fault and [it] is understandable but when this interrupts me, I think that the operating system is faulty or something" (P37)

For the field study, eight *UseAuth* participants had to use IA in addition to their EA scheme despite their preference to replace their current EA scheme with IA. Two of these participants mentioned that the cause of annoyance was "redundant authentications":

"I felt like it was a lot of work with two PINs. At times I would confuse which one was which and then had to re-enter it. That made it more annoying" (P20)

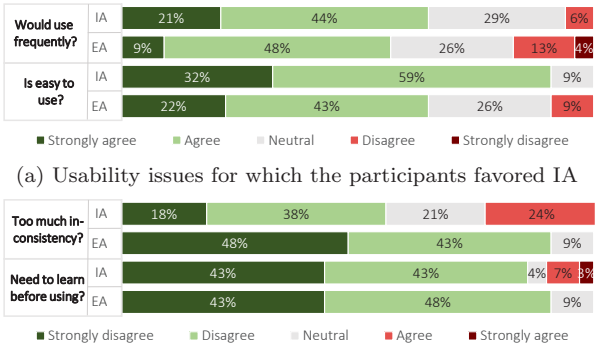
"I already use a password and after that it was quite annoying to use a second one. It seemed redundant" (P32)

5.1.4 U4: Overall usability of IA

The SUS scores for IA and EA are provided in Figure 8. Since we used nine questions from SUS, we report the usability scores out of a total of 90. A higher SUS score indicates that a system is more usable. A score is unavailable for the non-IA session of the *DontUseAuth* participants because they did not authenticate in that session. A t-test does not indicate any significant differences for the SUS scores between EA and IA (t = -2.4, p = 0.31). Similarly, a t-test does not indicate any significant differences between *DontUseAuth* and *UseAuth* participants for the IA session (t = 4.7, p = 0.1).

Discussion: While IA did not outperform EA on SUS, there are interesting differences for the individual SUS questions between IA and EA. Figure 9a shows that significantly more participants (8% more) indicated that they would like to use IA frequently as compared to EA ($\chi^2(1) = 5.1$, $p = 0.02$). Also significantly more participants (26%) thought that IA was easier to use compared to EA ($\chi^2(1) > 100$, $p < 0.001$). Supporting comments for the ease of use:

"I like that it really can be easy if your phone is effective at recognizing you and doesn't ask you to enter the password" (P21)



(a) Usability issues for which the participants favored IA

(b) Usability issues for which the participants favored EA

Figure 9: Responses to the individual SUS questions

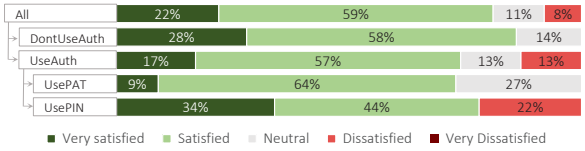


Figure 10: Responses for "How satisfied are you with the overall level of protection that is provided?"

"Actually I found it pretty easy to use and it wasn't bothersome" (P21)

As would be expected, significantly more participants (24% more) thought that IA was more inconsistent than EA ($\chi^2(1) = 64$, $p < 0.001$) (Figure 9b). Furthermore, significantly more participants (10% more) thought that they need to learn more about IA ($\chi^2(1) = 15$, $p < 0.001$). In § 5.2.3, we discuss the learnability issue in detail.

5.2 Security Perceptions of IA

5.2.1 S1: Perceptions of IA security properties

Participants were made aware of the security properties of IA including FAs, FRs and the detection delay through the briefing video and the lab-based experiment. We then asked participants how satisfied they were with the overall level of protection that was provided by IA (see Figure 10). Overall, 81% of the participants were satisfied (22% very satisfied and 59% were satisfied) with IA, 8% were not satisfied and the rest were undecided. *DontUseAuth* participants were significantly more (12% more) likely to be satisfied with the level of protection that was provided by IA as compared to *UseAuth* participants ($\chi^2(1) = 9.5$, $p = 0.001$). Two *UsePIN* participants were not satisfied while three *UsePAT* participants were undecided about the overall level of protection that was provided by IA.

We also asked participants regarding their perceived level of protection against different adversaries, device states and tasks (Figure 11). Overall, 12%, 6%, 3% and 15% of the participants were not satisfied with the level of protection that was provided against coworkers, spouse, friends and strangers, respectively. In terms of different device states, 21%, 6% and 3% of the participants were not satisfied with the provided level of protection if their device was lost in a public location, unattended at work and unattended at home, respectively. Finally, 33%, 12% and 9% of the par-

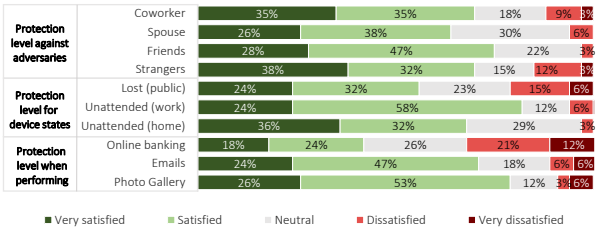


Figure 11: Security perception responses according to different adversaries, device states and tasks

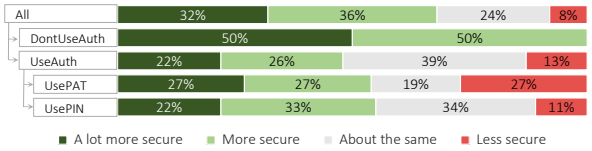


Figure 12: Responses for "How secure is this method as compared to your current authentication method?"

ticipants were not satisfied if IA was protecting their device while there was a banking app, email app and photo gallery app on their device, respectively.

Discussion: Overall, we found that participants were satisfied with the level of security that was provided by IA. However, 18/37 participants showed some concerns regarding FAs, detection delays and possible mimicry attacks in IA. We now shed some light on the concerns based on the participants' comments. The non-zero FA rate was a concern for 8/37 participants:

> *"I'm not sure what will happen when it is lost. It will depend on who picks it up and I may get unlucky if his behaviour is the same as mine"* (P30)

> *"No one can use the phone without entering the PIN but here someone 'can' use it"* (P25)

5.2.2 S2: Perceptions of IA security vs. current method

We asked participants how secure they thought IA was compared to the authentication method that they currently used (Figure 12). All *DontUseAuth* participants perceived IA to be more secure and 87% of *UseAuth* participants thought that IA was at least as secure as their current method or more secure. Only 13% of *UseAuth* participants perceived IA to be less secure due to the security concerns discussed in the previous section.

Discussion: Clarke et al. [5] and Crawford and Renaud [6] found that 92% and 73% of their participants considered IA to be more secure as compared to the traditional authentication schemes, respectively. On the other hand, only 48% of our *UseAuth* participants thought that IA was more secure as compared to their authentication schemes. Our results are not consistent with the previous findings. In the original papers, Clarke et al. [5] and Crawford and Renaud [6] do not mention briefing the participants regarding the IA limitations. We suspect that this difference in results is due to the increased knowledge of our participants about the limitations of IA.

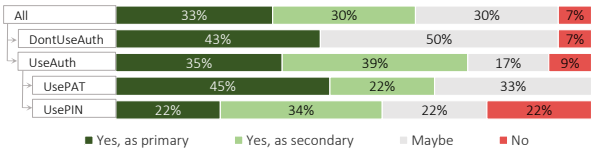


Figure 13: Responses for "Would you use IA?"

5.2.3 S3: Willingness to Adopt IA

We asked participants how willing they were to use IA with four choices: (i) Yes, I would replace my current scheme with IA; (ii) Yes, I would use it in addition to my current authentication scheme; (iii) I may use it; and (iv) No, I will not use it. The purpose of introducing a spectrum of answers was to understand the various types of authentication needs that IA might be able to satisfy. The response of participants is provided in Figure 13. Overall 63% of participants were interested in using IA either as a primary (33%) or a secondary (30%) authentication mechanism. On the other hand, 30% of participants were not sure whether they would use IA and 7% did not want to use IA. A further breakdown across authentication preferences indicates that *DontUseAuth* participants were significantly more likely (37% more) to be unsure about IA adoption as compared to *UseAuth* participants ($\chi^2(1) = 19,\ p < 0.001$). Interestingly, we found 4/7 participants who were not adequately satisfied with IA's level of protection (S2) were still interested in using it.

Discussion: Although 63% of our participants were willing to use IA as a primary or secondary authentication method, this is less than the 90% willingness to adopt IA that Crawford and Renaud [6] found. This difference is likely due to Crawford and Renaud providing a binary yes-or-no option rather than the spectrum of answers we provide. The interview provides rationale for the participants' choices. 6/14 *DontUseAuth* participants were interested in adopting IA because they thought that it provided convenience and protection which was better than no authentication:

> *"It seemed easier than entering a password and more secure than not using anything"* (P30)

> *" Instead of forcing me to enter the password every time, it offers me not to enter a password which is my current preferred level of security and it provides additional security on top"* (P14)

On the other hand, seven *DontUseAuth* participants who said that they may use IA wanted to test it before making up their minds, for example:

> *"I would give it a shot for a month and if I see that it is getting a lot better I will like using it."* (P12)

> *"I have only used it once so I am not sure. I will have to use it for a longer duration and would like to test it on other people too"* (P7)

One *DontUseAuth* participant who did not choose to use IA did so because she had nothing to protect on her device:

> *"If I had work related data or anything else on my device that needed protection, I would use it. Right now I don't have anything that needs protection"* (P36)

UseAuth participants who chose to replace their EA scheme with IA did so because they felt it was more convenient: saves time (6/8) or has fewer authentications (2/8). They seemed to understand the associated risks but thought the trade-off was reasonable:

> *"Because in past months I never had a non-approved access to my phone and in the seven months I have entered my PIN thousands of times and it will be less annoying to use IA"* (P11)

> *"Even with the disadvantages [perceived low level of protection], I think I like the less number of authentications, given that I carry my phone with me"* (P8)

UseAuth participants who said they would use IA as a secondary authentication did so to have an additional layer of security (4/9) or to test it further (5/9):

> *"I would be interested since I work at the community center and at times I have to leave it at places and then I have to worry about it."* (P6)

> *"It would be beneficial for spouse because they would know your password by asking it or see you type a lot but they won't know your pattern [behaviour]"* (P14)

> *"I think just because of my unfamiliarity with it, passwords, I am accustomed with it, but perhaps the more I use it, the more I will trust it"* (P32)

> *"I am so used to typing something in that I think it will take me a while to feel comfortable that authentication has occurred as oppose to you know when you turn it on and enter it"* (P19)

Finally, *UseAuth* participants who said no to using IA had concerns related to its detection delay (1/2) or they felt that EA was more suitable for them (1/2):

> *"Not the best for adoption because people would start looking at my photo gallery before it would lock them out."* (P16)

> *"I think the current system that I have is enough to deter strangers and for the cases when the phone gets stolen "* (P28)

6. DISCUSSION & DESIGN IMPLICATIONS

Our results suggest barriers for IA adoption and deployment along with associated design implications.

6.1 Mitigating the Effects of Interruptions

Overall, participants ranked IA higher than EA in terms of ease of use and the majority of participants reported that IA and interrupt-authenticates were at least tolerable. However, interrupt-authenticates did increase task completion time and several participants, especially those already using EA, felt the interruptions were annoying. Their comments suggest this was largely due to the unpredictability of interrupt-authenticates and the context-switch. Mitigating the negative effects of these necessary interrupt-authenticates remains a challenge for IA. Two participants suggested authentication without interruption by using the front camera of the device to perform facial-recognition. Similarly, a participant who was interested in using IA as a secondary mechanism to protect from misuse by friends or family members suggested IA took a picture of the perpetrator and email it to her. Two participants also suggested that instead of instantly locking the user out, the IA scheme could display an authentication screen in a smaller window on the screen and allow the user to choose the best moment to context-switch and authenticate:

> *"PINs were really annoying a lot of times. I would forget what I was about to say. I was wondering if there was a mechanism where it could indicate that it was locking on me in three, two, one and show me a small screen on the side to authenticate in parallel."* (P8)

It is important to understand that interruptions serve a purpose beyond authentication. Supporting the findings of Crawford and Renaud [6], some "annoyed" participants indicated that while interrupt-authenticates were annoying, they were necessary to indicate that IA was working and arguably contribute to a perceived sense of security:

> *"I guess it was frustrating that it kicked me out, but I could deal with that... and if it did ask for the PIN just knowing that the phone will be secure, it comes down to that"* (P22)

> *"Yeah the interruptions are annoying and I guess then I have to say to myself, practically it is not good but as soon as I see it, I know that it is protecting me"* (P5)

Balancing the need for interruptions with potential annoyance is a design challenge. The alternate authentication methods discussed above could be one approach, but the visual design of the authentication screen (e.g. choice of colour and language) as well as the timing of the interruption (e.g. postponing for non-sensitive tasks, slow fade in) are critical choices.

6.2 Opaque Deployment of IA

Our IA deployment was hidden as a background service, so participants were essentially unaware of its operation until a FR interrupt-authenticate occurred. While these interruptions currently serve to indicate that IA is working (see above), as IA detection algorithms evolve and FR rates decrease, these interruptions will become very infrequent and thus IA will be more opaque. Furthermore, operating at a relaxed operating threshold also reduces the number of FRs and the consequent visibility of IA. The background operations (opacity) of IA will raise concerns like:

> *"So looking at.. I see there is no lock. Sometimes I felt... the lock exists or not? When PIN is used, I know [it] every time. Now there is no way to discriminate if someone has hacked my phone and removed the lock. PIN is a secure feeling that the phone is safe"* (P31)

> *"The main problem for me is that if I am unaware that what I am doing is authenticated then I don't know if my device is secure. With this [IA] there really is no way of knowing that I have been authenticated when I open an app... How do I know it is not a fluke. Maybe it is no longer running and protecting, how do I know"* (P19)

The concerns of users regarding the background deployment of IA have never been raised in existing literature. Since these issues arise due to their inability to tell whether IA is protecting their device, simple UI changes may be able to address these. For example, an indicator on the status

bar can be used to indicate the current status of IA scheme. While such an indicator can keep the users up-to-date and act as a deflector against potential adversaries, it may also notify adversaries and enable them to launch highly focused attacks to gain access to the target data before being locked out. For IA deployment, the design and control of an IA status indicator needs to be studied closely and respective trade-offs need more exploration.

6.3 Operating Threshold Customization

The in-situ feedback screen (Figure 14d) provided participants with an option to adjust the operating threshold during the field study, and we asked them about their experiences with this functionality. 17/34 participants indicated that they found the customization capability to be useful. A common explanation was that they reduced the operating threshold when texting or when at home. 5/34 participants indicated that they always set it to high to get maximum protection (4/5 belonged to *DontUseAuth*). 12/34 participants never adjusted the threshold during the field study and relied on the value chosen by the IA scheme.

These specific results have some limitations since any operating threshold customization in our app was temporary – the threshold was set to a predetermined value each day of the three-day field study. Nevertheless, participant comments indicate that there is a need to explore the various customization options for IA (such as the trade-off between FA and FR; and the detection delay and FR). For example:

> *"I think threshold selection bar would be a useful function. I feel having that is more choice and useful."* (P21)

The threshold customization interface needs to communicate the security and annoyance trade-off for the chosen threshold so users can make informed decisions.

7. LIMITATIONS

Our study has reasonable limitations due to the inclusion of human subjects: the scope is limited to people willing to participate; it contains self reported and subjective views; participants might be inclined to provide favorable responses to the researchers; and the known limited duration of the field study might have made participants more optimistic about their annoyance. Since these are not easily preventable, we focus on limitations specific to our study:

1. We use a pseudo-IA scheme to strictly control FR rates and to circumvent restrictions on Android event data collection. As a result, it was possible for participants to witness unexpected behaviour of IA (for instance, they may get a FR or a TA for what they felt was the exact same sequence of touch input).
2. In the field study, all participants of the *UseAuth* group evaluated IA as a secondary authentication scheme (including those who indicated they would replace their EA scheme with IA). This resulted in multiple authentications during a single session (the system EA first, then an IA interrupt-authenticates), which may have contributed to feelings of annoyance. However, the deployment of IA as a primary authentication scheme in the field was not possible due to security and privacy issues.
3. When a FR occurred in the field study, participants had to authenticate and provide in-situ feedback. Although we designed the feedback pop-up to be simple to complete, it may have increased annoyance. Since participants had the option to dismiss the authentication and the feedback pop-up in extreme situations. Only 6% of the pop-ups were dismissed, but this may have slightly skewed results by underreporting annoyance.
4. In this study, we did not compare IA with biometric-based EA schemes, like fingerprint- or facial-recognition schemes. These alternative biometric-based authentication schemes have the same limitations of EA schemes discussed in § 1 and have usability limitations of their own [2]. A part of our future work is to validate our findings for these alternative EA schemes.

Limitations 1 and 2 are attributed to the pseudo-IA scheme, but this is a reasonable trade-off for the advantages of using pseudo-IA for strict control of FR rates and elimination of confounds from performance idiosyncrasies of specific IA algorithms. Limitation 3 is a commonly accepted trade-off for benefits from gathering in-situ feedback. Limitation 4 must be considered in light of the fact that our study compares IA to no authentication and dominant forms of EA (PIN and pattern): these arguably form lower and upper baselines for usability and security perception.

8. CONCLUSION

Our two-part study on IA usability and security perceptions provides empirical evidence for the "human side" of IA. In terms of performance, the interrupt-authenticate model may impose overhead for individual authentications, but it increases amortized task performance without affecting the error rate. For usability perceptions, there is no significant difference between IA and EA for SUS and 26% more of our participants agreed that IA was more convenient. However, annoyance is a potential issue with IA with 35% of the participants who found interrupt-authenticates annoying. For security perception, detection delay and FAs were issues for 27% and 22% participants respectively, and 11% of our participants were concerned about the feasibility of mimicry attacks. Yet, participants who currently use explicit authentication perceived IA to be more secure, or at least as secure as their current authentication method. Perhaps, most encouraging is that 63% of our participants were interested in adopting IA and a further 30% were interested in trying IA out with possibility of adoption. Based on insights gained from post-study interviews, we propose design implications that may reduce annoyance and increase security perception even more. Overall, our findings provide supporting evidence for earlier work's [25] postulation: IA is indeed a meaningful approach with a reasonable trade-off in terms of usability and security.

9. ACKNOWLEDGMENTS

We thanks the anonymous reviewers and Tao Wang for their valuable comments. We also thank Google and Ontario Research Fund for their support.

10. REFERENCES

[1] B. P. Bailey and J. A. Konstan. On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in Human Behavior*, 22(4), 2006.

[2] C. Bhagavatula, B. Ur, K. Iacovino, S. M. Kywe, L. F. Cranor, and M. Savvides. Biometric authentication on iphone and android: usability, perceptions, and influences on adoption. In *Workshop on Usable Security*, 2015.

[3] C. Bo, L. Zhang, X.-Y. Li, Q. Huang, and Y. Wang. Silentsense: silent user identification via touch and movement behavioral biometrics. In *19th Annual International Conference on Mobile Computing & Networking*. ACM, 2013.

[4] J. Brooke. SUS – a quick and dirty usability scale. *Usability Evaluation in Industry*, 189(194), 1996.

[5] N. Clarke, S. Karatzouni, and S. Furnell. Flexible and transparent user authentication for mobile devices. In *Emerging Challenges for Security, Privacy and Trust*. Springer, 2009.

[6] H. Crawford and K. Renaud. Understanding user perceptions of transparent authentication on a mobile device. *Journal of Trust Management*, 1(7), 2014.

[7] B. Draffin, J. Zhu, and J. Zhang. Keysens: Passive user authentication through micro-behavior modeling of soft keyboard interaction. In *Mobile Computing, Applications, and Services*. Springer, 2014.

[8] S. Egelman, S. Jain, R. S. Portnoff, K. Liao, S. Consolvo, and D. Wagner. Are you ready to lock? In *ACM SIGSAC Conference on Computer & Communications Security*. ACM, 2014.

[9] T. Feng, J. Yang, Z. Yan, E. M. Tapia, and W. Shi. Tips: Context-aware implicit user identification using touch screen in uncontrolled environments. In *15th Workshop on Mobile Computing Systems and Applications*. ACM, 2014.

[10] T. Feng, X. Zhao, B. Carbunar, and W. Shi. Continuous mobile authentication using virtual key typing biometrics. In *12th International Conference on Trust, Security and Privacy in Computing and Communications*. IEEE, 2013.

[11] J. Frank, S. Mannor, and D. Precup. Activity and gait recognition with time-delay embeddings. In *AAAI Conference on Artificial Intelligence*, 2010.

[12] M. Frank, R. Biedert, E. Ma, I. Martinovic, and D. Song. Touchalytics: On the applicability of touchscreen input as a behavioral biometric for continuous authentication. *IEEE Transactions on Information Forensics and Security*, 8(1), 2013.

[13] C. Giuffrida, K. Majdanik, M. Conti, and H. Bos. I sensed it was you: Authenticating mobile users with sensor-enhanced keystroke dynamics. In *Detection of Intrusions and Malware, and Vulnerability Assessment*. Springer, 2014.

[14] B. G. Glaser and A. L. Strauss. *The discovery of grounded theory: Strategies for qualitative research*. Transaction Publishers, 2009.

[15] M. Harbach, E. von Zezschwitz, A. Fichtner, A. De Luca, and M. Smith. It's a hard lock life: A field study of smartphone (un) locking behavior and risk perception. In *Symposium on Usable Privacy and Security*, 2014.

[16] E. Hayashi, O. Riva, K. Strauss, A. Brush, and S. Schechter. Goldilocks and the two mobile devices: going beyond all-or-nothing access to a device's applications. In *Symposium on Usable Privacy and Security*. ACM, 2012.

[17] H. Khan, A. Atwater, and U. Hengartner. A comparative evaluation of implicit authentication schemes. In *Recent Advances in Intrusion Detection*. Springer, 2014.

[18] H. Khan, A. Atwater, and U. Hengartner. Itus: an implicit authentication framework for android. In *20th Annual International Conference on Mobile Computing & Networking*. ACM, 2014.

[19] L. Li, X. Zhao, and G. Xue. Unobservable reauthentication for smart phones. In *20th Network and Distributed System Security Symposium*, volume 13, 2013.

[20] Lookout Blog. Sprint and lookout consumer mobile behavior survey. http://blog.lookout.com/blog/2013/10/21/sprint-and-lookout-survey/, Feb. 2015.

[21] M. Muaaz and R. Mayrhofer. An analysis of different approaches to gait recognition using cell phone based accelerometers. In *International Conference on Advances in Mobile Computing & Multimedia*. ACM, 2013.

[22] New Scientist. Touchscreen phones know it's you from taps and swipes. http://www.newscientist.com/article/dn24193-touchscreen-phones-know-its-you-from-taps-and-swipes.html, Feb. 2015.

[23] T. Patil, G. Bhutkar, and N. Tarapore. Usability evaluation using specialized heuristics with qualitative indicators for intrusion detection system. In *Advances in Computing and Information Technology*. Springer, 2012.

[24] M. Shahzad, A. X. Liu, and A. Samuel. Secure unlocking of mobile touch screen devices by simple gestures: you can see it but you can not do it. In *19th Annual International Conference on Mobile Computing & Networking*. ACM, 2013.

[25] E. Shi, Y. Niu, M. Jakobsson, and R. Chow. Implicit authentication through learning user behavior. In *Information Security*. Springer, 2011.

[26] W. Shi, J. Yang, Y. Jiang, F. Yang, and Y. Xiong. Senguard: Passive user identification on smartphones using multiple sensors. In *7th International Conference on Wireless and Mobile Computing, Networking and Communications*. IEEE, 2011.

[27] C. D. Spielberger. Manual for the State-Trait Anxiety Inventory STAI (Form Y). 1983.

[28] S. Trewin, C. Swart, L. Koved, J. Martino, K. Singh, and S. Ben-David. Biometric authentication on a mobile device: a study of user effort, error and task disruption. In *28th Annual Computer Security Applications Conference*. ACM, 2012.

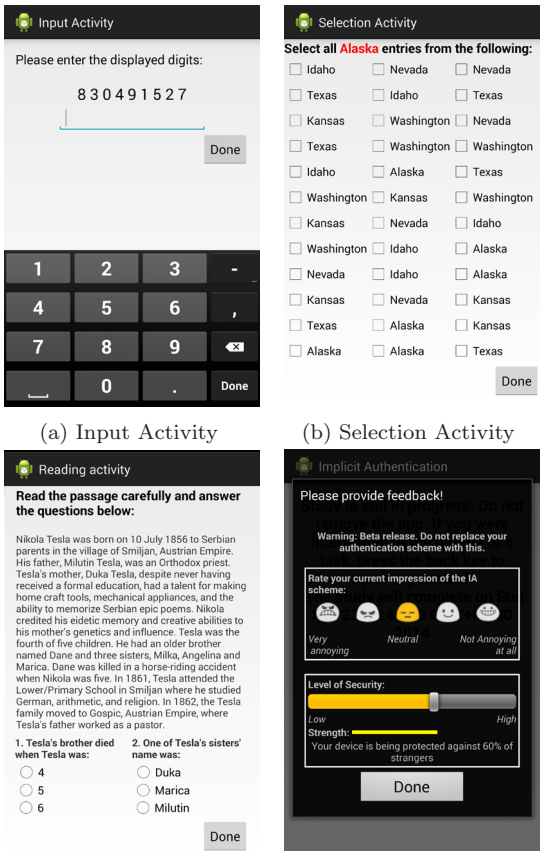
[29] E. Von Zezschwitz, P. Dunphy, and A. De Luca. Patterns in the wild: a field study of the usability of pattern and pin-based authentication on mobile devices. In *15th international conference on Human-computer interaction with mobile devices and services*. ACM, 2013.

[30] R. Werlinger, K. Hawkey, K. Muldner, P. Jaferian, and K. Beznosov. The challenges of using an intrusion detection system: is it worth the effort? In *Symposium on Usable privacy and security*. ACM, 2008.

[31] H. Xu, Y. Zhou, and M. R. Lyu. Towards continuous and passive authentication via touch biometrics: An experimental study on smartphones. In *Symposium On Usable Privacy and Security*, volume 14, 2014.

APPENDIX

A. APPS' ACTIVITY AND FEEDBACK SCREENS



(a) Input Activity (b) Selection Activity

(c) Reading Activity (d) In-situ feedback screen

Figure 14: Apps' screens showing different activities for the lab-based experiment and the feedback screen for the field study

B. PRE-STUDY

For the pre-study screening, in addition to collecting their name, email address, gender, age group, device that they owned (such as Nexus 4, Samsung SIII, LG G2), profession, domain (such as technology, medicine) and how long they have used an Android device, we asked the participants:

1. Which protection mechanism do you use on your smartphone:

   (a) None; (b) PIN (4 digit or more); (c) Password (characters and numbers); (d) Pattern Lock; (e) Face recognition; (f) Fingerprint recognition; (g) Other (please specify)

2. **IF NO AUTHENTICATION** Why do you not use any protection mechanism (choose all that apply):

   (a) It's too much of a hassle / takes time; (b) There is nothing on my phone that I need to hide; (c) No one would care about what's on my phone; (d) In an emergency, others can use my phone; (e) I've never thought about it; (f) Other (please specify)

3. **IF SOME AUTHENTICATION** Which of the following scenarios do you want to protect against (choose all that apply):

   (a) Phone protected if stolen; (b) Phone protected if lost (c) Phone protected if misplaced; (d) Phone protected if left unattended (e) Someone casually picking up the phone; (f) Unwanted disclosure, pranks; (g) Other (please specify)

4. **IF SOME AUTHENTICATION** Which of the following describes you (choose all that apply):

   (a) Unlocking my phone is annoying sometimes; (b) I like the idea that my phone is protected from unauthorized access; (c) It takes too much time; (d) Unlocking my phone is easy

5. **IF SOME AUTHENTICATION** Which of the following attackers do you want to protect from (choose all that apply):

   (a) Coworker; (b) Spouse; (c) Roommate; (d) Own children; (e) Other unwanted individual/stranger; (f) Friends; (g) Other (please specify)

6. Do you sometimes take additional measures to protect your smartphone (choose all that apply):

   (a) None; (b) I leave my phone in a safe place before going somewhere; (c) I conceal my smartphone in my clothes or in a bag (d) I have changed security settings on my device to improve security (such as reduced automatic lock time) (e) I have enabled device encryption on my smartphone (f) Other (please specify)

7. Do you share your smartphone with your friends or family members:

   (a) Never; (b) Rarely (once a month); (c) Occasionally (once a week); (d) Daily

C. POST-SURVEY

The following questions were asked in the post-survey conducted after each in-lab session in which a participant tested an authentication scheme (IA or EA). The questions that were only asked after the IA session are marked [**IA only**].

1. How satisfied are you with the level of protection that is provided against: (*5-point Likert scale "Very satisfied" - "Very dissatisfied"*)

   (a) Coworkers; (b) Spouse; (c) Roommate; (d) Own children; (e) Friends; (f) Strangers

2. How satisfied are you with the level of protection that is provided against the following phone states: (*5-point Likert scale "Very satisfied" - "Very dissatisfied"*)

   (a) Lost at public location; (b) Lost at work; (c) Unattended at work; (d) Unattended at home

3. How satisfied are you with the level of protection that is provided when you are performing following tasks on your Smartphone: (*5-point Likert scale "Very satisfied" - "Very dissatisfied"*)

(a) Online banking; (b) Online shopping; (c) Checking emails; (d) Checking texts; (e) Social networking (FaceBook); (f) Checking photo gallery

4. How satisfied are you with the overall level of protection that is provided? (*5-point Likert scale "Very satisfied" - "Very dissatisfied"*)
5. Do you agree with the statement "I think this method takes a lot of time"? (*5-point Likert scale "Strongly agree" - "Strongly disagree"*)
6. Do you agree with the statement "I think this method is annoying"? (*5-point Likert scale "Strongly agree" - "Strongly disagree"*)
7. Do you agree with the statement "I think this method is tiring"? (*5-point Likert scale "Strongly agree" - "Strongly disagree"*)
8. How annoying were the interruptions for authentication? (*5-point Likert scale "Very annoying" - "Not annoying at all"*)
9. [**IA only**] How annoying were the interruptions for authentication as compared to your current authentication method? (*5-point Likert scale "Very annoying" - "Not annoying at all"*)
10. [**IA only**] How secure this method is as compared to no authentication? (*5-point Likert scale "A lot more secure" - "A lot less secure"*)
11. [**IA only**] How secure this method is as compared to your current authentication method? (*5-point Likert scale "A lot more secure" - "A lot less secure"*)
12. [**IA only**] Would you use this authentication method?
    - Yes, I would replace my current scheme with IA
    - Yes, I would use it in addition to my current authentication scheme
    - I may use it
    - No, I will not use it.

D. SUS SURVEY

The modified SUS form that was completed by participants after each in-lab session in which they tested an authentication scheme (IA or EA). For each question, participants responded on a 5-point Likert scale ("Strongly agree" - "Strongly disagree").

1. I think that I would like to use this method frequently
2. I found this method unnecessarily complex
3. I thought this method was easy to use
4. I think that I would need the support of a technical person to be able to use this method
5. I thought there was too much inconsistency in this method
6. I would imagine that most people would learn to use this method very quickly
7. I found this method very cumbersome to use
8. I felt very confident using the system
9. I needed to learn a lot of things before I could get going with this system

E. SEMI-STRUCTURED INTERVIEWS

Participants were asked the following open-ended questions during the semi-structured interviews:

E.1 Lab-based Experiment

- What did you like about IA?
- What did you dislike about IA?
- Why did you perceive a specific protection level for IA?
- Why do you think IA will provide more/less/same protection as compared to your current scheme?
- Why (or why not) would you use IA?
- **IF NOT SATISFIED WITH IA PROTECTION LEVEL:** Why would you still use IA?
- **IF IA IS ANNOYING:** Why would you still use IA?
- Any particular scenarios where you think IA will be particularly useful/useless?
- **IF INTERRUPT-AUTHENTICATES ARE ANNOYING:** How do you think we can mitigate the annoyance?

E.2 Field Study

- How was your longer term experience of IA?
- Have you changed your opinion about IA? If yes, why?
- How annoying were the interruptions for authentication?
- Which apps were you using on your device when the interruptions were particularly annoying?
- Which apps were you using on your device when then interruptions were not annoying?
- Any particular scenarios where you think IA will be particularly useful/useless?
- Did you use the threshold adjustment bar useful? Why or why not?