Comparing Video Based Shoulder Surfing with Live Simulation

Adam J. Aviv
United States Naval Academy
aviv@usna.edu

Flynn Wolf
University of Maryland,
Baltimore County
flynn.wolf@umbc.edu

Ravi Kuber
University of Maryland,
Baltimore County
rkuber@umbc.edu

ABSTRACT

We analyze the claims that video recreations of shoulder surfing attacks offer a suitable alternative and a baseline, as compared to evaluation in a live setting. We recreated a subset of the factors of a prior video-simulation experiment conducted by Aviv et al. (ACSAC 2017), and model the same scenario using live participants ($n = 36$) instead (i.e., the victim and attacker were both present). The live experiment confirmed that for Android's graphical patterns video simulation is consistent with the live setting for attacker success rates. However, both 4- and 6-digit PINs demonstrate statistically significant differences in attacker performance, with live attackers performing as much 1.9x better than in the video simulation. The security benefits gained from removing feedback lines in Android's graphical patterns are also greatly diminished in the live setting, particularly under multiple attacker observations, but overall, the data suggests that video recreations can provide a suitable baseline measure for attacker success rate. However, we caution that researchers should consider that these baselines may greatly underestimate the threat of an attacker in live settings.

CCS CONCEPTS

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

KEYWORDS

Shoulder Surfing, Mobile Authentication

ACM Reference Format:
Adam J. Aviv, Flynn Wolf, and Ravi Kuber. 2018. Comparing Video Based Shoulder Surfing with Live Simulation. In *2018 Annual Computer Security Applications Conference (ACSAC '18), December 3–7, 2018, San Juan, PR, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3274694.3274702

1 INTRODUCTION

Biometric authentication mechanisms offer considerable promise to smartphone users. However, the protection of unlock authentication still relies on choosing hard to guess passcodes (e.g., PINs and unlock patterns), while not revealing those passcodes to untrusted parties. A common means of attack for gaining access to the passcode is via *shoulder surfing*. In a shoulder surfing attack, an observer attempts to view a victim in the process of entering his/her passcode with the intention of recreating that passcode after gaining possession of the device [26].

ACSAC '18, December 3–7, 2018, San Juan, PR, USA

ACM ISBN 978-1-4503-6569-7/18/12…$15.00
https://doi.org/10.1145/3274694.3274702

The area of shoulder surfing has been the subject of a great deal of work [4, 6–10, 12, 13, 18, 19], for both understanding the threat and proposing mechanisms to prevent it. Of particular relevance to this study (termed "current study"), is the work conducted by Aviv et al. [4] (termed: "prior study"). The prior study examined the shoulder surfing susceptibility of three commonly used unlock authentication mechanisms: 4- and 6-digit PINs, 4- and 6-length Android graphical patterns, and 4- and 6-length Android graphical patterns with the feedback display turned off (lines rendered by the interface between grid points as they are touched by the user). Due to the difficult nature of evaluating shoulder surfing attacks in the field, the goal of the prior study was to establish baselines for shoulder surfing vulnerability in controlled settings that can be used to compare across authentication types and used as baseline for evaluating authentication systems that are designed to defend against such attacks.

To control the analysis, the prior study was conducted using a video-based methodology where the researchers recorded a set of videos with highly controlled factors and then asked participants to view these videos as a simulated shoulder surfing scenario. The data was analyzed to determine shoulder-surfing susceptibility under each condition. The attack rate (how effectively the participant could recall the passcode entered in the video) was the primary metric.

In this paper, we seek to compare the video-based methodology to a similarly controlled live setting. In particular, we are interested in assessing the prior work's following findings relating to the attack success rate.

- Longer authentication lengths (e.g, 4-digit vs. 6-digit PINs) are less vulnerable.
- PIN authentication is less vulnerable to the attack compared to patterns with and without feedback lines.
- Removing the feedback lines from patterns decreases the vulnerability to shoulder surfing.
- Multiple observations increases vulnerability.
- Video based evaluation provides a baseline for live, in-person shoulder surfing vulnerability.

Using the raw results of the prior study, we compare the attacker success rates of the live setting to a comparable subset of the video study data. Testing for differences in proportionality, we are *unable* to reject the null hypothesis that the attacker success rate are the same for Android patterns as well as in many of the settings with patterns without feedback lines. This suggests that there is consistency between the results of the video and live simulations. However, the advantage of removing feedback lines previously observed in video simulation is considerably lessened in the live setting. For PINs, we observe significant difference between the video and the live settings, where live attackers performed up to 1.9x

better in some scenarios. Stereo vision seems to greatly improve the reliability of recalling the more complex motions of entering a PIN. Despite this discrepancy, the claim of Aviv et al. of these results forming a baseline is still supported: we never observed a situation by which the live simulation performed worse than a video study when significant differences exist.

We conclude that video studies do provide a reasonable approximation for live simulation of shoulder surfing in settings that involve graphical passwords (but not PINs), like the Android password pattern, and at least a lower-bound on the attack success rate for all tested authentication types (including PINs). However, researchers should consider that this lower-bound may be a significant underestimation compared to the true attack rate in live simulations.

2 RELATED WORK

Mobile authentication and observation attacks. Threats such as shoulder-surfing attacks have been well documented by researchers [4, 27]. Studies have been conducted examining experiences of users who had encountered observation attacks [11] where shoulder surfing was found to be "casual" and "opportunistic." Harbach et al. [14] found that participants only very rarely reported shoulder surfing (0.3% of 1134 sampled events) as an immediate high risk threat when authenticating.

In order to minimize the risk associated with observation attacks, users are known to modify their own usage behaviors when using a mobile device, hiding the device from sight and performing mobile interactions in the pocket or bag, or even shielding the screen [1]. Solutions also exist to obscure screens from third parties [8], to detect the presence of shoulder surfers in a nearby vicinity [20] or to deceive onlookers from data being entered [17, 24]. Attacks have also been simulated by having observers watch video footage of victims entering authentication sequences. Examples include [15] where attacks took place from top and side views. A range of solutions have also been proposed to minimize the likelihood of shoulder-surfing when entering authentication sequences [2]. However, as highlighted by Wiese and Roth [27], it can be difficult to compare the efficacy of these solutions, as the ways in which these systems are studied varies. Furthermore, the outcomes can be difficult to compare and interpret.

Evaluating resistance from shoulder surfing. Many evaluation studies have focused on observing unlock screen interactions where PINs and patterns are entered [4, 15, 22]. Wiese and Roth [27] suggest that conducting such studies are challenging because real-world adversaries are not available for study and must be simulated in one way or another. In contrast to live studies where participants and actors/researchers perform tasks together in person, video simulations have been used to identify susceptibility of on-screen threats [2, 21]. Video recordings offer consistency when presented to multiple users [27], and can also be accessed independent of location. However, research indicates that that the success of adversaries is lower when performing video observations compared to live settings [23, 27]; we make a similar observation here. Prior research also recommends that shoulder surfing attackers should be allowed a number of observations [27] as well as viewing interactions from a range of views [4, 21] and different properties of passcodes [4]. Additionally, the hand position [22] and interaction style when entering data into the device [4] should also be considered. We tested scenarios found to be significant in Aviv et al., following similar procedures.

Overview of Aviv et al. [4]. Aviv et al. considered the lack of a baseline for comparing common unlock authentication mechanisms under the threat of shoulder surfing. As a method of creating such a baseline, the authors used a series of controlled video simulations of a victim entering unlock authentications using several methods. These methods were PINs and Android's graphical pattern unlock, with and without feedback lines present. Additional factors were considered, including the angle of observation, number of observations, the number of recreation attempts by the observer, the hand posture of the victim, phone size, and spatial layout of the passcodes.

The methodology of that experiment was multi-factorial. Participants were selected into one of a number of independent factors (phone type, passcode choice, authentication type, hand posture) and then a set of randomized dependent factors (passcodes, observation angles, number of views, and attempts). For recruitment, the primary results were based off participants on Amazon Mechanical Turk (n = 1173) and participants recruited locally (n = 91), with both groups completing a web survey whereby they viewed videos of authentication and attempted to recreate the passcodes observed.

Using the results, the authors tested the following hypotheses (the **-p** indicates a prior work hypothesis):

- **H1-p**: The type of unlock authentication, PIN pattern with lines, patterns without lines, affects the shoulder surfing vulnerability.
- **H2-p**: Repeated viewing of user input increases the likelihood of a shoulder surfing vulnerability.
- **H3-p**: Multiple attempts to recreate the input affects the likelihood of a shoulder surfing vulnerability.
- **H4-p**: The angle of observations affects shoulder surfing vulnerability.
- **H5-p**: The properties of the unlock authentication, such as length and visual features, affect shoulder surfing vulnerability.
- **H6-p**: The phone size affects shoulder surfing vulnerability.
- **H7-p**: The hand position used to hold and interact with a device affects shoulder surfing vulnerability.

Of those hypotheses, **H1-p**, **H2-p**, **H3-p**, **H4-p**, and **H6-p** were accepted, while **H5-p** was partially accepted, and **H7-p** was rejected. The authors claim that the video studies, generally, can form a reasonable replacement for live simulation, and that at the very least a video study could provide a baseline for shoulder-surfing vulnerability.

3 METHODOLOGY

To investigate the efficacy of video-based recreations for evaluating observation attacks, we recreated the study conducted by Aviv et al. [4] with live participants in a controlled lab environment. We asked participants to position themselves in similar locations to where the cameras were positioned in the prior study. They then attempted to shoulder-surf a victim (played by a proctor). We varied

the type and length of authentication sequences, observation angle, and number of repeated viewing attempts, to determine if these factors impact the success of the attacker. The results were then compared with Aviv et al.'s findings using a comparable subset of the prior data. For simplicity of discussion, we refer to the prior work of Aviv et al. as the *video study* and the results here as the *live study*.

Hypotheses. In particular, we are interested in testing the following hypothesis related to the efficacy of video based shoulder surfing experiments as compared to live settings.

- **H1-r**: Live shoulder surfing confirms accepting prior hypotheses:
  - **H1-p**: The authentication type affects shoulder surfing vulnerability
  - **H2-p**: Repeated viewing affects shoulder surfing vulnerability
  - **H4-p**: The angle of observation affects shoulder surfing vulnerability
  - **H5-p**: The properties of the passcodes affects should surfing vulnerability
- **H2-r**: Video simulation forms a baseline of performance compared to live settings.

3.1 Study Design and Materials

Treatments. The study followed a mixed factorial design, similar to the video study. Independent variables included authentication type (PIN vs pattern) on the Nexus 5 device using the same hand posture/interaction style (one-handed, right thumb input). For dependent variables, we reduced the observation angle to two (left or right) as opposed to the five angles used in prior work. The video study used the variety of angles to simulate different heights, but height variation is naturally present in a live study. We kept the same variables for observations (single observation from one angle, two observations from the same angle, or two observations from different angles), and we used a lab environment for our live study very similar to the set-up to capture videos for the video study (Aviv et al.) (see Figure 1).

There were two notable differences between factors in the video study and the live study. First, we only allowed each participant a single attempt at recreating the passcode. This choice was motivated by results of the video study whereby participants, knowing they would have multiple attempts in advance, actually did worse at the tasks than those that knowingly had one attempt. It was conjectured that participants attempted to "game" the task knowing that they would have multiple attempts at recreating the passcode. As such, we only allowed participants to make one recreation attempt, and this fact was communicated during training.

Another difference in the live study was that passcode recreation occurred using pen-and-paper, as opposed to a simulation of the device used in the video study. This choice was made to simplify the data collection procedures for both proctors and participants.

Finally, as we only tested a subset of the treatments of the prior video study, we only performed our analytic comparisons on a relevant subset of the video study data. In particular, we removed data that included a top angle and reduced the two side angles into a single *left* or *right* setting. Additionally, as the video cannot

Auth. id	Patterns	PINs
0	0145	1328
1	014763	153525
2	1346	159428
3	136785	1955
4	3157	366792
5	4572	441791
6	642580	458090
7	6745	5962
8	743521	6702
9	841257	7272

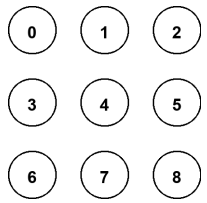


Table 1: Authentication identifiers for patterns and PINs. To the right, the numeric labeling for patterns to contact points.

control for monitor display size, which was a large factor in the prior results, we only used the most ideal viewing conditions, where the reported y-axis pixels were greater than 1800. We believe this restriction provided the *most fair* comparisons possible given the potential uncontrolled factors. We discuss limitations and realism further in Section 4.

Authentication types. We analyzed three authentication types with two different length settings, as used in the video study. These included:

- **PIN**: 4- or 6-length PINs consisting of a set of numbers.
- **PAT**: Android unlock patterns consisting of 4 or 6 contact points *with* the feedback lines present.
- **NPAT**: Android unlock patterns consisting of 4 or 6 contact points *without* the feedback lines present.

While the PIN interaction display is as one expects, the presence or absence of grid pattern feedback lines is less well known. When a pattern is entered with feedback lines (PAT), the display will show connecting lines on the screen between grid points touched by the user while entering their passcode shape. Alternatively, the connecting lines are not rendered on screen during passcode entry in the without feedback lines (NPAT) pattern display, although the user must still contact the appropriate points in the correct order. As identified by Aviv et al. [4] and von Zezschwitz et al. [25], the absence of feedback lines can make it more difficult for an observer to recreate the patterns. As part of **H1-r**, we will make a similar evaluation.

To maintain consistency, we used the same set of patterns and PINs as in prior work (Table 1 and Appendix B.1). The patterns were selected from an online study of self-reported patterns [3], and the PINs were obtained from sequences of digits in leaked password sets, similar to the analysis by Bonneau et al. [5]. Further, the set of passcodes were selected for physical properties, as the layout and sequence of gestures in entry may affect shoulder surfing attack rate. The patterns' spatial properties might affect surfing attacks because an attacker's view from some viewing angles might be obscured for some parts of the touchscreen.

Randomization and counterbalancing. One of the restrictions for performing the study using live participants as compared to video recreation is that the same level of randomization is nearly impractical for the target recruitment size and the set of factors

Order	Auth. id									
a	8	1	0	7	9	2	6	5	4	3
b	0	6	3	8	2	4	9	7	1	5
c	6	0	9	4	8	3	5	1	7	2

Table 2: Orderings of the patterns and PINs in the experiments.

Exp.	Angle(s)									
i	L	R	R	RR	L	LR	RL	R	L	LL
ii	RL	L	LR	R	LL	R	R	L	L	RR
iii	LL	RR	L	R	R	L	R	LR	RL	L
iv	LR	R	R	L	L	RL	RR	L	LL	R

Table 3: Angles used within each experiment, including multiple views with two angles indicated. L=view from left side, R=view from right side.

Auth. id	L	R	LL	RR	LR	RL	one	two-same	two-different
0	3	4	1	1	2	1	7	2	3
1	5	3	1	1	1	1	8	2	2
2	4	3	2	1	1	1	7	3	2
3	3	4	1	1	2	1	7	2	3
4	4	3	1	1	1	2	7	2	3
5	3	4	1	2	1	1	7	3	2
6	2	4	1	2	1	2	6	3	3
7	5	3	1	1	1	1	8	2	2
8	4	3	2	1	1	1	7	3	2
9	3	5	1	1	1	1	8	2	2
total	36	36	12	12	12	12	72	24	24

Table 4: Balancing of observation angles, number of views, for each authentication after 12 participants, Order × Exp.

being considered. As such, we designed a two stage randomization procedure, one for ordering the passcodes and one for ordering the observation angles.

In particular, Table 2 contains three different randomized orders across the passcode. These are labeled Order a, b, and c. Note that the authentication identifiers refer to Table 1. In Table 3 are four randomized orders for observation angles (i, ii, iii, and iv). For each participant, we randomly assigned them a passcode order and an observation angle, producing 12 different randomizations.

At this point, it is important to consider counterbalancing. Selecting randomized orders for passcodes or observations can weight the data improperly. This leads to an optimization problem, and we used a utility function to find a set of randomized orders that would provide (1) sufficient data in each factor for us to perform statistical tests, (2) a roughly equal ratio of data within each factor being compared (4- vs 6-length, auth-type, angle), (3) that each passcode only appears once per viewing, and (4) that within each viewing sequence, per participant, there are roughly an equal number of single and multiple observations. We found a case that nearly met these criteria, as displayed in Table 2 and 3. The weighting is then displayed based on 12 participants in Table 4, leaving us with 72 single-view observations and 48 multi-view observations, 24 from the same angle twice and 24 from two different angles. Additionally, there is equal weighting across angles and viewing (Table 3), and nearly equal weighting across passcodes.

We acknowledge that this counterbalancing is not a perfect weighting, and solving this particular optimization problem is challenging and may not have a solution. However, the resulting counterbalancing compares favorably to the subset of relevant video study data. For PINs, there is nearly an equal number of observations in the one-view and two-view conditions. For PAT/NPAT, there is 50% less observations in one-view condition with a significant proportion necessary for statistical testing, and the two-view conditions for PAT/NPAT are of the same magnitude as the video study (see Table 6).

In total, we were able to run complete trials for 18 participants each for PAT and NPAT, and all of those 36 participants also completed a PIN viewing. The order between PIN and PAT/NPAT for participants was randomized, so that half of the participants completed a PIN trial before doing a PAT/NPAT trial, and the other half completed the protocol in the reverse order, PAT/NPAT then PIN.

3.2 Live Simulation Setup and Coordination

We sought to recreate nearly the same scenario for shoulder surfing as the video study. Namely, we had our victim placed in a sitting position with the participant observer behind the victim, either standing to the right or the left, directed by one of two proctors. These were the same positions where the cameras were located

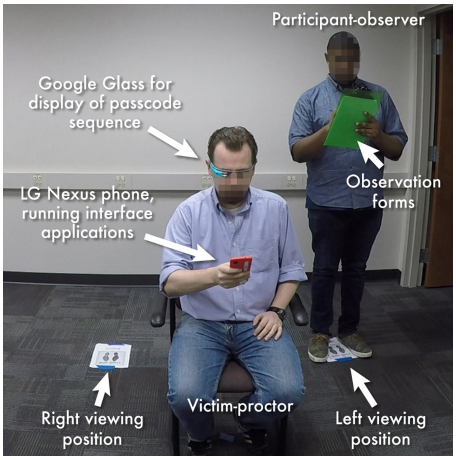


Figure 1: Experimental setup with an observer attacking a victim, a member of a research team. Note the Google Glass displaying the passcode to enter.

(near left and near right views) in the study by Aviv et al. [4]. See Figure 1 for a visual of this arrangement for the live study.

Additionally, for the phone application used to enter the passcodes, we used the same mobile applications as in the prior study, which includes a web-based platform for entering PINs and patterns. Screenshots of those applications are provided in Figure 2.

For patterns with feedback lines, the white tracing lines would follow the user gesture, and once the pattern was entered, it would remain visible on-screen for a half a second before disappearing. The same would be true for the patterns without feedback lines, however, neither the tracing lines nor the contact points of the grid would be rendered on the screen. For PINs, the layout allowed

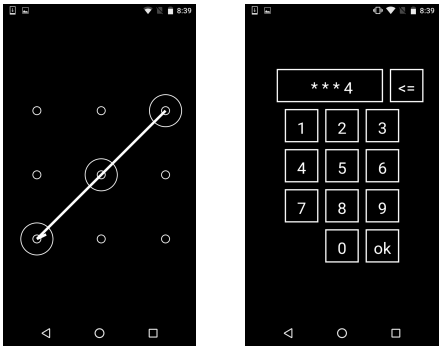


Figure 2: Screenshots of the web-based applications used by the victim entering the passcode. Note, that for the pattern without feedback lines setting, the white trace lines would *not* appear.

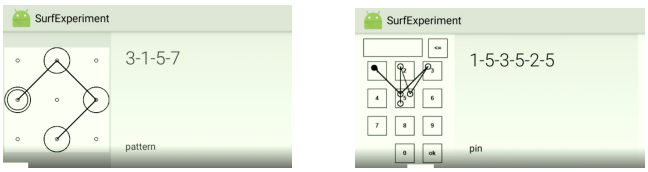


Figure 3: Screenshots of the Google Glass application as viewed by the victim and enter the correct PIN or pattern.

for numeric entry as expected. Once digit keys were selected, the corresponding digits were presented on the interface. These would then fade to a * after a half a second, similar to most mobile PIN entry interfaces.

For the participant observer to record their pattern entry, we used pen and paper. Examples of the observer forms are provided in the Appendix (A.3). The forms had text boxes and mini-diagrams of the application interfaces, so the participants could easily record the observed entry. Participants were asked not to write down the passcodes observed until directed following all observations, which was important for the multiple viewing scenario.

As shown in Figure 1, two pre-marked spots were placed on the floor to direct participants where to stand on the left or right side. The second proctor, following the randomized treatment order, would call out directions to the participant; for example, "one view, from the left" or "two views, first from left and then right." Once the participant was in place for each view, the second proctor would cue the first proctor (playing the victim) to enter a passcode.

At this point, a significant challenge we had to overcome was how to prompt the victim-proctor with the correct passcode to enter without tipping off the participant-observer. Due to the randomization procedures, requiring the victim-proctor to memorize the numerous orderings was not realistic. As a solution, we developed a Google Glass application to guide the victim-proctor through the various passcode orders. Google Glass is a wearable eyeglass display unit that runs on a modified Android OS. It enables one to scroll interactively through images projected onto a viewing screen built into the right eyepiece. Moreover, the small display screen on the Google Glass was not visible to the participant. A screenshot of the Google Glass application is provided in Figure 3.

3.3 Procedure

The replication experiment proceeded in four stages:

(1) Informed Consent and Ante Hoc Questionnaire: All participants were properly informed and consented, as we conducted an IRB approved experiment. Following consent, we asked participants to complete an ante hoc questionnaire that covered basic demographic questions, such as age and gender, as well as questions regarding the participants experience with smartphones, mobile authentication, and sense of risk from shoulder surfing. The subjective response questions were largely intended to orient participants to physical security issues related to the study. The ante hoc questions are found in the Appendix (A.1).
(2) Training: Depending on the set of authentications being observed in the trial run, a training session would include two basic passcodes, the L shape for patterns and the 1234 PIN, to help familiarize the participant with the procedures, how to record on the observation sheets, and where to stand for the trials (similar training was performed in the video study). Additional training on how to fill out the observation form was also provided, which is included in the Appendix (A.3).
(3) Trial: Under the direction of the proctor, the participant conducted 10 observations of either the PAT or NPAT pattern entry, and 10 observations of PIN entry.
(4) Post Hoc Questionnaire: Following the trials, the participant answered a series of post hoc questions related to the challenge of the task and his/her perceived performance thereon. See Appendix A.2 for the set of post hoc questions.

As each participant completed two trials, one for either PAT or NPAT and another for PIN, once the trial stage was over for the first authentication we would return to training for the second authentication. As a way to control for training effects, whereby observing PINs first could increase or decrease performance on observing PAT/NPAT, we ensured that there was an even ratio between the order of the trials. A guide was also followed to ensure that the researchers followed the same steps in the protocol (see Appendix A.4).

3.4 Recruitment

Participants were recruited from university student mailing lists, and paid $5 (USD). In total, we recruited 36 participants, including 10 females. The cohort was predominately aged between 18 to 24 years old. Almost two-thirds of participants used iOS mobile devices. 21 used a fingerprint reader to unlock their phones, and 6 used patterns (we did not ask if feedback lines were turned off). The demographic breakdown, as well as their choice in mobile device and authentication are presented in Table 5.

Additionally presented in Table 5 are the demographics of a comparable set of participants from the prior video study; these participants observed authentication on the Nexus 5 phone in the "in-person" lab setup or the on-line MTurk setup with a screen

		Live			*Video*			
		Male	**Female**	**Total**	**Male**	**Female**	**Neither**	**Total**
Age	18-24	16	4	10 (27.8%)	30	9	0	39 (39.4%)
	25-34	8	5	13 (36.1%)	24	10	1	34 (34.3%)
	35-44	1	1	2 (5.6%)	10	4	0	14 (14.1%)
	45-54	0	0	0 (0.0%)	4	3	0	7 (7.1%)
	55-64	1	0	1 (2.8%)	1	0	0	1 (1.0%)
	65+	0	0	0 (0.0%)	1	2	0	3 (3.0%)
	total	26 (72.2%)	10 (27.8%)	**36**	70 (70.7%)	28 (28.2%)	1 (1.0%)	**99**
Phone	iOS	15	7	22				
	Android	9	2	11				
	Windows	2	0	2				
Unlock	Fingerprint	15	6	21				
	PIN-6	8	5	13				
	PIN-4	10	2	12				
	Pattern	5	1	6				
	None	3	1	4				

Table 5: Demographic, phone usage, and unlock authentication types of participants. For the video study, the subset of comparable data that includes participants in both the "in-person" and "online" settings that had screen resolution greater than 1800px and observed patterns on the Nexus 5 phone.

resolution of at least 1800px in the y-axis, the most realistic setting of the prior work. The breakdown of these two groups are similar, slightly younger overall with about 70/30 gender breakdown.

4 REALISM AND LIMITATIONS

As described in the previous section, we attempted, as best as possible, to recreate the settings of the prior video study in live simulation. Due to the complexities of performing such a process, the study described in this paper had its own set of limitations.

Viewing angles. While we use a similar lab environment for the live simulation to that used in the video study, the participants could not stand in exactly the same position as the cameras due to height differences and the relatively close proximity of the *near* and *far* angles from a given side. We thus reduced the observations to simply *left* and *right* and relied on the fact that our participants naturally vary in height to compensate for the *near* and *far* setting of camera height placement in the prior study.

Victim entry speed. Another recreation challenge is that our victim (a proctor) must enter the authentication sequence many times over at a consistent speed. Clearly, a video ensures consistency here, and so we trained the victim-proctor on the original videos to maintain consistent timings of authentication entry. While there is no guarantee that every participant viewed the authentication at the same rate, we believe this training, and the total number of entries performed by the victim, ensures consistency. Further, the same victim-proctor was used in all data collection.

Subset of conditions. As summarized in Section 3, a subset of the original conditions were used in the live simulation. We kept factors that were shown to be significant in the video study, but also had to remove some that posed usability challenges for the proctor acting as the victim. While the selection process was done carefully to address conditions likely to be important, it was also done for a practical nature of conducting a study with live participants as compared to online. To ensure that we made a fair comparison, we selected a similar subset of the data from the prior study. In particular, we used results from the previous study from participants who had viewing screens of at least 1800px across, who viewed authentication attempts via the Nexus 5 phone with thumb input from the left or right side.

Pen-and-paper attacker recordings. As participants were using pen-and-paper to record their observations during the shoulder surfing attack, some participants were able to use this as an added aid to support recall of the passcodes. For example, some participants were viewed by the proctor mimicking the movements made by the victim-proctor between multiple-view conditions prior to writing down their final observation. While we directed participants to *not* do this during training, it was difficult to stop due to the nature of the task. In the video study, participants were also directed not to use additional aids, such as writing down observations while observing the passcodes, and were required to attest to this. However, it is possible that the attestations were not fully truthful, nor could the researchers verify this as the study was conducted online. As such, as neither study could fully control for this we believe that this provides for a fair comparison.

Ecological validity. Low levels of ecological validity are known to be commonplace among lab-based studies for mobile interactions [16]. Although the method and setting selected for our study cannot approximate the conditions by which shoulder surfing may take place in-the-wild, we designed the study to provide a sense of realism even in a lab-based environment (e.g. victim in seated position similar to attacks taking place while seated on public transport, while seated in a classroom, etc.). However, due to time constraints, conditions such as providing multiple attempts to observe and/or recreate entry, could not be examined. Further study would be needed to widen the range of factors examined, and to identify the applicability of these findings to other types of tasks (e.g. authenticating while ambulatory) or other types of settings (e.g. field-based).

		One-View		*Two-Same*		*Two-Different*	
Auth.	**Length**	**Live**	**Video**	**Live**	**Video**	**Live**	**Video**
PAT	4-len	50/53=94.3%	106/111=95.5%	15/17=88.2%	24/27=88.9%	19/20=95.0%	62/68=91.2%
		$\chi^2 = 0.00, p = 1.00, \delta_{95}[-0.10, 0.07]$		$\chi^2 = 0.00, p = 1.00, \delta_{95}[-0.21, 0.19]$		$\chi^2 = 0.01, p = 0.93, \delta_{95}[-0.11, 0.19]$	
	6-len	45/55=81.8%	74/95=77.9%	18/19=94.7%	23/24=95.8%	14/16=87.5%	78/84=92.9%
		$\chi^2 = 0.13, p = 0.72, \delta_{95}[-0.11, 0.19]$		$\chi^2 = 0.00, p = 1.00, \delta_{95}[-0.15, 0.13]$		$\chi^2 = 0.05, p = 0.82, \delta_{95}[-0.26, 0.15]$	
NPAT	4-len	**48/55=87.3%**	**74/104=71.2%**	18/19=94.7%	17/20=85.0%	16/16=100.0%	49/61=80.3%
		$\chi^2 = 4.37, p = 0.04*, \delta_{95}[0.02, 0.30]$		$\chi^2 = 0.22, p = 0.64, \delta_{95}[-0.14, 0.33]$		$\chi^2 = 2.38, p = 0.12, \delta_{95}[0.06, 0.34]$	
	6-len	35/53=66.0%	58/101=57.4%	**17/17=100.0%**	**15/24=62.5%**	18/20=90.0%	56/78=71.8%
		$\chi^2 = 0.75, p = 0.39, \delta_{95}[-0.09, 0.26]$		$\chi^2 = 6.13, p = 0.01*, \delta_{95}[0.13, 0.62]$		$\chi^2 = 1.95, p = 0.16, \delta_{95}[-0.01, 0.38]$	
PIN	4-len	**89/111=80.2%**	**48/94=51.1%**	**32/33=97.0%**	**12/21=57.1%**	**34/36=94.4%**	**37/65=56.9%**
		$\chi^2 = 18.17, p = 0.00*, \delta_{95}[0.16, 0.43]$		$\chi^2 = 10.98, p = 0.00*, \delta_{95}[0.14, 0.66]$		$\chi^2 = 13.88, p = 0.00*, \delta_{95}[0.21, 0.54]$	
	6-len	**46/105=43.8%**	**17/109=15.6%**	**25/39=64.1%**	**4/17=23.5%**	**31/36=86.1%**	**19/68=27.9%**
		$\chi^2 = 19.16, p = 0.00*, \delta_{95}[0.16, 0.41]$		$\chi^2 = 6.27, p = 0.01*, \delta_{95}[0.11, 0.70]$		$\chi^2 = 29.62, p = 0.00*, \delta_{95}[0.41, 0.76]$	

Table 6: Attacker accuracy results for the live experiment and the video experiment. The view type indicates if the participant provides a single (or one) view or multiple views (two), either from the same angle or different angles of observation. For the video study, only data where screen base resolution > 1800 with left or right views (no top) was considered. A 2-sample test for equality of proportions with continuity correction was used, and the χ^2 statistic, p-value, and 95% confidence interval (δ_{95}) of the difference between the proportions (*live - video*) are reported.

5 RESULTS

As the live simulation used a subset of the variables in the video study (see prior section), we in turn performed comparisons on an appropriate subset of the video study data. We used video data that met the following criteria: one-handed/thumb-input on the Nexus 5 (red) phone, viewed from the left or right angle, and a single recreation attempt. Additionally, we only included video data that was collected with a screen resolution > 1800 pixels, which was identified as the most ideal viewing condition in the prior study [4]. With these reductions, we compared 720 shoulder surfing attempts for the live simulation to a comparable 1,171 attempts in the video study.

5.1 Comparing Attack Rates Across Video and Live Studies

Statistical Procedures. As the results of the experiments for both the live and video study are proportional, either the participant succeeded in recreating the passcode or did not, we compare the results using a *proportionality test* for equality of proportions, which follows a χ^2 distribution. That is, we compare the attacker success rate for the video study to that of the live study using the same conditions, reporting the χ^2 statistic, the two-tailed p value, and the 95% confidence interval (δ_{95}) for the difference between proportions.

In the cases where $p \leq 0.05$, we can conclude that the live study was *not* well modeled by the video study because the proportions of attacker success are significantly different. Similarly if $p > 0.05$ we cannot reject the null hypothesis that the two proportions are the same and thus must conclude that the proportions are more likely measuring the same effect. The confidence interval reports the most likely range of difference between the attacker success rate for the video and live results, but is only relevant when a significant difference is found.

When comparing data across factors with greater than two conditions, we used a χ^2 test for goodness of fit to determine significant differences in attack success rates. Post-hoc analysis is conducted using pairwise comparisons with a Bonferonni correction.

Across tests, while the data is overlapping for some of the factors being examined, we do not normalize/correct p values as we are not attempting to control for type-1 errors across *all* tests. Instead, we are performing exploratory analysis and interested in determining if significant differences may exist and from where they may arise. In post-hoc analysis, as described above, we do correct p values as appropriate as this occurs within a single test with directly overlapping hypothesis.

Authentication Types (H1-r/H1-p). In the prior study, a key finding was that a statistical difference was identified in attacker performance across authentication type. We can perform the same tests by comparing vulnerability to shoulder surfing for the single view conditions; see the first column of Table 6.

We first compare each of the authentications between the video and the live study, irrespective of the authentication length. For patterns with feedback lines (termed: PAT) ($\chi^2 = 0.0, p = 1$), there is strong statistical similarity. However, for patterns without lines (termed: NPAT) ($\chi^2 = 4.54, p = 0.03$) we do see a significant difference between the live and video study, and an even more prominent difference for PINs ($\chi^2 = 37.76, p = 0.00$). Statistical differences for NPAT can be accounted for by an increase in the 4-length performance for attackers in the live setting (see Table 6), and for PINs, we consistently see performance increases for the live setting compared to the video setting. In this case, the success rate for PINs in the video setting is 65/208=32.0% compared to 135/216=62.5% for the live setting, an increase of 1.95x; however, the video study does provide a baseline.

We can also compare authentication types within collection method, as related to **H1-p**. Using a three-way χ^2 tests with pairwise comparisons, there are statistically significant differences between each of the success rates for each of the authentications for both the live ($\chi^2 = 24.8, p = 0.00$) and video ($\chi^2 = 133.4, p = 0.00$) settings. The residuals suggest the leading cause of this difference is the increased difficulty of shoulder surfing PINs, for both the video and live setting, but post-hoc, pairwise-analysis (with Bonferroni correction) suggest the benefits of removing feedback lines in NPAT is not consistent across studies. While there are statistical differences between PAT and NPAT in the video study, this effect disappears in the live study with $p = .147$ (under the correction). *This provides further evidence that removing traceback lines from pattern entry provides limited protection, and perhaps less than what was previously considered [25].*

Despite seeing a reduced benefit from NPAT as compared to PAT, we can **confirm H1-p** in the live setting. The authentication type has an impact on shoulder surfing performance as evident in the differences in attacker success rate for different authentication types, particularly for PINs.

Repeated Viewings (H1-r/H2-p). An important result of the video study was the finding that repeated viewings have significant impact on attacker performance (**H2-p**). By expanding our view of Table 6 to the *Two-Same* and *Two-Different* column, we can test for similar effects resulting from repeated viewings. As before, we observe the most consistency in the PAT and NPAT settings for the live and video study, and strong differences in the PIN setting. However, where we do see significant difference the confidence interval suggest that the video study does provide a baseline to the live setting.

We can further directly measure the impact of multiple viewings by performing within collection method χ^2 tests across viewing methods. For PAT, no effect could be identified for multiple views in both the video and live settings. There is an effect for NPAT in the live ($\chi^2 = 12.0, p < 0.01$) but not in the video setting ($\chi^2 = 5.1, p = 0.08$). Post-hoc analysis revealed that, for NPAT in the live setting, having the same viewing angles twice compared to a single viewing angle or two difference angles drives this difference ($p = 0.03$, corrected), *suggesting that two-different viewing conditions for NPAT is most advantageous to an attacker.*

The case is similar for PINs. In the live setting, a statistically significant difference occurs for conditions of repeated views ($\chi^2 = 23.1, p = 0.00$). However, this was not the case for the video setting ($\chi^2 = 4.1, p = 0.14$). *Post-hoc analysis suggests that gaining any repeated viewing, the same angle twice or two different, benefits the attacker significantly in the live setting.* The lack of significance for the video setting may be due to using this particular subset of video data, but we conjecture that it more likely reflects the high difficulty of shoulder surfing PINs, generally, which was further exacerbated by the video observation setting without stereo vision.

Overall, we can **confirm H2-p** in the live setting, that repeated viewings have an impact on performance. Where there were previous significant differences in the video study, these persisted in the live setting, except for NPAT. While there is consistency in viewing the same angle twice, observing the entry from multiple angles seems to play a larger role in the live setting compared to the video setting. However, the larger hypothesis that repeated views impacts performance of shoulder surfing is confirmed.

Observation Angle (H1-r/H4-p). To assess the impact of observation angle, we use only single-view conditions so as not to conflate the results with the impact of multiple observations. These results are presented in Table 7 with pairwise comparisons between the live and video study for different passcode lengths.

While we continue to see significant differences for PIN and a lack thereof for PAT, we see significant improvements in the live setting for NPAT viewed from the right angle. We conjecture that this improved attacker performance relates to being able to stereoscopically determine touch locations that are more challenging to see from the same angle via video simulation. However, depth of touch events continue to be more challenging when viewed from the left angle. The difference between the observations angles here may also explain other statistical differences in the previously presented results for NPAT.

However, overall, we *do not* see significant differences when comparing within a collection method and authentication when comparing left vs. right angle. This is in conflict with prior work; however, recall that the top observation angle was removed and the two near and far angles were reduced to a single side angle (L or R). As the two comparable subsets are consistent, we can **confirm** that under **H4-p** the live settings are well predicted by a comparable subset of video data.

Passcode Properties (H1-r/H5-p). In Table 8, again using single view data, a direct comparison between each of the passcodes used in the study is displayed, with findings from proportionality tests between the live and video setting. We find that no significant differences exist for the PAT and NPAT codes, and only three of the PIN codes show differences. These include the following PINs: 5962, 159428, and 366792 with the live setting attacker performance being significantly better in each case. The spatial properties of these codes (see Appendix B.2) does not suggest that a single factor played a role. Although both 5962 and 3669722 are both right shifted PINs, there are too many other features at play to draw conclusions.

We can perform a within-collection method analysis across the passcodes using a χ^2 test, and we find that significant differences exist for the attacker success rate within both the live and video study, for all authentication types. However, post-hoc analysis suggest that none of the NPAT pairwise comparisons are significant, and only one set of PAT pairwise comparisons are significant (743521 vs. 3157) — 743521 was the most difficult of the patterns to shoulder surf. For PINs in post-hoc analysis, again 159428 and 366792 have significant comparisons, particularly with PINs 7272 and 1955, which were two of the easiest PINs to shoulder surf in comparison to 159428 and 366792, two of the most difficult to shoulder surf.

Finally, we can compare the impacts of length. For PAT, we do not see significant differences between success rate for 4- vs. 6-length patterns ($\chi^2 = 2.9, p = 0.09$), but we do for the video study ($\chi^2 = 12.83, p < 0.001$). We find the reverse for NPAT, where there is a significant difference in length for the live setting ($\chi^2 = 5.7, p = 0.02$) and not for the video study ($\chi^2 = 3.64, p = 0.06$). Finally, we see significant differences for PIN for both live ($\chi^2 = 28.9, p = 0.00$) and video ($\chi^2 = 27.6, p = 0.00$). This suggests that, yes, the length

		Left		*Right*	
Auth.	**Length**	**Live**	**Video**	**Live**	**Video**
PAT	4-len	27/28=96.4%	47/49=95.9%	23/25=92.0%	59/62=95.2%
		$\chi^2 = 0.00, p = 1.00, \delta_{95}[-0.09, 0.10]$		$\chi^2 = 0.00, p = 0.95, \delta_{95}[-0.18, 0.12]$	
	6-len	22/26=84.6%	36/48=75.0%	23/29=79.3%	38/47=80.9%
		$\chi^2 = 0.44, p = 0.51, \delta_{95}[-0.12, 0.31]$		$\chi^2 = 0.00, p = 1.00, \delta_{95}[-0.22, 0.19]$	
NPAT	4-len	23/29=79.3%	34/46=73.9%	**25/26=96.2%**	**40/58=69.0%**
		$\chi^2 = 0.07, p = 0.80, \delta_{95}[-0.17, 0.28]$		$\chi^2 = 6.11, p = 0.01*, \delta_{95}[0.10, 0.44]$	
	6-len	15/25=60.0%	29/52=55.8%	20/28=71.4%	29/49=59.2%
		$\chi^2 = 0.01, p = 0.92, \delta_{95}[-0.22, 0.31]$		$\chi^2 = 0.69, p = 0.41, \delta_{95}[-0.12, 0.37]$	
PIN	4-len	**44/54=81.5%**	**22/45=48.9%**	**45/57=78.9%**	**26/49=53.1%**
		$\chi^2 = 10.31, p = 0.00*, \delta_{95}[0.13, 0.53]$		$\chi^2 = 6.86, p = 0.01*, \delta_{95}[0.06, 0.45]$	
	6-len	**25/54=46.3%**	**6/45=13.3%**	**21/51=41.2%**	**11/64=17.2%**
		$\chi^2 = 10.91, p = 0.00*, \delta_{95}[0.14, 0.52]$		$\chi^2 = 6.98, p = 0.01*, \delta_{95}[0.06, 0.42]$	

Table 7: Effects on angle on attacker accuracy. The angle is either an observation from the left or right with a single view (no repeat viewings). For video-based results, no top views were considered. The prior "far" type angles for each side are reduced to simply, left or right, and only data where screen base resolution > 1800 was considered. A 2-sample test for equality of proportions with continuity correction was used, and the χ^2 statistic, p-value, and 95% confidence interval (δ_{95}) of the difference between the proportions (*live - video*) are reported.

	Passcode	**Live**	**Video**	
PAT	0145	8/10=80.0%	23/23=100.0%	$\chi^2 = 2.01, p = 0.16, \delta_{95}[-0.52, 0.12]$
	1346	11/11=100.0%	26/28=92.9%	$\chi^2 = 0.01, p = 0.92, \delta_{95}[-0.09, 0.23]$
	3157	10/10=100.0%	28/29=96.6%	$\chi^2 = 0.00, p = 1.00, \delta_{95}[-0.07, 0.14]$
	4572	11/11=100.0%	17/19=89.5%	$\chi^2 = 0.13, p = 0.72, \delta_{95}[-0.10, 0.32]$
	6745	10/11=90.9%	12/12=100.0%	$\chi^2 = 0.00, p = 0.96, \delta_{95}[-0.35, 0.17]$
	014763	8/10=80.0%	15/19=78.9%	$\chi^2 = 0.00, p = 1.00, \delta_{95}[-0.31, 0.33]$
	136785	11/11=100.0%	14/17=82.4%	$\chi^2 = 0.72, p = 0.40, \delta_{95}[-0.08, 0.43]$
	642580	9/9=100.0%	21/22=95.5%	$\chi^2 = 0.00, p = 1.00, \delta_{95}[-0.09, 0.18]$
	743521	5/12=41.7%	11/21=52.4%	$\chi^2 = 0.05, p = 0.82, \delta_{95}[-0.52, 0.31]$
	841257	12/13=92.3%	13/16=81.2%	$\chi^2 = 0.10, p = 0.75, \delta_{95}[-0.20, 0.42]$
NPAT	0145	10/11=90.9%	12/16=75.0%	$\chi^2 = 0.29, p = 0.59, \delta_{95}[-0.19, 0.51]$
	1346	8/10=80.0%	14/18=77.8%	$\chi^2 = 0.00, p = 1.00, \delta_{95}[-0.31, 0.36]$
	3157	11/11=100.0%	15/21=71.4%	$\chi^2 = 2.22, p = 0.14, \delta_{95}[0.02, 0.55]$
	4572	7/10=70.0%	12/25=48.0%	$\chi^2 = 0.65, p = 0.42, \delta_{95}[-0.19, 0.63]$
	6745	12/13=92.3%	21/24=87.5%	$\chi^2 = 0.00, p = 1.00, \delta_{95}[-0.20, 0.29]$
	014763	12/14=85.7%	15/21=71.4%	$\chi^2 = 0.33, p = 0.57, \delta_{95}[-0.18, 0.47]$
	136785	5/10=50.0%	7/19=36.8%	$\chi^2 = 0.08, p = 0.77, \delta_{95}[-0.32, 0.59]$
	642580	7/9=77.8%	16/26=61.5%	$\chi^2 = 0.23, p = 0.63, \delta_{95}[-0.24, 0.57]$
	743521	3/9=33.3%	12/18=66.7%	$\chi^2 = 1.52, p = 0.22, \delta_{95}[-0.79, 0.13]$
	841257	8/11=72.7%	8/17=47.1%	$\chi^2 = 0.90, p = 0.34, \delta_{95}[-0.17, 0.69]$
PIN	1328	15/21=71.4%	16/28=57.1%	$\chi^2 = 0.53, p = 0.47, \delta_{95}[-0.17, 0.45]$
	1955	18/21=85.7%	11/19=57.9%	$\chi^2 = 2.60, p = 0.11, \delta_{95}[-0.04, 0.60]$
	5962	22/24=91.7%	6/18=33.3%	$\chi^2 = 13.23, p = 0.00*, \delta_{95}[0.29, 0.88]$
	6702	15/21=71.4%	6/17=35.3%	$\chi^2 = 3.61, p = 0.06, \delta_{95}[0.01, 0.71]$
	7272	19/24=79.2%	9/12=75.0%	$\chi^2 = 0.00, p = 1.00, \delta_{95}[-0.29, 0.38]$
	1328	15/21=71.4%	16/28=57.1%	$\chi^2 = 0.53, p = 0.47, \delta_{95}[-0.17, 0.45]$
	153525	8/24=33.3%	4/13=30.8%	$\chi^2 = 0.00, p = 1.00, \delta_{95}[-0.31, 0.37]$
	159428	12/21=57.1%	1/23=4.3%	$\chi^2 = 12.27, p = 0.00*, \delta_{95}[0.25, 0.80]$
	366792	8/21=38.1%	2/26=7.7%	$\chi^2 = 4.72, p = 0.03*, \delta_{95}[0.03, 0.58]$
	441791	10/21=47.6%	5/27=18.5%	$\chi^2 = 3.40, p = 0.07, \delta_{95}[-0.01, 0.59]$
	458090	8/18=44.4%	5/20=25.0%	$\chi^2 = 0.84, p = 0.36, \delta_{95}[-0.16, 0.55]$

Table 8: Comparison of attacker accuracy per-passcode. Only single view conditions were considered in both live and video, and for the video results, only data with resolution > 1800 was included and the "top" angle was excluded. A 2-sample test for equality of proportions with continuity correction was used, and the χ^2 statistic, p-value, and 95% confidence interval (δ_{95}) of the difference between the proportions (*live - video*) are reported.

of the passcode can have an impact, **confirming H1-r** for the **H5-p** condition; however, other properties of the passcode were not significant, but were so in similar ways between the two studies under the subset being evaluated.

Hypothesis H1-r. Based on the results presented previously, in each case we are able to find confirmation of each of the previous hypotheses, although, we also find that PINs are the least consistent. This suggests that researchers should be more skeptical of results related to PIN based authentication in the video setting. In particular, the true values may be much higher. Additionally, we find strong evidence that the differences between PAT and NPAT may be greatly dimensioned (although still different) in the live setting.

Hypothesis H2-r. We can confirm **H2-r** that video based recreations do provide a baseline for live simulation. Observe that in all cases where there is significant differences between a video and live measurement in Tables 6, 7, and 8, the confidence interval suggests that the live setting has *higher* proportionality than that of video setting. In essence, yes, the video study provides a baseline, but the baseline may be much lower than one may expect, as much as 1.7x.

5.2 Post-Hoc Participant Feedback

One advantage of the live study is that the researchers can directly observe the strategies of the participants and the relative difficulties encountered, as well as via post hoc questions (the precise questions are found in the Appendix A.2). There is no direct comparison to the Aviv et al. prior work here, but we believe that the strategies likely mirror those used by participants in the video study, to some extent.

The most commonly reported strategy for the observation task (n=16) was simply focusing on memorizing the passcode as it appeared and then, after it was completely entered, writing it down immediately without delay. Only three participants reported strategies involving writing or physically mirroring the input gesture while it was happening. Other participants (n=2) described "chunking" PINs into larger numbers (e.g. "seventeen" versus "one-seven")

Figure 4: Grid patterns with crossing and knightmove (4572) features that challenged some observers, and patterns that were deemed more memorable by some observers because they offered easy symbolic associations.

in their first languages (Farsi and Chinese) to make quick memorization easier. Five participants mentioned that they watched the readout field in the PIN conditions, while others preferred to watch only the finger gesture as it was performed.

Participants mentioned several factors that could make PIN and grid passcodes challenging to accurately record. These included grid pattern shapes that crossed over themselves or contained knightmoves (n=11, e.g. 743521 and 4572, Figure 4), as well as both long physical jumps between sequential PIN digits (n=3) and sequential digits physically close together (n=7). Ten participants reported that viewing from the right was harder because their view of the phone screen was partially blocked by the victim-proctor's thumb in his right-handed grip, which is supported in the data, particularly for NPAT results. Six participants also felt that glare from overhead lighting was sometimes an issue.

Other passcode features and conditions were described as helpful by observers. Four participants mentioned that it was easier to memorize shapes that they could easily associate with a visual image, such as 136785 as a house, or 842157 as a picnic table (Figure 4).

Finally, multiple observations of the same passcode were commonly deemed helpful for confirming or piecing together sequences, although one participant stated that it was easier to do this if both observations were made from the same side. This is supported by the quantitative data.

6 IMPLICATIONS

Importance of evaluating in appropriate settings. Researchers often favor performing studies examining observational attacks with video-based stimuli presented to participants. While likely simpler to coordinate and easier to control compared to studies conducted in live settings, video studies can lack realism and are considered a methodological substitute only when necessary [27]. While findings from video studies can be helpful to determine attack rate, our findings suggest that researchers evaluating authentication interfaces should be aware that there is no substitute for testing in live settings, as the video baseline may greatly underestimate the threat of an attacker. The video baseline may serve as a method for a preliminary assessment.

Factors which should be taken into account when performing observational attack studies. While factors such as authentication type and repeated views can impact attack rate, as evidenced through our study, other factors are worthy of further investigation. Examples include examination of the impact of observational angle and spatial properties of passcodes and device screen sizes. While significant differences in some of these factors could not always be detected, subjective feedback gathered from participant observers suggested that these factors could make a difference to attacker success. Examining these in more detail, alongside gathering subjective data for purposes of identifying reasoning, is suggested to researchers, as these may play a greater role than once thought.

Care in selection of passcode. Our results suggest that specific types and properties of passcode may be more susceptible to observational attack, as identified through the comparison with live settings. As a result, users should be aware that removing the feedback lines from pattern unlock interfaces may not provide the security benefits that users expect. Secondly, PINs are more susceptible to attack than previously identified by researchers performing video-based studies. This is also supported in our qualitative feedback where participants noted that PINs with larger jumps were harder to attack, and for PAT/NPAT, those that are less "shape like" (e.g. resembling a house-like shape) are harder for participants.

Need for training. As our findings have highlighted that observational attacks are more successful under specific conditions, security training for mobile device users can be developed to better understand the nature of observational threats, encouraging them to make better security choices. Some users may need to better understand what methods and parameters would provide resilience against high-probability multiple-view observation attacks mounted by "insider threats" [28]. Others might want those authentication factors tilted towards greater ease of use if they perceive less risk of observational attack. Better informing these choices could come in the form of interactive guidance/prompting when setting-up devices.

7 CONCLUSIONS

In this paper, we have described a study comparing video recreations of shoulder surfing to live simulation. We recreated a subset of the factors explored in the video study and attempted to confirm prior findings in this setting. We were able to confirm many of the prior claims regarding the video study, that authentication type, repeated viewings, observation angle, and passcode properties can affect attacker performance. We were also able to confirm that video study does form a baseline for the live simulation; however, this baseline may be much less than desired, as much as 1.9x difference. From these findings we suggest, for researchers conducting shoulder surfing studies with video components, that data can form a baseline and be representative, in many situations, of what would occur in a live simulation. However, when possible, those results should be compared to a live simulation to get a fuller picture of the data and results.

ACKNOWLEDGMENTS

This work was supported by the Office of Naval Research. The authors wish to thank Chukwuemeka KC Marume and John T. Davin for their assistance conducting the study.

REFERENCES

[1] Ali Abdolrahmani, Ravi Kuber, and Amy Hurst. 2016. An empirical investigation of the situationally-induced impairments experienced by blind mobile device users. In *Proceedings of the 13th Web for All Conference*. ACM, 21.

[2] Abdullah Ali, Adam J Aviv, and Ravi Kuber. 2016. Developing and evaluating a gestural and tactile mobile interface to support user authentication. *IConference 2016 Proceedings* (2016).

[3] Adam J. Aviv, Devon Budzitowski, and Ravi Kuber. 2015. Is Bigger Better? Comparing User-Generated Passwords on 3x3 vs. 4x4 Grid Sizes for Android's Pattern Unlock. In *Proceedings of the 31st Annual Computer Security Applications Conference (ACSAC 2015)*. ACM, New York, NY, USA, 301–310. https://doi.org/10.1145/2818000.2818014

[4] Adam J. Aviv, John T. Davin, Flynn Wolf, and Ravi Kuber. 2017. Towards Baselines for Shoulder Surfing on Mobile Authentication. In *Proceedings of the 33rd Annual Computer Security Applications Conference (ACSAC 2017)*. ACM, New York, NY, USA, 486–498. https://doi.org/10.1145/3134600.3134609

[5] Joseph Bonneau, Sören Preibusch, and Ross Anderson. 2012. A birthday present every eleven wallets? The security of customer-chosen banking PINs. In *International Conference on Financial Cryptography and Data Security*. Springer, 25–40.

[6] Alexander De Luca, Martin Denzel, and Heinrich Hussmann. 2009. Look into My Eyes!: Can You Guess My Password?. In *Proceedings of the 5th Symposium on Usable Privacy and Security (SOUPS '09)*. ACM, New York, NY, USA, Article 7, 12 pages. https://doi.org/10.1145/1572532.1572542

[7] Alexander De Luca, Alina Hang, Frederik Brudy, Christian Lindner, and Heinrich Hussmann. 2012. Touch Me Once and I Know It's You!: Implicit Authentication Based on Touch Screen Patterns. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 987–996. https://doi.org/10.1145/2207676.2208544

[8] Alexander De Luca, Marian Harbach, Emanuel von Zezschwitz, Max-Emanuel Maurer, Bernhard Ewald Slawik, Heinrich Hussmann, and Matthew Smith. 2014. Now You See Me, Now You Don'T: Protecting Smartphone Authentication from Shoulder Surfers. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 2937–2946. https://doi.org/10.1145/2556288.2557097

[9] Alexander De Luca, Katja Hertzschuch, and Heinrich Hussmann. 2010. ColorPIN: Securing PIN Entry Through Indirect Input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 1103–1106. https://doi.org/10.1145/1753326.1753490

[10] Serge Egelman, Sakshi Jain, Rebecca S. Portnoff, Kerwell Liao, Sunny Consolvo, and David Wagner. 2014. Are You Ready to Lock?. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security (CCS '14)*. ACM, New York, NY, USA, 750–761. https://doi.org/10.1145/2660267.2660273

[11] Malin Eiband, Mohamed Khamis, Emanuel von Zezschwitz, Heinrich Hussmann, and Florian Alt. 2017. Understanding shoulder surfing in the wild: Stories from users and observers. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 4254–4265.

[12] Alain Forget, Sonia Chiasson, and Robert Biddle. 2010. Shoulder-surfing Resistance with Eye-gaze Entry in Cued-recall Graphical Passwords. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 1107–1110. https://doi.org/10.1145/1753326.1753491

[13] H. Gao, Z. Ren, X. Chang, X. Liu, and U. Aickelin. 2010. A New Graphical Password Scheme Resistant to Shoulder-Surfing. In *2010 International Conference on Cyberworlds*. 194–199. https://doi.org/10.1109/CW.2010.34

[14] Marian Harbach, Emanuel Von Zezschwitz, Andreas Fichtner, Alexander De Luca, and Matthew Smith. 2014. It's a hard lock life: A field study of smartphone (un) locking behavior and risk perception. In *Symposium on usable privacy and security (SOUPS)*. 213–230.

[15] Hassan Khan, Urs Hengartner, and Daniel Vogel. 2018. Evaluating Attack and Defense Strategies for Smartphone PIN Shoulder Surfing. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, New York, NY, USA, Article 164, 10 pages. https://doi.org/10.1145/3173574.3173738

[16] Jesper Kjeldskov and Mikael B Skov. 2014. Was it worth the hassle?: ten years of mobile HCI research discussions on lab and field evaluations. In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services*. Acm, 43–52.

[17] Katharina Krombholz, Thomas Hupperich, and Thorsten Holz. 2016. Use the Force: Evaluating Force-Sensitive Authentication for Mobile Devices. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*. USENIX Association, Denver, CO, 207–219. https://www.usenix.org/conference/soups2016/technical-sessions/presentation/krombholz

[18] Manu Kumar, Tal Garfinkel, Dan Boneh, and Terry Winograd. 2007. Reducing Shoulder-surfing by Using Gaze-based Password Entry. In *Proceedings of the 3rd Symposium on Usable Privacy and Security (SOUPS '07)*. ACM, New York, NY, USA, 13–19. https://doi.org/10.1145/1280680.1280683

[19] Shushuang Man, Dawei Hong, and Manton M Matthews. 2003. A Shoulder-Surfing Resistant Graphical Password Scheme-WIW. , 105–111 pages.

[20] Hee Jung Ryu and Florian Schroff. 2017. Electronic Screen Protector with Efficient and Robust Mobile Vision. In *Demos section, Neural Information Processing Systems Conference*.

[21] Alireza Sahami Shirazi, Peyman Moghadam, Hamed Ketabdar, and Albrecht Schmidt. 2012. Assessing the vulnerability of magnetic gestural authentication to video-based shoulder surfing attacks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2045–2048.

[22] Florian Schaub, Ruben Deyhle, and Michael Weber. 2012. Password Entry Usability and Shoulder Surfing Susceptibility on Different Smartphone Platforms. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia (MUM '12)*. ACM, New York, NY, USA, Article 13, 10 pages. https://doi.org/10.1145/2406367.2406384

[23] Florian Schaub, Marcel Walch, Bastian Könings, and Michael Weber. 2013. Exploring the design space of graphical passwords on smartphones. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*. ACM, 11.

[24] Emanuel Von Zezschwitz, Alexander De Luca, Bruno Brunkow, and Heinrich Hussmann. 2015. SwiPIN: Fast and secure pin-entry on smartphones. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1403–1406.

[25] Emanuel von Zezschwitz, Alexander De Luca, Philipp Janssen, and Heinrich Hussmann. 2015. Easy to Draw, but Hard to Trace?: On the Observability of Grid-based (Un)Lock Patterns. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 2339–2342. https://doi.org/10.1145/2702123.2702202

[26] Susan Wiedenbeck, Jim Waters, Leonardo Sobrado, and Jean-Camille Birget. 2006. Design and Evaluation of a Shoulder-surfing Resistant Graphical Password Scheme. In *Proceedings of the Working Conference on Advanced Visual Interfaces (AVI '06)*. ACM, New York, NY, USA, 177–184. https://doi.org/10.1145/1133265.1133303

[27] Oliver Wiese and Volker Roth. 2015. Pitfalls of Shoulder Surfing Studies. In *NDSS Workshop on Usable Security*. 1–6.

[28] Oliver Wiese and Volker Roth. 2016. See you next time: A model for modern shoulder surfers. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 453–464.

A SURVEY MATERIAL

A.1 Ante Hoc Demographic Questionnaire

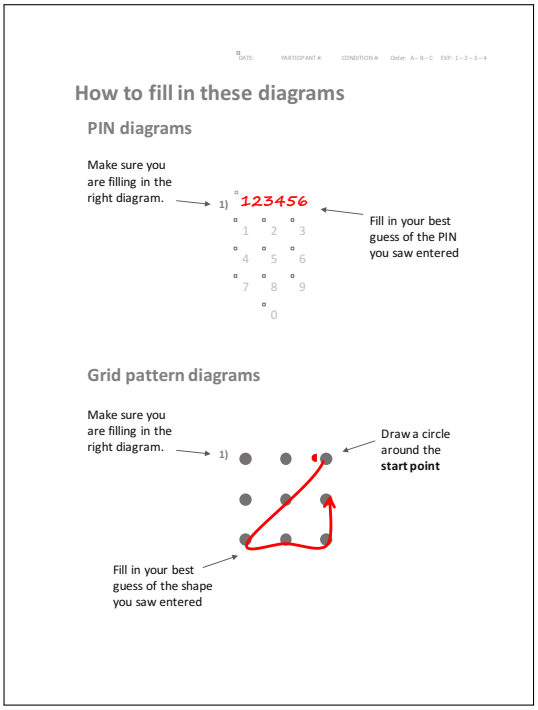
(1) What is your age? (18-24, 25-34, 35-44, 45-54, +65, NA)?
(2) What is your identified gender?
(3) Do you have any physical conditions that might prevent you from observing authentication gestures performed on a mobile phone?
(4) Do you use a smartphone currently? If so, what is its operating system?
(5) Why did you select that phone and OS?
(6) If you currently use an authentication method to lock your phone, what is the method (i.e. PIN, TouchID, grid, etc.), and why did you select it?
(7) What types of mobile phone authentication have you used? (i.e. PIN, grid pattern, password, fingerprint, face, voice, other)
(8) Without telling me your current passcode, how do you select the passcodes you use?
(9) How concerned are you with keeping your phone secure (1, not at all concerned, to 5, highly concerned)?
(10) What experiences can you recall involving people either trying to steal or use your phone without permission?
(11) What experiences can you recall involving people trying to observe your passcodes without permission?
(12) How concerned are you with the threat of someone watching you authenticate and collecting your passcodes (1, not at all concerned, to 5, highly concerned)?
(13) If you had any of these experiences, how did it affect your behavior?

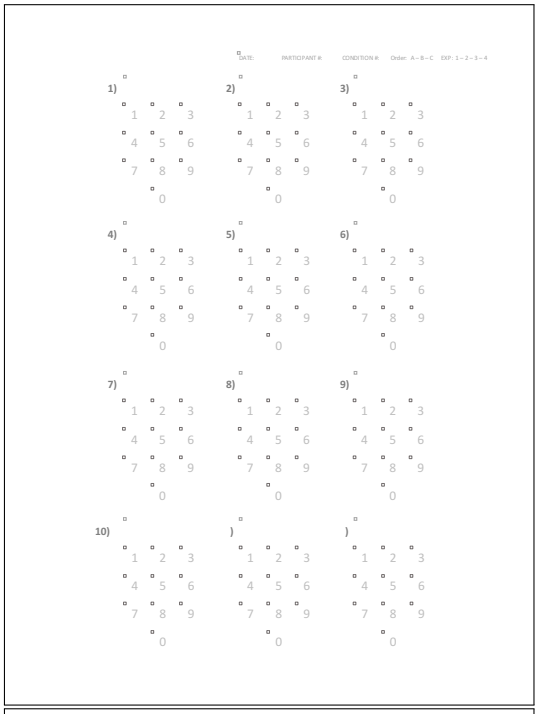
(14) Have any other experiences or concerns affected your authentication?

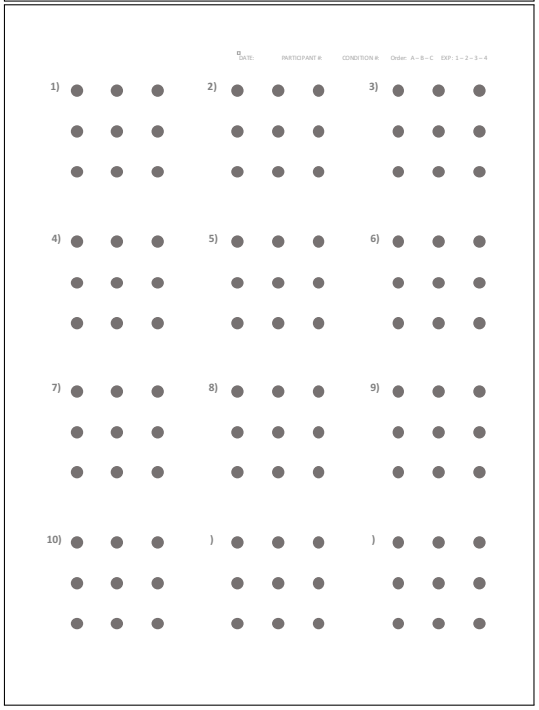
A.2 Post Hoc Participant Strategies Questionnaire Questions

(1) What strategies did you employ to collect the passcodes?
(2) Do you have any ideas for additional strategies?
(3) How challenging was it to collect PIN passcodes (1, not at all challenging, to 5, very challenging)?
(4) How challenging was it to collect grid passcodes (1, not at all challenging, to 5, very challenging)?
(5) What features of the passcodes made it easier or more difficult to collect the passcodes you saw?
(6) How did the number of views you were given make a difference?
(7) How did which side you stood on make any difference?

A.3 Observation Forms







A.4 Guide/Script for Administering Study

(1) Verify current participant number, exp (1-4, order (a-c). Record this.
(2) Introduction - "Welcome, thanks for participating. Our study deals with the security of different types of passcodes for mobile phones. Your help today will be pretty straightforward. We will record some basic demographic information about

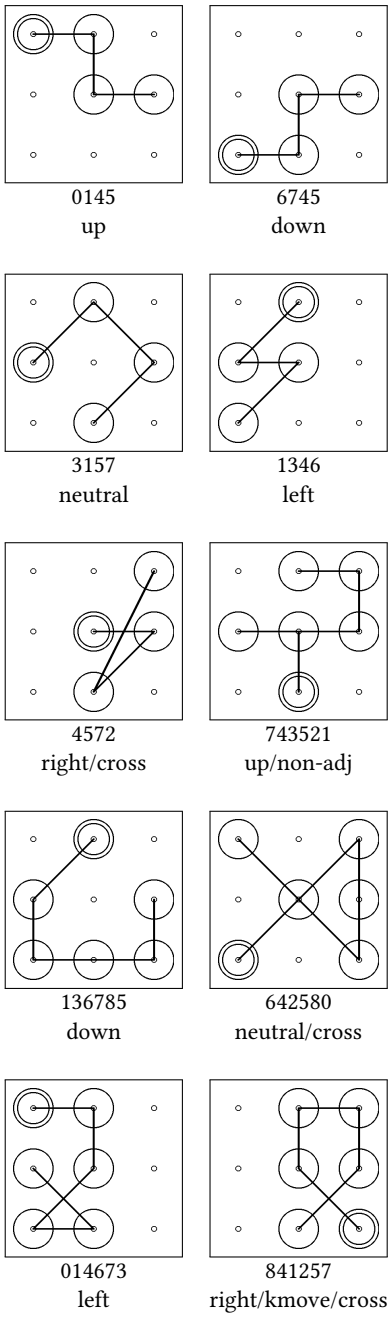
you. You will observe someone entering in passcodes from a few feet away, and write down your best guess of what you have seen. We will go over the steps required to make sure you are comfortable and understand your role. We will record the session to verify the results. All data collected will be anonymized for publication. Your part in the study should take about 20-30 minutes."

(3) Payment - "The study pays 5 USD."
(4) Observer disclosure - "Are there any issues, such as corrective glasses or contacts, which might interfere with performing the role I have described?"
(5) IRB Introduction - "This study has been reviewed by the University's review board, the IRB, and approved as safe and ethical. Here is a copy of that form that describes the study that you can read. Please ask any questions you may have, and sign the form if you would like to participate."
(6) Demographic questionnaire - "Please fill out the demographic questionnaire."
(7) Training: Overview - "Here is the process your role as an observer. You will stand to the left or right behind our researcher, who will sit in the same position entering in passcodes. We will specify where to stand for each attempt. You will watch each attempt, and then draw on the form we provide your best guess of the passcode you just saw being entered on the phone. Passcodes may vary in length. We will repeat this ten times."
(8) Training: Filling out the form diagrams - "Look at the form we have provided. It has blank PIN and grid pattern diagrams for you to enter your guesses. For the grid patterns passcodes, draw the shape you saw entered, and circle the starting point of the shape. For PIN passcodes, write out the sequence, like "1234", and draw the shape you saw entered." [Demonstrate drawing a diagram, then allow the participant to practice drawing 2-3 times, based on a practice code for their prescribed passcode condition (PIN, grid, no-line grid) that you show them slowly, up close, on a phone. Confirm for grid shapes that they are circling the starting point. Correct any issues that appear, and repeat until ready.]
(9) Training: Taking position for each attempt - "We will call out a position, LEFT or RIGHT, for you to stand in for each attempt. Move to the corresponding marker on the floor, figure out which diagram you are going to fill in, and when set to begin, say "Ready." Sometimes, we may also call out MULTIPLE if you are allowed to view the passcode being entered twice before making your final guess." "Any questions?"
(10) Post hoc questionnaire - [Conduct post hoc interview]

B VISUALIZATION OF AUTHENTICATION

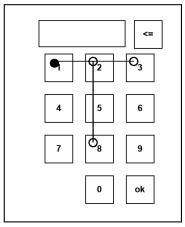
Below are the set of patterns and PINs as visualized in prior work [4], as well as the description of the visual properties.

B.1 Patterns

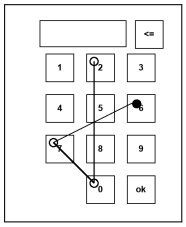


B.2 PINs

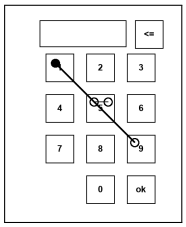
Note that filled circle is the start point, multiple circles on a number indicate multiple touches.



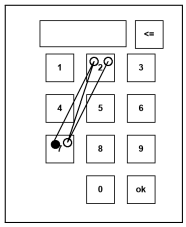
1328
up/non-adj



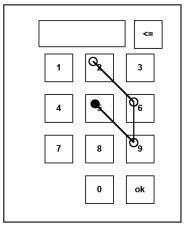
6702
down/kmove/cross



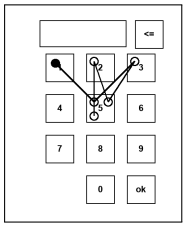
1955
neutral/non-adj/repeats



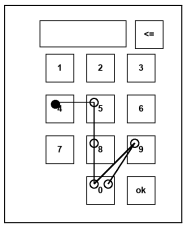
7272
left/kmoves/repeats



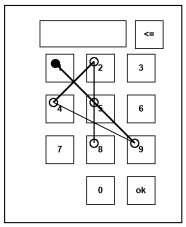
5962
right



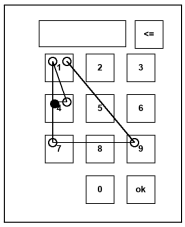
152525
up/repat



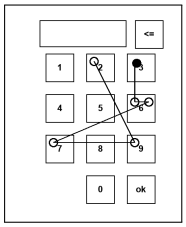
458090
down/repeat



159428
neutral/cross/non-adj



441791
left/kmove/repeat



366792
right/repeat/kmove/cross