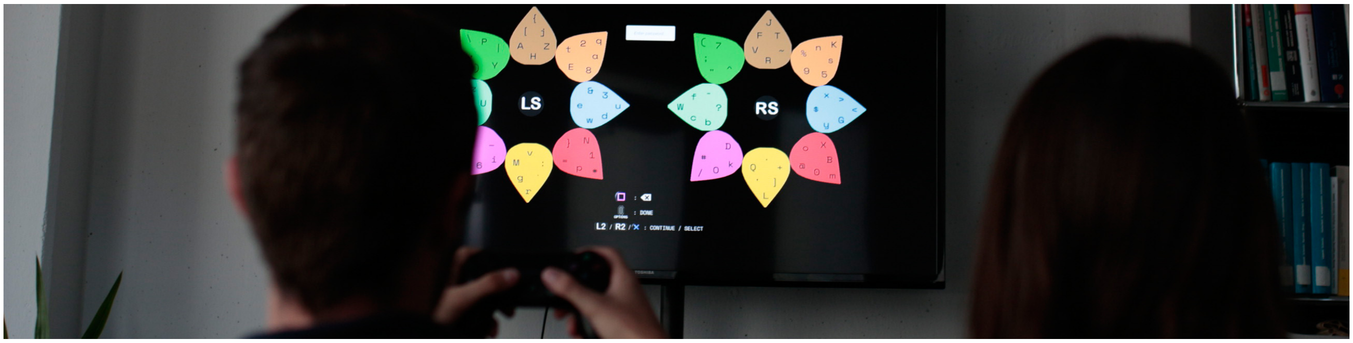
I (Don't) See What You Typed There! Shoulder-surfing Resistant Password Entry on Gamepads

Peter Mayer[*], **Nina Gerber**[*], **Benjamin Reinheimer**[*], **Philipp Rack**[†], **Kristoffer Braun**[†], **Melanie Volkamer**[*]

[*]Karlsruhe Institute of Technology
SECUSO - Security, Usability, Society
firstname.lastname@kit.edu

[†]Technische Universität Darmstadt
Department of Computer Science



ABSTRACT

Using gamepad-driven devices like games consoles is an activity frequently shared with others. Thus, shoulder-surfing is a serious threat. To address this threat, we present the first investigation of shoulder-surfing resistant text password entry on gamepads by (1) identifying the requirements of this context; (2) assessing whether shoulder-surfing resistant authentication schemes proposed in non-gamepad contexts can be viably adapted to meet these requirements; (3) proposing "Colorwheels", a novel shoulder-surfing resistant authentication scheme specifically geared towards this context; (4) using two different methodologies proposed in the literature for evaluating shoulder-surfing resistance to compare "Colorwheels", on-screen keyboards (the de facto standard in this context), and an existing shoulder-surfing resistant scheme which we identified during our assessment and adapted for the gamepad context; (5) evaluating all three schemes regarding their usability. Having applied different methodologies to measure shoulder-surfing resistance, we discuss their strengths and pitfalls and derive recommendations for future research.

CCS CONCEPTS

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

KEYWORDS

Authentication, Shoulder-Surfing, User Study, Gamepads

ACM Reference format:
Peter Mayer, Nina Gerber, Benjamin Reinheimer, Philipp Rack, Kristoffer Braun, Melanie Volkamer. 2019. I (Don't) See What You Typed There! Shoulder-surfing Resistant Password Entry on Gamepads. In *Proceedings of CHI Conference on Human Factors in Computing Systems Proceedings, Glasgow, Scotland, UK, May 4–9, 2019 (CHI 2019)*, 12 pages.
https://doi.org/10.1145/3290605.3300779

1 INTRODUCTION

Gamepad-driven devices such as gaming consoles are an important part of many people's lives. It is common to use accounts for e.g. video streaming like Netflix, music streaming like Spotify, or game networks like Playstation Network on these devices. A 2017 report by the entertainment software association found that about half of all American households own a dedicated gaming console [14]. For many users time

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300779

spent on their consoles is a social activity and therefore occurs in shared spaces: 53% of the users play on average five hours with others in person per week (as opposed to online multiplayer games) [14]. Accounts used on these consoles are protected by text passwords, and the passwords are entered almost exclusively using *on-screen keyboards*. Considering that Renaud et al. [24] found in their survey that 90.9% of their participants would authenticate when not alone, opportunistic shoulder-surfing [34] is a real threat, leaving users in a dilemma: either show mistrust of people by asking them to look away [12], behave insecurely by letting them observe, or store the password on the device, which enables purchases by every person with access to the device.

In this work, we present the first investigation of shoulder-surfing resistant authentication using gamepads. Our main contributions are: (1) We identify the requirements of authentication using gamepads as application scenario (section 2). One of the defining requirements of this scenario is the resistance to opportunistic shoulder-surfing. (2) Based on these requirements, we analyse both, authentication schemes currently deployed for gamepads and shoulder-surfing resistant schemes proposed in non-gamepad contexts (section 3). While most do not meet the gamepad scenario specific requirements identified in section 2, the *grid-based* scheme proposed by Kim et al. [21] can be easily adapted to meet these requirements. (3) We propose a novel authentication scheme called *Colorwheels* which is designed to meet the requirements identified before, i.e. in particular to be resistant to opportunistic shoulder-surfing (section 4). Its design is specifically geared towards usage with gamepads. (4) To evaluate its efficacy in mitigating shoulder-surfing attacks, we conducted two user studies (section 5): an online study (section 6) and a lab study (section 7). In both studies, we compared *Colorwheels* to two other solutions: (a) an *on-screen keyboard* which represents the de facto standard of authentication using gamepads and (b) the *grid-based* scheme [21] which we adapted for the usage with gamepads. To evaluate the shoulder-surfing resistance, participants were asked to recover a password by observing video recordings of its entry. (5) To gauge the usability, we let the participants of the lab study use our implementations of the three authentication schemes themselves and measured their performance with respect to the metrics efficiency, effectiveness and satisfaction (section 7). Additionally, we captured the participants' thoughts on the schemes using qualitative questions.

Our results confirm that the commonly used *on-screen keyboard* provides only little protection even against opportunistic shoulder-surfing: It is significantly more susceptible to shoulder-surfing than the other two schemes in both studies. Both other schemes fare better, but our proposal *Colorwheels* seems to exhibit a more robust shoulder-surfing resistance. Usability-wise, the *on-screen keyboard* fares best. It performs significantly better in terms of efficiency and satisfaction than the other two schemes as well as significantly better in terms of effectiveness than *Colorwheels*. *Colorwheels* scores significantly better in terms of efficiency and satisfaction than the *grid-based* scheme. No significant difference between the two could be found in terms of effectiveness. We discuss these results as well as the strength and limitations of the two methodologies we used to evaluate shoulder-surfing resistance (section 8).

2 REQUIREMENTS

In this section we describe the requirements that specifically apply to the scenario of password entry on gamepads[1].

Security Requirements. Using gamepad-driven devices is for many users a social activity: 53% of the users play on average five hours with others in person per week (as opposed to online multiplayer games) [14]. Thus, usage of these devices occurs in a so-called shared space [24]. Another defining aspect of authentication using gamepad-driven devices such as game consoles is that they are usually used in conjunction with large displays such as TVs. Tan and Czerwinski [29] found that users were more likely to read sensitive content on large screens and note that since such devices are usually outside a user's "personal zone", they might be perceived as less private. Together, these two aspects indicate a large potential for shoulder-surfing threats. However, several types of shoulder-surfing attackers must be differentiated. Wiese and Roth [34] propose four types:

1. *Single Recording, SR:* Attacker gets a small number of recorded authentication procedures.
2. *Multiple Recording, MR:* Attacker gets a huge number of recorded authentication procedures.
3. *Opportunistic observer, OO:* Attacker observes a small number of authentication procedures.
4. *Insider Observer, IO:* Attacker observes a huge number of authentication procedures.

In the first two categories (SR & MR) the human ability, e.g. memory retention, plays a subordinate role, as the password entries are recorded and can be played back and paused at will. In the second two categories (OO & IO) the attackers observe the whole process and try to remember the most important details. Afterwards they depend on their memory retention to try to log in with the user data. Due to the threat model of usage in shared spaces, the opportunistic observer is the most likely attacker in the gamepad scenario. Any recording of the authentication procedure by a friend sitting on the user's couch right next to them is likely to draw attention. Therefore, the first requirement is:

R1: *Authentication schemes used on gamepad-driven devices must resist shoulder-surfing attacks by opportunistic observers.*

[1] Note, this does not lift general requirements of the authentication domain.

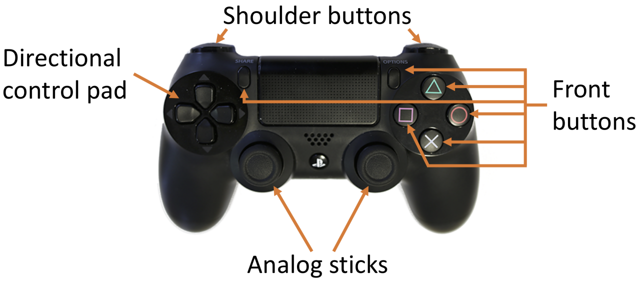
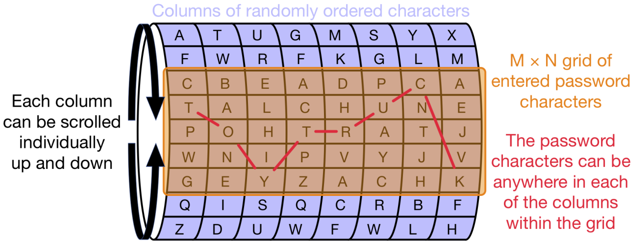


Figure 1: The controls available on a typical gamepad.


Columns of randomly ordered characters
Each column can be scrolled individually up and down
M × N grid of entered password characters
The password characters can be anywhere in each of the columns within the grid

A	T	U	G	M	S	Y	X
F	W	R	F	K	G	L	M
C	B	E	A	D	P	C	A
T	A	L	C	H	U	N	E
P	O	H	T	R	A	T	J
W	N	I	P	V	Y	J	V
G	E	Y	Z	A	C	H	K
Q	I	S	Q	C	R	B	F
Z	D	U	W	F	W	L	H

Figure 2: The grid-based scheme.

Technical Requirements. In comparison to a keyboard, gamepads offer far less buttons and in comparison to a mouse, far less precise analog movement input. The input capabilities of a typical gamepad (see Figure 1) are reflected by:

R2: *Authentication schemes used on gamepad-driven devices must not require as controls more than eight freely programmable buttons, one directional control pad, two analog sticks, and two analog triggers if they are to be compatible with the gamepads of most modern gamepad-driven devices.*

The controls used for directional input on gamepads (i.e. analog sticks) are generally less precise than a mouse [15]. The second technical requirement reflects this:

R3: *Authentication schemes used on gamepad-driven devices must not require high-precision input.*

The accounts used on gamepad-driven devices usually require text password authentication (e.g. common services such as Amazon Video or Xbox Live). Therefore, remaining compatible to these accounts is of the essence:

R4: *Authentication schemes used on gamepad-driven devices must be compatible with text passwords.*

The operating systems of gamepad-driven devices do usually not allow installation of drivers for additional hardware. Thus, the following requirement arises:

R5: *Authentication schemes used on gamepad-driven devices must not require support for additional hardware such as biometric readers or token devices.*

Usability Requirement. The layout of the gamepads in conjunction with human anatomy poses restrictions on the controls which can be used simultaneously to enter a password. Thus, we need the following requirement:

R6: *Authentication schemes used on gamepad-driven devices must not require two or more controls on the front to be operated at the same time.*

3 ASSESSMENT OF EXISTING SCHEMES

In this section, we first briefly introduce the deployed schemes and explain why they are not likely to meet the security requirement *R*1. Thereafter, we introduce existing shoulder-surfing resistant schemes (i.e. schemes meeting *R*1), and assess their suitability for usage with gamepads.

Schemes Deployed in the Gamepad Context. All current game consoles offer *on-screen keyboards* for all text entries, including text passwords. The *on-screen keyboards* consist of a grid of buttons. To enter a character, the user moves a visible cursor to the desired character and confirms the input by pressing a specific (platform dependent) button on the gamepad. Due to the cursor highlighting the current input the whole time, this scheme is highly prone to shoulder-surfing threats. The only other currently deployed scheme to enter text passwords is *Daisy Wheel*. However, it is only available on the Steam platform with a special gamepad. Its interface contains eight petals, each displaying four different characters. First the petal is selected and then one of the four characters in this petal. Since the selected character in the petal briefly blinks as visual feedback, this scheme is highly prone to shoulder-surfing threats. Some devices also offer PIN code protection when being switched on (analogously to smartphone PIN code locks), where the numbers 0 to 9 are mapped to different gamepad controls. At the PIN security level these schemes are, however, inadequate for other accounts (e.g. video or music streaming) and observing which buttons are pressed on the gamepad is easily possible.

Existing Shoulder-Surfing Resistant Schemes. To counteract shoulder-surfing threats, several techniques can be employed. To comply with *R*5, only knowledge-based authentication methods are considered. We present an overview of the assessment of existing schemes in table 1. In summary, only three schemes seem to meet all requirements: (1) the grid-based scheme [21]: It uses a $M \times N$ grid of characters, where N is the length of the password, i.e. there is one column in the grid for each character in the password, and M is the number of characters visible at the same time in each column, i.e. the number of rows. Figure 2 depicts the interface and concept of this authentication scheme. The characters of the password do not have to be aligned in the same row. Thus, the number of rows determines the shoulder-surfing resistance of the scheme. (2) PairPasswordChar [23]: It uses a grid in which all characters available for input are randomly placed and the user has to click into certain areas of the grid determined by the placement of the characters used in their

Table 1: Assessment of whether the existing shoulder-surfing resistant authentication schemes from non-gamepad contexts fulfill the remaining requirements R2-R6.

Proposal	R2	R3	R4	R5	R6
Secure Haptic Keypad [4]	yes	yes	yes	somewhat, haptic feedback not for individual buttons	yes
Glass Unlock [35]	yes	yes	yes	no, requires external private display	yes
ForcePIN [22]	no, requires more force sensitive controls	yes	yes	yes	yes
Undercover [27]	no, requires concealed placement of trackball	yes	somewhat, requires extension of the input grid	yes	yes
Tetrad [24]	yes	yes	no, graphical	yes	yes
DAS variants [36]	yes	yes	no, graphical	yes	yes
Convex Hull Click [32]	yes	somewhat, designed for mouse input	no, graphical	yes	yes
Grid-based scheme [21]	yes	yes	yes	yes	yes
S3PAS [37]	yes	somewhat, designed for mouse input	yes	yes	yes
PairPasswordChar [23]	yes	yes	yes	yes	yes
Xside [11]	no, requires touch interface on back of device	yes	yes	yes	yes
Magnetic Gestures [26]	yes	yes	yes	no, requires magnetic sensor	yes
Cognitive Trapdoor [25]	yes	yes	somewhat, requires extension of character grid	yes	yes
SwiPIN [30]	no, requires too many directional controls when scaled up from PIN to text password	yes	yes	yes	yes
Behavioural biometrics [10]	yes	yes	yes	yes	yes

password and a set of rules. (3) Behavioural biometrics based on the user's specific gesturing patterns [10].

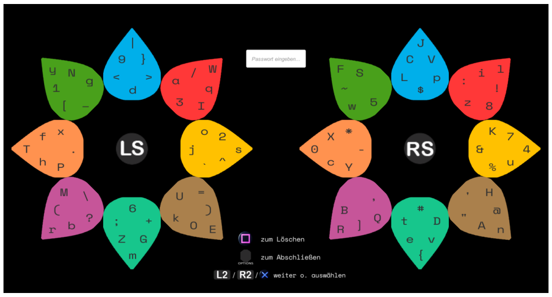


Figure 3: Colorwheels.

4 NEW PROPOSAL: COLORWHEELS

While some of the shoulder-surfing resistant schemes described in the last section can be adapted for use on gamepads, none of them were developed specifically for the usage on gamepads. Thus, we propose in this section a novel password entry scheme: *Colorwheels*. *Colorwheels* is specifically designed for shoulder-surfing resistant input of text passwords on gamepads. The general design of the scheme is based on pie menu structures [6], similar to the Daisy Wheel scheme described in section 3. Its interface consists of two pie menu "flowers" with eight petals each. These overall 16 petals contain all possible characters (i.e. uppercase and lowercase letters, numbers, and special characters) for the password entry. This design is depicted in figure 3. *Colorwheels* is designed specifically for text entry with a gamepad, thus meeting requirements *R*4 and *R*5. Due to the two flowers, *Colorwheels*'s operation necessitates the availability of two analog sticks on the gamepad. Each stick is used to select petals in one of the flowers: the left stick to select petals in the left flower and the right stick to select petals in the right flower. Each petal holds either 6 or 5 characters, allowing the placement of all 94 printable ASCII-characters (excluding the white space) on the petals of the two flowers.

To introduce shoulder-surfing resistance into the scheme, the entry of each password character is performed with the following procedure: (1) The characters appear randomly distributed on the petals of the two flowers. (2) The user locates the petal with the desired character and presses the ×-button to confirm that they have found it. Upon pressing the button, all the characters vanish from the petals. (3) The user selects the petal. Each flower corresponds input-wise to one of the two analog sticks. To select a petal in the left or right flower, the left or right analog stick has to be used respectively. Since there are only eight different positions for each analog stick, the scheme meets $R3$. Then the selection of the petal is confirmed using any of the shoulder buttons.

At any time during the procedure, the scheme requires at most two concurrent controls on the front and less than the number of overall available buttons, meeting requirements $R2$ and $R6$. Using the □-button, the last entered character can be deleted. To obtain the password, a shoulder-surfing attacker would have to memorize the random distribution of all characters in the time the user locates the petal with the character they want to enter. We believe this task is cognitively demanding enough to protect against opportunistic shoulder surfing, i.e. meeting requirement $R1$.

5 METHODOLOGY OF THE USER STUDIES

We evaluate our proposal *Colorwheels* described in the last section against the following two existing approaches: *Firstly,* we compare it against the grid-based scheme. Of the three schemes identified in section 3, the grid-based scheme seems to be the most suitable for a comparative study. Behavioural biometrics have not yet been applied in the gamepad context and getting a baseline is beyond the scope of this work. PairPasswordChar requires the memorization of multiple rules for its correct operation and is thus not easily understood. *Secondly,* we compare it against the on-screen keyboard: While it likely does not meet requirement $R1$ as explained in section 3, it is the current de facto standard in the gamepad context and hence a useful baseline for our comparison. We measure the actual level of shoulder-surfing resistance for all three schemes as well as their general usability according to ISO 9241-11:2018-03 [18]. The methodology conforms to all requirements of our university's ethics commission.

Hypotheses. Since the *on-screen keyboard* (as our control condition) does not employ any measures to counter shoulder-surfing attacks, we expect it to perform worst in this aspect. The respective hypotheses are:

H_{1a}*: The Colorwheels scheme is more resistant to shoulder-surfing attacks than the on-screen keyboard.*
H_{1b}*: The grid-based scheme is more resistant to shoulder-surfing attacks than the on-screen keyboard.*

While the *grid-based* scheme is geared towards being shoulder-surfing resistant, the full entry is gradually revealed during password entry and old input is visible until the complete password is entered. Thus, a simple shoulder-surfing strategy is to memorize the characters in one column each time and then check during the next observation which one of the five characters appears again. In contrast, *Colorwheels* shows the randomly distributed characters only before the individual character selection is performed (cf. section 4). Therefore, the attacker would have to memorize the complete (randomized) character layout (i.e. the positions of all 94 characters) for each character in the password. Therefore, we expect *Colorwheels* to be more resistant:

H_{1c}*: The Colorwheels scheme is more resistant to shoulder-surfing attacks than the grid-based scheme.*

Design Decisions for the Methodology. Most studies regarding shoulder-surfing resistance are conducted as lab studies (e.g. [4, 5, 11, 13, 16, 19, 20, 26, 30, 35]). In contrast, only very few user studies have been conducted online [1, 31]. Yet, both types of studies have their advantages and disadvantages in the domain of password entry on gamepads.

A lab study allows our participants to use the schemes themselves to get familiar with the potentially unknown scheme. At the same time, a lab study also allows testing the usability of the schemes. In contrast, an online study would have to rely on explanatory videos and texts for the familiarization with the schemes and testing the usability would not be possible. Also, users might be more motivated in a lab study since they are observed. Online studies, on the other hand, are unobserved and therefore allow controlling less confounding variables. Attention check questions can help mitigate this problem by asking whether the participant cheated [1]. However, such attention checks always rely on self-reported data and must thus be complemented by other metrics. Yet, participants can engage the online study whenever they have time, without the need of supervision by an experimenter, facilitating the collection of large samples.

Due to the different advantages and disadvantages, we decided to conduct an online study and a lab study. The design of our two studies is based on the methodology of Aviv et al. [1], who conducted a blend of an online study and a lab study to gain a baseline for the shoulder-surfing resistance of smartphone PINs and the Android pattern lock. In the following we only give a brief overview of their methodology and how we needed to adapt it for the gamepad context. The detailed procedures are described in section 6 for the online study and section 7 for the lab study.

The actual evaluation of the shoulder-surfing resistance by Aviv et al. [1] is based on 10 attack trials using "video recordings of a single expert user being attacked by participants". We decided to follow this methodology. As in [1], the videos for the grid-based and on-screen keyboard schemes were varied regarding the entry speed of the password and the interaction, i.e. different paths taken on the *on-screen keyboard* from one character to the next and different scrolling direction for the columns in the *grid-based* scheme. For *Colorwheels* only the speed was varied, since the interaction cannot be changed. Due to the mobile scenario, Aviv et al. [1] also varied the size of the device and the viewing angle in the videos and treated both as within-factors. We felt that this was not in line with the gamepad scenario, since it can be assumed that the observer has a clear and unobstructed view of the full GUI on the screen (e.g. from the couch in front of the television), when the goal is to watch a movie or play a game together. Also, observing the gamepad in addition to the screen does not give an advantage, since all input is directly reflected in the authentication schemes' GUIs. In the *grid-based* scheme, the GUI shows which column is selected and the letters are shifted in the direction pressed on the analog stick. Only which letters are visible on the screen at the end is important to an observing attacker, the input at the gamepad is not. For *Colorwheels*, the GUI highlights the petal which is selected using the analog sticks, also directly reflecting the input. Consequently, the videos showed only the authentication schemes as displayed on the screen.

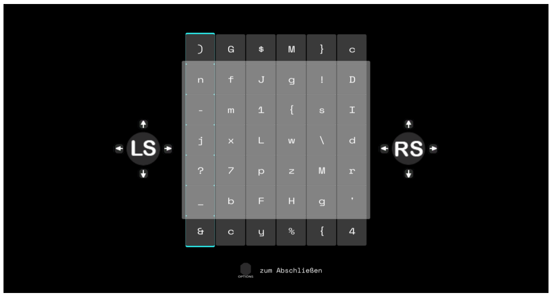
The password in the videos was chosen to not introduce bias into the study. We wanted to use a password that required usage of the different character sets in the *on-screen keyboard*, since we did not want to render the shoulder-surfing task unnecessarily easy for this scheme and therefore favour the other two schemes in the study. Also, using a dictionary word might have put the *grid-based* scheme in a disadvantage and favoured the other two schemes. Therefore, the password in the videos was chosen at random to include uppercase letters, lowercase letters, numbers and symbols. The random password entered in all the videos for the *on-screen keyboard* and the *grid-based* scheme is *W8@b=L*. The length of six characters was chosen according to the NIST recommendation for random memorized secrets [17]. To align the guessing probability of *Colorwheels* with that of the *grid-based* scheme, we had to increase the length of the password by one character to *W8@b=Lx*.

Like Aviv et al. [1], we treated the different schemes as between-subjects factor and participants were not allowed to take notes. For our lab study, we recruited participants locally. For the online study, we used the German panel "Clickworker" since this allowed us to use the exact same videos as for the locally recruited (German) participants.

Apparatus. We implemented all three schemes using the Unity game engine. During development, all implementations were tested in informal pre-test sessions with people recruited on campus. We already described the implementation of *Colorwheels* in section 4.

On-Screen Keyboard. The implementation was designed to resemble the layout and functionality commonly found on gamepad-driven devices.

Grid-Based Scheme. This scheme was designed to resemble the original depictions in [21] as much as possible. The interface comprises a grid of 6 by 7 cells, as depicted in figure 4. However, the centre 6 by 5 cells (visually highlighted in the interface) are considered for the password entry. To operate the scheme only the analog sticks are needed. Pushing any stick to the left or right lets the user select the column. Pushing up or down scrolls the characters in that column. Following the pre-tests, we also added the functionality to scroll the characters in the selected column up with the L1/R1-buttons and to scroll the characters in the column down with the L2/R2-buttons. The scroll buttons in the original depictions, which were used for the mouse input, were not needed anymore and therefore removed.



)	G	$	M	}	c
n	f	J	g	!	D
-	m	1	{	s	I
j	x	L	w	\	d
?	7	p	z	M	r
_	b	F	H	g	'
&	c	y	%	{	4

Figure 4: Our implementation of the *grid-based* scheme. The currently selected column is highlighted in turquoise.

6 ONLINE STUDY

Procedure. The study consisted of the following phases:

Introduction and Informed Consent. The participants first received information about the study scenario, the remainder of the study, the anonymity of their data, and explanations in case they want to withdraw from the study. Furthermore, participants were asked to not complete the study on a mobile device and to provide their consent for participation and processing. On the next site, they were asked whether their eye sight was normal (with or without corrections). Participants who reported to have a bad eye sight were told that they could unfortunately not participate in our study.

Familiarization with the Scheme. Participants were then randomly assigned to one of the three schemes and shown an explanatory text and illustrations describing the assigned scheme. They were also asked to watch a video similar to those used later in the study but with a different password (they were told which password was entered in this video).

Shoulder-surfing the Scheme. Participants had to play the role of an attacker performing a shoulder-surfing attack. To that end, they watched the videos as outlined in section 5. Participants were told that they are not allowed to pause or rewind the video, take pictures or videos of the online study and its contents, or use pen and paper to take notes. We further hid the control elements of the video player to prevent participants from pausing, rewinding, or replaying the video. Videos could be started by the participant by clicking on it and were played in full-screen mode. We also hid the button for proceeding to the next page of the questionnaire until the video was finished. Participants watched the

Table 2: Participants' demographics (F=female, M=male) and the shoulder-surfing results of the online study.

	Gender		Age						Number of observations										Failed to obtain password
Authentication scheme	F	M	< 20	20–30	31–40	41–50	51–50	> 60	1	2	3	4	5	6	7	8	9	10	
On-screen keyboard	8	11	0	6	5	5	2	1		4	1	1	2		1	3			7
Grid-based scheme	6	13	0	4	8	4	3	0							1		1		17
Colorwheels scheme	14	12	1	9	7	6	3	0											26
Total	*28*	*36*	*1*	*19*	*20*	*15*	*8*	*1*											

videos for the scheme they were assigned to one after the other until they either had successfully recovered the password from the input shown in the video, or had watched all ten videos without successfully recovering the password. In between the videos they could either enter a guess for the password in a free text field or indicate that they have no idea about the password at all. In case the participants successfully recovered the password they were complimented on their performance. In case they failed to recover the password they were told that they should not bother as the aim of the study was to evaluate the respective scheme's resistance against shoulder-surfing. Participants who were assigned *Colorwheels* or the *grid-based* scheme were also told that this scheme was specifically developed to be resistant.

Attention Checks. We asked them whether they had used any aids to help them guess the password, whether they had found and applied a possibility to pause or rewind the video, and whether they had completed the study on a mobile device. These questions served both as attention check and to check whether participants had followed our instructions.

Demographics and Debriefing. Finally, participants were asked to provide information about their demographics. On the last page, we thanked the participants and provided the code they needed to receive their compensation as well as contact details in case any questions would arise.

Participants. We recruited our participants using the German panel "Clickworker". Participants required on average 13 minutes and received a compensation of 3€. 93 participants completed our study. 10 had to be excluded from the analysis because they failed attention checks. In addition, we excluded 10 participants who stayed less than 14 seconds on the page introducing the scheme, since this is insufficient to familiarize themselves with the scheme[2]. Four participants had to be excluded because the completion times for the video pages were shorter than the length of the video and five due to technical problems. The final sample thus includes 64 participants (see table 2).

Results. To test H_{1a} - H_{1c}, we ran three Mann-Whitney U tests to account for the ordinal scale level of our data (i.e., number of observations needed to obtain the password, whereas participants who failed to obtain the password after having watched all ten videos were coded with "11"). We used a Bonferroni-adjusted alpha-level of .0167. Table 2 lists the values for all three authentication schemes. The analysis showed that, in accordance to our assumptions, the *on-screen keyboard* is least resistant to shoulder-surfing attacks, with participants needing more observations to obtain the password compared to *Colorwheels* (Z=-4.61, p<.001, r=.687). Thus, H_{1a} is supported. Likewise, more observations are needed to obtain the password entered with the *grid-based* scheme than with the *on-screen keyboard* (Z=-3.50, p=.002, r=.567), thereby supporting H_{1b}. The analysis did not reveal significant differences between the number of observations needed to obtain the password entered with the *Colorwheels* scheme and with the *grid-based* scheme (Z=-1.67, p=.094). Thus, H_{1c} is not supported.

7 LAB STUDY

Procedure. The study consisted of the following phases:

Introduction and Informed Consent. The participants first received a short briefing and signed the consent form.

Familiarization with the Scheme. Participants were randomly assigned to one of the schemes and received an explanation of the scheme. To familiarize themselves with the assigned scheme they used it three times to enter a password.

Usability Assessment. All participants had to enter three different randomly generated passwords which we provided to them. Effectiveness is measured using the portion of successful password entries among the three. Efficiency is assessed using the mean of the average time needed to enter the password across the three password entries. Satisfaction was measured with the SUS.

Shoulder-surfing the Scheme. All participants watched the videos as outlined in section 5. Participants watched the videos for the scheme they had used before, one after the other until they either had successfully recovered the password, they had watched all ten videos, or they asked to stop the experiment since they felt they would never be able to recover the password. In between the videos they noted their guess for the password on a paper provided to them.

[2] 14 seconds was the shortest time spent on the introductory page by a participant who successfully guessed the password.

Table 3: Participants' demographics in the lab study

	Gender		Age						Experience with game consoles		
Authentication scheme	Female	Male	< 20	20–30	31–40	41–50	51–50	> 60	low	medium	high
On-screen keyboard	14	15	3	16	3	5	1	1	10	10	9
Grid-based scheme	12	17	6	17	6	0	0	0	10	9	10
Colorwheels scheme	11	18	7	13	8	0	1	0	10	10	9
Total	37	50	16	46	17	5	2	1	30	29	28

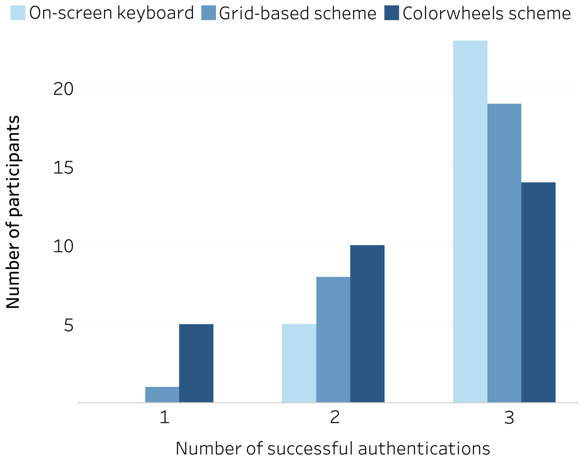


Figure 5: Effectiveness measured as the number of successful authentications performed in the first part of the study.

Demographics and Debriefing. Participants provided information about their demographics, received a short debriefing, were thanked, and received their compensation.

Participants. A total of 87 individuals (37 female, 50 male) participated in our study (see table 3 for participants' demographics). For our study, we wanted a diverse mix of participants having varying degrees of prior experience with gamepads and game consoles. Therefore, all potential participants had to fill out a short online signup-questionnaire asking them about their experience with game consoles and gamepads (low, medium, or high) and their email address, so that we could inform them in case they were selected for the study. Links to this signup-questionnaire were distributed on campus using flyers and mailinglists. Additionally, postings were made in several Facebook groups and online forums relating to console gaming to recruit participants outside the university. Participants received a compensation of 5€.

Shoulder-Surfing Resistance Results. Since the participants in our lab study had the opportunity to stop before having watched all ten videos, we cannot analyze how many observations participants needed to obtain the password. Participants might have been able to successfully guess the password if they had continued. We therefore consider how many participants succeeded in obtaining the password, independently of how many videos they watched.

To account for the nominal scale level of our data, we use Fisher's exact test to investigate H_{1a} - H_{1c}. We used a Bonferroni-adjusted alpha-level of .0167. Similar to the results from the online study, the analysis showed that the *on-screen keyboard* provides little protection against shoulder-surfing, with 27 out of 29 participants successfully obtaining the password entered with this scheme, which is significantly more than those entered with the *Colorwheels* scheme (0 out of 29, FET: p<.001) and the grid-based scheme (11 out of 29, FET: p<.001). H_{1a} and H_{1b} are thus supported by our results. Finally, a third Fisher's exact test revealed that significantly more participants succeeded to obtain the password entered with the grid-based scheme than with the *Colorwheels* scheme (FET: p<.001), providing support for H_{1c}.

Usability Results. The usability was assessed in terms of effectiveness, efficiency, and the participants' satisfaction.

Effectiveness. Figure 5 shows an overview of the successful authentication attempts. We used non-parametric tests for the analysis since a Kolmogorov-Smirnov test indicated non-normally distributed scores for all three schemes (p<.001). A Kruskall-Wallis test revealed significant differences in effectiveness between the three schemes (χ^2(2)=7.54, p=.023, η^2=.066). Pairwise comparisons using Mann-Whitney U tests with a Bonferroni-adjusted alpha-level of .0167 revealed a significantly higher rate of successful password entries with the *on-screen keyboard* compared to the *Colorwheels* scheme (Z=-2.67, p=.008, r=.35). There were no significant differences between *Colorwheels* and the *grid-based* scheme (Z=-1.59, p=.113) or the *grid-based* scheme and the *on-screen keyboard* (Z=-1.22, p=.222). Looking at the errors participants made during the three authentication attempts reveals that the similarity of characters is a major problem of the *grid-based* scheme: 9 out of 11 failed authentication attempts using the *grid-based* scheme can be attributed to participants confusing the target character with a similar character, whereas only 3 out of 6 failed authentication attempts arise from this problem for the *on-screen keyboard* and 9 out of 20 for *Colorwheels*. Other errors include participants mixing up uppercase and lowercase letters (1 for the *on-screen keyboard* and *Colorwheels* each, 2 for the *grid-based* scheme), forgetting to enter a character or entering an extra character (2 for the *on-screen keyboard*), and entering completely wrong characters (10 for *Colorwheels*).

Efficiency. Figure 6 shows an overview of the times needed to enter the password. Outliers deviating more than 1.5-times the interquartile range from the mean were excluded from the analysis, resulting in an exclusion of four data points (two falling below the threshold for the *on-screen keyboard*, and one each falling below and exceeding the threshold for the *Colorwheels* scheme). All assumptions for conducting an ANOVA were met. Thus, we ran an ANOVA with the used scheme as the independent variable and the mean of the overall time needed to enter the password across the three password entries as the dependent variable. The analysis revealed significant differences in the mean time needed to enter the password (F(2,80)=93.78, p<.001, partial η^2=0.701). Pairwise comparisons with a Bonferroni-adjusted alpha-level of .0167 showed that participants need significantly less time to enter the password using the *on-screen keyboard* compared to the *Colorwheels* scheme (t(52)=-16.81, p<.001, r=.92) and the *grid-based* scheme (t(32.39)=-12.10, p<.001, r=.91). However, authenticating themselves also took participants significantly less time using the *Colorwheels* scheme compared to the *grid-based* scheme (t(34.36)=-4.88, p<.001, r=.62).

Satisfaction. Figure 7 shows an overview of the SUS scores. Again, we excluded outliers deviating more than 1.5-times the interquartile range from the mean from the analysis, resulting in an exclusion of four data points (three falling below the threshold for the *on-screen keyboard* and one falling below the threshold for the *Colorwheels* scheme). All assumptions for conducting an ANOVA were met. Thus, we ran an ANOVA with the authentication scheme used to enter the password as the independent variable and the SUS scores as the dependent variable. The analysis revealed significant differences in the SUS scores (F(2,80)=33.40, p<.001, partial η^2=0.455). Pairwise comparisons with a Bonferroni-adjusted alpha-level of .0167 showed that the SUS scores were significantly higher for the *on-screen keyboard* compared to the *Colorwheels* scheme (t(40.01)=5.71, p<.001, r=.67) and the *grid-based* scheme (t(39.30)=8.85, p<.001, r=.82). However, the SUS scores also indicate that participants were significantly more satisfied with the *Colorwheels* scheme than with the *grid-based* scheme (t(55)=2.92, p<.005, r=.37).

8 DISCUSSION

Study Results. Unsurprisingly, the *on-screen keyboard* does not fare well in terms of security. Hypotheses H_{1a} and H_{1b} are supported in both studies, indicating that the *grid-based* scheme and our own proposal *Colorwheels* are more shoulder-surfing resistant than the *on-screen keyboard*. The results regarding H_{1c}, i.e. the differences between the *grid-based* scheme and *Colorwheels* are more ambiguous. While the online study does not indicate a difference, the lab study does. This discrepancy might be due to differences in the study setting. The attackers in the lab study were stronger: they (1) had expertise with gamepads (cf. table 3), (2) used the schemes before the actual "attack" and (3) saw their previous guesses when writing down the next one (since all guesses were written down on the same sheet of paper). We believe this points towards *Colorwheels*'s shoulder-surfing resistance being more robust, even in the face of stronger attackers.

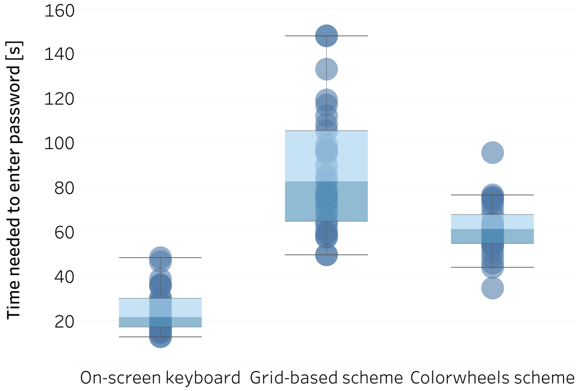


Figure 6: Boxplots of the mean overall time needed to enter the password.

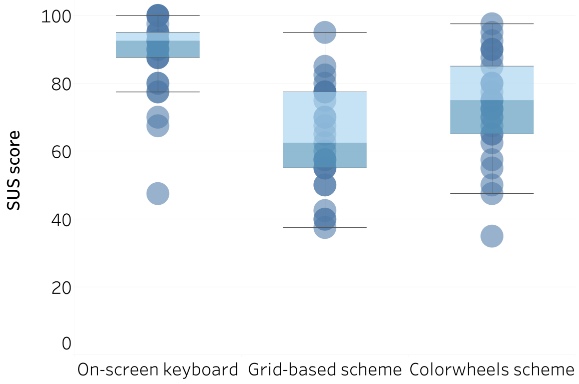


Figure 7: Boxplots of the SUS scores.

Usability-wise, the *on-screen keyboard* fares overall best, outperforming the other two schemes. This is unsurprising, since the scheme is deployed in the wild and therefore participants are likely to know it. In contrast, *Colorwheels* and the *grid-based* scheme were unknown to them. Regarding the differences between these two schemes, *Colorwheels* fares slightly better. Efficiency-wise, the combination of visual search and interaction required to enter each character in the *grid-based* scheme seems to take longer than the alternating visual search and interaction tasks in *Colorwheels*. Also, Shiffrin and Schneider [28] showed that people can learn to search in parallel for a particular set of targets, indicating that entry times with *Colorwheels* might decrease over time. In the gamepad context passwords are usually not entered very frequently, rendering efficiency less important. Effectiveness-wise there does not seem to be a significant difference. Yet, both schemes can be improved: despite the font in the schemes being explicitly chosen to be monospaced, with serfis, and so that usually similar looking characters (e.g. the letter "O" and the number "0") could be distinguished, it lead to a number of mistakes, where similar characters were confused with one another. Satisfaction-wise, *Colorwheels* and the *grid-based* scheme fare worse than the *on-screen keyboard*. However, *Colorwheels*'s SUS score (75) is still in the "good" range [3] and exceeds the grid-based scheme (64).

Limitations. We discuss the limitations of our studies along the metrics validity, replicability, and reliability [7–9].

Validity. The online study setting did not allow using a large screen (e.g. television) as monitor for the videos. Hence, to remain compatible, the lab study was performed on a smaller 13" screen as well. Thus, participants might have performed differently in settings with large screens. We argue, however, that this limitation does not negatively impact our study. Participants had the explicit instruction to try and observe the password entry. Thus, the privacy-diminishing effects of large displays as outlined in section 2 lose importance. Furthermore, both studies used videos. We had to compromise and use videos due to the intention of conducting an online study. Yet, it must be acknowledged that in a live observation study more attackers might have successfully observed the password [2, 33]. Also, all videos were watched directly one after the other and there was no other interaction during password entry. Usually the victim would enter the password only once and in case multiple people are present, the attacker might be distracted by e.g. a conversation. In the lab study the setting was easily controlled. It was ensured that participants followed all instructions. To compensate the lack of direct control over the participants, the online study used additional self-reported attention check questions (e.g. whether participants took notes) and technical measures (e.g. to prevent controlling the video playback and prevent skipping of videos by hiding the button to get to the next page until the video had finished playing). Yet, all measures ran in the participant's browser. Some participants managed to circumvent these measures and had to be excluded from our study. These exclusions were based on the participants' time spent on the pages showing the videos. We also excluded participants who spent very little time familiarizing themselves with the authentication schemes in the online study. Thereby, we used the same time limit for all schemes, despite participants being potentially already familiar with the *on-screen keyboard*, to prevent favouring the other schemes, by making sure people spent more time with their descriptions. Moreover, analogously to Aviv et al. [1] we did not allow participants to take any notes during the attack. We believe this increases the consistency between the two studies and among the participants of the online study. Yet, in a real "attack", an attacker might use their smartphone to take text notes, when recording a video is too obvious. Lastly, the discrepancy in the shoulder-surfing resistance between *Colorwheels* and the *grid-based* scheme in the two studies might be due to the differences in the study setting leading to a stronger attacker in the lab study.

Replicability. Using videos for the evaluation of the shoulder-surfing resistance poses the challenge of selecting representative videos. Thus, we used recordings of one single expert user, but varied the speed and interaction between the videos. The participants saw the videos in a randomized order to mitigate ordering bias. To increase the replicability of our results, the videos are available as supplemental material. Another aspect potentially impacting replicability is the choice of the password used in the shoulder-surfing studies. As we outlined in section 5 the password was chosen as to not favor any scheme over the others. However, users seldom choose random passwords for their accounts. Thus, the shoulder-surfing results might differ, when another password is used. In particular, dictionary words might impair the resistance of the *grid-based* scheme.

Reliability. Our implementations recorded the effectiveness and efficiency metrics automatically and therefore reliably without errors. The satisfaction was measured using the standardized SUS. With respect to the recording of the shoulder-surfing metric, i.e. the password guesses, the online study recorded the password entry in a text field not obfuscating the input (i.e. not hiding the password behind symbols such as "*") and in the lab study the experimenter always clarified any legibility issues. Thus, we argue that the results of our studies are as reliable as possible. However, contrary to our expectations, participants asked to be allowed to stop the shoulder-surfing task prematurely, if they believed it to be futile. We let those participants stop out of ethical considerations and since we believe that this might reflect the sentiment of an opportunistic observer in the real world. This introduced the subjective perception of success probability into our shoulder-surfing metric for the lab study. Thus, a direct comparison of the two studies is not possible.

ACKNOWLEDGEMENTS

Many thanks to our reviewers for their valuable comments and to Andreas Koch for his input regarding the PIN protection on gaming consoles. This work has been co-funded by the DFG as part of project D.1 within the RTG 2050 "Privacy and Trust for Mobile Users". This work was further supported by the German Federal Ministry of Education and Research (BMBF) in the Competence Center for Applied Security Technology (KASTEL) and the Center for Research in Security and Privacy (CRISP).

REFERENCES

[1] Adam J. Aviv, John T. Davin, Flynn Wolf, and Ravi Kuber. 2017. Towards Baselines for Shoulder Surfing on Mobile Authentication. In *the 33rd Annual Computer Security Applications Conference*. ACM, 486–498.

[2] Adam J. Aviv, Flynn Wolf, and Ravi Kuber. 2018. Comparing Video Based Shoulder Surfing with Live Simulation. In *Annual Computer Security Applications Conference*. ACM, 453–466.

[3] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining what individual SUS scores mean: adding an adjective rating scale. *Journal of usability studies* 4, 3 (May 2009), 114–123.

[4] Andrea Bianchi, Ian Oakley, and Dong Soo Kwon. 2010. The secure haptic keypad: a tactile password system. In *Conference on Human Factors in Computing Systems*. ACM, 1089–1092.

[5] Andreas Bulling, Florian Alt, and Albrecht Schmidt. 2012. Increasing the security of gaze-based cued-recall graphical passwords using saliency masks. In *Conference on Human Factors in Computing Systems*. ACM, 3011–3020.

[6] Jack Callahan, Don Hopkins, Mark Weiser, and Ben Shneiderman. 1988. An empirical comparison of pie vs. linear menus. In *International Conference on Human Factors in Computing Systems*. ACM, 95–100.

[7] Donald T. Campbell and Julian C. Stanley. 1963. Experimental and quasi-experimental designs for research. *Handbook of research on teaching* (1963), 171–246.

[8] Edward G. Carmines and Richard A. Zeller. 1979. *Reliability and validity assessment*. Vol. 17. Sage publications.

[9] Dato N. M. de Gruijter and Leo J. Th. van der Kamp. 2007. *Statistical test theory for the behavioral sciences*. Chapman and Hall/CRC.

[10] 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 *Conference on Human Factors in Computing Systems*. ACM, 987–996.

[11] 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 *ACM Conference on Human Factors in Computing*. ACM, 2937–2946.

[12] Alexander De Luca, Marc Langheinrich, and Heinrich Hussmann. 2010. Towards Understanding ATM Security: A Field Study of Real World ATM Use. In *Symposium on Usable Privacy and Security*. ACM.

[13] Alexander De Luca, Emanuel von Zezschwitz, Ngo Dieu Huong Nguyen, Max-Emanuel Maurer, Elisa Rubegni, Marcello Paolo Scipioni, and Marc Langheinrich. 2013. Back-of-device authentication on smartphones. In *International Conference on Human Factors in Computing Systems*. ACM, 2389–2398.

[14] Entertainment Software Association. 2017. *Essential Facts About the Computer and Video Game Industry*. Technical Report.

[15] Brian W. Epps. 1987. A Comparison of Cursor Control Devices on a Graphics Editing Task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 31, 4 (Sept. 1987), 442–446.

[16] Alain Forget, Sonia Chiasson, and Robert Biddle. 2010. Shoulder-surfing resistance with eye-gaze entry in cued-recall graphical passwords. In *Conference on Human Factors in Computing Systems*. ACM, 1107–1110.

[17] Paul A. Grassi, James L. Fenton, Elaine M. Newton, Ray A. Perlner, Andrew R. Regenscheid, William E. Burr, Justin P. Richter, Naomi B. Lefkovitz, Jamie M. Danker, Yee-Yin Choong, Kristen K. Greene, and Mary F. Theofanos. 2017. *Digital Identity Guidelines: Authentication and Lifecycle Management*. Technical Report.

[18] ISO 9241-11:2018-03 2018. *Ergonomics of human-system interaction - Part 11: Usability: Definitions and concepts*. Standard. International Organization for Standardization, Geneva, Switzerland.

[19] Hassan Khan, Urs Hengartner, and Daniel Vogel. 2018. Evaluating Attack and Defense Strategies for Smartphone PIN Shoulder Surfing. In *Conference on Human Factors in Computing Systems*. ACM, 164–10.

[20] David Kim, Paul Dunphy, Pam Briggs, Jonathan Hook, John Nicholson, James Nicholson, and Patrick Olivier. 2010. Multi-touch authentication on tabletops. In *Conference on Human Factors in Computing Systems*. ACM, 1093–1102.

[21] Sung-Hwan Kim, Jong-Woo Kim, Seon-Yeong Kim, and Hwan-Gue Cho. 2011. A new shoulder-surfing resistant password for mobile environments. In *International Conference on Ubiquitous Information Management and Communication*. ACM, 27.

[22] Katharina Krombholz, Thomas Hupperich, and Thorsten Holz. 2016. Use the Force: Evaluating Force-Sensitive Authentication for Mobile Devices. In *Symposium on Usable Privacy and Security*.

[23] M. Kameswara Rao and Sushma Yalamanchili. 2012. Novel Shoulder-Surfing Resistant Authentication Schemes using Text-Graphical Passwords. *International Journal of Information Network Security* 1, 3 (2012), 163–170.

[24] Karen Renaud and Joseph Maguire. 2009. Armchair authentication. In *British HCI Group Annual Conference on People and Computers*. British Computer Society.

[25] Volker Roth, Kai Richter, and Rene Freidinger. 2004. A PIN-entry method resilient against shoulder surfing. In *ACM conference on Computer and communications security*. ACM, 236–245.

[26] 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 *Conference on Human Factors in Computing Systems*. ACM, 2045–2048.

[27] Hirokazu Sasamoto, Nicolas Christin, and Eiji Hayashi. 2008. Undercover: authentication usable in front of prying eyes. In *Conference on Human Factors in Computing Systems*. ACM, 183–192.

[28] Richard M. Shiffrin and Walter Schneider. 1977. Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review* (1977), 127–190.

[29] Desney S. Tan and Mary Czerwinski. 2003. Information voyeurism: social impact of physically large displays on information privacy. In *CHI Extended Abstracts on Human Factors in Computing Systems*. 748–749.

[30] Emanuel von Zezschwitz, Alexander De Luca, Bruno Brunkow, and Heinrich Hussmann. 2015. SwiPIN: Fast and Secure PIN-Entry on Smartphones. In *Conference on Human Factors in Computing Systems*. ACM, 1403–1406.

[31] 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 *Conference on Human Factors in Computing Systems*. ACM, 2339–2342.

[32] Susan Wiedenbeck, Jim Waters, Leonardo Sobrado, and Jean-Camille Birget. 2006. Design and evaluation of a shoulder-surfing resistant graphical password scheme. In *AVI '06: Proceedings of the working conference on Advanced visual interfaces*. ACM, 177.

[33] Oliver Wiese and Volker Roth. 2015. Pitfalls of Shoulder Surfing Studies. In *Workshop on Usable Security*.

[34] 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.

[35] Christian Winkler, Jan Gugenheimer, Alexander De Luca, Gabriel Haas, Philipp Speidel, David Dobbelstein, and Enrico Rukzio. 2015. Glass Unlock: Enhancing Security of Smartphone Unlocking through Leveraging a Private Near-eye Display. In *Conference on Human Factors in Computing Systems*. ACM, 1407–1410.

[36] Nur Haryani Zakaria, David Griffiths, Sacha Brostoff, and Jeff Yan. 2011. Shoulder surfing defence for recall-based graphical passwords. In *Symposium on Usable Privacy and Security*. ACM.

[37] Huanyu Zhao and Xiaolin Li. 2007. S3PAS: A scalable shoulder-surfing resistant textual-graphical password authentication scheme. In *International Conference on Advanced Information Networking and Applications Workshops (AINAW)*. IEEE, 467–472.