Project Report
HSF Research Area 4

Graphical Authentication on Augmented Reality

Principle Investigator of the Research Area: Melanie Volkamer
Involved Project Staff: Reyhan Düzgün, Tobias Hilt, Philipp Matheis, Peter Mayer

1 Summary

Authenticating at Augmented Reality head-mounted displays (HMD) usually requires users to select their (6 digit) PIN on the virtual PIN pad once they start using the AR glasses. Unfortunately, the PIN entry can easily be observed. Past research has proposed fully shoulder surfing-resistant authentication schemes and some of them have also been applied and evaluated in the AR context. We had a closer look at "Things", a recognition-based graphical password scheme and (a) identified several shortcomings both with the scheme as well as with the methodology of previous evaluations; and (b) noticed that due to the virtual screen in AR HMD, it is worth studying whether there are grid sizes that fit better for this context than others. Consequently, we performed a between-subject lab study (N=126) evaluating three different combinations of grid size and length of the secret.We found that a grid of 10 images displayed in two rows showed small advantages but from the qualitative data, we conclude that the best overall usability can be reached by offering personal choice. Thus, users should be able to decide on their preferred grid size and length of secret.

2 Introduction

When authenticating at Augmented Reality (AR) glasses, currently, users see on their virtual screen a PIN pad and are asked to enter their six digit PIN by selecting the corresponding numbers with their finger. While observers cannot see the screen, they can observe the hand movement (either manually or by video recording) and use this information to deduce the PIN (see Fig. 1). As it is not possible to hide the hand movement similarly to when entering a PIN at an ATM, shoulder-surfing resistant authentication schemes [20] are required for AR glasses.

This problem was already identified by Düzgün et al. [9]. The authors propose to use Things - a graphical recognition-based authentication scheme for which the secret consists of several images which are displayed among distractors in a randomized order. The Things scheme solves the shoulder surfing issue, but also allows taking advantage of the benefits of graphical passwords being more memorable than textual passwords such as PINs[1]. The authors conducted a first user study evaluating the schemes' usability with promising results. However, the prior proposal as well as the conducted user study come with some shortcomings which we address with our adopted Things AR scheme and our usability evaluation.

With respect to the authentication scheme, the main differences are (for a complete list and justifications see Section 4): (1) Three mandatory training rounds during the registration phase (and optionally more if needed) to ensure users understand how to use the scheme; (2) Ten tries before the trial ended (both during the third training round as well as during the authentication); (3) Different approach for image selection: only one image per semantic group on one grid rather than per grid only images from the same semantic group; (4) Option to display the selected images before confirming that this secret should be considered which is possible as only users see the virtual display; (5) The images in the grid are arranged more horizontally instead of vertically to better fit the aspect ratio of the users field of view.

With respect to the methodology, the main differences are (for a complete list and justifications see Section 5.6): (1) All participants got the same set of images as secret to increase the internal validity; (2) Questions about users' perception were added.

[1] Graphical authentication leverages the pictorial superiority effect [24].

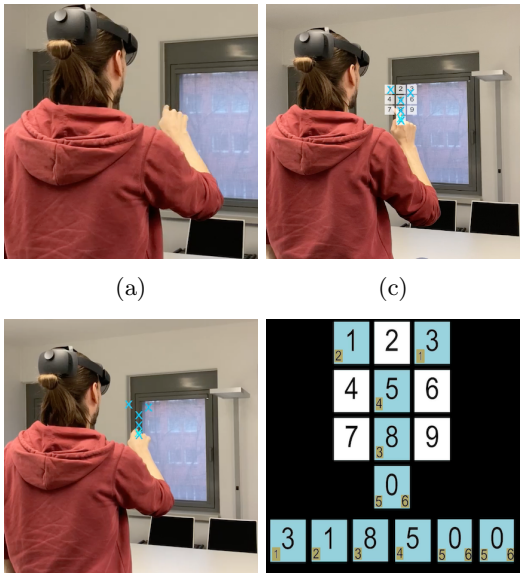

Figure 1: Shoulder-surfing a password in AR: a) observe the input, b) register the position and order of inputs c) match the registered input to the standard pin pad and d) reconstruct the password.

Düzgün et al. [9] decided to go with a five image secret and a grid size of 16 images (to reach a similar security level as the default six number PIN). However, given the shape of the AR virtual screen it is not obvious that this is the best fit for the AR context - which was also acknowledged by the authors. With the aspect ratio of the HMD being different to standard displays, we aim to find the best fit for a grid size. Accordingly, as we wanted to keep the security level of a six-digit PIN, the length of the secret had to be altered along side the grid size. Therefore, besides fixing some issues with the scheme and the methodology, we see our main contribution in studying different combinations of secret-length and grid-size (i.e., 8 secrets with grid size 6, 6 secrets with grid size 10, 4 secrets with grid size 32). Our main research question therefore is:

- Which of the three settings performs best regarding effectiveness, efficiency, and satisfaction?

For potential users to adopt the scheme, their perception of it is equally important as the captured metrics, henceforth, we also aimed to examine the general perception of graphical authentication schemes [34, 33].

Our results show that all three grid sizes perform very similar in terms of effectiveness and satisfaction. Regarding efficiency, grid size 10 performs best. Analyzing the perceived usability, we could not detect big differences between the grid sizes, indicating that grid size and length of secret is objective to personal preferences of the users. Consequently, to reach maximum usability, different grid sizes and secret lengths with a similar security level should be made available to users to select their preferred choice.

3 Related Work

In order for an authentication scheme on augmented reality (AR) devices to be fully shoulder-surfing resistant, it is necessary, according to Länge et al., [20] that no information is leaked from observing. Therefore it does not matter how often the entry of the secret is observed and how it is captured, either by direct observation or by video recording. We first discuss existing proposals for authentication on augmented reality glasses as well as virtual reality (VR) head mounted displays. Note, we also consider those for VR because it is likely that a scheme proposed for VR can be easily adopted for AR.

Numerous authentication schemes have been developed for augmented reality (AR) [28], each with distinct features and considerations in terms of resilience to shoulder surfing and usability. These authentication schemes can be categorized in knowledge-based, biometric and token-based schemes. Since token-based schemes require a second device to be carried around, and biometrics have drawbacks in terms of privacy and reliability , we focus on knowledge-based schemes, which are also often required as a fallback, even when another type of authentication is used. Among knowledge-based schemes we identified two approaches to achieve full shoulder-surfing resistance, (1) using randomization and (2) schemes based on challenge-response.

Schemes based on challenge-response [8, 30] are rather complex and time-consuming to authenticate, as demonstrated in the user study conducted by Wang et al. [30] where the lowest average authentication time was 24.46 seconds.

A simpler approach is to use schemes that use randomization like randomized pin pads. There are multiple examples where the input mechanism varies slightly [3, 14, 19, 26, 31, 32]. For instance, Li et al. [18] proposed a scheme utilizing the touchpad of Google Glass, head movements, or verbal commands for PIN input. Another approach involves shuffling keyboards into random positions for text-based passwords [3, 19]. Although these schemes are fully resistant to shoulder surfing, they do have usability shortcomings, particularly in terms of password memorization, as they are all based on either numbers or text. Moncur et al.[22] conducted a comparison study between PIN and graphical passwords, with graphical passwords demonstrating superior performance.

There are some shoulder-surfing resistant graphical authentication schemes for AR and VR. Funk et al. [11] presented a recognition-based scheme for AR, where virtual objects are randomly positioned in a room and selected by looking and dwelling at the object to select it. Länge et al. [20] analyzed knowledge-based VR schemes to determine whether they meet the requirements for full shoulder-surfing resistance and conducted a usability study of the schemes that fulfill the requirements. Only a few of them are fully shoulder-surfing resistant. One of these schemes is Roomlock [12], which employs 3D objects positioned in a virtual room, similar to Funk et al. [11] but with a laser-pointer for selecting the objects. Länge et al. [20] proposed own fully shoulder-surfing resistant schemes for VR based on randomization. One of these schemes, Passimoji, is a graphical scheme similar to a randomized PIN field but emoji are used instead of numbers. Their study showed that 2D schemes like Passimoji have a significantly shorter authentication time compared to 3D schemes like RoomLock.

Thus we see most potential in 2D graphical authentication schemes. An extensive study by Mayer et al. [21] in 2014 compared several graphical authentication schemes, revealing "Things" to be the most effective among all examined schemes. The Things implementation used by Düzgün et al. [9] mentioned in the introduction is a fully shoulder-surfing resistent AR adaptation of the original Things.

4 Scheme Description

The scheme we used for this study builds upon the scheme [9] used for their study. Both schemes are based on "Things"[21], a recognition-based graphical authentication scheme which leverages the benefits of the pictorial superiority effect. The underlying idea of "Things" evolves around the fact, that instead of passwords, users authenticate with the help of a secret, consisting of different images. Users need to recognize the images from their secret on several grids containing various distractor images as well as one of the images of the secret, as illustrated in Figure 2. The images for each grid are different and for each sign-in their order is randomized making the scheme fully resistant to shoulder-surfing.

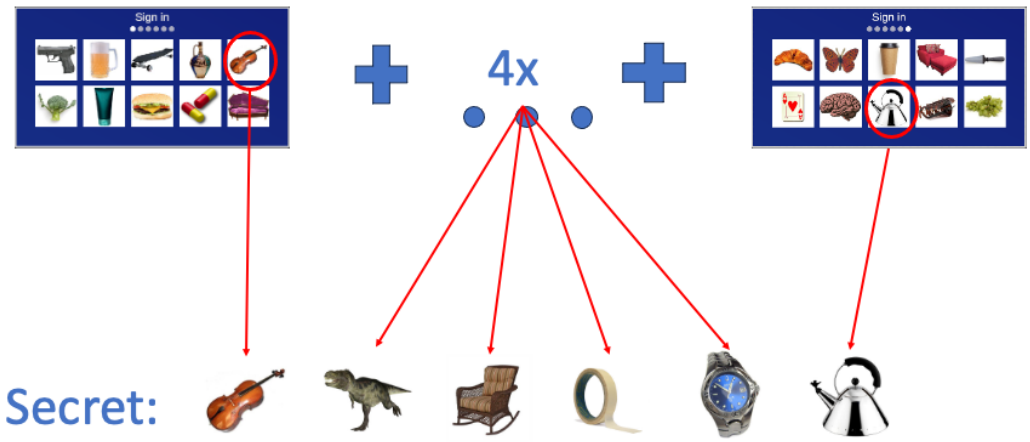

Figure 2: The secret being formed by different input from grid

For Things AR, the scheme we used in the study, we enhanced the scheme from [9] by implementing the changes as outlined in the following:

First, we implemented a mandatory training phase to make sure users understood how the scheme worked. Additionally, this training phase served as a step to let participants get used to the interaction with the Hololens. Current literature on AR refer to different number of training sessions ranging from three to ten rounds [27, 31, 16]. We opted to mandate at least three training rounds and gave participants the option to repeat the training as often as they liked. In the authentication rounds, participants needed to submit the correct secret within ten tries, as recommended by [7], to lock users out after ten failed authentication attempts. We also changed the images in one grid so that no two images look similar on the same grid, to reduce the risk of confusing participants by too similar images. In the original scheme the images on each grid were thematically related to each other, e.g. a grid of fruit images with the secret being an apple. The option to display the input before submitting was also added, to provide feedback to the participants to confirm their selection.

5 Methodology

5.1 Research Goal

Recognition-based graphical authentication schemes can be configured in different ways to reach the default security level of a 6-digit PIN by using a larger or smaller grid and correspondingly a shorter or longer number of secrets. To the authors' knowledge, the effect of different combinations of grid size and number of secrets has not yet been studied. As the virtual screen in the AR context and the interaction with it is different from well-known desktop and mobile contexts, we believe it is worth investigating this effect for the AR context. Therefore, our first research question is as follows:

RQ_{main}: How do different grid sizes influence the usability of recognition-based graphical authentication on the Hololens?

There are clear advantages of using recognition-based graphical authentication schemes, independently of the used settings, in the AR context (shoulder-surfing resistance; the better memorability of images compared to numbers). However, it might be that potential users have subjective doubts which one would need to address before introducing such a scheme for the AR context. Therefore, our second research question is as follows:

$RQ_{context}$: How do users perceive the recognition based graphical authentication in the AR context?

To answer these research questions, we designed and conducted a user study comparing three different grid sizes (32 images, 10 images, and 6 images). The details of the study are described in the remainder of this section.

5.2 Study Design

Recruitment We required our participants to meet two requirements, i.e., being at least 18 years old and not suffering from poor eyesight. If participants chose to counter poor eyesight by using contact lenses, they were also allowed to participate. The usage of glasses was not allowed to prevent potential injuries from using personal glasses and the Microsoft Hololens on top. We recruited participants through various channels. Firstly, we used snowball sampling by making announcements to fellow researchers at our university, through social media channels, and reached out to several organizations associated with innovative and digital technologies in proximity to our institutions campus. Secondly, we distributed leaflets at several public places, such as cafes, restaurants, small stores, or public transportation stations.

Data Protection & Ethics In cooperation with the data protection officer of our university, we created information about the usage of the collected data, conforming to recent GDPR, which we presented to participants at the beginning of the study to inform participants about their rights and the usage of their collected data.

Our institution does not require approval by an IRB, but we took the most care when designing the study. Fatigue or general discomfort when using any AR device can occur, but we informed participants that in such a case they were allowed to put down the device and would still be paid their participation fee. Due to the COVID-19 pandemic, we took

several hygiene measures, such as providing gloves, disinfecting the study devices after each participant and wrapping the glasses into plastic, which was renewed for each participant. Participants were offered a compensation of 15 euros for their efforts, which included the time for participating in the study and potential travel to the study destination. The compensation represents a payment above minimum wage in Germany. The participants' compensation was paid using bank transfer, as our institute mandates this way of paying participants. The necessary process was approved by the data protection officer of our institution. Participants were also given the option to waive the compensation, which two participants did. One due to privacy concerns (we needed to collect their IBAN in order to transfer the money) and one stated personal reasons.

5.3 Study procedure

A between-subject design was used and the full study procedure is illustrated in Figure 3. In the subsequent paragraphs, we provide a more elaborate explanation of the key steps and major processes comprising the study procedure.

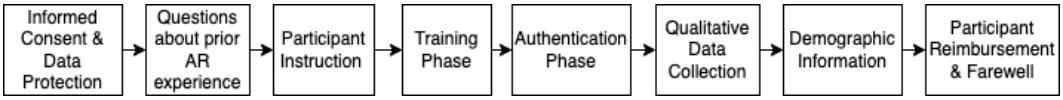

Figure 3: Flowchart illustrating the study procedure

Informed Consent & Data Protection Information. The study commenced with participants being greeted and seated, after which they started the study on a computer using an online questionnaire. Participants were presented with an informed consent for and the data protection regulations.

Questions about prior AR experience. The participants' prior experience with augmented reality/virtual reality (AR/VR) was assessed.

Participant Instructions. Participants received detailed instructions on paper, specifically tailored to the grid size of their group. These instructions covered general information about the Microsoft Hololens 2 and AR, as well as information about graphical passwords and their security. The scheme and study processes were extensively explained using both text and pictures. This was done, so that all participants had a good understanding of the scheme's working principles and differences in this regard between participants were minimized.

Training Phase. To familiarize participants with AR and the input process on the Hololens, a training phase was conducted. Participants were shown their password and told to memorize it. Overall there were three training rounds. During the first two training rounds, participants were allowed to make errors, such as inputting the wrong password. Moreover, the correct image in each grid was highlighted with a red framing (see 4). In these two rounds, the correctness of the input was not checked, as these rounds were intended for the participants to familiarize with the AR input motion.

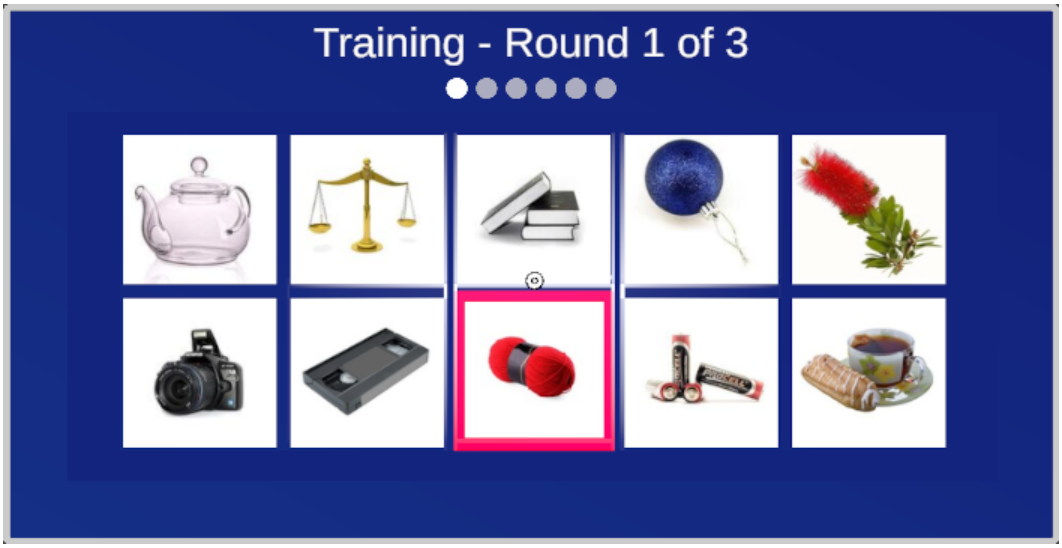

Figure 4: Highlighting of correct image in the first two training rounds.

However, in the third round, participants were required to input the correct password to proceed and no highlighting was present. Participants had a total of ten attempts for this round, in order to advance to the actual authentication phase. If participants failed to authenticate within ten tries, they were excluded from the study. After completing the training phase, participants were given the option to repeat the training if they desired.

Authentication Phase. The authentication phase also consisted of three authentication rounds[2] and followed the same regulation as the final round of training. Therefore, no highlighting of the correct image was given and participants had ten attempts for each of the authentication rounds.

SUS and open questions. Once all authentication rounds were completed, participants were asked to continue with the online questionnaire. The System Usability Scale (SUS) was employed to assess the participants' perception of the usability of the scheme. Additionally, open-ended questions were posed to gather feedback on participants' likes, dislikes, and their willingness to use the scheme further in AR and other contexts.

Demographics. We also collected some demographic information from our participants, namely their age, preferred gender, questions related to their technical expertise, if they are enrolled in university and if they need glasses.

Participant Reimbursement & Farewell. Finally, participants were thanked for their participation and asked to fill out the forms confirming their participation with their banking details (IBAN). Shortly after that, these forms were handed over to the finance department and the participants were paid by bank transfer.

[2] An extensive graphic illustration of the practical phase of participants with the Hololens is shown in the appendix A.

5.4 Pre-Study

We decided to test the implementation of the scheme and the study procedure in a pre-study. We recruited 21 participants in total to test the implementation and procedure for each of the grid sizes (seven participants for each grid size). During the pre-study, we collected feedback, which led to the adjustment of several settings in the application and the Hololens itself, such as brightness, hologram-to-eye distance or the point at which the Hololens registered user input (so to speak the depth at which one has to position their finger). We also implemented the option for participants to show their input before submitting by clicking on a dedicated button[3]. As a result of the pre-study we reworked the participant instructions to clarify the authentication steps, as many participants perceived the initial version of this instruction to be rather extensive and not intuitive. Finally, we replaced some of the images in the grids, to avoid possible misunderstandings and confusion as a few of the depicted objects were initially quite similar.

5.5 Data Analysis

The data analysis was performed using a mixed methods approach. The quantitative data was collected using the Hololens and the SUS questionnaire. From the Hololens, we collected several metrics of data namely the duration of each step, how often each button was pressed, which images were selected by the participants and if that image matched the password or was adjacent to the correct image. The data was extracted from the Hololens after each participant, saved in a unique file and then deleted from the Hololens. Qualitative data was collected using an online questionnaire as described before.

The mapping of data collected with the Hololens and from the online questionnaire was performed after the data collection was completed. We mapped both sources of data based on the timestamps related to each point of data to correctly combine the usage data from each participant's interaction with the Hololens and their collected data from the online questionnaire. Once the mapping was complete the timestamps were deleted and the combined data was shuffled to further anonymize the data. Once all participants were paid, the list containing their personal information was also deleted, so that no possibility of connecting one's personal data with the collected study data was possible.

The three major categories to evaluate usability, *effectiveness*, *efficiency*, and *satisfaction* were captured as follows. The effectiveness was measured by (a) the number of incorrect images selected, (b) the number of corrections made, and (c) the success rate. Efficiency was captured as (a) the time required to successfully authenticate and (b) how many repetitive displays of the given secret were needed for participants to memorize it, captured by inspecting how often participants pressed the repeat button during sign-up. If participants choose to correct their input or submitted a wrong password, the time for each attempt was added until they successfully authenticated. The system usability scale was used to assess the satisfaction rate of participants.

The same process was used in the analyses of the quantitative data for each captured variable. First outliers that were 1.5 times the interquartile range (IQR) over the third quartile (Q3) or below the first quartile (Q1) were inspected manually using the screen recorded video of the user interacting with the Hololens. If it was apparent that the participant didn't understand the study instructions, their complete recorded data (both from

[3]In the course of the main study we also examined the perception of the addition of this feature, which was proven to be well perceived by 72% of participants.

the Hololens and the online questionnaire) was excluded from the analysis. After removing these outliers, we tested the variable for normal distribution by performing a Shapiro Wilk test and manually checking the histograms. If the results were normally distributed we performed an ANOVA analysis, if not then we performed a Kruskal-Wallis test. If either ANOVA or Kruskal-Wallis reported significance a post-hoc t-test or Mann Whitney U-test with Bonferroni-Holm correction was performed to make pair-wise comparisons between the grid sizes.

The qualitative data was analysed using an inductive coding approach [25, 29], performed by two members of the research team. Initially, both reviewed 30% of the collected data and created codes to capture the themes in the content. Afterwards, they met and discussed the identified codes, their criteria of application, as well the explanation of each code. They also discussed the categorization of codes into several thematically distinct categories into which the codes were orchestrated. This process was conducted separately for each question that was analyzed to maintain clarity and structure. The full codebook can be found in appendix B. Subsequently, both coders re-applied the revised codebook to the whole data. The coded data was afterwards compared to detect disagreements in the application of codes. All disagreements were then solved by discussion between the two coders. These results were then discussed and approved by the remaining members of the research team.

5.6 Differences to the study conducted by Duezguen et al.

Although this work is based on the study conducted by Düzgün et al. [8], there are notable differences in both methodology and the implementation of the scheme. This section provides a detailed explanation of these differences and the rationale behind them. A major difference in terms of methodology lies in the provision of participant instructions, that were handed out to the participants during the study. These instructions contained detailed step-by-step explanations of the scheme and the study procedure, aiming to enhance participant understanding. We assigned every participant for each grid size the same password, to enhance comparability and reduce possible effects caused by different images being easier or harder to memorize. To aid participants in memorizing their password, in the first two rounds of training the correct image on each grid was highlighted with a red frame, which was inspired by [23, 2, 17]. By mandating at least one successful authentication we could guarantee that participants were able to memorize their assigned password, which was also implemented in similar studies such as [13, 1, 23, 6]. Instead of using a dedicated app, in the case of [9] "Hololens Tips", we mandated the participants to perform three training rounds as already explained. This provided both the possibility of participants to accommodate themselves with the interaction with the Hololens as well as helping them memorize the secret. We also opted to show the participants their assigned password and removed the buttons allowing them to go back and fourth during sign-up, to reduce cluttering the interface. Another modification was made to the grid sizes and password length to accommodate different conditions in our study. In [9], the grid size was fixed at 16 with a password length of 5. The authors of [9] pointed out, that in order to optimize the usability of the scheme the number of displayed images per grid should be changed. As we wanted to examine the influence of different grid sizes and secret lengths on the perceived usability, we decided to create different versions of the proposed scheme. In a first step we created versions with extreme values in regard to the grid size and secret length, while still maintaining the same theoretical password space. As a result of the pre-study an optimized version, settled in between these two extreme versions was created. Consequently, the study consisted of the

following three versions: 32 images per grid with a password length of 4, 10 images per grid with a password length of 6, and 6 images per grid with a password length of 8. To compare the three variations with each other, all the combinations of grid size and password length are chosen in a way that the password space is 10^6 like in the original things scheme. This corresponds to a 6-digit PIN and is according to Florencio et al. [10] enough to protect against online attacks. Due to the online-offline gap, a significantly larger password space would be required to further increase the security of the scheme. Additionally, we adjusted the grids to be more rectangular instead of squared. This change was made to better fit the field of view of the Hololens, enhancing the user experience (see Figure 5).

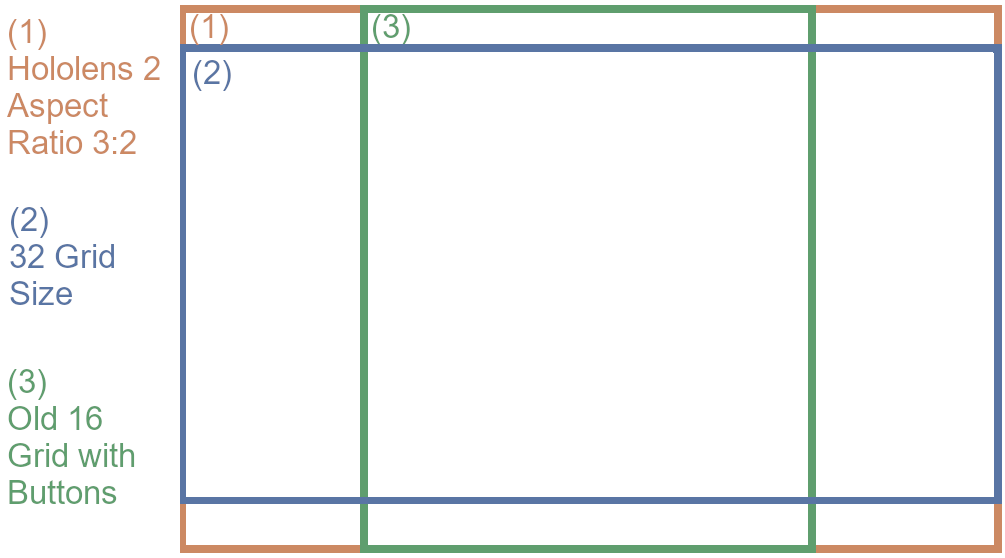

Figure 5: Comparison of aspect ratios of the Hololens 2, the grid of size 32 of this work and the grid size of 16 with buttons below the images from [9].

Moreover, we eliminated the option to go back to the previous grid or restart the authentication process at any point. Instead, users needed to input the whole password again. This way the graphical user interface was less cluttered, as no additional buttons were needed.

6 Results

6.1 Sample description

In total 126 people participated in the main study, from which 121 were included for the analysis[4]. Five were excluded due to the following reasons: in two instances the provided study equipment, namely the Microsoft Hololens 2, crashed, rendering the results from these participants unusable. One of the participants failed to authenticate in the training phase and two participants misunderstood the study instructions, which both led to exclusion.

[4]We calculated the minimal amount of participants to be able to draw statistical sound conclusions as 111, using G*Power, expecting an effect size of 0.3, with $1 - \beta$ as 0.8.

From the remaining 121 participants 80 identified as men, 40 as women and one preferred not to say. 62 participants were between 18-25, 43 between 26-35, nine between 36-45, five between 46-55, one between 56-65, and one above 65 years old. Every participant was from the city where our institution is located and 73 were enrolled at a university. 73 stated that they have had experience with AR or VR devices in the past, whereas 49 did not have any prior contact with these devices before to the study. From the participants that have had prior experience with AR or VR the vast majority (43%) stated that they no longer use AR or VR, only use it very rarely (34%) or rarely (16%). In total just five people stated to use AR or VR occasionally to regularly. Regarding their technical expertise 88 participants self-categorized themselves as rather helping others with their computer problems, 26 stated they would rather receive help from others with computer problems, and seven did not specify.

6.2 RQ_{main}: How do different grid sizes influence the usability of graphical authentication on the Hololens

Regarding RQ_{main}, our findings indicate that there were only minor differences concerning the usability across the grid sizes. As mentioned in section 5.5, to objectively assess usability we analyzed the metrics *effectiveness*, *efficiency* and *satisfaction*, as demonstrated in Table 1.

Table 1: Distribution of key metrics to assess usability.

	Effectiveness			**Efficiency**		**Satisfaction**
Grid Size	Success rate	Wrong images	Corrections made	Authentication time	Repeat button pressed Sign-up	System Usability Score
6 grid	0.99 (0.05)	0.02 (0.08)	0 (0)	15.2s (2.85)	0.68 (0.65)	86.16 (7.97)
10 grid	1 (0)	0 (0)	0.02 (0.07)	13.47s (2.64)	0.68 (0.53)	84.25 (9.59)
32 grid	0.99 (0.04)	0.006 (0.04)	0.03 (0.09)	16.42s (6.78)	0.38 (0.58)	87.56 (7.26)
Total	0.99 (0.04)	0.008 (0.06)	0.01 (0.07)	15.03s (4.62)	0.57 (0.6)	85.99 (8.37)

Regarding the metrics for *effectiveness* we observed minor differences between the grid sizes, but we found no statistical significant differences. For grid size 10 the success rate was 100%, therefore no image was selected wrong but two corrections were made before submitting the password. The authentication in grid size 6 and 32 had a success rate of 99% with three and one wrong entered images, respectively. In grid size 6 no corrections were made before submitting and in grid size 32 three corrections were made.

To investigate the *efficiency*, we analyzed the time until successful authentication (see Figure 6). A Shapiro Wilk test and histogram analysis indicated non-normally distributed data for grid size 32 ($W = 0.828, p = .001$). A subsequent Kruskal-Wallis test suggested significance ($H(2) = 6.010$, $p = .047, \eta^2 = 0.0347$), and a post-hoc analysis using a two-sided Mann-Whitney U test with Bonferroni-Holm correction found significant differences between grid sizes 6 and 10 ($u = 548.0, p = .031, \eta^2 = 0.081,$). This indicates grid size 10 has significantly lower authentication time than grid size 6.

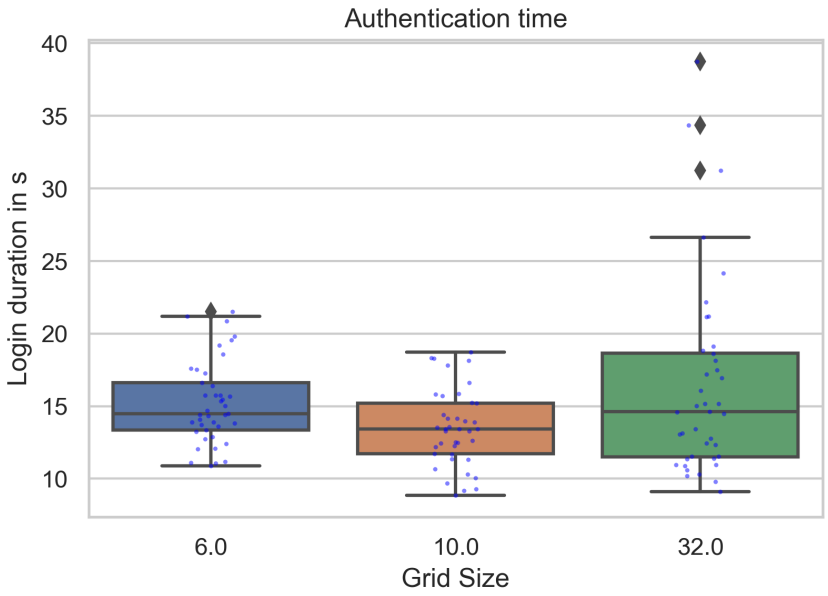

Figure 6: Authentication time

Regarding the amount of users that pressed the "repeat" button during sign-up to view the password again at least once, a Shapiro Wilk test ($W = 0.591 - 0.619, p < .001$) and histogram indicated that this data is not normally distributed (see Figure 7), leading to a Kruskal-Wallis test that showed significance ($H(2) = 10.021, p = .007, \eta^2 = 0.06797$). A post-hoc analysis with the Mann-Whitney U test and Bonferroni-Holm correction revealed significant differences between grid sizes 32 and 10 ($u = 540, p = .012, \eta^2 = 0.078$) as well as between grid sizes 32 and 6 ($u = 586.5, p = .022, \eta^2 = 0.060$). On average users of grid size 32 pressed the "repeat" button during sign-up 0.38 times in comparison to users from the groups grid size 6 and 10, both pressing the button on average 0.68 times.

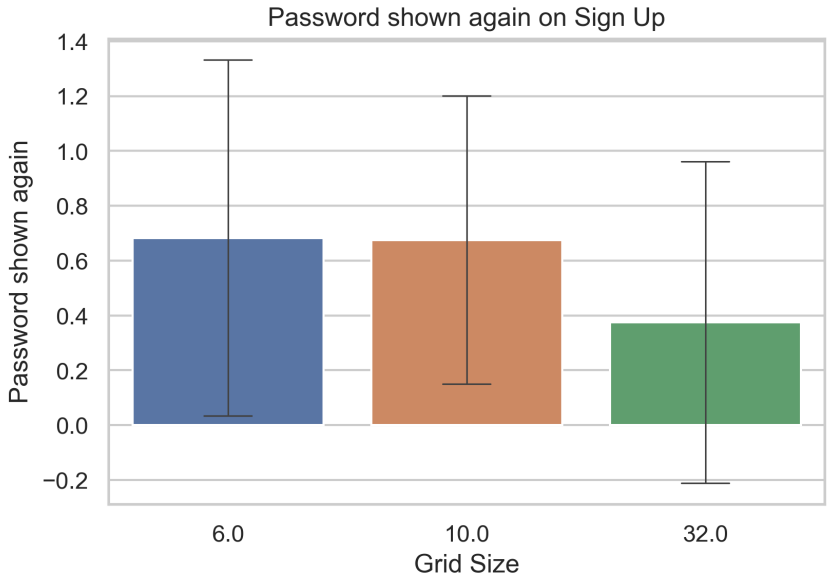

Figure 7: How often the participants looked at the password again during sign-up for each grid size

In terms of *satisfaction*, captured by SUS scores all grids were rated as "excellent", with a score ranging from 84.25 to 87.56, according to [5]. After a Shapiro Wilk test indicated that the data for grid size 6 ($W = 0.941, p = 0.035$) and grid size 10 ($W = 0.918, p = .007$) are non-normally distributed, a Kruskal-Wallis test was performed and found no significant differences between the grid sizes ($p = .322$).

Summary $\mathbf{RQ_{main}}$. In summary, the quantitative data displays only statistical differences regarding the metric *efficiency*. Metrics for *effectiveness* and *satisfaction* only show minor differences, none of them achieving statistical significance. The quantitative data suggests grid size ten as the fastest having the shortest authentication time and grid size 32 as the easiest to memorize as users needed the least amount of repetitive displays of their secret during sign-up.

6.3 $\mathbf{RQ_{context}}$: How do users perceive graphical authentication on the Hololens?

Answering $RQ_{context}$ we assessed the qualitative data captured in the online questionnaire using an open-coding approach.First, participants were asked, what they liked and disliked about the scheme. Afterwards, they were given the option to leave further remarks.

Table 2 displays the distribution of the most relevant code-categories created to capture positive aspects stated by the participants, for a complete overview please refer to the code-book in the appendix (Appendix B). 54 of the participants directly attributed good usability to the scheme and 44 described it to be easy to learn. A high perceived feeling of security was reported by 17 participants and 14 also reported the scheme to be efficient. Only a few participants addressed the aspects of gamification (12), inclusivity (10) or novelty (7). Not

Table 2: Distribution of positive code categories.

Usability	Learnability	Security	Efficiency	Gamification	Inclusivity	Novelty
54 (43.9%)	44 (35.77%)	17 (13.8%)	14 (11.3%)	12 (9.75%)	10 (8%)	7 (5.6%)

related to graphical authentication in general but rather specifically the version of the things scheme they used, eight participants liked the image selection and respective nine participants stated positive aspects about the study design decisions in general or the graphical design decisions in the application. Interesting to highlight is the fact, that 55 participants either directly named or described the concept of graphical authentication as a beneficial aspect of the scheme, as it was perceived helpful in memorizing one's password. The understanding of the graphical authentication concept was made clear through phrases like *"An image or object sequence is easier to remember than a number sequence. "* (participant 33) or *"I think it is good to work with visual images, because it is sometimes easier to remember a coherent context of images than a sequence of numbers and/or letters."* (participant 120).

Table 3: Distribution of negative code categories.

Exclusivity	Inefficiency	Insecurity	Uncomfortable
30 (24%)	28 (22.7%)	18 (14.63%)	13 (10.5%)

In total we captured a lot less negative feedback from the participants. The negative aspects that were mentioned most frequently is "Exclusivity" indicating graphical authentication to be difficult for different demographics, such as elderly people. The most mentioned aspect about exclusivity refers to the cognitive demand involved with GA. Cognitive demand includes aspects such as requiring more attention to find the correct images than entering a password or PIN where users may become accustomed to certain patterns. Another prominent factor we captured regarding exclusivity are the required motor skills involved with selecting the image, which was named by eight participants. The authentication time needed was mentioned by 28 participants, which is double the amount of the mentions of the positive counterpart ("Efficiency"). 18 participants expressed a perception of "insecurity" and 13 expressed the process to be "uncomfortable". The UI design decisions were criticized by three participants and 15 participants highlighted negative aspects about the study designs, such as the use of a random password instead of user-selected images, and that corrections could only be made by re-entering the entire password. Nine participants criticized grid-size related aspects. One participant from group grid size 10 and four participants from group grid size 32 expressed, that the grids contained too many images. In contrast, three participants from group grid size six and one participant from group grid size 32, the largest grid size, expressed a desire for more images on the grid.

Despite the critique, when asked whether they would use graphical authentication in the future in an AR/VR context the vast majority (87.8%) expressed willingness to do so. Even more impressively, 64% expressed interest in using it outside of the AR/VR domain, namely on their smartphone (13.5%) in general (12.7%) or on a computer (9.5%). However these results might be affected by social desirability.

Examining the qualitative data with regards to the specific grid sizes, we were unable to detect a trend for one of the grid sizes. Grid size 32 was attributed both negative and positive aspects more often than the other grid sizes with regards to efficiency. Similarly, grid size 10 was attributed to "inclusivity" the most, while simultaneously being attributed to "exclusivity" also the most of all the grid sizes. The following overview displays the distribution of the various codes to the grid sizes:

- "Usability" was attributed more often to grid size 32 (23 times) compared to grid size six (17 times) and ten (14 times).
- "Learnability" was attributed to grid sizes six and 32 more frequently, with 16 and respective 17 mentions in comparison to grid size ten with 11 mentions.
- "Efficiency" favored grid sizes ten (5) and 32 (7) in comparison to grid size six (2).
- "Inefficiency" was also more frequently attributed to larger grid sizes (grid 6: 7; grid 10: 9; grid 32: 12).
- "Inclusivity" was very evenly distributed with three to four participants mentioning this aspect for each of the grid sizes (grid 6: 3; grid 10: 4; grid 32: 3).
- "Exclusivity" was addressed more frequently in grid sizes ten (12), grid size six (11) in comparison to grid size 32 (7)
- "Security" was attributed to grid size 10 (10) much more frequently than to grid 6 (3) and grid 32 (4).

Summary $\mathbf{RQ_{context}}$. In summary, the analysis of the qualitative data shows a diverse perception of the graphical authentication scheme, indicating that the perception is shaped by personal preference. Participants didn't agree on the scheme being inclusive or exclusive, with each side offering understandable arguments. Similarly, there was also no consensus towards the authentication time being fast or slow. Although more participants expressed negative stances towards the two before mentioned metrics, no overall conclusion can be drawn as less than a quarter of participants expressed these believes. In terms of perception towards security there is also no consensus, as nearly the same amount of participants expressed positive and negative feelings towards perceived security. Regarding the effect of grid size on the qualitative data, no effect could be detected.

6.4 Further Findings

Regarding the authentication time we could find a small learning effect for grid size 32, as seen in figure 8.

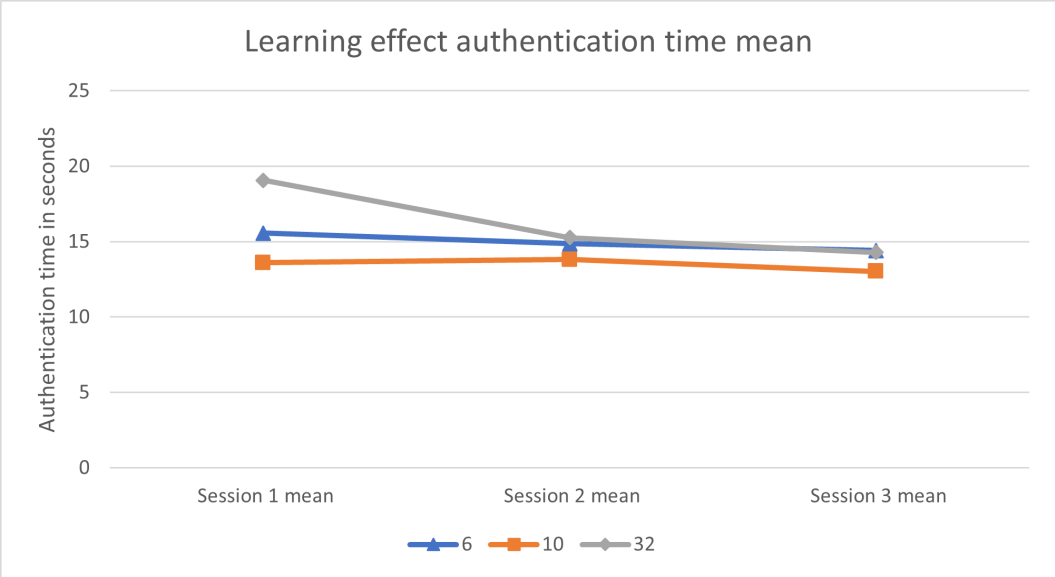

Figure 8: Learning effect for different grid sizes regarding authentication time

From the general oral feedback we received during and after the study, we realized that there was a strong disagreement between participants regarding the settings we set for the Hololens. While some people liked the distance at which the grids were projected others noted different opinions on that matter, with some people stating they would prefer the grid to be closer and some preferring it to be further away. Another point of critique that was mentioned several times, was the depth at which the Hololens recognizes an interaction as submitting information, e.g. selecting an image from the grid. While some participants stated that they would prefer this point to be closer to them, other expressed discontent as the current setting was too close for their preference and the interaction felt too "sensitive" for their liking.

With our sample consisting of 80 men and 40 women, the influence of gender was also examined[5]. In terms of *effectiveness* (success rate) we found no significant differences, with means of 0.994 (standard deviation: 0.039) for men and 0.994 (0.04) for women. The means for the System Usability Scale were also very similar, ranging from 85.03 (8.79) for men, to 87.63 (7.14) for women. Regarding the authentication time we noticed slight differences, with an average authentication time of 15.64s (5.18) for men and 13.87 (3.01) for women, that proved to be not significant according to a Kruskal-Wallis test ($H(2) = 3.486, p = .175, \eta^2 = 0.01259$).

We also examined the potential influence of prior experience with AR/VR but found no significant differences with a Kruskal-Wallis test ($p > .335$). 73 participants stated to have had prior experience with AR/VR, while the remaining 48 did not. No effect on *effectiveness* (0.997 (0.03) vs. 0.990 (0.05)), *efficiency* (15.06s (4.39) vs. 14.99s (5.01)) or *satisfaction* (86.03 (9.04) vs. 85.94 (7.32)) could be determined. The effect of technical expertise [6] was

[5]For the influence of gender we excluded people who preferred not to state their gender, as it was only one person who did so.

[6]Captured by a self-rating to one of the two options: (1) I rather help others with their computer problems (2) I rather receive help from other with my computer problems

also examined with a Kruskal-Wallis test, but no significant differences could be found in any of the three metrics ($p > .155$). The differences in *effectiveness* (0.997 (0.03) vs. 0.981 (0.07)), *efficiency* (14.26s (6.92) vs. 15.38s (4.49)) and *satisfaction* ((1) 85.60 (8.64) vs. 87.41 (7.37)) are only minimal.

The effect of education[7] also proved to have no influence. As a Kruskal-Wallis test found no significant differences ($p > .118$). *Effectiveness* for people enrolled in university (0.997) is similar to the ones currently not enrolled (0.989), students authenticated slightly slower than non-students with an average authentication time of 15.58s (5.21) compared to 14.10s (3.50) and gave slightly higher SUS-scores – 86.25 (8.14) vs. 85.59 (8.96).

7 Discussion

The results of the study show that all three variants of the things authentication scheme perform similarly in terms of *effectiveness* (success rate $\geq 99\%$) and user *satisfaction* (average SUS ≥ 84.25). However, we found a significant difference for *efficiency* (the average authentication time). The group using a grid size of 10 needed only $13,47s$ which was significantly faster than grid size 6 ($15,19s$). This suggests grid size 10 should be preferred over grid size 6. Although there is no significant difference in efficiency between grid size 32 and grid size 10, grid size 32 performed much slower with an average of 16.42s but it's standard deviation is also much higher (6.78 vs. 2.64). The reason for this might be, that authentication time can vary a lot, depending on the time users need to find the correct image on the large grid[8]. Keeping this in mind, one should also prefer grid size 10 over grid size 32 in terms of authentication time. On the other hand, the participants in the group using grid size 32 had to look significantly less often at the images to memorize them compared to grid size 10, as the secret is only four digits long compared to six. This indicates that grid size 32 has advantages in terms of memorability and therefore it might be preferred depending on the use case, since we could not show any significance in the other metrics between grid sizes 32 and 10. Comparing the grid sizes with objective measurements along side the three dimensions of usability, *effectiveness*, *efficiency* and *satisfaction*, we can conclude, that the detected differences were minimal with only *efficiency* bearing significant differences.

Analysing the qualitative data, no clear preference for one of the grid sizes could be detected. Although several codes were assigned more often to grid size 32 in comparison to the other grid sizes no preference can be determined here, as both negative and positive aspects for the same measurement were attributed to grid size 32 the most. Similarly, users from the group grid size 32 mentioned usability most, but simultaneously grid size 32 received the most grid size related critique. Condensing down the findings from the qualitative data we conclude, that the preference for a grid size is highly depended on personal preferences, as some users seem to prefer larger grid sizes and shorter passwords, so they have to memorize less and have to search longer to find the correct image and vice versa. However as the study was between-subject and not within-subject, no precise statement about preference can be made.

Grid size 10 being perceived as the most secure, may be due to the general familiarity from users from the standard PIN-pad consisting of ten options. This in combination with

[7] Captured by asking participants if they were enrolled in university.

[8] The correct image is randomly placed on one of the 32 options on the grid, so the authentication time greatly depends on where the image is and where the users starts searching.

a password length of six digits, in contrast to the most common used PIN-length of four digits [4], may lead to participants perceiving this grid size as the most secure.

As most of the measurements are very similar, only showing slight differences apart from authentication time proven to have significant differences we conclude, that the grid size has only a small influence on usability of graphical authentication on the Hololens. The differences mainly seem to be due to personal preferences. To cater to these preferences we advise potential adopters and developers to do one of the following:

(1) Adjust the grid size according to the expected user base and their preferences, which may be difficult to determine, as technical expertise, prior experience with AR/VR and gender, all don't seem to have an influence on usability, as described in Section 6.4.

(2) A more generally applicable solution would be to give user the option to choose their preferred grid size layout. Important to highlight in the second case would be that all of these grid sizes are equally secure as they all have the same password space of 10^6.

Regarding the overall perception of graphical authentication on the Hololens the results prove to be very positive. An average SUS-score of 85.99 indicating excellent usability and more than half of the participants addressing good usability directly in the qualitative data shows graphical authentication is well received. Nearly half of the participants directly and indirectly addressing the underlying concept of graphical authentication as beneficial in comparison to standard passwords further cements that claim.

Nearly 90% of participants expressing they would use graphical authentication in the future in AR/VR and nearly two thirds also in a different context outside of AR/VR (mainly smartphone) also shows broad acceptance of graphical authentication outside of our specific use case.

Interestingly thrice as many participants assess graphical authentication as exclusive over inclusive, with the most given arguments being "cognitive demand" (21) and "motor skill requirements" (9). The relative large amount of participants mentioning "cognitive demand" may be due to the unfamiliarity and novelty of graphical authentication for them. They are probably more familiar with standard passwords and have developed strategies to memorize these passwords already, whereas they have not done so with graphical passwords. At least some of this critique will most likely decrease as users become more familiar with AR in general and graphical passwords as well. Regarding "motor skill requirements" is not exclusive to graphical authentication but rather to the interaction with the Hololens itself, different input methods, which are also available on the Hololens, such as eye-gaze, voice-input or eye tracking could help with this. In summary, graphical authentication was very good perceived by the participants of our study with the positive codes being attributed far more often than the negative codes. We were able to detect some points of critique regarding the exclusivity, which at least partly could be due to the AR device we used in the study and the general novelty of the graphical authentication concept. Nonetheless, there is still room for improvement.

The comparison to the study Duezguen et al. conducted is interesting, although is difficult to compare both studies directly due to methodological difference. The average authentication time in their study amounted to $32.2s$ (9.39), experiencing a steep learning effect reducing the average authentication time from 40 seconds in the first round to 25 seconds in the third. As all of our examined grids performed far better in terms of efficiency, we conclude that the inclusion of a mandatory training phase in which participants were already exposed to the scheme was a beneficial addition. Moreover, the success rate for all

our proposed grid sizes was also improved compared to the study by Duezguen. We argue, that this might be due to our proposed change in the selection of images, reducing possible confusion due to similar images on one grid. Finally, all our proposed grid sizes were rated remarkably higher regarding user satisfaction by comparing their respective SUS-scores. We argue, that matching the grid sizes to the aspect ratio to the Hololens (changing the grid arrangement from vertically stacked to horizontally aligned) and reducing the buttons in the interfaces themselves to a minimum may have aided the usability, resulting in higher SUS scores.

7.1 Limitations and Future Work

The majority of participants were quite young, with age groups 18-25 (62) and 26-35 (43) representing 87% of participants, is a limitation to generalizability of this study. While these age groups are most likely the target group of AR it would still be interesting to examine the usability and perception with elderly people, especially wrt. to the claims of exclusivity. Sadly the small sample size of participants older than 36 (16 in total) doesn't allow for generalization. As one of the main points of critique was found to be about motor skill requirements and one possible solution we proposed to cope with that is to try different input methods such as gaze or eye tracking it would be interesting to see if elderly people more frequently report (a) motor difficulties, (b) slower authentication time, (c) worse success rate and (d) lower SUS-scores.

No effects based on education, prior experience with AR/VR or technical expertise could be determined. On this regard we need to highlight, that these effects were not the primary objective of this study and were only captured using simple questions from the online questionnaire. For future work it would be interesting to investigate some of this effects in greater detail.

We also excluded people wearing glasses from participating, due to ethical and legal reasons. Nonetheless it would be interesting to examine if perception and usability changes when the Hololens is used while also wearing glasses.

Participants being assigned a password and not given the option to self select a password was done for security reasons, as people tend to select similar images ([15]) when given the option. This may have had an influence on memorability and could explain the different amount of times people pressed repeat during sign-up to be shown their password once more. To confirm this, the aspect of memorability would have to be further investigated. On this regard it would also be interesting to investigate the effect of long-term usage towards usability and perception, as we could already detect a small learning effect wrt. authentication time.

Our design decision to not group images based on object type together on one grid also forms a limitation to security, as this makes the scheme more vulnerable to social engineering attacks, as users can more easily communicate their password to others [21].

Lastly, the AR device used, Microsoft Hololens 2, and the settings we decided on for this study represent several limitations. Using a different device may have yielded different results, as well as implementing different settings inside the application. Due to comparability reasons we opted for specific settings alongside the recommendation provided by Microsoft[9] but changing these settings or even making it possible for participants to adjust them to their liking may produce different results.

[9]https://learn.microsoft.com/en-us/windows/mixed-reality/design, Last accessed on 14 Feb. 2024

8 Conclusion

This research aims to investigate the influence of different grid sizes on the usability of graphical authentication on the Hololens. Therefore we adapted the *Things* scheme, proposed by [21], and created three different versions with different sized grids and password lengths, while still maintaining the same password space 10^6. To investigate potential differences between these versions and the overall perception we conducted a between-subject lab study with 126 participants. We logged the participant's usage data from the Hololens and also captured qualitative feedback by using an online questionnaire. To assess potential influences on usability we performed several statistical tests on the collected data in the usability categories *effectiveness*, *efficiency* and *satisfaction*. The qualitative data was analysed using an open coding approach. We found that the grid size only has a small influence on the usability of graphical authentication on the Hololens, mainly wrt. to *efficiency*, preferring grid size 10. We found that personal preference of the users plays a central role in usability and should be catered to in future implementations. A potential solution for that would be to offer user a personal choice of preferred grid size and length of the secret, while still maintaining the same security level. The general concept of graphical authentication was very well received with an excellent SUS-score of 85.99. Participants strongly expressed willingness to use graphical authentication in future AR/VR context and even in different contexts, such as on their smartphone. We also detected some complaints mainly wrt. to graphical authentication excluding certain demographic such as elderly people. We argue that this perception might be due to the AR context in which participants experienced graphical authentication in this study. Finally, we proposed potential directions for future research, such as investigating different input methods or addressing a specific target group such as elderly people.

In summary, while graphical authentication in general was very well perceived we did only find minor influences of grid size on the usability. We believe that personal preference of users plays a central use in usability. That's why we encourage researchers to replicate this study and investigate the influence of different personal preference metrics in more detail.

References

[1] Yomna Abdelrahman, Florian Mathis, Pascal Knierim, Axel Kettler, Florian Alt, and Mohamed Khamis. Cuevr: Studying the usability of cue-based authentication for virtual reality. In *Proceedings of the 2022 International Conference on Advanced Visual Interfaces*, pages 1–9, 2022.

[2] Hala Assal, Ahsan Imran, and Sonia Chiasson. An exploration of graphical password authentication for children. *International Journal of Child-Computer Interaction*, 18:37–46, 2018.

[3] Daniel V Bailey, Markus Dürmuth, and Christof Paar. "Typing" passwords with voice recognition: How to authenticate to Google Glass. *Proc. of the Symposium on Usable Privacy and Security*, 2014.

[4] Joseph Bonneau, Sören Preibusch, and Ross Anderson. A birthday present every eleven wallets? the security of customer-chosen banking pins. In Angelos D. Keromytis, editor, *Financial Cryptography and Data Security*, pages 25–40, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[5] John Brooke. *SUS – a quick and dirty usability scale*, pages 189–194. Taylor & Francis, 01 1996.

[6] Sacha Brostoff and M Angela Sasse. Are passfaces more usable than passwords? a field trial investigation. In *People and computers XIV—usability or else! Proceedings of HCI 2000*, pages 405–424. Springer, 2000.

[7] Sacha Brostoff and M Angela Sasse. "ten strikes and you're out": Increasing the number of login attempts can improve password usability. *Proceedings of CHI 2003 Workshop on HCI and Security Systems*, 2003.

[8] Reyhan Duezguen, Peter Mayer, Sanchari Das, and Melanie Volkamer. Towards Secure and Usable Authentication for Augmented and Virtual Reality Head-Mounted Displays. *arXiv*, 2020. arXiv:2007.11663 [cs].

[9] Reyhan Düzgün, Peter Mayer, and Melanie Volkamer. Shoulder-Surfing Resistant Authentication for Augmented Reality. *Nordic Human-Computer Interaction Conference*, pages 1–13, 2022.

[10] Dinei Florêncio, Cormac Herley, and Paul C. Van Oorschot. An administrator's guide to internet password research. In *Proceedings of the 28th USENIX Conference on Large Installation System Administration*, LISA'14, page 35–52, USA, 2014. USENIX Association.

[11] Markus Funk, Karola Marky, Iori Mizutani, Mareike Kritzler, Simon Mayer, and Florian Michahelles. LookUnlock: Using Spatial-Targets for User-Authentication on HMDs. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–6, Glasgow Scotland Uk, 05 2019. ACM.

[12] Ceenu George, Mohamed Khamis, Daniel Buschek, and Heinrich Hussmann. Investigating the third dimension for authentication in immersive virtual reality and in the real world. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 277–285, Piscataway, New Jersey, 2019. IEEE.

[13] Ceenu George, Mohamed Khamis, Emanuel von Zezschwitz, Marinus Burger, Henri Schmidt, Florian Alt, and Heinrich Hussmann. Seamless and secure vr: Adapting and evaluating established authentication systems for virtual reality. In *Network and Distributed System Security Symposium (NDSS 2017)*. NDSS, 2017.

[14] Gabriela Gheorghe, Nicolas Louveton, Benoît Martin, Benjamin Viraize, Louis Mougin, Sébastien Faye, and Thomas Engel. Heat is in the eye of the beholder: Towards better authenticating on smartglasses. In *2016 9th International Conference on Human System Interactions (HSI)*, 2016 9th International Conference on Human System Interactions (HSI), pages 490–496. IEEE, 07 2016.

[15] Maximilian Golla, Dennis Detering, and Markus Dürmuth. Emojiauth: quantifying the security of emoji-based authentication. In *Proceedings of the usable security mini conference (USEC)*, 2017.

[16] Jonathan Gurary, Ye Zhu, and Huirong Fu. Leveraging 3d benefits for authentication. *International Journal of Communications, Network and System Sciences*, 10:324–338, 01 2017.

[17] Saranga Komanduri and Dugald R Hutchings. Order and entropy in picture passwords. In *Proceedings of graphics interface 2008*, pages 115–122. Citeseer, 2008.

[18] Yan Li, Yao Cheng, Weizhi Meng, Yingjiu Li, and Robert Deng. Designing Leakage-Resilient Password Entry on Head-Mounted Smart Wearable Glass Devices. *IEEE Transactions on Information Forensics and Security*, 16:1–1, 07 2020.

[19] Yingjiu Li, Qiang Yan, and Robert Deng. ShadowKey: A Practical Leakage Resilient Password System. In *SpringerBriefs in Computer Science*, pages 53–64. Springer International Publishing, 01 2015. Journal Abbreviation: SpringerBriefs in Computer Science DOI: 10.1007/978-3-319-17503-4_3.

[20] Tobias Länge, Philipp Matheis, Reyhan Düzgün, Melanie Volkamer, and Peter Mayer. Vision: Towards fully shoulder-surfing resistant and usable authentication for virtual reality. In *Accepted for USEC2024*, 2024.

[21] Peter Mayer, Melanie Volkamer, and Michaela Kauer. Authentication Schemes - Comparison and Effective Password Spaces. *Information Systems Security (ICISS), Hyderabad, India, December 16-20, 2014. Ed.: A. Prakash*, page 204, 00 2014. ISBN: 9783319138404.

[22] Wendy Moncur and Grégory Leplâtre. Pictures at the ATM: exploring the usability of multiple graphical passwords. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, volume 63 of *CHI '07*, page 887–894, New York, NY, USA, 04 2007. Association for Computing Machinery.

[23] Dan R Olsen, Richard B Arthur, Ken Hinckley, Meredith Ringel Morris, Scott Hudson, Saul Greenberg, Katherine M Everitt, Tanya Bragin, James Fogarty, and Tadayoshi Kohno. A comprehensive study of frequency, interference, and training of multiple graphical passwords. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 889–898, 2009.

[24] Allan Paivio and Kalman Csapo. Picture superiority in free recall: Imagery or dual coding? *Cognitive Psychology*, 5(2):176–206, 1973.

[25] Johnny Saldaña. *The coding manual for qualitative researchers.* Sage, 2009. OCLC: ocn233937452.

[26] Hwajeong Seo, Jiye Kim, Howon Kim, and Zhe Liu. Personal identification number entry for Google glass. *Computers & Electrical Engineering*, 63:160–167, 10 2017.

[27] Ravi S. Sharma, Aijaz A. Shaikh, and Eldon Li. Designing Recommendation or Suggestion Systems: looking to the future. *Electronic Markets*, 31(2):243–252, 2021.

[28] Sophie Stephenson, Bijeeta Pal, Stephen Fan, Earlence Fernandes, Yuhang Zhao, and Rahul Chatterjee. Sok: Authentication in augmented and virtual reality. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 267–284, 2022.

[29] David R. Thomas. A general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation*, 27(2):237–246, 2006.

[30] Jiawei Wang, BoYu Gao, Huawei Tu, Hai-Ning Liang, Zitao Liu, Weiqi Luo, and Jian Weng. Secure and memorable authentication using dynamic combinations of 3d objects in virtual reality. *International Journal of Human–Computer Interaction*, 0(0):1–19, 2023.

[31] Dhruv Kumar Yadav, Beatrice Ionascu, Sai Vamsi Krishna Ongole, Aditi Roy, Nasir Memon, Michael Brenner, Nicolas Christin, Benjamin Johnson, and Kurt Rohloff. Design and Analysis of Shoulder Surfing Resistant PIN Based Authentication Mechanisms on Google Glass. In *Financial Cryptography and Data Security*, Lecture Notes in Computer Science, pages 281–297, Berlin, Heidelberg, 00 2015. Springer.

[32] Ruide Zhang, Ning Zhang, Changlai Du, Wenjing Lou, Y. Thomas Hou, and Yuichi Kawamoto. AugAuth: Shoulder-surfing resistant authentication for augmented reality. In *2017 IEEE International Conference on Communications (ICC)*, pages 1–6, 05 2017. ISSN: 1938-1883.

[33] Verena Zimmermann and Nina Gerber. The password is dead, long live the password – a laboratory study on user perceptions of authentication schemes. *International Journal of Human-Computer Studies*, 133, 08 2019.

[34] Verena Zimmermann, Paul Gerber, and Alina Stöver. That depends – assessing user perceptions of authentication schemes across contexts of use, 2022.

A Scheme Procedure

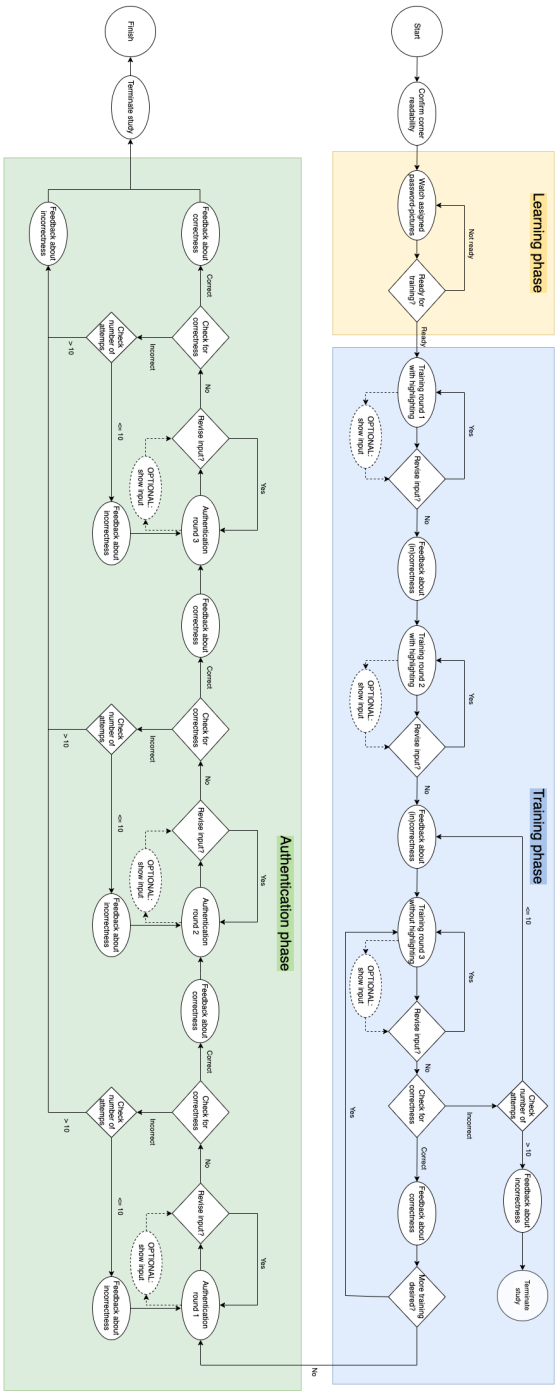

Figure 9: Extensive Flowchart of the *things* scheme used in the study.

B Codebook

Code	Description	6 grid	10 grid	32 grid	Total
POSITIVE-Codes					
Lernability	All cases, where somebody describes that they perceived the procedure as fast/easy/quick to learn	16 (39%)	11 (26,8%)	17 (41,47%)	44 (35,77%)
Usability	All cases, where a good usability is described	17 (41,47%)	14 (34,1%)	23 (56,01%)	54 (43,9%)
Picture Selection	All cases, that emphasize on positive attributes of selected pictures	1 (2%)	4 (9%)	3 (7,3%)	8 (6,5%)
Graphic Recognition	All cases, that mention the better memorizability of graphical passwords versus classic passwords / good memoriabilty in general	17 (41,47%)	21 (51%)	17 (41,47%)	55 (44%)
Security	All cases that describe a high perceived security (either in general or in comparison to PIN/Passwords)	3 (7,3%)	10 (24%)	4 (9%)	17 (13,8%)
Inclusivity	All cases that describe inclusive factors which come into play using this scheme	3 (7,3%)	4 (9%)	3 (7,3%)	10 (8%)
Gamification	All references to gamification attributes	4 (9%)	3 (7,3%)	5 (12%)	12 (9,75%)
Good Graphic Design	All references regarding positive aspects of graphic design decisions	4 (9%)	2 (4,8%)	3 (7,3%)	9 (7,3%)
Good Study Design	All references to good aspects about general study design decisions (e.g. explanations, setup, ...)	5 (12%)	1 (2%)	3 (7,3%)	9 (7,3%)
Efficiency	All references stating the efficient usage of procedure (general and in comparison)	2 (4,8%)	5 (12%)	7 (17%)	14 (11,3%)
Novelty	All references regarding novelty of procedure (can be in comparison to existing schemes)	2 (4,8%)	4 (9%)	1 (2%)	7 (5,6%)
NEGATIVE - Codes					
Exlusivitiy	All cases where it is mentioned that the scheme may not be usable for different demographic (e.g. old people) - cotains motoric difficulty, mental requirements, ...	11 (26,9%)	12 (29%)	7 (17%)	30 (24%)
Bad Graphic Design	All references about negative graphic design decisions	2 (4,8%)	1 (2%)	0	3 (2%)
Inefficiency	All references stating the inefficient usage of procedure (general and in comparison)	7 (17%)	9 (21,9%)	12 (29%)	28 (22,7%)
Uncomfortable	All cases that described experienced inconvenience and uncomfortableness.	2 (4,8%)	6 (14,6%)	5 (12%)	13 (10,5%)
Bad study Design	All cases that highlight negative aspects of study parameters (technical and logical nature)	7 (17%)	3 (7,3%)	6 (14,6%)	15 (12,1%)
Grid Size Related	All references to negative aspects noted that are related to the grid size the interviewee used	3 (7,3%)	1 (2%)	5 (12%)	9 (7,3%)
Insecurity	All references to bad feelings/Perception regarding the security	6 (14,6%)	6 (14,6%)	6 (14,6%)	18 (14,63%)
Willingness to use					
Usage in Future AR/VR context- yes	All references that expressed willingness to use graphical authentication in the future in an AR/VR context	37 (90%)	33 (80%)	38 (92%)	108 (87,8%)
Usage in Future AR/VR context- no	All references that expressed no willingness to use graphical authentication in the future in an AR/VR context	4 (9%)	8 (20%)	3 (7,3%)	15 (12,2%)
Usage in Future other context - yes	All references that expressed willingness to use graphical authentication in the future in a different context	27 (65,9%)	26 (63,4%)	25 (60,9%)	78 (63,4%)
Usage in Future other context - no	All references that expressed no willingness to use graphical authentication in the future in a different context	14 (34,1%)	15 (36,6%)	16 (39%)	45 (36,6%)

Figure 10: Codebook with all captured codes, their description and frequency.