Stay Home! Conducting Remote Usability Evaluations of Novel Real-World Authentication Systems Using Virtual Reality

Florian Mathis
University of Glasgow
Glasgow, United Kingdom
florian.mathis@glasgow.ac.uk

Joseph O'Hagan
University of Glasgow
Glasgow, United Kingdom
j.ohagan.1@research.gla.ac.uk

Kami Vaniea
University of Edinburgh
Edinburgh, United Kingdom
kvaniea@inf.ed.ac.uk

Mohamed Khamis
University of Glasgow
Glasgow, United Kingdom
mohamed.khamis@glasgow.ac.uk

Figure 1: We propose *Remote Virtual Reality for simulating Real-world Research* (RVR^3) to evaluate novel real-world prototype systems. We implemented two real-world authentication systems for automated teller machines (ATMs) (i.e., *Hand Menu* (❸) and *Tap* (❹)) and compared their usability against *Traditional 4-digit PIN authentication* (❶) and *Glass Unlock* (❷) [59].

ABSTRACT

Evaluating interactive systems often requires researchers to invite user study participants to the lab. However, corresponding evaluations often lack realism and participants are usually recruited from a local area only. In this work, we propose *Remote Virtual Reality for simulating Real-world Research* (RVR^3) to evaluate novel real-world authentication prototypes. A user study (N=25) demonstrates the feasibility of using VR for remote usability research on simulated real-world prototypes. Our remote VR user study provides a glimpse into the usability and social acceptability of two novel authentication systems: *Hand Menu* and *Tap*. We build on prior research in this space and discuss the impact RVR^3 studies have on the range of possible studies. In summary, our remote VR research method to design, implement, and evaluate interactive real-world prototypes is a next step towards moving human-centred research out of the lab and potentially reaching a more diverse and larger participant sample over time.

AVI 2022, June 6–10, 2022, Frascati, Rome, Italy

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Virtual reality**; • **Security and privacy** → **Usability in security and privacy**.

KEYWORDS

Remote Research, Virtual Reality, Authentication, Usability

ACM Reference Format:
Florian Mathis, Joseph O'Hagan, Kami Vaniea, and Mohamed Khamis. 2022. Stay Home! Conducting Remote Usability Evaluations of Novel Real-World Authentication Systems Using Virtual Reality. In *Proceedings of the 2022 International Conference on Advanced Visual Interfaces (AVI 2022), June 6–10, 2022, Frascati, Rome, Italy*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3531073.3531087

1 INTRODUCTION

Lab studies form a popular research method when conducting human-centred research [22, 49]. However, they often do not present participants with a close-to-reality environment in which prototype systems are eventually deployed [11, 30]. Participants are usually also recruited within a specific population [25], which can impact the resulting research. While many research methods exist, such as interviews, online surveys, or field studies, to learn about users and their behaviour [8, 13], these methods are often not suitable to evaluate prototype systems that involve hardware.

In this work, we propose virtual reality (VR) as a research method to remotely evaluate real-world prototypes, hereafter referred to as *Remote Virtual Reality for simulating Real-world Research* (RVR^3). Providing researchers with complementary research methods beyond traditional lab and field studies can be advantageous in many ways [1, 49]. RVR^3 enables researchers to evaluate virtual replicas of real-world prototypes in an affordable way. Virtual replicas do not require physical storage space, are easy to maintain, and do not require access to hardware prototypes or contexts that are hard to reach (e.g., security and safety-critical environments [27, 31]). RVR^3 is also capable of targeting user study subjects from multiple countries and with different backgrounds, eventually contributing to a sample's diversity. Finally, RVR^3 can be beneficial when direct interaction between researchers and participants is challenging or even prohibited (e.g., due to COVID-19) [33, 54].

Our work is the first that applies remote VR research to replicate and evaluate real-world prototype systems. We combine *traditional remote VR research* with using *VR as a proxy for real-world research* [33] and evaluate the usability and social acceptability of two novel real-world authentication systems: *Hand Menu* and *Tap*. We describe the prototype systems and the role of augmented reality (AR) for advanced real-world authentication in more detail in Section 3.

A user study with 25 participants provides promising insights into the usability and social acceptability of novel touch-less authentication systems. Our work has long-lasting implications for the Human-computer Interaction (HCI), Usable Security, and the VR communities by moving typical research on real-world prototypes out of the lab and potentially contributing to large-scale, diverse, and cross-cultural study samples over time.

Contribution Statement. Our contribution is three-fold: **(1)** We apply VR to investigate novel real-world authentication prototypes in a remote user study (i.e., RVR^3). While prior work moved traditional VR research online [36, 45] and investigated VR's feasibility for real-world authentication research in the lab [31, 32], we extend research in this space by conducting a VR-based usability study of real-world systems in a fully remote VR setup. **(2)** We propose two authentication methods for shared and social spaces (i.e., *Hand Menu* and *Tap*) and contribute results of a user study with 25 participants from 9 different countries. **(3)** We discuss our findings and conclude with promising future research directions when applying VR to remotely evaluate real-world prototypes.

2 RELATED WORK

2.1 Virtual Reality as a Research Platform

Researchers have recently begun investigating the use of immersive VR for behavioural research and studying virtual replicas of real-world prototype systems in the lab. Voit et al. [56] compared online, VR, AR, lab, and in situ evaluations when evaluating smart artefacts. They found a VR-based evaluation achieves similar results as an in-situ evaluation [56]. Mäkelä al. [27] studied the differences between a user study in front of a real-world public display and one in VR. Mathis et al. [31] evaluated a security system's resistance against observations and its usability in VR to compare its findings to the results previously reported in the original real-world study [20]. Savino et al. [48] compared pedestrian navigation methods in VR and in the real world. Although these works found some differences between VR-based and real-world investigations (e.g., navigation performance and landmark recognition differed significantly between real life and VR [48], people are "more interested in virtual environments than in real public spaces" [27]), they also highlighted many similarities between the different study types. Mäkelä et al. [27] observed largely similar user behaviour between their two real-world and VR settings and there was an increase in attention towards the public displays in the presence of an audience (i.e., honey pot effect) in both environments. Mathis et al. [31] found their VR users' perceptions of the usability and security of touch, mid-air, and eye gaze authentication matched to a great extent with the perceptions of the real-world study participants [20]. There are additional works that used VR to study expensive or safety-critical situations in the real world [16, 32] or to simulate AR [42, 57].

2.2 Authentication Systems for Public Spaces

Providing users with private and secure interaction in public can roughly be divided into three categories [9]: **(1)** research that aims to provide software solutions (e.g., [55]), **(2)** research that utilises additional input hardware (e.g., [6]), and **(3)** research that uses users' private hardware (e.g., using a mobile device as a physical token [41]), on which we focus in the following.

Guerar et al. [17] proposed an authentication system that relies on the user's smartphone and makes use of a QR code to match colors to digits. Sharp et al. [52] made use of the user's personal device to view a one-time password for authentication at public displays. Nyang et al. [37] proposed the use of the user's smartphone to obtain a random permutation of a keyboard layout to authenticate in public. Work by De Luca et al. [9] used the user's mobile device for secure authentication based on shared lies. Their prototype utilises the user's mobile device for tactile feedback that provides secret information when to add an overhead of "lies" to the input [9]. Khan et al. [21] used a mobile device to allow for obfuscated PIN template input. To authenticate, users receive a PIN template (e.g., *[48**29**]*) on their device that they then combine with their PIN. Winkler et al. [59] used a private near-eye display to communicate keypad layouts to users when authenticating on a mobile device.

2.3 Summary and Research Gap

Previous works successfully moved traditional VR research out of the lab [35, 36, 45]. There is also a growing interest in using VR to simulate real-world environments and prototypes to then evaluate user behaviour in otherwise hard-to-reach locations [27, 28, 32]. However, there is a gap in research that moved research on real-world prototypes out of the lab, with the majority of user studies on real-world systems being conducted in the lab (e.g., [7, 9, 11, 59]). We fill this gap by combining *traditional remote VR research* [36, 45] with using *VR as a proxy for real-world research* [27, 31, 32] and conduct a remote VR user study to evaluate the usability and social acceptability of novel real-world authentication systems.

3 STUDIED PROTOTYPES AND CONTEXT

We studied two novel authentication systems to investigate VR's feasibility to conduct remote research on real-world systems: *Hand Menu* and *Tap*. Both systems make use of augmented reality to

present users during their authentication with a unique (and private) PIN layout (Figure 2). This makes authentications **resilient against shoulder surfing** – the act of observing other people's information without their consent [13]. Both methods allow for **touch-less user authentication**, avoiding touching public surfaces, e.g., keypads, which pose a considerable risk in the transmission of bacteria and viruses [58]. In line with previous work, we randomised the keypad layout once at the start of each 4-digit PIN authentication (e.g.,*"1234"*) due to security [59]. We describe the prototypes and the implemented baselines in more detail below.

3.1 Traditional Authentication + Glass Unlock

We implemented two authentication systems as our baselines: **(1)** *Traditional* 4-digit PIN authentication and **(2)** Glass Unlock [59], an AR-based authentication system. Using traditional 4-digit PIN authentication as a baseline condition is a common approach in authentication research (e.g., [2, 7]). We added Glass Unlock (10Key) [59] as a second baseline because 1) we made use of the underlying concept of using AR for advanced authentication in both *Hand Menu* and *Tap* and 2) we aimed to study participants' behaviour and *Glass Unlock* 's usability when simulated in VR and used for a different context than initially proposed for (i.e., secure smartphone unlocking [59]). In our implementation of *Glass Unlock*, the user provides input on a traditional keypad, but this time with unlabeled buttons (i.e., the keypad has no digits). The randomised keypad layout is instead presented using the user's AR glasses (Figure 2–❷).

3.2 Hand Menu Authentication (*Hand Menu*)

Instead of entering a PIN on a physical keypad, a one-time randomised keypad layout is augmented next to the user's wrist on which they provide mid-air input. The combination of augmenting a keypad and applying a one-time randomisation to the digits ensures the system's resilience against observations [59] and allows for touch-less authentication. In summary, *Hand Menu* is a hand-attached user interface that follows Microsoft's HoloLens 2 *"Hand menu"* implementation [34] and allows for quick input (Figure 2-❸).

3.3 Tap Authentication (*Tap*)

In *Tap*, digits are augmented above the finger tips of the user's left hand in random order (Figure 2-❹). Only the user can see this mapping through their AR glasses. This allows for the underlying mapping of the digits (i.e., number assigned to finger) to be unknown by a bystander. To provide input, the user taps with their right index finger on their left-hand finger tips. Each finger of the user's left hand allows input of two digits depending on the current mode (A or B). Each mode covers five digits (e.g., mode A: 6,2,0,5,9; mode B: 1,3,7,8,4). To switch between the modes and enable users to correct and submit their PIN, *Tap* makes use of pinch gestures in the user's dominant hand, which is a common interaction method in AR. The user can switch between the two modes by performing a pinch gesture between thumb and index finger of the right hand. By using these two modes the user can access all ten digits (e.g., in Figure 2-❹ the little finger allows input of the digit "9" in mode A). To delete the last digit entered, the user performs a pinch gesture with their thumb and middle finger of their right hand. This gesture can be repeated multiple times to continually delete digits.

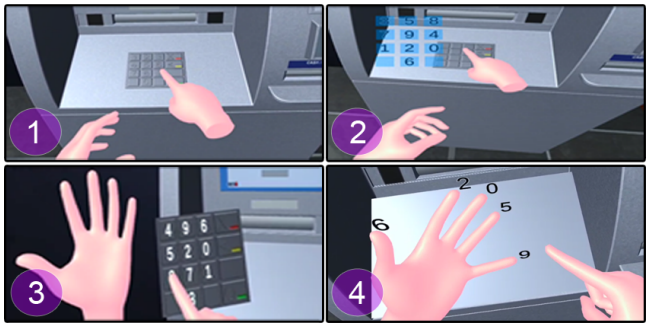

Figure 2: In ❶, users use a traditional keypad to authenticate. *Glass Unlock* [59] (❷) makes use of AR to present users with a private keypad layout. In *Hand Menu* (❸) and *Tap* (❹), users provide input on a hand-attached AR keypad (❸) or on augmented digits that are attached to their finger tips (❹).

Confirming the entered PIN is implemented using a pinch gesture between the thumb and ring finger of the user's right hand.

3.4 Replicated Real-World Context

We replicated an automated teller machine (ATM) authentication scenario to evaluate the different real-world authentication systems in a remote VR study. Conducting research in such a sensitive context is often challenging due to the lack of resources and the ethical and legal constraints that often hinder in-depth investigations [8, 30]. While simulating the sensory experience of ATM interaction in the lab is possible, these setups often lack realism [11, 17].

4 IMPLEMENTATION

We implemented our prototype systems in VR using Unity 3D (C#). We used Oculus Integration and the Oculus hand tracking SDK [10] and modified Oculus' sample hand tracking implementation to provide participants with a keypad that comes with auditory and visual feedback when providing input. For *Glass Unlock* [59], we present users with a virtual keypad layout in an egocentric view (i.e., the virtual keypad mapping is linked to the user's head movements). This simulates the situation where the user wears AR glasses (e.g., the Microsoft HoloLens 2). We made use of Oculus' OVRHand and OVRSkeleton [10] to a) augment a keypad next to the user's wrist when performing an open palm hand gesture in *Hand Menu*, and b) map the digits to the user's finger tips and allow for mode switches in *Tap* (Figure 2). The augmented keypad in *Hand Menu* is aligned to the right side of the user's wrist (non-dominant hand). In *Tap*, we added a 0.35 s delay between subsequent digit entries to avoid accidental inputs, determined through pilot tests. In summary, *Hand Menu* and *Tap* simulate AR-based authentication systems that augment a) a virtual keypad next to users' wrist (*Hand Menu*) or b) virtual digits on top of users' finger tips (*Tap*).

For the environment, we used a 3D ATM model [15] and replicated a real-world ATM scenario using *Snaps Prototype* [40]. The replicated scenario consists of five ATMs, one customer who is chatting with a bank employee, and one customer who is interacting with one of the ATMs (Figure 1). We added environmental sound (i.e., people chatting) to further enrich the environment. The virtual ATM is fully functional and allows to input a credit card

and navigate through a traditional European ATM user interface. After successful PIN input, the ATM outputs the credit card and the cash. We implemented additional in-VR menus and questionnaires (e.g., for NASA-TLX [19], SUS [4]) to guide participants through the study and to not break their VR experience [43].

5 METHODOLOGY

Our research combines traditional out-of-the-lab VR research via participant-owned VR headsets [36, 45] with immersive VR to conduct (simulated) real-world research in the lab [27, 31]. Ethical approval was sought and received from the University of Glasgow. Participants were recruited using social media and word of mouth. We paid participants according to their local standard (e.g., £15 for participants from the UK). Participants used an Oculus Quest 1 or 2 headset to participate. We distributed the VR application (.apk) in advance of the study and provided participants with an installation guide describing how to install the study environment on their VR headset. We then scheduled a Zoom meeting for a 1.5 h user study session. Participants‘ demographics were collected prior to the user study session using Qualtrics [44]. Our framework follows one of the primary approaches of conducting remote VR studies [45] and can be defined as a remotely conducted immersive VR study to evaluate (simulated) real-world prototype systems (i.e., RVR^3).

5.1 Independent and Dependent Variables

We compare *Hand Menu* and *Tap* against *Traditional* and *Glass Unlock* (10Key) [59] authentication. We have one independent variable (IV) with four levels: (1) *Traditional*, (2) *Glass Unlock*, (3) *Hand Menu*, and (4) *Tap*. We measured common usability metrics when evaluating novel authentication systems [7, 20], such as participants‘ authentication time, the number of PIN corrections, and the number of incorrect PIN entries. We asked participants about their perceived workload using the NASA-TLX questionnaire [19] and their user experience using the UEQ [24]. We measured the prototypes‘ usability using the SUS questionnaire [4] and asked participants additional 5-point Likert scale questions that we borrowed from prior work (e.g., *“Input using this method is [usable in public].”*) [20]. We concluded with a usability, security, and combined usability and security ranking, and with a semi-structured interview (see Appendix A). To allow for better comparisons in future experiments and support replication studies, we report our sample‘s security behaviour using the Security Behavior Intentions Scale (SeBIS) [12] and its technological affinity using the Affinity for Technology Interaction (ATI) scale [14]. We also report participants‘ sense of presence using the IPQ [51] and the TPI (dimension: social realism) questionnaire [26]. This was done in the beginning of the user study, after participants experienced the VR environment for roughly one minute, and provides us with insights into participants‘ sensation of being in a real place and if the scenario is perceived as actually occurring (i.e., place and plausibility illusion [53]).

5.2 Study Design and Task

Our user study followed a within-subject experiment. Each participant experienced all authentication systems. The order was counterbalanced using a Latin Square. We first introduced participants to the different authentication systems using a slide deck. Participants then experienced the virtual environment and filled in the IPQ [51] and TPI questionnaire [26]. We then had a training session where participants entered three PINs (e.g., “1234”) using the first authentication system (e.g., *Glass Unlock*). Training participants at the beginning of a user study is a common approach to ensure participants are familiar with the authentication mechanisms [7, 20]. The ATM‘s user interface provided an in-VR authentication video and all necessary details for the training session and for the subsequent data collection session (e.g., the PINs to use). After each training, participants went through a series of five PIN authentications. We refer to this as *isolated* authentications as the authentication itself was not part of a production task [32, 47]. This means participants only had to enter 4-digit PINs using the corresponding authentication method. After five successful authentications, participants experienced the authentication system as a supporting task (i.e., in situ) [32], similar to how authentication happens in reality [47]. Participants had to a) take the (virtual) credit card, b) put the credit card into the ATM, c) authenticate using the corresponding authentication method, d) select the amount of money to withdraw, and e) take the card and the cash out of the ATM. Participants then reported their perceived workload [19] and their user experience [24], rated the prototype‘s usability [4], and filled in 5-point Likert scale questions [20]. The same procedure, including training, was repeated for the other prototypes. We concluded with a ranking by usability, security, and combined usability and security, with an interview, and with the SEBIS [12] and ATI questionnaire [14].

5.3 Demographics

Our results are based on 25 participants (15 male, 9 female, 1 non-binary) who participated from overall 9 different countries: 12 from the UK, 4 from France, 3 from the USA, and one each from Spain, Belgium, Finland, Czech Republic, Canada, and Singapore. Participants were on average 25.76 years old (min=17, max=35, SD=4.36). All participants where right-handed, except one who had no marked preference for the use of the right or left hand. To participate, 19 participants used an Oculus Quest 2, six an Oculus Quest 1. Our participants have VR experience up to 5 years 11 months (M=20.58 months, SD=23.036). All participants mentioned they have used an ATM before. Their security behaviour score [12] was M=3.37 (Md=4.0, SD=1.47) on a scale ranging from 1 to 5 (Device Securement (M=4.25, SD=1.34), Password Generation (M=3.19, SD=1.48), Proactive Awareness (M=2.66, SD=1.36), and Updating (M=3.63, SD=1.1)), and their technology affinity score [14], ranging from 1 to 6, was M=4.20 (SD=1.43).

5.4 Data Collection and Analysis

Data was stored locally on participants‘ headsets (.csv). Participants uploaded their files to a shared folder at the end of the study. We used participant IDs (e.g., P1) to ensure anonymity.

Unless otherwise stated, we ran one-way repeated-measures ANOVAs (for parametric data) and Friedman tests (for non-parametric data). Post-hoc tests were Bonferroni corrected to correct for multiple comparisons. The semi-structured interviews were audio recorded and literally transcribed. The lead researcher went through all interviews to split participants‘ statements into meaningful excerpts.

Table 1: Authentications in *Traditional* and *Hand Menu* were significantly faster than *Glass Unlock* and *Tap*, but *Glass Unlock* and *Tap* did not necessarily result in significantly more PIN corrections and entry errors. Statistical analysis follows the description in Section 5.4. $p < 0.05$ highlighted. The $p < 0.05$ columns show pairwise comparisons.

	Isolated						In Situ					
	(1) *Traditional*	(2) *Glass Unlock*	(3) *Hand Menu*	(4) *Tap*	Statistical Analysis	p<0.05	(1) *Traditional*	(2) *Glass Unlock*	(3) *Hand Menu*	(4) *Tap*	Statistical Analysis	p<0.05
Authentication Times	3.70 (1.31)	5.29 (1.75)	3.17 (0.95)	7.10 (1.64)	$F(3,69)=67.33$, $p<0.05$, $\eta_p^2=0.745$	1-2;1-4;2-4;3-4;2-3	3.85 (1.91)	4.35 (1.75)	3.46 (1.63)	6.65 (1.89)	$F(3,27)=12.67$, $p<0.05$, $\eta_p^2=0.585$	1-4;3-4
PIN Corrections	0.35 (0.44)	0.30 (0.37)	0.06 (0.14)	0.70 (0.76)	$\chi^2(3)=13.45$, $p<0.05$	3-4	0.24 (0.66)	0.40 (0.82)	0.12 (0.44)	0.38 (0.77)	$\chi^2(3)=3.16$, $p=0.367$	n/a
PIN Entry Errors	0.11 (0.20)	0.07 (0.13)	0.07 (0.13)	0.22 (0.28)	$\chi^2(3)=7.16$, $p=0.067$	n/a	0.16 (0.37)	0.04 (0.20)	0.08 (0.40)	0.04 (0.20)	$\chi^2(3)=3.86$, $p=0.277$	n/a

A group of researchers (N=5) then conducted an initial affinity diagram using Miro, an online collaborative whiteboard platform. The lead researcher first introduced the prototypes and the interview questions. The team then grouped the participant statements into themes. All researchers were instructed to divide participant statements into two (or more) statements if required. The lead researcher finalised the affinity diagram based on the initial 2h session with the other researchers. This process resulted in an affinity diagram of 778 participant statements. The main findings are reported in Section 6.7. Reporting the number of participants who shared certain opinions would be inaccurate due to the use of semi-structured interviews. Thus, we only report frequencies where appropriate.

6 RESULTS

6.1 Authentication Metrics

We report the authentication times from the first digit entry to the last input. To ensure internal consistency and a fairer comparison between the prototypes, we count only successful authentications w/o corrections for this analysis. Corrections and PIN input error rates are reported in Section 6.1.2 and 6.1.3. Table 1 shows the measures and statistical analysis. We first report results from *isolated* authentications and then from authentications that were part of an ATM interaction experience (*in situ*), as described in Section 5.2.

6.1.1 Authentication Time (in seconds). There was a significant difference of authentication times between the authentication systems ($F(3,69)=67.33$, $p<0.05$, $\eta_p^2=0.745$). *Traditional* (M=3.70, SD=1.31) and *Hand Menu* (M=3.17, SD=0.95) were significantly faster than *Glass Unlock* (M=5.29, SD=1.75) and *Tap* (M=7.10, SD=1.64) ($p<0.05$). *Glass Unlock* was also significantly faster than *Tap* ($p<0.05$). For *in situ*, there was also a significant main effect ($F(3,27)=12.67$, $p<0.05$, $\eta_p^2=0.585$). Authentication times differed significantly between *Traditional* (M=3.85, SD=1.91) and *Tap* (M=6.65, SD=1.89), between *Hand Menu* (M=3.46, SD=1.63) and *Glass Unlock* (M=4.35, SD=1.75), and between *Hand Menu* and *Tap* ($p<0.05$).

6.1.2 Number of Corrections. The number of corrections differed significantly between the prototypes ($\chi^2(3)=13.45$, $p<0.05$). *Tap* resulted in significantly more digit corrections (M=0.70, SD=0.76) than *Hand Menu* (M=0.06, SD=0.14). There was no significant difference between the other pairs (*Glass Unlock*: M=0.30 (SD=0.37), *Traditional*: M=0.35 (SD=0.44)). For *in situ*, there was no evidence that the number of corrections differed significantly ($\chi^2(3)=3.16$, $p=0.367$). The values were M=0.24 (SD=0.66) for *Traditional*, M=0.40 (SD=0.82) for *Glass Unlock*, M=0.12 (SD=0.44) for *Hand Menu*, and M=0.38 (SD=0.77) for *Tap*.

6.1.3 Number of Incorrect PIN Entries. There was no evidence that the number of incorrect PIN entries differed significantly between the prototypes ($\chi^2(3)=7.16$, $p=0.067$). There was also no evidence of a significant difference for *in situ* ($\chi^2(3)=3.86$, $p=0.277$). Table 1 provides an overview of all values.

Table 2: The table shows the NASA-TLX, 5-point Likert scale, UEQ, and SUS scores. $p < 0.05$ highlighted.

	(1) *Traditional*	(2) *Glass Unlock*	(3) *Hand Menu*	(4) *Tap*	Friedman Test	p<0.05
NASA-TLX [19]						
Mental Demand	15.80 (20.34)	54.00 (28.10)	20.00 (25.21)	58.00 (31.72)	$\chi^2(3)=33.61$, $p<0.05$	1-2;1-4;2-3;3-4
Physical Demand	23.40 (25.11)	26.80 (24.74)	20.40 (20.61)	49.80 (30.63)	$\chi^2(3)=18.32$, $p<0.05$	1-4;3-4
Temporal Demand	25.20 (21.77)	21.40 (17.23)	19.80 (23.21)	33.20 (26.88)	$\chi^2(3)=15.42$, $p<0.05$	3-4
Performance	14.40 (16.91)	15.40 (13.53)	9.40 (9.50)	30.80 (29.46)	$\chi^2(3)=10.14$, $p<0.05$	3-4
Effort	19.80 (21.04)	46.80 (28.83)	19.00 (16.89)	61.80 (30.17)	$\chi^2(3)=38.74$, $p<0.05$	1-2;1-4;2-3;3-4
Frustration	24.20 (26.95)	33.00 (28.39)	12.00 (18.37)	54.80 (31.61)	$\chi^2(3)=28.10$, $p<0.05$	1-4;2-3;3-4
Overall	20.47 (16.12)	32.90 (16.75)	16.77 (12.80)	48.07 (22.36)	$\chi^2(3)=35.98$, $p<0.05$	1-2;1-4;2-3;3-4
5-point Likert Scale						
Ease	4.48 (0.92)	3.32 (1.03)	4.72 (0.54)	2.48 (1.19)	$\chi^2(3)=47.62$, $p<0.05$	1-2;1-4;2-3;3-4
Naturalness	4.36 (0.81)	2.68 (1.18)	3.88(1.20)	2.36 (1.22)	$\chi^2(3)=34.39$, $p<0.05$	1-2;1-4;2-3;3-4
Pleasantness	3.20 (1.22)	3.16 (1.14)	4.36 (0.81)	2.88 (1.20)	$\chi^2(3)=18.65$, $p<0.05$	1-3;2-3;3-4
Speed	4.00 (1.15)	2.68 (1.44)	4.24 (0.72)	2.64 (1.22)	$\chi^2(3)=22.91$, $p<0.05$	1-2;1-4;2-3;3-4
Error-proneness	2.48 (1.33)	3.80 (1.26)	2.44 (0.82)	4.04 (1.24)	$\chi^2(3)=34.18$, $p<0.05$	1-2;1-4;2-3;3-4
Usable in Public	3.36 (1.41)	4.36 (0.81)	4.40 (0.76)	3.64 (1.29)	$\chi^2(3)=10.32$, $p<0.05$	(not confirmed)
Comfortable in Public	2.96 (1.43)	3.96 (1.02)	4.16 (0.85)	2.96 (1.65)	$\chi^2(3)=16.06$, $p<0.05$	1-3;3-4
UEQ [24]						
Attractiveness	0.29 (0.72)	0.39 (1.21)	1.94 (0.78)	-0.21 (1.39)	$\chi^2(3)=34.21$, $p<0.05$	1-3;2-3;3-4
Perspicuity	2.74 (0.31)	1.44 (1.11)	2.39 (0.57)	0.44 (1.28)	$\chi^2(3)=47.71$, $p<0.05$	1-2;1-4;2-3;3-4
Efficiency	1.62 (1.05)	0.46 (1.29)	2.11 (0.67)	-0.31 (1.54)	$\chi^2(3)=40.556$, $p<0.05$	1-2;1-4;2-3;3-4
Dependability	1.06 (0.76)	1.33 (0.93)	1.90 (0.62)	0.32 (1.13)	$\chi^2(3)=31.31$, $p<0.05$	1-3;2-4;3-4
Stimulation	-0.82 (0.91)	0.81 (1.01)	1.68 (0.89)	1.24 (1.00)	$\chi^2(3)=50.32$, $p<0.05$	1-2;1-3;1-4;2-3
Novelty	-2.45 (0.97)	0.91 (0.99)	1.47 (1.06)	2.16 (0.71)	$\chi^2(3)=58.26$, $p<0.05$	1-2;1-3;1-4;2-4
Hedonic Quality	-1.64 (1.24)	0.86 (0.99)	1.58 (0.98)	1.70 (0.98)	$\chi^2(3)=53.65$, $p<0.05$	1-2;1-3;1-4
Pragmatic Quality	1.81 (1.03)	1.08 (1.19)	2.13 (0.64)	0.15 (1.35)	$\chi^2(3)=46.07$, $p<0.05$	1-4;1-2;2-3;3-4
SUS [4]	84.5 (11.39)	70.2 (17.45)	90.5 (7.64)	50.3 (21.06)	n/a	n/a

6.2 Perceived Workload (NASA-TLX)

Participants' perceived workload differed significantly between the prototypes ($\chi^2(3)=35.98$, $p<0.05$). *Glass Unlock* (M=32.90, SD=16.75) and *Tap* (M=48.07, SD=22.36) resulted in a significantly higher perceived workload than *Traditional* (M=20.47, SD=16.12) and *Hand Menu* (M=16.77, SD=12.80) ($p<0.05$). A more nuanced analysis on the level of each dimension is reported in Table 2.

6.3 System Usability Scale (SUS)

We report the SUS as a standard metric for calculating the relative usability of the authentication schemes [46]. *Hand Menu* yielded an "excellent" SUS score [3] of M=90.5 (SD=7.64), followed by *Traditional* with M=84.5 (SD=11.39). *Glass Unlock* and *Tap* yielded an average SUS score between "OK" and "GOOD" [3], with M=70.2 (SD=17.45) for *Glass Unlock* and M=50.3 (SD=21.06) for *Tap*.

6.4 User Experience Questionnaire (UEQ)

Hand Menu received a positive user experience evaluation (> 0.8 [50]) in all dimensions (i.e., attractiveness, perspicuity, efficiency, dependability, stimulation, novelty). *Tap* received a neutral evaluation (-0.8 < score < 0.8 [50]) except for stimulation and novelty (>0.8).

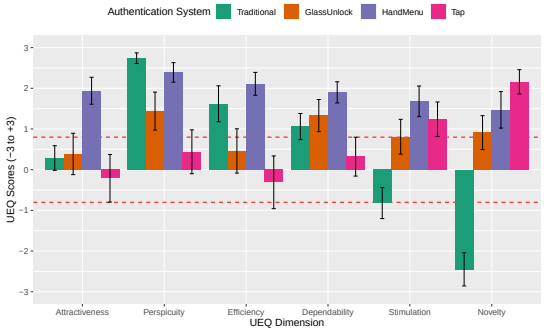

Figure 3: The visualisation shows all dimensions of the UEQ questionnaire. Black error bars denote 95% CI. Dotted red lines denote UEQ's +/-0.8 threshold [50].

The UEQ dimensions for all authentication systems are visualised in Figure 3, with the statistical analysis reported in Table 2.

6.5 Usability/Security Ranking + Likert Scales

We calculated weighted scores (i.e., rank 1×4, rank 2×3, etc.) to report a usability, security, and combined usability and security score. *Hand Menu* achieved the highest usability score (82), followed by *Traditional* (72), *Glass Unlock* (56), and *Tap* (40). *Tap* was perceived as most secure (80), followed by *Hand Menu* (79), *Glass Unlock* (65), and *Traditional* (26). For combined usability and security, *Hand Menu* achieved the highest score (92), followed by *Glass Unlock* (65), *Tap* (52), and *Traditional* (41). This means that participants liked *Hand Menu* the most and *Traditional* the least. While *Tap* was perceived as secure, its usability impacted participants' preference. We discuss this further in Section 7.1.

There was a significant difference between the prototypes' ease ($\chi^2(3)=47.62$, $p<0.05$), naturalness ($\chi^2(3)=34.39$, $p<0.05$), pleasantness ($\chi^2(3)=18.65$, $p<0.05$), speed ($\chi^2(3)=22.91$, $p<0.05$), error-proneness ($\chi^2(3)=34.18$, $p<0.05$), and the extent to which they were perceived as usable ($\chi^2(3)=10.32$, $p<0.05$) and comfortable to use in public ($\chi^2(3)=16.06$, $p<0.05$). Table 2 shows all pairwise comparisons.

6.6 Sense of Presence (IPQ+TPI)

Paticipants' sense of presence (IPQ from 0 to 5 [51]) was M=3.59 (SD=1.80), and their perceived social realism (TPI [26], from 0 to 6) was M=4.89 (SD=1.27). The values for IPQ's dimensions were M=4.2 (SD=1.55) for sense of being part, M=4.61 (SD=1.33) for spatial presence, M=2.99 (SD=1.76) for involvement, and M=2.66 (SD=1.75) for experienced realism.

6.7 Semi-structured Interview

The affinity diagram (Section 5.4) resulted in six main themes.

6.7.1 ***Theme 1: Differences to Real-world Authentication.*** Participants mentioned that *"most of the techniques are really similar to how [they] would imagine they are being implemented in the real world"* (P19) and that their interaction experience and behaviour was *"fairly similar to reality"* (P14), *" quite the same as I would experience in real"* (P17), and *"realistic in how it worked"* (P8). However, P22 mentioned *"it is always easier with a real [ATM] machine"* (P22) and that they missed the physical keypad in *Glass Unlock* and *Traditional*. The lack of haptic feedback when providing input on the (physical) keypad was brought up by many participants. P13 voiced that *Traditional* and *Glass Unlock* use a physical keypad in reality, therefore, they *"expect the [physical] buttons to be there"* (P13). Another topic that came up frequently was the hand tracking's accuracy and that this may be different in reality: *"most errors I did would disappear in the real world"* (P22); *"maybe I did more mistakes because of VR"* (P9); *"usually in real life I use multiple fingers and [authenticating] is a bit quicker"* (P2).

6.7.2 ***Theme 2: Prior Real-world ATM Experience.*** There was a general consensus our setting provided participants with a good simulation of an ATM interaction scenario and that the *"ATM was very convincing [and] felt like authenticating on the machine"* (P22). P7 mentioned that *"in virtual reality, I had the same mindset as if I were to be in front of a real ATM"* (P7). P3 referred back to the training session and stated that *"if I had to do a training session in the real world, I would expect what I did [...] everything was matching my expectations"* (P3). However, some participants mentioned that they were less aware of the (virtual) context around them : *"I have this tunnel effect in VR that I do not have in the real world"* (P22), and that they could tell *"it is not real"* (P12). P5 voiced the graphic fidelity gave the impression it was only a simulation. P18 mentioned although the scenario, sound, and setting was good, the visual representation impacted their perceived realism. There were also differences between participants' previous experience of ATM interaction. P23 voiced their ATM experience differs significantly from our user study environment: *"i don't usually see people in the bank [...] I don't even see the receptionist for the bank"*(P23). P14 mentioned that they usually go to a drive-by ATM where they drive up with their car, open the car window, and reach out of the window to withdraw cash with their smartphone. Similarly, P25 mentioned that *"this is very much down to my personal setup [but] I can use my phone to withdraw money from an ATM"* (P25). P21 mentioned that ATMs are sometimes located at open spaces, which makes them more nervous when withdrawing cash. P2 voiced they *"usually get [money] from like a hole in the wall [...] usually they have them on the street"*(P2).

6.7.3 ***Theme 3: Authentication Systems' Usability.***

Traditional. Participants found *Traditional* as usable and mostly referred to their familiarity with the system. P15 voiced that *Traditional* is *"sort of intuitive [...] partly because that's something we've used for ages"* (P15). P19 mentioned that *"people are already familiar with such a system"* (P19) and that it is a *"common model"* for them because of prior exposure. Participants' overall comments suggest that they were already familiar with *Traditional*, which reinforces our decision of treating *Traditional* as a baseline to simulate 4-digit PIN authentication on a keypad.

Glass Unlock. In line with the original *Glass Unlock* study [59], participants reported that *Glass Unlock* requires an attention switch between the AR content and the actual keypad: *"I've tried to switch between the two [layouts], but I am terrible at memorising numbers [...] memorising the keypad layout was not possible for me"* (P14). The fact that input in *Glass Unlock* still used a "traditional" keypad influenced some participants as they had, due to habituation, already established a mental model of the original keypad layout:

"sometimes, because of habits [...] I have the mapping of 1,2,3,4,5,6 in my mind, which causes a bit of confusion" (P24). Participants perceived *Glass Unlock* as socially acceptable *"because it's very similar to traditional [authentication]"* (P19) and because *"it is quite discreet and not embarrassing [when authenticating]"* (P8).

Hand Menu. Participants perceived *Hand Menu* as fast and easy to use. P21 voiced it takes the advantages of AR but is *"still not too complicated, easy to understand, and easy for people to adapt from the real world to AR, which is important from a product perspective"* (P21). P24 mentioned *Hand Menu "is really close to what we are used to in real life, but provides a little bit more flexibility"* (P24). Some participants found that *Hand Menu* is slightly less acceptable in public because of *"poking the air, which is a little bit weird right now [...] less conventional input, maybe in like 30 years that's not the case anymore then, but we're not there yet."* (P6). P14 stated *"if someone was to walk by me as I'm at the ATM they think I'm a freak because I'm just tapping in the air and I'm wearing these crazy glasses"* (P14).

Tap. *Tap* received negative usability comments due to hand tracking issues and its complexity to recall gestures: *"people have to learn a lot [...] I wasn't able to remember what combination it was to delete or confirm"* (P11). Some participants voiced that *Tap* is comparable slow and does not feel "natural" (P8). P5 voiced they sometimes have to hold something in their hand when using an ATM, which makes *Tap* inconvenient to use. Participants also questioned *Tap* 's social acceptability. P20 mentioned that *"each hand gesture has a different meaning in a different culture [...] a **pinch gesture between middle finger and thumb** in the Buddhism culture has [the meaning of] trying to mediate"* (P20). P25 brought up that *Tap* "is a little bit awkward and I might feel a bit stupid doing that in public" (P25). P6 mentioned that showing the middle finger was the easiest way for them to provide input and that they would be less inclined using *Tap* in public due to such (offensive) gestures.

6.7.4 ***Theme 4: Perceived Security of AR-based Authentication***. Participants perceived all three AR-based authentication systems as more secure than *Traditional*. While shoulder surfing [13] was frequently mentioned by participants when using *Traditional*, they mentioned for the others they *"probably need to get some view of the AR experience"* (P6) or perform a man-in-the-middle attack because *"[the systems] require some network communication"* (P20). P18 mentioned they are usually aware of bystanders when authenticating in public, but that the use of AR could influence their awareness of the real-world surrounding: *"with AR, I think that breaks a little bit of the reality, so you are more prone to the security problem."* (P18). Overall, participants mentioned to either get access to the user's AR view by a) trying to catch a glimpse of what is rendered on the glasses (P4) or b) hacking the system (P7), or conducting a man-in-the-middle attack (P20) to capture information transferred from the user's AR glasses to the ATM.

6.7.5 ***Theme 5: VR-based Real-world Studies: Pros and Cons***. We received overall positive feedback on our remote VR user study on (simulated) real-world authentication systems. P4 mentioned using a VR-based introduction to novel real-world systems can be particularly promising for someone who *"is nervous about doing this on the street or in a (real) bank"* (P4) and that it could be particularly helpful to *"teach kids who have their first experience using these interfaces, like getting their first bank card"* (P4). P20 mentioned that *"AR is still hard to use in real life and [our VR setting] gives a very good setup for reconstructing [an ATM] situation"*. Others mentioned *"it is easier to get more people to try it, because you can have multiple people using it at the same time"* (P25), that implementing and evaluating all different systems in reality would be expensive (P21), and that a remote VR study allowed them to participate, despite being in another country (P16). P23 mentioned that experiencing those systems in VR changed their initial preference: *"I didn't expect* Glass Unlock *to be good [...] I thought* Hand Menu *would be my favorite, but it turns out* Glass Unlock *was actually my favorite; so I'm happy that I get to experience all three in virtual reality, before I can apply it to real life."* (P23). Participants raised some concerns about the lack of interaction fidelity, specifically, that it might not be an accurate representation of how input on a physical keypad works in reality: *"this is still simulated, maybe in a real use case you would have different opinions"* (P19). P25 further voiced that such a VR-based method comes with *"lower fidelity than if you actually build four machines [in reality]"* (P25).

6.7.6 ***Theme 6: Participants' Real-world Study Environment***. Participants participated in our study from a variety of locations: from their living room (n=8), their home office (n=7), their bedroom (n=5), from a research lab (n=4), and from their private gym (n=1). All participants voiced there was nothing that significantly impacted them during the study. Although some participants mentioned minor issues with their Oculus Guardian [38] when configuring their Quest prior to the actual user study or during the training session, they *"did not pay attention to the [real-world] surrounding at all"* (P15). However, P22 brought up the problem of bumping into real-world obstacles: *"I just have to be careful not bumping into my desk when approaching the ATM"* (P22).

7 DISCUSSION

We showcased RVR^3's potential to move traditional lab-based research on prototypes out of a physical lab. Our results provide a glimpse into the usability and social acceptability of two novel AR-based authentication systems: *Hand Menu* and *Tap*. Authentications using *Tap* take significantly longer and are more demanding than *Traditional*, but there is no notable difference between *Traditional* and *Hand Menu* with respect to perceived workload, input speed, number of digit corrections, PIN entry error rate, and pragmatic quality (Table 1 and Table 2). *Hand Menu* resulted in an "excellent" SUS score, achieved the highest usability score, and received an overall positive UEQ evaluation. However, the perceived social acceptability is less prevalent for *Hand Menu* and *Tap* due to the use of AR glasses and mid-air input which some participants perceived as inappropriate in public. As put by P14: *"if AR glasses can get a better form factor and when apple comes out with [their AR glasses], that's when we get more of that acceptance."* (P14). We received mixed comments about participants' perceived realism of the authentication context, which we discuss further in Section 7.2.2. While our findings imply users are reluctant in adopting *Hand Menu* and *Tap*, the results of our study can be decisive and trend-setting for the future of usable and secure authentication.

7.1 RVR^3: A Complementary Research Method

Schmidt et al. [49] and Alt [1] highlighted the HCI community's recent interest in moving human-centred research out of the lab. We believe evaluating real-world prototypes using remote VR can notably advance human-centred research. So far, evaluating hardware prototypes outside of a research lab is often infeasible due to deployability issues [30]. While prototypes can be evaluated in traditional labs, corresponding studies often lack realism [11] and exhibit small and homogenous samples [5, 23, 25]. Using VR as a proxy for real-world research opens a whole new world of opportunities for the research community. Researchers can scale up their sample sizes, recruit user study subjects from different countries, and adjust their systems without being required to purchase or build special hardware. However, despite prior works that validated VR's use for empirical real-world research [27, 29, 31, 56], it is important to 1) acknowledge potential technical limitations and 2) have a clear vision of what can be expected from evaluations that are conducted on virtual artefacts. For example, investigating *Tap* 's usability in the lab using better hand tracking technology (e.g., an OptiTrack system [39]) may impact participants' usability perception. At this point *"we as a community just need to be a little bit more open to what sort of solutions/evaluations we are expecting out of something that has not actually been deployed in the real world"* [30]. RVR^3 forms **a promising research method to implement and evaluate real-world prototypes** and to **move user-centred research out of the lab**, but it is important to be careful when interpreting results from virtually conducted user studies as, similar to lab and organised field studies, they still do not necessarily achieve the often desired high ecological validity.

7.2 Next Steps for RVR^3

7.2.1 ***Direction #1: Longitudinal Studies.*** We successfully conducted a remote VR study to evaluate the usability and social acceptability of novel real-world prototypes. However, we noticed that users' familiarity with prior systems can impact research findings. This was apparent in our study as follows: P24 mentioned their prior experience impacted them when using *Glass Unlock*: *"because of habits [...] I have the mapping of 1,2,3,4,5,6 in my mind"*. At this point, RVR^3 can be particularly valuable by tasking participants to authenticate using *Glass Unlock* once every day (for several months). This would allow researchers to obtain learning effects and receive closer-to-reality usability assessments. Longitudinal studies are rare in HCI, with over 85% of studies lasting a day or less [23]. RVR^3 can address this shortcoming as it opens the door for the community to conduct longitudinal user studies without much resources and effort (e.g., unsupervised [36]).

7.2.2 ***Direction #2: Cross-cultural Studies.*** RVR^3 enabled us to recruit participants from nine different countries. However, we did not formally conduct cross-cultural comparisons of the prototypes' usability and social acceptability. Based on participants' qualitative feedback (Section 6.7.2), conducting large-scale geographically agnostic comparisons is one promising future research direction that can further contribute to the transition of research findings into practice. This would also further highlight VR's strengths for remote investigations on (simulated) real-world prototypes and allows identifying the impact of cultural differences on prototype evaluation findings. This is particularly interesting because technology is often designed and evaluated using western, educated, industrialised, rich, and democratic *"WEIRD"* samples [25]. RVR^3 provides an excellent opportunity for researchers to broaden the international representation of participant samples and work towards impactful systems that are universally useful and engaging. However, in the same breath, are researchers required to be careful when implementing VR-based real-world prototypes and environments as their mental model (e.g., how an environment should look like) may not necessarily align well with the expectations of an international participant sample. While RVR^3 allows researchers to conduct cross-cultural evaluations, further research is required to identify the challenges associated with cross-cultural research using RVR^3.

8 LIMITATIONS

Some specific decisions we made are worth discussing. First, we did not empirically assess the prototypes' security. Although Winkler et al. [59] argued that private near-eye displays (i.e., AR glasses) allow for secure authentication by design, future work may want to conduct exhaustive security evaluations before widely deploying such systems in the wild. Furthermore, to understand the prototypes' usability, we used a large number of common usability measures (e.g., NASA-TLX [19], SUS [4], UEQ [24]) and in-depth qualitative data from interviews. However, some studies [8, 18] also consider participants' preparation times (i.e., the mental time it takes until a user performs input), but as our study was not set up to precisely measure these, we do not include them in our analysis. Finally, we did not evaluate the prototypes with users whose preferred dominant hand is their left hand. However, both *Hand Menu* and *Tap* can accommodate for a switch in users' preference (e.g., left-handed users may want to augment *Hand Menu* next to their right hand).

9 CONCLUSION

In this work, we introduced *Remote Virtual Reality for simulating Real-world Research* (RVR^3). We conducted a remote VR user study with 25 participants to evaluate the usability and social acceptability of two novel real-world authentication systems: *Hand Menu* and *Tap*. Our user study provided a glimpse into the usability and social acceptability of AR-based authentication systems. Authentications were moderately fast in both *Hand Menu* (up to M=3.17 s, SD=0.95 s) and *Tap* (up to M=6.65 s, SD=1.89 s). However, participants criticised their social acceptability and mentioned users might feel reluctant to use AR-based authentication systems these days. Our work highlighted VR's affordances to move traditional lab-based research on real-world systems to participants' home. In summary, we demonstrated RVR^3's potential to complement traditional lab research and how it can open the door for the HCI, Usable Security, and the VR communities to evaluate real-world prototypes out of the lab.

ACKNOWLEDGEMENT

We thank Karola Marky, Alina Störer, and Habiba Farzand for helping with the affinity diagram. Special thanks to Ilyena Hirskyj-Douglas for her feedback on the manuscript. This work was supported by the University of Edinburgh and the University of Glasgow jointly funded PhD studentships, the EPSRC (EP/V008870/1), and PETRAS, which is funded by the EPSRC (EP/S035362/1).

REFERENCES

[1] Florian Alt. 2021. *Out-of-the-Lab Research in Usable Security and Privacy.*

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

[3] Aaron Bangor, Philip Kortum, and James Miller. [n. d.]. Determining what individual SUS scores mean: Adding an adjective rating scale. *J. of usability studies.*

[4] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* (1996).

[5] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proc. of the 2016 CHI Conf. on Human Factors in Comp. Systems.* ACM.

[6] Lynne Coventry, Antonella De Angeli, and Graham Johnson. 2003. Usability and Biometric Verification at the ATM Interface. In *Proc. of the SIGCHI Conf. on Human Factors in Comp. Systems.* ACM.

[7] Alexander De Luca, Katja Hertzschuch, and Heinrich Hussmann. 2010. ColorPIN: Securing PIN Entry through Indirect Input. In *Proc. of the SIGCHI Conf. on Human Factors in Comp. Systems.* ACM.

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

[9] Alexander De Luca, Emanuel von Zezschwitz, and Heinrich Hußmann. 2009. Vibrapass: Secure Authentication Based on Shared Lies. In *Proc. of the SIGCHI Conf. on Human Factors in Comp. Systems.* ACM.

[10] Oculus Developers. 2021. *Oculus Integration SDK: Hand Tracking in Unity.*

[11] Paul Dunphy, Andrew Fitch, and Patrick Olivier. 2008. Gaze-contingent passwords at the ATM. In *4th Conf. on Communication by Gaze Interaction.*

[12] Serge Egelman and Eyal Peer. 2015. *Scaling the Security Wall: Developing a Security Behavior Intentions Scale (SeBIS).* ACM.

[13] Malin Eiband, Mohamed Khamis, Emanuel von Zezschwitz, Heinrich Hussmann, and Florian Alt. 2017. Understanding Shoulder Surfing in the Wild: Stories from Users and Observers. In *Proc. of the 2017 CHI Conf. on Human Factors in Comp. Systems.* ACM.

[14] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *Int. J. of Human–Computer Interaction* (2019).

[15] Free3D. 2019. *3D ATM Model.* https://free3d.com/3d-model/atm-57251.html

[16] Andrzej Grabowski and Jarosław Jankowski. 2015. Virtual Reality-based pilot training for underground coal miners. *Safety Science* (2015).

[17] Meriem Guerar, Mohamed Benmohammed, and Vincent Alimi. 2016. Color wheel pin: Usable and resilient ATM authentication. *J. of High Speed Networks* (2016).

[18] Marian Harbach, Alexander De Luca, and Serge Egelman. 2016. The Anatomy of Smartphone Unlocking: A Field Study of Android Lock Screens. In *Proc. of the 2016 CHI Conf. on Human Factors in Comp. Systems.* ACM.

[19] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proc. of the human factors and ergonomics society annual meeting.*

[20] Mohamed Khamis, Ludwig Trotter, Ville Mäkelä, Emanuel von Zezschwitz, Jens Le, Andreas Bulling, and Florian Alt. 2018. CueAuth: Comparing Touch, Mid-Air Gestures, and Gaze for Cue-Based Authentication on Situated Displays. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* (2018).

[21] Rasib Khan, Ragib Hasan, and Jinfang Xu. 2015. SEPIA: Secure-PIN-Authentication-as-a-Service for ATM Using Mobile and Wearable Devices. In *IEEE Int. Conf. on Mobile Cloud Computing, Services, and Engineering.*

[22] Jesper Kjeldskov and Connor Graham. 2003. A review of mobile HCI research methods. In *Int. Conf. on Mobile Human-Computer Interaction.*

[23] Lisa Koeman. 2021. *HCI/UX Research: What methods do we use?*

[24] Bettina Laugwitz, Theo Held, and Martin Schrepp. 2008. Construction and evaluation of a user experience questionnaire. In *Symp. of the Austrian HCI and usability engineering group.* Springer.

[25] Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, Klaus Opwis, and Katharina Reinecke. 2021. How WEIRD is CHI?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* ACM.

[26] Matthew Lombard, Theresa B Ditton, and Lisa Weinstein. 2009. Measuring presence: the temple presence inventory. In *Proc. of the Int. workshop on presence.*

[27] Ville Mäkelä, Rivu Radiah, Saleh Alsherif, Mohamed Khamis, Chong Xiao, Lisa Borchert, Albrecht Schmidt, and Florian Alt. 2020. Virtual Field Studies: Conducting Studies on Public Displays in Virtual Reality. In *Proc. of the 2020 CHI Conf. on Human Factors in Comp. Systems.* ACM.

[28] Florian Mathis. 2021. [DC] VirSec: Virtual Reality as Cost-Effective Test Bed for Usability and Security Evaluations. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW).*

[29] Florian Mathis, Joseph O'Hagan, Mohamed Khamis, and Kami Vaniea. 2022. Virtual Reality Observations: Using Virtual Reality to Augment Lab-Based Shoulder Surfing Research. In *2022 IEEE Virtual Reality and 3D User Interfaces (VR).*

[30] Florian Mathis, Kami Vaniea, and Mohamed Khamis. 2021. Prototyping Usable Privacy and Security Systems: Insights from Experts. *Int. J. of Human–Computer Interaction* (2021).

[31] Florian Mathis, Kami Vaniea, and Mohamed Khamis. 2021. RepliCueAuth: Validating the Use of a Lab-Based Virtual Reality Setup for Evaluating Authentication Systems. In *Proc. of the 2021 CHI Conf. on Human Factors in Comp. Systems.* ACM.

[32] Florian Mathis, Kami Vaniea, and Mohamed Khamis. 2022. Can I Borrow Your ATM? Using Virtual Reality for (Simulated) In Situ Authentication Research. In *2022 IEEE Virtual Reality and 3D User Interfaces (VR).*

[33] Florian Mathis, Xuesong Zhang, Joseph O'Hagan, Daniel Medeiros, Pejman Saeghe, Mark McGill, Stephen Brewster, and Mohamed Khamis. 2021. *Remote XR Studies: The Golden Future of HCI Research?*

[34] Microsoft. 2019. *Hand menu - Mixed Reality | Microsoft Docs.*

[35] Aske Mottelson and Kasper Hornbæk. 2017. Virtual Reality Studies Outside the Laboratory. In *Proc. of the Symp. on VR Software and Technology.* ACM.

[36] Aske Mottelson, Gustav Bøg Petersen, Klemen Lilija, and Guido Makransky. 2021. Conducting Unsupervised Virtual Reality User Studies Online. *Frontiers in Virtual Reality* (2021).

[37] DaeHun Nyang, Aziz Mohaisen, and Jeonil Kang. 2014. Keylogging-resistant visual authentication protocols. *IEEE Transactions on Mobile Computing* (2014).

[38] Oculus. 2022. *Oculus Guardian.* https://support.oculus.com/guardian

[39] OptiTrack. 2022. *OptiTrack System.* https://optitrack.com/

[40] Asset Store Originals. 2020. *Snaps Prototype | Office.*

[41] Shwetak N. Patel, Jeffrey S. Pierce, and Gregory D. Abowd. 2004. A Gesture-Based Authentication Scheme for Untrusted Public Terminals. In *Proc. of the 17th Annual ACM Symp. on User Interface Software and Technology.* ACM.

[42] Arnaud Prouzeau, Yuchen Wang, Barrett Ens, Wesley Willett, and Tim Dwyer. 2020. Corsican Twin: Authoring In Situ Augmented Reality Visualisations in Virtual Reality. In *Proc. of the Int. Conf. on Advanced Visual Interfaces (AVI '20).*

[43] Susanne Putze, Dmitry Alexandrovsky, Felix Putze, Sebastian Höffner, Jan David Smeddinck, and Rainer Malaka. 2020. Breaking The Experience: Effects of Questionnaires in VR User Studies. In *Proc. of the 2020 CHI Conf. on Human Factors in Comp. Systems.* ACM.

[44] Qualtrics. 2005. *Qualtrics: Experience-Design.* https://www.qualtrics.com

[45] Rivu Radiah, Ville Mäkelä, Sarah Prange, Sarah Delgado Rodriguez, Robin Piening, Yumeng Zhou, Kay Köhle, Ken Pfeuffer, Yomna Abdelrahman, Matthias Hoppe, Albrecht Schmidt, and Florian Alt. 2021. Remote VR Studies: A Framework for Running Virtual Reality Studies Remotely Via Participant-Owned HMDs. *ACM Trans. Comput.-Hum. Interact.* (2021).

[46] Scott Ruoti and Kent Seamons. 2016. Standard metrics and scenarios for usable authentication. In *Twelfth Symp. on Usable Privacy and Security.*

[47] M Angela Sasse and Ivan Flechais. 2005. Usable security: Why do we need it? How do we get it? O'Reilly.

[48] Gian-Luca Savino, Niklas Emanuel, Steven Kowalzik, Felix Kroll, Marvin C. Lange, Matthis Laudan, Rieke Leder, Zhanhua Liang, Dayana Markhabayeva, Martin Schmeißer, Nicolai Schütz, Carolin Stellmacher, Zihe Xu, Kerstin Bub, Thorsten Kluss, Jaime Maldonado, Ernst Kruijff, and Johannes Schöning. 2019. Comparing Pedestrian Navigation Methods in Virtual Reality and Real Life. In *2019 Int. Conf. on Multimodal Interaction.* ACM.

[49] Albrecht Schmidt, Florian Alt, and Ville Mäkelä. 2021. *Evaluation in Human-Computer Interaction – Beyond Lab Studies.* ACM.

[50] Martin Schrepp. 2019. *User Experience Questionnaire Handbook.*

[51] Thomas Schubert, Frank Friedmann, and Holger Regenbrecht. 2001. The experience of presence: Factor analytic insights. *Presence: Teleoperators & Virtual Environments* (2001).

[52] Richard Sharp, James Scott, and Alastair R. Beresford. 2006. Secure Mobile Computing Via Public Terminals. In *Pervasive Computing*, Kenneth P. Fishkin, Bernt Schiele, Paddy Nixon, and Aaron Quigley (Eds.).

[53] Mel Slater. 2009. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences* (2009).

[54] Anthony Steed, Francisco R. Ortega, Adam S. Williams, Ernst Kruijff, Wolfgang Stuerzlinger, Anil Ufuk Batmaz, Andrea Stevenson Won, Evan Suma Rosenberg, Adalberto L. Simeone, and Aleshia Hayes. 2020. Evaluating Immersive Experiences during Covid-19 and Beyond. *Interactions* (2020).

[55] Desney S Tan, Pedram Keyani, and Mary Czerwinski. 2005. Spy-resistant keyboard: more secure password entry on public touch screen displays. In *Proc. of the 17th Australia Conf. on Computer-Human Interaction: Citizens Online: Considerations for Today and the Future.*

[56] Alexandra Voit, Sven Mayer, Valentin Schwind, and Niels Henze. 2019. Online, VR, AR, Lab, and In-Situ: Comparison of Research Methods to Evaluate Smart Artifacts. In *Proc. of the 2019 CHI Conf. on Human Factors in Comp. Systems.* ACM.

[57] Kieran Watson, Robin Bretin Mohamed Khamis, and Florian Mathis. 2022. The Feet in Human-Centred Security: Investigating Foot-Based User Authentication for Public Displays. In *Proc. of the 2022 CHI Conf. on Human Factors in Comp. Systems* (New Orleans, LA, USA) *(CHI EA '22).*

[58] Kieran Waugh and Judy Robertson. 2021. Don't Touch Me! A Comparison of Usability on Touch and Non-touch Inputs. In *INTERACT 2021.*

[59] 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.*

A SEMI-STRUCTURED INTERVIEW QUESTIONS

Our semi-structured interviews were guided by the following questions. Additional questions were asked when appropriate.

(1) **Virtual Reality and Perceived Presence and Realism**
   (a) Could you please walk us through the different authentication methods and tell us what differences may appear when using the methods in the real world rather than in VR as just experienced?
   (b) Do you think the virtual environment impacted your behaviour when providing input with the corresponding authentication method? If so, how?
   (c) Could you please tell us why (or why not) you felt being part of the environment where the authentication happened?
   (d) What (if any) is the difference between withdrawing cash at a real-world bank ATM and what you have just experienced?
   (e) How did the virtual environment (+ virtual bystanders) impact your ATM interaction behaviour?
   (f) Please think about your last ATM withdrawal in the real world. What was different to what you have just experienced?
   (g) Could you please tell us how realistic the ATM experience was for you? Please briefly justify your response.

(2) **Perceived Usability and Ranking of the Prototypes**
   (a) Please justify the your ranking of the methods in terms of (a) usability, (b) security, and (c) usability + security.
   (b) Have you used such an authentication method previously in any other context?
   (c) Please tell us how you felt using this authentication method to withdraw cash on an ATM.
   (d) Please tell us (a) what you particularly liked, and (b) what you did not like when using this method to authenticate on an ATM.
   (e) *(only for Glass Unlock)* Did you constantly switch between the private near-eye display and the keypad on the ATM or rather stayed on either of them?

(3) **Perceived Security of the Prototypes**
   (a) How secure do you think is this authentication method against observations where a bystanders observes your authentication?
   (b) Can you think of any attacks that could break the security of this authentication system?
   (c) Consider you want to attack a user‘s ATM authentication when using this method. How would you try to access their PIN?

(4) **Enhancements of the Prototypes**
   (a) Is there anything in particular that you would like to improve in this authentication method?
   (b) Do you have any other ideas on how authentication in front of public displays like ATMs could look like?

(5) **Impact of the Real-world Environment and the Experimenter**
   (a) Could you please describe your real-world surrounding and how it looks like? Please note that we do not expect a detailed description of your personal space, but it would be great if you could give a rough overview of the room you are currently in.
   (b) Could you please tell us to what extent the real-world surrounding impacted you while performing the authentications?
   (c) Could you please tell us how the experimenter on the Zoom call impacted you while performing the authentications?