IEEE CONFERENCE ON VIRTUAL REALITY + 3D USER INTERFACES | 12-16 MARCH 2022 | VIRTUAL EVENT

Can I Borrow Your ATM? Using Virtual Reality for (Simulated) In Situ Authentication Research

In Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces (VR)

Christchurch, New Zealand, March 2022 (IEEE VR 2022)

Florian Mathis, University of Glasgow, United Kingdom
florian.mathis@glasgow.ac.uk
Kami Vaniea, University of Edinburgh, United Kingdom
kvaniea@inf.ed.ac.uk
Mohamed Khamis, University of Glasgow, United Kingdom
mohamed.khamis@glasgow.ac.uk

Can I Borrow Your ATM? Using Virtual Reality for (Simulated) In Situ Authentication Research

Florian Mathis*
University of Glasgow
University of Edinburgh

Kami Vaniea†
University of Edinburgh

Mohamed Khamis‡
University of Glasgow

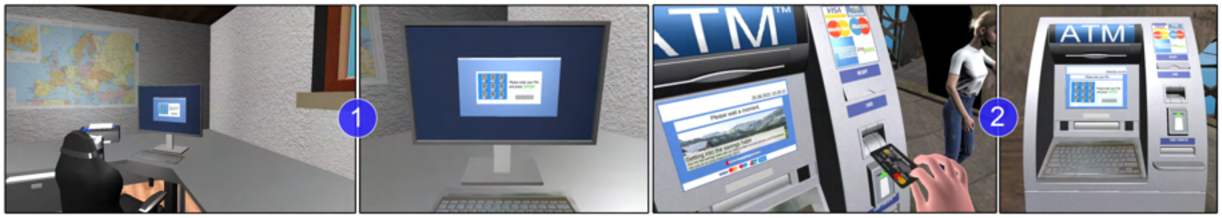

Figure 1: We propose the use of VR for in situ authentication evaluations in private and sensitive contexts. We investigated the impact of *isolated* authentications, where users authenticate in a lab environment (❶), and *in situ* authentications, where users' authentication precedes a primary task (❷), on a system's usability evaluation results. Our work highlights the importance of *in situ* authentication evaluations, demonstrates how the use of VR advances common authentication system evaluations, and enables researchers to research contexts that are close to impossible to study in the wild and challenging to replicate in the lab.

Abstract

In situ evaluations of novel authentication systems, where the system is evaluated in its intended usage context, are often infeasible due to ethical and legal constraints. Consequently, researchers evaluate their authentication systems in the lab, which questions the ecological validity. In this work, we explore how VR can overcome the shortcomings of authentication studies conducted in the lab and contribute towards more realistic authentication research. We built a highly realistic automated teller machine (ATM) and a VR replica to investigate through a user study (N=20) the impact of in situ evaluations on an authentication system's usability results. We evaluated and compared: Lab studies in the real world, lab studies in VR, in situ studies in the real world, and in situ studies in VR. Our findings highlight 1) VR's great potential to circumvent potential restrictions researchers experience when evaluating authentication schemes and 2) the impact of the context on an authentication system's usability evaluation results. In situ ATM authentications took longer (+24.71% in the real world, +14.17% in VR) than authentications in a traditional (VR) lab environment and elicited a higher sense of being part of an ATM authentication scenario compared to a real-world and VR-based evaluation in the lab. Our quantitative findings, along with participants' qualitative feedback, provide first evidence of increased authentication realism when using VR for in situ authentication research. We provide researchers with a novel research approach to conduct (simulated) in situ authentication research, discuss our findings in the light of prior works, and conclude with three key lessons to support researchers in deciding when to use VR for in situ authentication research.

Keywords: Virtual Reality, Authentication, In Situ Research

1 Introduction

Usable security researchers experience significant challenges when conducting research in sensitive and private contexts. A classic example is research on automated teller machines (ATMs), an area of research that is challenging to conduct due to ethical and legal constraints [21,51,76]. While a plethora of novel ATM authentication methods has been proposed (e.g., [11,20,25]), there is a shortcoming in research that evaluates these systems in their corresponding environment (in situ[1]), creating uncertainty in the value and validity of authentication research conducted in the lab. For example, while researchers proposed a significant amount of authentication methods that can potentially outperform the widely used 4-digit PINs, the vast majority of those were not evaluated in realistic scenarios in the wild. Consequently, novel ATM authentication methods have not yet found their way into the real world with ATMs using traditional 4-digit PINs, which are subject to many vulnerabilities such as shoulder surfing[2], since 1967. One potential reason is the fact that researching such private and sensitive contexts is near impossible because researchers often do not have the required resources and links to industry partners [51]. It is also not ethically and legally feasible to video record users' actual PIN input on a real-world ATM [7,21,76]. As a result, researchers either evaluate their security systems in a setup that is isolated[3] from an actual authentication scenario (e.g., [20,42]) or aim to create "realistic" authentication scenarios in the lab (e.g., [25]). However, it remains unclear to what extent the conclusions drawn from such lab experiments match with the findings from an in situ evaluation where the authentication scheme is part of an actual production task [67] (e.g., withdrawing cash from an ATM). In situ investigations are particularly interesting because authentication is usually not users' primary task, but a secondary task that precedes a production task [67]. This raises the following questions: Does a usability evaluation of an authentication scheme differ when the scheme is evaluated in situ rather than isolated from users' production task? And do users behave in a VR-simulated authentication scenario similarly as they do in the wild? To address these questions, we conducted a user study (N=20) in which we exposed participants to an ATM authentication scenario in the real world and in VR. We compared a) users' per-

*e-mail: florian.mathis@glasgow.ac.uk
†e-mail: kvaniea@inf.ed.ac.uk
‡e-mail: mohamed.khamis@glasgow.ac.uk

[1] *In situ* refers to a (simulated) environment for which the authentication scheme is intended to be used (similar to [75,78]).

[2] *Shoulder surfing* refers to the act of observing other people's information without their consent [27].

[3] *Isolated* refers to a scenario where participants experience the authentication independent from an actual production task [67].

formance and behaviour when interacting with two different ATM replications and b) how embedding an authentication scheme into its actual usage context impacts the experiment's usability results. Our results show that in situ authentications take longer than isolated ones with an increase in the authentication times by 24.71% in the real world (RW) and by 14.17% in VR. We also investigated a more vivid VR environment (see Fig. 3) to investigate the impact of external factors such as social density [47] on in situ authentication evaluations, which increased authentication times by 22.31% compared to an evaluation in a virtual lab environment. Our VR research approach contributed to some sense of authentication realism, even in a fully controlled user study. We found that (simulated) in situ VR evaluations elicited a higher sense of presence (M=3.77, SD=1.67) compared to a VR authentication evaluation in the lab (M=3.05, SD=1.80) and resulted in a decreased feeling of being part of a user study (M=3.15, SD=1.11) compared to a traditional lab study in the real world (M=3.70, SD=1.23). These results highlight the potential of using VR for authentication research that is closer to reality. Despite the effort put into simulating such an ATM scenario in the lab and recreating it in VR, there still remains a gap between research in the wild and what is possible to simulate. For example, our simulations were not able to elicit PIN entry shielding behaviour, which happens in reality as observed by De Luca et al. [21].

Contribution Statement. The contribution of our work is threefold: **(1)** While prior work showed some first evidence of the use of VR for real-world authentication research in a lab setting [52], we propose VR as a research method for **in situ authentication research** and evaluate its methodology through a replication and comparison study of ColorPIN [20], whose characteristics we describe in more detail in Sect. 3.1. **(2)** We compare users' authentication performance and behaviour in two different scenarios (i.e., isolated and in situ) using two different means: a real-world evaluation and a VR evaluation. We contribute to the validation of applying VR in the real-world authentication research domain, show how in situ (VR) evaluations lead to a sense of realism, and enable researchers to study novel authentication methods in their intended usage contexts. **(3)** We discuss the importance of (simulated) in situ authentication evaluations, link our findings to prior real-world ATM research, and provide three key lessons to support researchers in their future (in situ) authentication research using VR.

2 Related work

We review previous works in the authentication domain and works that used VR as a research platform for human-centred research.

2.1 Authentication Research

The authentication research domain is considerably large with works ranging from novel prototype systems aiming to provide users with usable and secure authentication schemes (e.g., [1, 13, 34, 77]) to qualitative research that aims to understand users' authentication behaviour and preference (e.g., [19,43,56,63]). While a significant part of usable security research draws on out-of-the-lab research methods [4] (e.g., interviews [10, 68], online studies [8, 9, 19]), there is a considerable amount of research that still relies on lab-based investigations. Corresponding user studies are suitable for the evaluation of novel authentication schemes for mobile devices (e.g., [14, 22]) and desktop computers (e.g., [1, 71, 79]), but conducting authentication research on devices that are not easy to access (e.g., ATMs) is particularly challenging. Despite the challenges, there is research in (or about) such sensitive and private contexts. Eiband et al. [27] conducted an online survey to better understand shoulder surfing in the wild by synthesising stories from users and observers. Instead of observing real-world shoulder surfing, they applied the critical incident technique [30] to "generate a comprehensive and detailed description" of shoulder surfing in the wild [27]. De Luca et al. [21] aimed to understand real-world ATM use and discussed several ethical and legal constraints that played an important role in their study design. De Luca et al. [21] relied on observational data and emphasised the importance of knowing the limitations of such observational research (e.g., some findings are based on speculative reasoning). Dunphy et al. [25] exposed participants to a low-fidelity ATM prototype environment and explored the strengths of gaze-based ATM authentication. Building upon the previously mentioned approaches to conduct more realistic security research, Krol et al. [45] emphasised the importance of providing participants with a primary task, which is usually not authentication [65], to work towards robust usable security research. Fahl et al. [29] showed that findings of roleplaying scenarios to create passwords in the lab match to a great extent with users' real-world behaviour. However, roleplaying to improve ecological validity of password studies [29] still comes with limitations that need to be considered when drawing conclusions. Dunphy et al. [25] argued that future work is required to increase the ecological validity of their lab study's ATM environment. Although De Luca et al. [20] made use of "typical" ATM hardware in their lab study, they argued that lab experiments "can never mirror completely the real situation when using an ATM" [21].

In summary, conducting in situ authentication research is an underexplored research field due to the ethical and legal constraints and the resource-intensive nature of this type of research. Mathis et al. [51] argued that a future direction of usable security should be to "investigate alternative platforms for conducting research" to result in noticeable real-world impact, to which we aim to contribute with our work.

2.2 VR Studies for Real-World Research

There is a large body of work that investigated how *VR studies* can complement existing research paradigms. Mäkelä et al. [48] investigated the feasibility of VR to evaluate deployments of public displays. Their comparison between a real-world setting and a VR replica showed that users behave similarly in front of a public display in both realities [48]. Savino et al. [69] showed that VR can offer a promising simulation environment to test pedestrian navigation techniques, but that using VR also comes with challenges that impact its validity compared to an equivalent real-world investigation. For example, the limited field of view and legibility can significantly impact users' experience in VR. Voit et al. [75] explored the differences and similarities when using a set of five different research methods (i.e., online, VR, AR, lab setup, in situ) to evaluate smart artefacts. Their user study showed that the selected research method can negatively impact the study outcome with regards to usability ratings [75]. However, VR and in situ provided similar ratings for usability, attractiveness, pragmatic and hedonic qualities [75]. Mathis et al. [52] conducted a replication study of a real-world authentication system to find that many findings collected in VR match with the previously conducted real-world study [77]. However, in a similar vein as Voit et al. [75] and Savino et al. [69], Mathis et al. argued that the technology can have a negative impact on the evaluation (e.g., touch input was significantly slower in VR than in reality) [52]. Others investigated the feasibility of a non-immersive VR ATM environment to relearn the use of ATMs for people with acquired brain injury, showing that a VR-powered ATM presents participants with a valid assessment and training tool [31]. Ragan et al. [64] proposed the use of virtual environments to simulate augmented reality systems for usability evaluations to allow for complete control of the study environment.

In summary, previous works provided some initial evidence of the use of VR as a substitute technology for empirical evaluations and argued that using VR as an empirical research method can be particularly promising in situations where evaluations are challenging to conduct (e.g., in safety-critical, expensive scenarios [6, 48, 54, 62, 78]).

2.3 Research Gap and Research Questions

Novel authentication schemes are often evaluated in a way where participants do not interact with the authentication scheme in its intended context (e.g., [20, 42]). To overcome the existing shortcomings of authentication research in the lab, we build upon the success of prior *VR studies* and set out to answer the following two research questions: (**RQ_1**) How does a VR-based in situ authentication evaluation impact a real-world authentication system's usability results? (**RQ_2**) To what extent does the use of VR contribute towards a sense of authentication realism, even in a controlled user study?

We carry out in-depth **empirical (simulated) in situ authentication research** by applying VR as a research method to replicate real-world authentication scenarios. This builds upon Mathis et al.'s work that provided some initial evidence of the use of VR for real-world authentication research [52], but did not conduct a comparison between *in situ* and *isolated* authentications. Understanding the impact of in situ evaluations on an authentication systems usability evaluation results is important because understanding users' behaviour is a key factor in human-centred security research [5, 65]. Our work is the first that applies the proposed idea of virtual in situ authentication research by Mathis et al. [50–52], with a particular focus on VR-powered *in situ* authentications, which is a novel application domain for VR studies as security is best studied when users experience it as a secondary task like in reality [67].

3 Method

While a common approach to evaluate novel authentication schemes is to compare their usability to traditional authentication systems [18, 39, 41, 60], we investigate to what extent the context in which such novel systems are evaluated impacts the system's usability results and users' behaviour. Mathis et al. [52] voiced that significantly more validation work is required before the research community can treat a VR-based research approach as a valid alternative to real-world studies. As detailed in Sect. 2.3, prior work did not investigate a comparison between *in situ* and *isolated* authentications, which is particularly interesting for human-centred authentication research because 1) conducting in-depth in situ authentication research in the real world is close to impossible [21], and 2) authentication is usually not considered to be users' primary task [44, 45, 67]. As a result, and to further contribute to the validation of VR studies for real-world authentication research, we compare in situ evaluations to lab-based evaluations in both the real and virtual world. We, therefore, built a real-world ATM prototype and a virtual replica of the same ATM. The study followed a within-subject design with the order of the conditions being counter-balanced using a Latin Square. The ATM environments were set up as realistic as possible (see Fig. 3 and Sect. 3.7) to contribute to high ecological validity. Skarbez et al. [72] showed that a realistic scale of the space is the most important factor for generating a "feeling of reality" [72]. While our setting depicts a "realistic" ATM scenario, there is still a gap to an ATM experience in the wild, which we discuss further in Sect. 5 based on our findings and prior research in the wild [21, 76].

3.1 Studied Authentication Scheme: ColorPIN

We decided to replicate ColorPIN [20] to achieve our main goal of **investigating the suitability of VR for in situ authentication research** and **assessing the impact of VR and in situ evaluations on users' authentication performance and behaviour**. Our motivation behind replicating ColorPIN [20] is manyfold. First, ColorPIN is proposed as an authentication scheme for ATMs, but its original evaluation took place in an isolated way, which means that users' authentications were not part of an actual ATM interaction scenario [20]. Through our investigation, we aim to close the gap between such isolated usability evaluations and in situ evaluations of authentication schemes. Second, ColorPIN's intended application context, ATM authentication, received significant attention in the past that highlighted the challenges researchers experience in such contexts [21, 76]. Furthermore, ColorPIN's underlying concept (i.e., one-to-one relationship between PIN length and required input) is commonly used in authentication research (e.g., [24, 42, 53, 77]). In summary, the ethical and legal barriers when conducting authentication research in the wild (e.g., [21, 76]), the widespread use of ATMs [59], and ColorPIN's characteristics [20] make it a suitable candidate for our investigation.

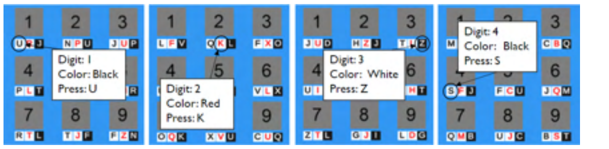

Figure 2: Exemplary ColorPIN [20] entry. To input the PIN *"1(black) 2(red) 3(white) 4(black)"* the user inputs the letters "UKZS".

ColorPIN: A Brief Overview. ColorPIN is a highly secure and usable ATM authentication scheme, initially proposed by De Luca et al. [20] and further studied by Bianchi et al. [12] and Lee [46]. A user enters a ColorPIN using a commercial keyboard by selecting a letter that corresponds to a digit, see Fig. 2. For example, instead of entering *1-2-3-4* on a keypad, users map their ColorPIN to coloured letters that are displayed below the digits on the authentication interface. To input 1(black) in Fig. 2 the user would need to press "U" on the keyboard. Letters are randomly assigned after each input.

3.2 Independent Variables

We investigated the impact of two independent variables on user authentications: the **authentication context** (isolated from a primary task vs integrated into a primary task) and **authentication environment** (real world vs virtual reality). In all conditions (see Fig. 1 and Fig. 3), we use ColorPIN [20] as authentication method.

3.2.1 Authentication Context (IV1)

We investigate the extent to which the authentication context impacts users' authentication performance and behaviour.

Isolated Authentication (Lab). *Isolated* refers to a traditional lab setting where the authentication scheme is not evaluated in the intended usage context (e.g., on a desktop PC instead of an ATM). This presents participants with an authentication isolated from a production task, the de facto standard when evaluating authentication scheme (e.g., [1, 39, 41, 42, 77]), which is inline with the original real-world study context [20]. We use this condition as our baseline in the real world (*RW Lab*) and in VR (*VR Lab*).

Integrated Authentication (ATM). We embedded ColorPIN into an actual ATM system for which the scheme has initially been built [20]. We did this to increase the realism of the authentication scenario by preceding the authentication by a production task [65]. By doing this, we do not artificially draw participants' attention to the authentication itself, but rather to the production task, which depicts a scenario closer to how authentication usually happens in reality. We had two identical scenarios: one in the real world (*RW ATM*) and one in VR (*VR ATM*). Due to the required resources in the real world to simulate realistic ATM environments (e.g., additional bystanders, access to a public space), we further aimed to demonstrate how VR can be used as an effective and affordable research method. We included an additional condition (*VR ATM Public*) to investigate if participants change their authentication behaviour based on external factors such as additional bystanders (e.g., social density, location [47]).

3.2.2 Authentication Environment (IV2)

We investigate the extent to which the environment impacts users' authentication performance and behaviour.

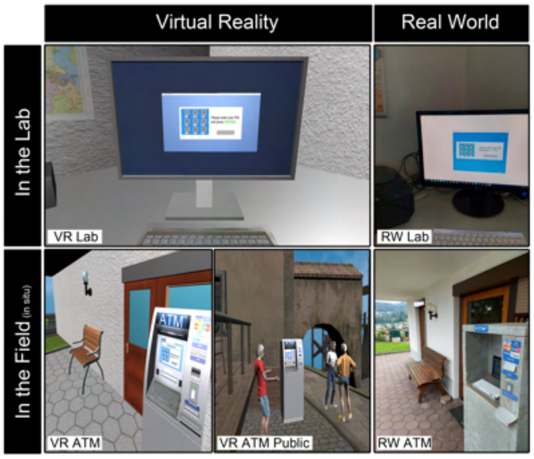

Figure 3: We studied five authentication scenarios: Two in the real world and three in VR. We had virtual replicas of both real-world environments and treat *in the lab* as our baseline in the real world (*RW Lab*) and in VR (*VR Lab*).

Real World (RW). Depending on the authentication context (isolated or integrated), the real-world condition depicts a traditional lab environment or a replicated ATM authentication scenario. We aimed to compare how such a simulated ATM scenario matches a virtual replication of the same environment and to what extent our results match the findings from ColorPIN's original study [20].

Virtual Reality (VR). We created replicas of the real-world environments (i.e., lab and outdoor) to explore users' authentication performance and behaviour when using ColorPIN [20] in a virtual environment. This allows us to compare participants' authentication performance and behaviour in VR to our real-world setup and pinpoint differences. We created an additional virtual replica of a public space to further demonstrate the strengths of VR for in situ research as described in Sect. 3.2.1. Fig. 3 shows all conditions.

3.3 Participant Instructions

We leveraged storytelling to present users with a realistic authentication scenario – a method where researchers introduce plausible authentication scenarios to increase the ecological validity of lab-based user studies [29,45]. While in *RW Lab* and *VR Lab* participants were told to imagine they would need to use their credit card to withdraw money, in *integrated* (e.g., *RW ATM*) participants had to take out their credit card (a fake one which we provided) and navigate through the ATM user interface (UI) before authenticating. Consequently, the ATM interaction steps in *RW ATM*, *VR ATM*, and *VR ATM Public* consisted of a) inserting the credit card, b) interacting with the ATM according to the given scenario, c) authenticating using ColorPIN, and d) taking the credit card and the money out of the ATM. For *RW Lab* and *VR Lab*, participants authenticated using ColorPIN in front of a (virtual) desktop screen, which depicts a traditional usability evaluation of authentication methods [20,42]. Here, participants were directly exposed to ColorPIN and their only task was to authenticate. We used the following ATM scenario: *"Your PIN for your credit card is: [predefined ColorPIN]. As a customer, you now want to login to your account using card and PIN code so that you can withdraw [amount of cash]. After entering your PIN, you expect that the system provides you with the requested cash and spits out the money. Please withdraw [amount of cash] now."*. The story remained the same across the conditions, but the amount of cash the participants had to withdraw and their ColorPIN changed. Participants were asked to perform the ATM withdrawal task in a way most similar to how they would do it in the wild. We did this to collect insights into their input and shielding strategies when interacting with the (virtual) ATM.

3.4 Study Procedure

Participants' task was to authenticate with ColorPIN using each of the five conditions, which means that participants went through 5 authentication sessions (authentication context × authentication environment = 4 + *VR ATM Public* = 5). We introduced participants first to the scenarios (lab and ATM) and environments (real world and VR). Participants then went through a training phase where they were introduced to ColorPIN prior to the data collection. This is a common approach in authentication research to ensure participants are familiar with the system [20,53,77]. Participants then authenticated using ColorPIN. After each authentication session, they reported their sense of presence using the IPQ questionnaire [70] and their perceived workload using the raw NASA-TLX questionnaire [35] (for both the authentication and the overall ATM interaction). Although the use of presence questionnaires for real-world experiences is debatable [74], we treat the reported sense of presence in the RW conditions as an indication of the user‘s experience and sense of being part of an ATM authentication scenario, which we further discuss along with qualitative feedback. After filling in the IPQ and NASA-TLX, participants were asked to verbally walk us through their interactions and tell us about their perceived primary and secondary task (structured interview, see Sect. 4.4). We were interested in participants' task perception, i.e., if they perceived the authentication as their primary or secondary task. We also asked participants to fill in a set of 5-point Likert scale questions.

We concluded with a ranking on the realism of the different authentication contexts and environments and with a semi-structured interview (reported in Sect. 4.6). We also collected participants' security knowledge and attitude using the Security Behavior Intentions Scale (SeBIS) [26] and their technological affinity using the Affinity for Technology Interaction (ATI) scale [32] to allow for better comparisons in future studies and support replications.

3.5 Statistical Analysis and Qualitative Data Analysis

Unless otherwise stated, we used repeated-measures ANOVAs for our statistical analysis. We performed an aligned rank transformation on our data to correct for violations of normalcy using ART by Wobbrock et al. [80]. For post-hoc pairwise comparisons we used ART-C [28], which were corrected using Bonferroni correction. We report η_p^2 (*partial eta square*) as an effect size statistic for our ART analysis (0.01 = small, 0.06 = medium, 0.14 = large [15,16]). We had two baselines in our work: *RW Lab* and *VR Lab*; therefore, we ran two-way repeated-measures ANOVA where the independent variables were: Context (Lab vs ATM) and Environment (RW vs VR); this covered *RW Lab*, *VR Lab*, *RW ATM*, and *VR ATM*. We ran additional one-way repeated measures ANOVAs when comparing *VR Lab* to *VR ATM* and *VR ATM Public*. There were no outliers that had to be removed (e.g., measurement errors, data entry errors) – we kept those data points that are suspected of being legitimate to be representative of the population as a whole [61]. Previous work on ColorPIN showed that such outliers can be expected [46], especially when studying a sample not recruited within a university environment [36,45,51]. The structured interviews after each condition were transcribed and coded, with the most common themes discussed in Sect. 4.4. For the semi-structured interviews at the end of the study we split participants' statements into meaningful excerpts, which resulted in overall N=280 participant statements. The lead researcher systematically clustered all participant statements using an affinity diagram. A second researcher performed a review of the clustering and added tags to clusters that required another iteration. Two researchers met to discuss the clustering and to resolve any discussion points that came up during the review process. Through this process, we identified five themes: 1) Perceived Realism: Reasoning, 2) Perceived Differences: ATM Authentication in the Wild, 3) Input Behaviour: The Keyboard, 4) ColorPIN Recall Strategy, 5) General Comments. We discuss those that are particularly relevant for the scope of our research in Sect. 4.6. Note that reporting the number of participants who shared certain opinions would be inaccurate due to the use of a semi-structured interview approach; thus, we only report frequencies where appropriate. Quotes were translated from German where necessary.

3.6 Call for Participation and Demographics

We recruited 20 participants through local societies and word of mouth after receiving ethical approval from the College of Science and Engineering Ethics Committee at the University of Glasgow. The study was conducted in Austria, with participants being paid according to the local standard *(€10/hour)*. We used a single-blinded research approach to ensure that results better reflect real-world behaviour when interacting with authentication systems [45]. We did not disclose our experimental motive because doing so may impact participants' behaviour and responses. Although blinded experiments have already been conducted in the VR, security, and HCI field (e.g., [2, 37, 49, 58, 81]), there are ethical considerations with not telling the entire truth to participants. We disclosed the aim of our study at the end of each study session.

Demographics. Participants were on average 35.45 years old (SD=9.46). 13 participants self identified as male, 7 as female. All participants have used an ATM before, with M=2.33 (SD=2.03) ATM cash withdrawals a month. Almost all participants (n=17) had previous VR experience (a couple of times at a friend‘s house (n=6), briefly at a demonstration (n=10), in their job (n=1)). The sample‘s security knowledge and attitude score was M=3.18 (SD=1.57) on a scale ranging from 1 to 5 (*Device Securement*: M=4.21,SD=1.36; *Password Generation*: M=3.3, SD=1.59; *Proactive Awareness*: M=2.44,SD=1.28; *Updating*: M=2.87,SD=1.45) and its technological affinity ranging from 1 to 6 was M=3.88 (SD=1.63).

3.7 Apparatus and Implementation

We implemented two software elements (in Unity, C#) to evaluate the strengths of VR for (simulated) in situ research and compare its findings to a state-of-the-art evaluation in the lab. We implemented a fully functional ATM UI for a real-world and virtual ATM (Fig. 3). For the real-world ATM that employs ColorPIN, we used a touch screen, cardboard, styrofoam, and metallic spray paint. Participants interacted with the ATM in an outdoor environment (see Fig. 3). We did this to further increase the realism of such an ATM interaction experience. We attached a commercial keyboard (as done in the original study [20]) to the ATM‘s touch screen. Both ATMs (RW and VR) enable participants to navigate through the UI as they wish. We simulated a sensory behaviour for the real-world ATM to ensure high internal validity between the two ATMs. This means that once participants put in the credit card, the ATM‘s UI changed based on an external trigger initiated by the experimenter (i.e., Wizard of Oz [17]). This behaviour was fully implemented in VR. For the virtual ATM, we used an ATM that matches our prototype in the real world [33]. We replicated the study environment as close as possible and used the branding of a local bank to increase the realism of our ATMs. Due to our baseline conditions, we also replicated the lab in the real world (*RW Lab*) to present participants with the same environment in VR (*VR Lab*). We used an Oculus Quest 2 and a Logitech C920 to bring the real-world keyboard into virtuality, which we mounted on a mini tripod and on a flexible camera holder (see Fig. 4). For the transition of users' virtual hands (rendered through the Oculus Integration SDK [23]) to users' real hands (rendered through the camera feed), we used an inferred partial blending. This means that a view of the keyboard and users‘ hands only were blended into virtuality using a chroma key shader and a green screen (similar to [55]). We checked the position of users' virtual hands and if they do not overlap with the physical keyboard we render the virtual hands, otherwise users' real-world hands. We used Adobe Mixamo [3] for the virtual avatars in *VR ATM Public* (see Fig. 1 and Fig. 3) and added environmental sound (e.g., people chatting, birds twittering). We did this to improve the fairness of the comparison between *RW ATM* and *VR ATM* and to immerse participants into a more vivid environment in *VR ATM Public*.

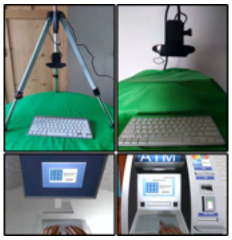

Figure 4: For each setup, we had a physical keyboard, a greenscreen, and a camera to blend the keyboard and users‘ hands into virtuality, similar to McGill et al. [55] and Oculus‘ Passthrough API [23].

4 Results

4.1 Authentication Time (in seconds)

We measured participants‘ authentication time from the first character entry until the last character entry. This depicts the overall authentication time that has also been reported in the original ColorPIN study [20], and it is a common approach when evaluating authentication systems and their usability [22, 40, 53]. We only consider successful ColorPIN entries for the analysis. There was a significant main effect of environment $(F(1,49) = 27.00, p < 0.05, \eta_p^2 = 0.36)$ on users' authentication time. Authentications were significantly faster in the real world than in VR ($p < 0.05$), with *RW Lab* (M=13.28,SD=7.76,Md=10.04) being significantly faster than *VR Lab* (M=20.89,SD=8.33,Md=20.50), and *RW ATM* (M=16.57,SD=14.01,Md=12.36) being significantly faster than *VR ATM* (M=23.85,SD=25.32,Md=16.45). There was no main effect of context and no interaction effect ($p > 0.05$). When comparing *VR ATM Public* (M=25.55,SD=13.73,Md=22.57) to *VR Lab* and *VR ATM*, there was a significant effect of context $(F(2,33) = 3.676, p < 0.05, \eta_p^2 = 0.18)$. Post-hoc pairwise comparisons did not confirm these significant differences ($p > 0.05$). Despite the absence of significance, we noticed an increase of the mean authentication times in both environments: from *RW Lab* to *RW ATM* (+24.71%), from *VR Lab* to *VR ATM* (+14.17%), and from *VR Lab* to *VR ATM Public* (+22.31%). Results are visualised in Fig. 5. Authentication times in *RW Lab* are roughly the same as reported in the original ColorPIN paper (M=13.28,SD=7.76 vs M=13.33,SD=1.74) [20].

4.2 Error Rate (Corrections and Incorrect Entries)

We distinguish between *corrections*, the number of corrections before submitting a ColorPIN, and *errors*, the number of incorrect ColorPIN entries. Note that there was a maximum of three tries to authenticate correctly. There was no main effect of the environment, the context, and no interaction effect on participants' number of corrections. Appendix D in our supplementary material shows the F-ratios, together with effect sizes, means, and standard deviations (stdevs). There was also no significant effect of the context between *VR Lab*, *VR ATM*, and *VR ATM Public* on participants' number of corrections. Corrections were lowest in *RW Lab* with no corrections at all, followed by *VR ATM Public* (M=0.20,SD=0.68), *VR ATM* (M=0.30,SD=0.90), *VR Lab* (M=0.40,SD=0.73), and *RW ATM* (M=0.45,SD=1.07). There is also no evidence that the number of incorrect entries differs significantly between the conditions. The values are *RW Lab* (M=0.60,SD=1.11), *RW ATM* (M=0.65,SD=1.07), *VR Lab* (M=0.55,SD=0.92), *VR ATM* (M=0.40,SD=0.92), and *VR ATM Public* (M=0.75,SD=0.99). Results are visualised in Fig. 5.

Incorrect Cash Withdrawals. We collected participants‘ cash recall performance (i.e., to what extent they could recall the amount

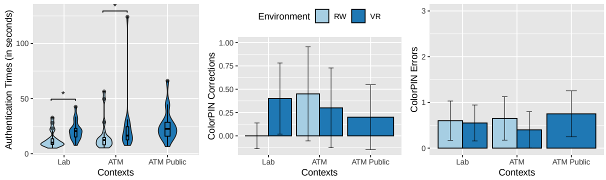

Figure 5: Authentications were significantly faster in reality than in VR. Authentications were slower in both environments when they were performed in advance of a production task (i.e., withdrawing cash on an ATM). There is no evidence that the number of ColorPIN corrections and errors differ significantly between the conditions. Error bars denote adjusted 95% CIs [57].

of cash they were supposed to withdraw). We ran a Cochran‘s Q test to investigate if the differences between our five conditions are significant. Participants‘ primary task performance, i.e. correctly recalling the amount of cash they had to withdraw, was not statistically significant between the conditions *($\chi^2(4) = 2.194, p = 0.70$)*. There were two participants in *RW Lab*, four in *RW ATM*, and five each in *VR Lab*, *VR ATM*, and *VR ATM Public* who were not able to correctly recall the amount of cash they were supposed to withdraw.

4.3 Sense of Presence and Perceived Workload

Table 1 and Table 2 provide an overview of all means, stdevs, and statistical analyses of the IPQ and NASA-TLX values, featuring the subscales 1) sense of being there (PRES), 2) spatial presence (SP), 3) involvement (INV), 4) experienced realism (REAL), and 5) the raw NASA-TLX scores. Participants‘ sense of presence was significantly higher in *RW ATM*, *VR ATM*, and *VR ATM Public* than *RW Lab* and *VR Lab* ($p < 0.05$), and significantly higher in *VR ATM Public* than in *VR Lab* ($p < 0.05$). A more nuanced analysis on the level of each subscale is reported in Table 1 and Table 2, including means, stdevs, and our statistical analysis, which followed the approach described in Sect. 3.5. There were statistically significant main effects in all IPQ‘s subscales, with participants‘ sense of being part of an ATM authentication scenario, spatial presence, and realism being statistically significantly higher in the in situ ATM experiences than in *RW Lab* and *VR Lab*.

For the NASA-TLX values, there is no evidence that participants‘ perceived workload differed significantly between the conditions ($p > 0.05$). Fig. 6, Table 1, and Table 2 provide an overview of the values and our statistical analysis. Appendix C in our supplementary material shows all NASA-TLX subdimensions.

4.4 Structured Interview

We were interested in learning which part of the task participants perceived as their primary task and the differences between their user study experience and ATM interactions in the wild. The questionnaire is available in our supplementary material (Appendix A).

4.4.1 Primary and Secondary Task Perception

When isolating ColorPIN from an actual production task (*RW Lab* and *VR Lab*), we noticed that a notable amount of participants perceived entering the correct ColorPIN as their primary task (*RW Lab*: n=12; *VR Lab*: n=17) rather than withdrawing cash at an ATM. When ColorPIN was part of an overall production task where participants had to perform a task before and after the actual authentication (i.e., *RW ATM*, *VR ATM*, *VR ATM Public*), they mentioned less often that they perceived entering the correct ColorPIN as their primary task (*RW ATM*: n=9; *VR ATM*: n=8; *VR ATM Public*: n=8). This means that while our three ATM conditions did a slightly better job in providing participants with a realistic authentication context compared to *RW Lab* and *VR Lab*, we were still not capable of providing them with a fully realistic authentication experience. Some participants still perceived the authentication as their primary task, which is usually not the case in the real world [45, 68].

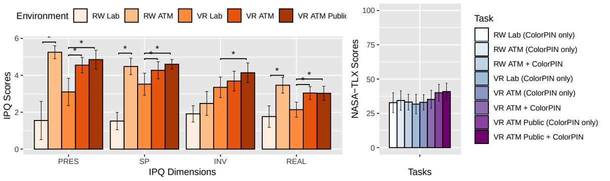

Figure 6: Participants‘ sense of presence was significantly higher in *RW ATM*, *VR ATM*, and *VR ATM Public* compared to *RW Lab* and *VR Lab*. The mean raw NASA-TLX values did not differ significantly between the conditions. Note that we requested participants‘ perceived workload two times for each ATM condition: 1× for their overall ATM interaction experience and 1× for their ColorPIN authentication.

4.4.2 Perceived Differences to ATM Interaction in the Wild

When asked about participants‘ perceived differences compared to a real-world ATM withdrawal there was one comment that appeared frequently across all conditions. Participants mentioned that ColorPIN notably differs from the authentication scheme they are familiar with (i.e., traditional 4-digit PIN authentication, n=9 for *RW Lab*, n=15 for *RW ATM*, n=12 for *VR Lab*, n=14 for *VR ATM*, n=10 for *VR ATM Public*). This is an interesting finding because it implies that at the point where researchers study novel real-world authentication schemes it is challenging to present participants with a highly realistic scenario. We discuss this further in Sect. 5.2.

In *RW Lab* and *VR Lab* participants mentioned that they were sitting in front of a PC (n=8 for *RW Lab*, n=4 for *VR Lab*) and that this leads to a different experience than being part of an ATM interaction scenario (n=10 for *RW Lab*, n=12 for *VR Lab*). About half of our participants (n=11) mentioned that in *RW ATM* the fidelity and location of the ATM deviated from an ATM withdrawal scenario in the wild (n=7 for *VR ATM*, n=2 for *VR ATM Public*). There were some participants who mentioned that using VR (e.g., putting on the headset) is something they would not do in the real world (n=6 for *VR Lab*, n=5 for *VR ATM*, n=4 for *VR ATM Public*) and n=7 mentioned that they would usually take precautions when they see other people next to them, which they did not do in our study.

4.4.3 5-Point Likert Scales

We asked participants on 5-Point Likert scales (1=strongly disagree, 5=strongly agree) a) if they felt being part of a laboratory study during the authentication, b) if they were aware of the experimenter, c) if the experimenter‘s presence impacted their performance and behaviour, and d) if recalling the PIN made it more challenging to complete the other cash withdrawal steps, and vice versa. A Friedman test with post-hoc Wilcoxon signed-rank tests resulted in a significant difference between the conditions and participants‘ feeling of being part of a laboratory study *($\chi^2(4) = 12.670, p < .05$)*. Participants' feeling of being part of a laboratory study was statistically significantly less in *RW ATM* (Z = -2.375, p = 0.018), *VR ATM (Z = -2.484, p = 0.013)*, and *VR ATM Public* (Z = -2.365, p = 0.018) compared to *RW Lab*. No other comparisons were statistically significant. Table 1 shows the means and stdevs.

4.5 Perceived Realism

Participants were asked to rank the different conditions based on their perceived realism (1=best, 5=worst). Raw scores were multiplied by a weight factor (×5 for rank 1, ×4 for rank 2, etc.) and then summed up to compute weighted scores (based on [73]). *RW ATM* achieved the highest score (85) with *VR ATM Public* (80) on rank two and *VR ATM* on rank three (70). Both baseline conditions were perceived as the least realistic ATM contexts with *RW Lab* being perceived as slightly more realistic (37) than *VR Lab* (31). The ranking tells us

Table 1: The table shows the means and the stdevs of the IPQ scores, the NASA-TLX scores, and the participants‘ responses on the 5-point Likert scale questions (1=strongly disagree, 5=strongly agree). Statistical analysis follows the description in Sect. 3.5. $p < 0.05$ highlighted. The $p < 0.05$ column shows pairwise comparisons against our two baselines (*RW Lab* and *VR Lab*).

IPQ Scores (Two-way RM ANOVA)	**(1)** *RW Lab*	**(2)** *RW ATM*	**(3)** *VR Lab*	**(4)** *VR ATM*	Context (Lab/ATM)	Environment (RW/VR)	Context×Environment	p<0.05
Sense of being there (PRES)	1.55 (2.16)	5.25 (0.70)	3.1 (1.04)	4.55 (1.02)	F(1,57) = 47.89, p < 0.05, $\eta_p^2 = 0.46$	F(1,57) = 5.323, p < 0.05, $\eta_p^2 = 0.09$	F(1,57) = 18.03, p < 0.05, $\eta_p^2 = 0.24$	1-2;3-4
Spatial Presence (SP)	1.52 (1.49)	4.48 (1.72)	3.52 (1.78)	4.27 (1.46)	F(1,57) = 56.07, p < 0.05, $\eta_p^2 = 0.50$	F(1,57) = 13.11, p < 0.05, $\eta_p^2 = 0.19$	F(1,57) = 20.11, p < 0.05, $\eta_p^2 = 0.26$	1-2
Involvement (INV)	1.91 (1.91)	2.48 (1.97)	3.35 (1.70)	3.69 (1.72)	F(1,57) = 3.05, p = 0.09, $\eta_p^2 = 0.05$	F(1,57) = 40.80, p < 0.05, $\eta_p^2 = 0.42$	F(1,57) = 0.19, p = 0.66, $\eta_p^2 = 0.003$	n/a
Realism (REAL)	1.76 (1.96)	3.46 (2.04)	2.14 (1.57)	3.04 (1.71)	F(1,57) = 45.058, p < 0.05, $\eta_p^2 = 0.44$	F(1,57) = 0.0017, p = 0.966, $\eta_p^2 = 0.00003$	F(1,57) = 6.830, p < 0.05, $\eta_p^2 = 0.11$	1-2;3-4
Overall Presence Score	1.70 (1.82)	3.67 (2.06)	3.05 (1.80)	3.77 (1.67)	F(1,57) = 69.403, p < 0.05, $\eta_p^2 = 0.55$	F(1,57) = 18.151, p < 0.05, $\eta_p^2 = 0.24$	F(1,57) = 18.147, p < 0.05, $\eta_p^2 = 0.24$	1-2;3-4
NASA-TLX Scores (Two-way RM ANOVA)	**(1)** *RW Lab*	**(2)** *RW ATM*	**(3)** *VR Lab*	**(4)** *VR ATM*	Context (Lab/ATM)	Environment (RW/VR)	Context×Environment	p<0.05
ColorPIN only	31.79 (30.48)	34.33 (32.03)	31.71 (29.49)	33.04 (30.91)	F(1,57) = 0.491, p = 0.49, $\eta_p^2 = 0.009$	F(1,57) = 0.803, p = 0.37, $\eta_p^2 = 0.014$	F(1,57) = 0.0002, p = 0.99, $\eta_p^2 = 0.0000033$	n/a
ATM + ColorPIN	n/a	33.21 (28.60)	n/a	35.17 (30.29)	n/a	F(1,19) = 0.33, p = 0.57, $\eta_p^2 = 0.017$	n/a	n/a
5-Point Likert Scale Questions	**(1)** *RW Lab*	**(2)** *RW ATM*	**(3)** *VR Lab*	**(4)** *VR ATM*	**(5)** *VR ATM Public*	Friedman	p<0.05	
Feeling of being part of a user study	3.70 (1.23)	2.95 (1.16)	3.35 (1.31)	3.15 (1.11)	3.05 (1.16)	$\chi^2(4) = 12.670$, p < .05	1-2;1-4;1-5	
Awareness of the experimenter	2.85 (1.31)	2.70 (1.31)	2.55 (1.36)	2.65 (1.28)	2.50 (1.43)	$\chi^2(4) = 0.874$, p = 0.928	n/a	
Impact of experimenter‘s presence on performance	1.50 (0.59)	1.60 (0.92)	1.40 (0.58)	1.45 (0.74)	1.55 (0.74)	$\chi^2(4) = 1.538$, p = 0.82	n/a	
Impact of experimenter‘s presence on behaviour	1.30 (0.46)	1.55 (0.74)	1.40 (0.58)	1.60 (0.92)	1.55 (0.74)	$\chi^2(4) = 4.155$, p = 0.385	n/a	
Impact of the secondary task on the primary	2.55 (1.24)	2.95 (1.40)	2.85 (1.35)	2.75 (1.22)	2.55 (1.24)	$\chi^2(4) = 1.957$, p = 0.744	n/a	
Impact of the primary task on the secondary	1.9 (1.18)	1.95 (1.07)	1.75 (0.89)	1.90 (0.83)	1.75 (0.77)	$\chi^2(4) = 1.784$, p = 0.775	n/a	

Table 2: Results of one-way RM ANOVA. $p < 0.05$ highlighted. The $p < 0.05$ column shows pairwise comparisons.

IPQ Scores *(One-way RM ANOVA)*	**(1)** *VR Lab*	**(2)** *VR ATM*	**(3)** *VR ATM Public*	Context(Lab/ATM/Public)	p<0.05
Sense of being there (PRES)	3.10 (1.04)	4.55 (1.02)	4.85 (1.24)	F(2,38) = 22.41, p < 0.05, $\eta_p^2 = 0.54$	1-2;1-3
Spatial Presence (SP)	3.52 (1.78)	4.27 (1.46)	4.60 (1.18)	F(2,38) = 8.880, p < 0.05, $\eta_p^2 = 0.32$	1-2;1-3
Involvement (INV)	3.35 (1.70)	3.69 (1.72)	4.14 (1.61)	F(2,38) = 3.822, p < 0.05, $\eta_p^2 = 0.17$	1-3
Realism (REAL)	2.14 (1.57)	3.04 (1.71)	3.03 (1.77)	F(2,38) = 8.71, p < 0.05, $\eta_p^2 = 0.31$	1-2;1-3
Overall Presence Score	3.05 (1.80)	3.77 (1.67)	4.04 (1.64)	F(2,38) = 19.275, p < 0.05, $\eta_p^2 = 0.50$	1-2;1-3
NASA-TLX Scores *(One-way RM ANOVA)*	**(1)** *VR Lab*	**(2)** *VR ATM*	**(3)** *VR ATM Public*	Context (Lab/ATM/Public)	p<0.05
ColorPIN only	31.71 (29.49)	33.04 (30.91)	40.04 (30.03)	F(2,38) = 2.65, p = 0.084, $\eta_p^2 = 0.12$	n/a
ATM + ColorPIN	n/a	35.17 (30.29)	40.88 (27.78)	F(1,19) = 1.48, p = 0.24, $\eta_p^2 = 0.07$	n/a

that in both the real world and in VR the replicated ATMs improved participants‘ perceived ATM authentication realism.

4.6 Semi-structured Interview

We conducted a semi-structured interview at the end of the study to capture the rich nuances of participants‘ experiences in a more qualitative way. The data has been analysed as described in Sect. 3.5 and our roughly used questionnaire can be found in Appendix B in our supplementary material.

4.6.1 Perceived Realism: Reasoning

As reported in Sect. 4.5, participants perceived *RW ATM* as most similar to an ATM withdrawal experience in the wild. P1 voiced that *“the real-world ATM was the most realistic because there was something in front of me, I could really feel the card.”* (P1). Others voiced that they perceived the real-world ATM as the most realistic one because *“you cannot get closer to that where you have the ATM 1:1 in front of you”* (P11). We also noticed discussions around the realism of an ATM scenario where other people are relatively close to the user authenticating. P19, for example, perceived *RW ATM* and *VR ATM* as more realistic than *VR ATM Public* because he is not familiar with a situation where other people are relatively close to the ATM. Others perceived *VR ATM Public* as more realistic than *VR ATM* and explained this around the fact that *”ATMs are usually at locations where much more is going on“* (P4). For both *RW Lab* and *VR Lab*, participants mentioned that they felt “like playing a game; you sit in front of a keyboard and enter a PIN” (P14) and that this does not represent an ATM withdrawal scenario: *“I never withdraw cash in front of a desktop monitor”* (P1).

4.6.2 Perceived Differences: ATM Authentication in the Wild

When asked about any differences to an ATM authentication in the wild, participants voiced that they were familiar with the actions they had to do: *“the actions I had to do were very similar to the real world; take the card, put in the card, take it out – it is the same mechanism”* (P18). However, participants frequently brought up the authentication scheme and that this is not the way in which they authenticate in the real world: *“it was quite realistic, I mean you enter a different PIN - the ColorPIN - which is different to the real world“* (P14). This was mentioned for both our staged real-world ATM scenario and for the two VR ATM scenarios. For *VR ATM* and *VR ATM Public* participants further voiced that there was a lack of haptic feedback: *“in VR all the haptics were missing, and also to identify the distance when interacting with the touch screen”* (P12). Some participants (e.g., P2, P6) mentioned that they probably need more exposure to VR and that the novelty led to a different feeling compared to an ATM withdrawal in the wild.

4.6.3 Input Behaviour: The Keyboard

About half of our participants used touch typing when authenticating using ColorPIN, independent of the environment. Some participants voiced that they usually use touch typing when providing keyboard input, but this was different in our study: *“[I] only used one finger because that is how I do it when interacting with an ATM”* (P10) and *“like I‘d type on a traditional ATM, there wasn‘t much difference”* (P6). Interestingly, P1 mentioned that she only used touch typing in *RW Lab* and *VR Lab* because these two environments provided her with a feeling of being part of a workplace rather than an ATM environment. We discuss the importance of the context and users‘ behaviour for authentication prototype designs further in Sect. 5.1.

5 Discussion

Our findings confirm the previously reported differences between in-VR and real-world authentications [52] and keyboard-based input in general [55]: user authentications took significantly longer in VR than in the real world. When conducting, for the first time, *in situ* authentication research in VR and in a staged real-world setting, we found that both RW and VR exhibit similar patterns when comparing *isolated* with *in situ* authentications. Authentication times increased by 24.71% from *RW Lab* to *RW ATM*, by 14.17% from *VR Lab* to *VR ATM*, and by 22.31% from *VR Lab* to *VR ATM Public*. This indicates that in situ evaluations can have an impact on a real-world authentication system‘s usability results (**RQ_1**). In both the real and virtual environment, participants‘ sense of presence increased significantly from the laboratory settings (*RW Lab*, *VR Lab*) to the ATM environments (*RW ATM*, *VR ATM*, *VR ATM Public*). In all ATM conditions, participants felt less being part of a user study compared to *RW Lab*, the de facto standard evaluation of novel authentication systems. This means, together with participants‘ qualitative feedback, that applying VR for (simulated) authentication research contributes towards a considerable high authentication realism (**RQ_2**) and shows that the use of VR to conduct **simulated in-the-wild evaluations** in the authentication domain achieved promising early results. There is no evidence that authentications using ColorPIN are more/less demanding in VR than in the real world (Table 1). This finding also supports the use of VR replicas for real-world authentication research and further contributes to the validation of VR for human-centred security research [52], particularly **through the lens of (simulated) in situ evaluations**.

Although we achieved promising results when comparing a staged real-world ATM scenario with a virtual replica, some typical user behaviours in the wild (e.g., shielding PIN entries [7,21]) were not present in our user study, which we discuss further in Sect. 5.2.

KEY LESSON #1

The use of **VR enables conducting (simulated) in situ real-world authentication research in private and sensitive contexts** that are often infeasible to research in the wild. Yet, it is important to acknowledge potential differences to in-the-wild observations.

5.1 There Is More to Context Than Authentication

Authentication systems that are proposed for specific contexts (e.g., public displays [20,40,42], mobile devices [13,39,41]), should, if possible, be evaluated in their intended usage scenario. Johnson [38] argued that the conventional usability laboratory is not able to adequately simulate conditions in the wild and cannot *"provide for the wide range of competing activities and demands on users that might arise in a natural setting"* [38]. Researchers need to understand how novel technology is being used to more accurately tailor the security mechanisms to users' behaviour [5]. Lab-based evaluations are necessary for early and fast iterations of prototype authentication systems, but often there are no follow-up evaluations in the wild [51]. Therefore, it remains unclear how these systems perform in their intended usage context. In our study, there were several participants who voiced that they used touch typing in *RW Lab* and *VR Lab* but not in *RW ATM*, *VR ATM*, and *VR ATM Public*, and explained this around the fact that they perceived *RW Lab* and *VR Lab* as "sitting in front of a PC at work". We encourage future work to further contribute to usability evaluations in different contexts as they can, as evidenced through our work, impact users' performance, sense of presence, and behaviour. This is particularly important because user behaviour is a key factor in security failures [65], but if authentication systems are studied in traditional lab settings only (*RW Lab*, *VR Lab*), we as a community will not be able to identify and capture the causes of undesirable user behaviour because users' behaviour in the lab might not depict their behaviour in the wild.

KEY LESSON #2

Context is a key factor when evaluating authentication schemes and can impact a system's usability evaluation results and how users interact and behave. Leveraging VR to replicate real-world space that is hard to research contributes to more realistic, affordable, and effective authentication research.

5.2 Achieving High Realism is Hard

Although our VR ATMs (i.e., *VR ATM*, *VR ATM Public*) outperformed our baseline (*VR Lab*), e.g., resulted in a higher level of presence and perceived realism (Sect. 4.3, Sect. 4.5), eliciting in-the-wild authentication behaviour using VR still remains a challenge. This is apparent in our study as follows. Participants mentioned that both the real-world ATM and the two VR ATM replicas provided them with a high level of realism and that all three setups came close to in-the-wild ATM authentication. However, we noticed that the authentication scheme - ColorPIN [20] - impacted their perceived realism and the extent to which their behaviour in our study matched with reality [21]. There was a general consensus that our lab setup did a good job in replicating an ATM scenario, but that the novelty of the authentication scheme made them realise they are a) still in a user study and b) that there is a mismatch between an ATM authentication in the wild and our simulations (Sect. 4.4.2). While it can be argued that the novelty effect in this context could be reduced by replacing ColorPIN [20] with traditional authentication, doing this would hinder researchers from drawing any conclusions on the usability of novel systems and would restrict such user studies to already deployed systems. VR can contribute to more realistic authentication research, especially in private and sensitive contexts that are otherwise challenging to study, but we learned that at the point where novel systems are introduced users are likely to not behave like they would do in the wild. For example, none of our participants shielded their PIN entry, whereas observational studies showed that about a third of ATM users shield their PIN entry [21]. This implies that **using VR for in situ authentication research cannot fully replace studies in the wild**, but as evidenced through our user study, it can advance state-of-the-art authentication research in the lab and enables researchers to study scenarios that are challenging to study in the wild. The aim of VR-based in situ studies should not be to replace field research or traditional lab studies, but to complement existing methodologies and provide researchers with an additional research approach to conduct (simulated) in situ research.

KEY LESSON #3

Staged real-world environments and VR replicas contribute to a high sense of authentication realism. However, evaluating (novel) authentication systems in a highly ecologically valid context still remains a challenge due to the nature of user studies.

6 Limitations

We decided to study ATM authentication, a context that is challenging to study in the real world [21]. Other contexts and authentication systems, for example, biometric airport systems [66], novel authentication systems for doors [56], or gaze-based ATM authentication [25] are worth investigating to exploit the full potential of VR for usable security research. Furthermore, Volkamer et al. [76] highlighted significant differences in users' ATM interaction behavior between different countries. Future work may want to run a cross-country study of a VR-based in situ research approach to compare the results with our findings and ColorPIN's original study [20]. Finally, as technology improves, more advanced VR headsets may increase participants' perceived realism when interacting with virtual replicas of real-world authentication systems. Our user study was conducted in 2021 using the Oculus Quest 2, which has to be noted and considered when aiming to replicate our findings. Increased display resolutions and larger field of views (e.g., Pimax Vision 8k) may contribute to even more realistic (virtual) study environments.

7 Conclusion

We studied the use of VR for in situ authentication research, a promising application for VR to contribute to the transition of usable and secure authentication methods into the real world [51]. We aimed to understand the extent to which VR can be used for advanced authentication research and how the evaluation context (isolated vs integrated) impacts an authentication system's usability evaluation results. We compared how ColorPIN [20] performs when evaluated in a lab environment (*RW Lab* and *VR Lab*) and as part of an ATM interaction experience (*RW ATM*, *VR ATM*, *VR ATM Public*). Our findings showed that presenting users with an ATM in the real world and a virtual replica in VR contributes to more realistic authentication research, improves participants' sense of being part of an ATM authentication scenario, and impacts a system's usability evaluation results with similar patterns in reality and VR. Our findings have implications for VR, HCI, and security researchers by providing them with a novel research approach to conduct research in contexts that are otherwise infeasbile to research in the wild. We concluded with three key lessons and hope that the use of VR for in situ authentication research finds widespread adoption, complements existing research methods, and results in advanced usable and secure authentication methods in the long run.

Acknowledgments

We thank all participants for taking part in our research, all reviewers for their valuable feedback, and Adam Jenkins for helpful discussions and proofreading. This publication was supported by the University of Edinburgh and the University of Glasgow jointly funded PhD studentships, and partially by EPSRC (EP/V008870/1) and PETRAS National Centre of Excellence (EP/S035362/1).

References

[1] Y. Abdrabou, M. Khamis, R. M. Eisa, S. Ismail, and A. Elmougy. Just gaze and wave: Exploring the use of gaze and gestures for shoulder-surfing resilient authentication. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, ETRA '19. ACM, New York, NY, USA, 2019.

[2] E. Adar, D. S. Tan, and J. Teevan. Benevolent deception in human computer interaction. In *Proceedings of Human Factors in Computing Systems*, CHI '13. ACM, New York, NY, USA, 2013.

[3] Adobe. Mixamo: Animated 3d characters, 2021.

[4] F. Alt. Out-of-the-lab research in usable security and privacy. In *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. ACM, New York, NY, USA, 2021.

[5] F. Alt and E. von Zezschwitz. Emerging trends in usable security and privacy. *Journal of Interactive Media*, 2019.

[6] T. Amano, S. Kajita, H. Yamaguchi, T. Higashino, and M. Takai. Smartphone applications testbed using virtual reality. In *Proceedings of the 15th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services*, MobiQuitous '18. ACM, New York, NY, USA, 2018.

[7] M. P. Ashby and A. Thorpe. Self-guardianship at automated teller machines. *Crime Prevention and Community Safety*, 2017.

[8] A. J. Aviv, D. Budzitowski, and R. Kuber. Is bigger better? comparing user-generated passwords on 3x3 vs. 4x4 grid sizes for android's pattern unlock. In *Proceedings of the 31st Annual Computer Security Applications Conference*, ACSAC. ACM, New York, NY, USA, 2015.

[9] A. J. Aviv and D. Fichter. Understanding visual perceptions of usability and security of android's graphical password pattern. In *Proceedings of the 30th Annual Computer Security Applications Conference*, ACSAC '14. ACM, New York, NY, USA, 2014.

[10] S. Azenkot, K. Rector, R. Ladner, and J. Wobbrock. Passchords: Secure multi-touch authentication for blind people. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '12. ACM, New York, NY, USA, 2012.

[11] A. Bianchi. Authentication on public terminals with private devices. In *Proceedings of the Fifth International Conference on Tangible, Embedded, and Embodied Interaction*. ACM, New York, NY, USA, 2010.

[12] A. Bianchi and I. Oakley. Multiplexed input to protect against casual observers. In *Proceedings of HCI Korea*. Seoul, KOR, 2014.

[13] A. Bianchi, I. Oakley, V. Kostakos, and D. S. Kwon. The Phone Lock: Audio and Haptic Shoulder-Surfing Resistant PIN Entry Methods for Mobile Devices. In *Proceedings of the Fifth International Conference on Tangible, Embedded, and Embodied Interaction*, TEI '11. ACM, New York, NY, USA, 2010.

[14] A. Bianchi, I. Oakley, and D. S. Kwon. Spinlock: A single-cue haptic and audio pin input technique for authentication. In *Haptic and Audio Interaction Design*. Springer, Berlin, Heidelberg, 2011.

[15] J. Cohen. Eta-squared and partial eta-squared in fixed factor anova designs. *Educational and Psychological Measurement*, 1973.

[16] J. Cohen. *Statistical power analysis for the behavioral sciences*. 2013.

[17] N. Dahlbäck, A. Jönsson, and L. Ahrenberg. Wizard of oz studies—why and how. *Knowledge-based systems*, 1993.

[18] J. T. Davin, A. J. Aviv, F. Wolf, and R. Kuber. Baseline measurements of shoulder surfing analysis and comparability for smartphone unlock authentication. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '17. ACM, New York, NY, USA, 2017.

[19] A. De Luca, A. Hang, E. von Zezschwitz, and H. Hussmann. I feel like i'm taking selfies all day! towards understanding biometric authentication on smartphones. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 2015.

[20] A. De Luca, K. Hertzschuch, and H. Hussmann. Colorpin: Securing pin entry through indirect input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI. ACM, New York, NY, USA, 2010.

[21] A. De Luca, M. Langheinrich, and H. Hussmann. Towards understanding atm security: A field study of real world atm use. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, SOUPS '10. ACM, New York, NY, USA, 2010.

[22] A. De Luca, E. von Zezschwitz, N. D. H. Nguyen, M.-E. Maurer, E. Rubegni, M. P. Scipioni, and M. Langheinrich. Back-of-device authentication on smartphones. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13. ACM, New York, NY, USA, 2013.

[23] O. Developers. Oculus integration sdk. accessed 04 November 2021.

[24] G. Dhandapani, J. Ferguson, and E. Freeman. Hapticlock: Eyes-free authentication for mobile devices. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. ACM, New York, NY, USA, 2021.

[25] P. Dunphy, A. Fitch, and P. Olivier. Gaze-contingent passwords at the atm. In *Communication by Gaze Interaction (COGAIN)*, 2008.

[26] S. Egelman and E. Peer. Scaling the security wall: Developing a security behavior intentions scale (sebis). In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 2015.

[27] M. Eiband, M. Khamis, E. von Zezschwitz, H. Hussmann, and F. Alt. Understanding shoulder surfing in the wild: Stories from users and observers. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 2017.

[28] L. A. Elkin, M. Kay, J. J. Higgins, and J. O. Wobbrock. An aligned rank transform procedure for multifactor contrast tests, 2021.

[29] S. Fahl, M. Harbach, Y. Acar, and M. Smith. On the ecological validity of a password study. In *Proceedings of the Ninth Symposium on Usable Privacy and Security*, SOUPS '13. ACM, New York, NY, USA, 2013.

[30] J. C. Flanagan. The critical incident technique. *Psychological bulletin*.

[31] K. N. Fong, K. Y. Chow, B. C. Chan, K. C. Lam, J. C. Lee, T. H. Li, E. W. Yan, and A. T. Wong. Usability of a virtual reality environment simulating an automated teller machine for assessing and training persons with acquired brain injury. *Journal of neuroengineering and rehabilitation*, 2010.

[32] T. Franke, C. Attig, and D. Wessel. A personal resource for technology interaction: Development and validation of the affinity for technology interaction (ati) scale. *International Journal of Human–Computer Interaction*, 2019.

[33] Free3D. Free3d: 3d atm model, 2019. https://free3d.com/3d-model/atm-57251.html, accessed 04 November 2021.

[34] J. Gugenheimer, A. De Luca, H. Hess, S. Karg, D. Wolf, and E. Rukzio. Colorsnakes: Using colored decoys to secure authentication in sensitive contexts. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '15. ACM, New York, NY, USA, 2015.

[35] S. Hart and L. Staveland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Human mental workload*.

[36] J. Henrich, S. J. Heine, and A. Norenzayan. The weirdest people in the world? *Behavioral and Brain Sciences*, 2010.

[37] L. Hodges, P. Anderson, G. Burdea, H. Hoffmann, and B. Rothbaum. Treating psychological and phsyical disorders with vr. *IEEE Computer Graphics and Applications*, 2001.

[38] P. Johnson. Usability and mobility; interactions on the move. In *Proceedings of the First Workshop on Human-Computer Interaction with Mobile Devices*, 1998. http://www.dcs.gla.ac.uk/ johnson/papers/mobile/HCIMD1.html, accessed 09 January 2022.

[39] M. Khamis, F. Alt, M. Hassib, E. von Zezschwitz, R. Hasholzner, and A. Bulling. Gazetouchpass: Multimodal authentication using gaze and touch on mobile devices. In *Proceedings of the 34th Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16. ACM, New York, NY, USA, 2016.

[40] M. Khamis, R. Hasholzner, A. Bulling, and F. Alt. Gtmopass: Two-factor authentication on public displays using gaze-touch passwords and personal mobile devices. In *Proceedings of the 6th ACM International Symposium on Pervasive Displays*, PerDis '17. ACM, New York, NY, USA, 2017.

[41] M. Khamis, M. Hassib, E. von Zezschwitz, A. Bulling, and F. Alt. Gazetouchpin: Protecting sensitive data on mobile devices using secure multimodal authentication. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI 2017. ACM, New York, NY, USA, 2017.

[42] M. Khamis, L. Trotter, V. Mäkelä, E. v. Zezschwitz, J. Le, A. Bulling, and F. Alt. Cueauth: Comparing touch, mid-air gestures, and gaze for cue-based authentication on situated displays. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, Dec. 2018.

[43] H. Khan, J. Ceci, J. Stegman, A. J. Aviv, R. Dara, and R. Kuber. Widely reused and shared, infrequently updated, and sometimes inherited: A holistic view of pin authentication in digital lives and beyond. In *Annual Computer Security Applications Conference*, ACSAC '20. ACM, New York, NY, USA, 2020.

[44] I. Kirlappos and M. A. Sasse. What usable security really means: Trusting and engaging users. In *International Conference on Human Aspects of Information Security, Privacy, and Trust*. Springer, 2014.

[45] K. Krol, J. M. Spring, S. Parkin, and M. A. Sasse. Towards robust experimental design for user studies in security and privacy. In *The LASER Workshop: Learning from Authoritative Security Experiment Results (LASER 2016)*. USENIX Association, San Jose, CA, May 2016.

[46] M.-K. Lee. Security notions and advanced method for human shoulder-surfing resistant pin-entry. *IEEE Transactions on Information Forensics and Security*, 2014.

[47] L. Little. Attitudes towards technology use in public zones: The influence of external factors on atm use. In *EA on Human Factors in Computing Systems*. ACM, New York, NY, USA, 2003.

[48] V. Mäkelä, S. R. R. Rivu, S. Alsherif, M. Khamis, C. Xiao, L. M. Borchert, A. Schmidt, and F. Alt. Virtual Field Studies: Conducting Studies on Public Displays in Virtual Reality. In *Proceedings of the 38th Annual ACM Conference on Human Factors in Computing Systems*, CHI '20. ACM, New York, NY, USA, 2020.

[49] M. Malheiros, S. Brostoff, C. Jennett, and M. A. Sasse. Would you sell your mother's data? personal data disclosure in a simulated credit card application. In *The economics of information security & privacy*. 2013.

[50] F. Mathis. [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)*, 2021.

[51] F. Mathis, K. Vaniea, and M. Khamis. Prototyping usable privacy and security systems: Insights from experts. *International Journal of Human–Computer Interaction*, 2021.

[52] F. Mathis, K. Vaniea, and M. Khamis. Replicueauth: Validating the use of a lab-based virtual reality setup for evaluating authentication systems. In *Proceedings of the 39th Annual ACM Conference on Human Factors in Computing Systems*, CHI '21. ACM, New York, NY, USA, 2021.

[53] F. Mathis, J. H. Williamson, K. Vaniea, and M. Khamis. Fast and secure authentication in virtual reality using coordinated 3d manipulation and pointing. *ACM Trans. Comput.-Hum. Interact.*, Jan. 2021.

[54] F. Mathis, X. Zhang, J. O'Hagan, D. Medeiros, P. Saeghe, M. McGill, S. Brewster, and M. Khamis. Remote xr studies: The golden future of hci research? In *Proceedings of the CHI 2021 Workshop on XR Remote Research*, 2021.

[55] M. McGill, D. Boland, R. Murray-Smith, and S. Brewster. A dose of reality: Overcoming usability challenges in vr head-mounted displays. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 2015.

[56] L. Mecke, K. Pfeuffer, S. Prange, and F. Alt. Open sesame! user perception of physical, biometric, and behavioural authentication concepts to open doors. In *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia*. ACM, New York, NY, USA, 2018.

[57] R. D. Morey et al. Confidence intervals from normalized data: A correction to cousineau (2005). 2008.

[58] A. Naiakshina, A. Danilova, C. Tiefenau, and M. Smith. Deception task design in developer password studies: Exploring a student sample. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*. USENIX Association, Baltimore, MD, Aug. 2018.

[59] NationalCash. Atm statistics, 2021. accessed 04 November 2021.

[60] I. Olade, H.-N. Liang, C. Fleming, and C. Champion. Exploring the vulnerabilities and advantages of swipe or pattern authentication in virtual reality (vr). In *Proceedings of the 2020 4th International Conference on Virtual and Augmented Reality Simulations*, ICVARS 2020. ACM, New York, NY, USA, 2020.

[61] J. M. Orr, P. R. Sackett, and C. L. Dubois. Outlier detection and treatment in i/o psychology: A survey of researcher beliefs and an empirical illustration. *Personnel Psychology*, 1991.

[62] S. Pedram, R. Skarbez, S. Palmisano, M. Farrelly, and P. Perez. Lessons learned from immersive and desktop vr training of mines rescuers. *Frontiers in Virtual Reality*, 2021.

[63] S. Prange, L. Mecke, A. Nguyen, M. Khamis, and F. Alt. Don't use fingerprint, it's raining! how people use and perceive context-aware selection of mobile authentication. In *Proceedings of the International Conference on Advanced Visual Interfaces*, AVI '20. ACM, New York, NY, USA, 2020.

[64] E. Ragan, C. Wilkes, D. A. Bowman, and T. Hollerer. Simulation of augmented reality systems in purely virtual environments. In *2009 IEEE Virtual Reality Conference*, 2009.

[65] A. M. Sasse, S. Brostoff, and D. Weirich. Transforming the 'weakest link' — a human/computer interaction approach to usable and effective security. *BT Technology Journal*, Jul 2001.

[66] M. A. Sasse. Red-eye blink, bendy shuffle, and the yuck factor: A user experience of biometric airport systems. *IEEE S &P*, 2007.

[67] M. A. Sasse and I. Flechais. Usable security: Why do we need it? how do we get it? O'Reilly, 2005.

[68] M. A. Sasse, M. Steves, K. Krol, and D. Chisnell. The great authentication fatigue – and how to overcome it. In P. L. P. Rau, ed., *Cross-Cultural Design*. Springer International Publishing, Cham, 2014.

[69] G.-L. Savino, N. Emanuel, S. Kowalzik, F. Kroll, M. C. Lange, M. Laudan, R. Leder, Z. Liang, D. Markhabayeva, M. Schmeißer, N. Schütz, C. Stellmacher, Z. Xu, K. Bub, T. Kluss, J. Maldonado, E. Kruijff, and J. Schöning. Comparing pedestrian navigation methods in virtual reality and real life. In *2019 International Conference on Multimodal Interaction*, ICMI '19. ACM, New York, NY, USA, 2019.

[70] T. Schubert, F. Friedmann, and H. Regenbrecht. The experience of presence: Factor analytic insights. *Presence: Teleoperators & Virtual Environments*, 2001.

[71] T. Seitz, F. Mathis, and H. Hussmann. The bird is the word: A usability evaluation of emojis inside text passwords. In *Proceedings of the 29th Australian Conference on Computer-Human Interaction*, OZCHI '17. ACM, New York, NY, USA, 2017.

[72] R. Skarbez, J. Gabbard, D. A. Bowman, T. Ogle, and T. Tucker. Virtual replicas of real places: Experimental investigations. *IEEE Transactions on Visualization and Computer Graphics*, 2021.

[73] W. G. Stillwell, D. A. Seaver, and W. Edwards. A comparison of weight approximation techniques in multiattribute utility decision making. *Organizational Behavior and Human Performance*, 28(1), 1981.

[74] M. Usoh, E. Catena, S. Arman, and M. Slater. Using presence questionnaires in reality. *Presence*, 2000.

[75] A. Voit, S. Mayer, V. Schwind, and N. Henze. Online, VR, AR, Lab, and In-Situ: Comparison of Research Methods to Evaluate Smart Artifacts. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA.

[76] M. Volkamer, A. Gutmann, K. Renaud, P. Gerber, and P. Mayer. Replication study: A cross-country field observation study of real world {PIN} usage at atms and in various electronic payment scenarios. In *Fourteenth Symposium on Usable Privacy and Security*, 2018.

[77] E. von Zezschwitz, A. De Luca, B. Brunkow, and H. Hussmann. Swipin: Fast and secure pin-entry on smartphones. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15. ACM, New York, NY, USA, 2015.

[78] M. Weiß, K. Angerbauer, A. Voit, M. Schwarzl, M. Sedlmair, and S. Mayer. Revisited: Comparison of empirical methods to evaluate visualizations supporting crafting and assembly purposes. *IEEE Transactions on Visualization and Computer Graphics*, 2020.

[79] S. Wiedenbeck, J. Waters, L. Sobrado, and J.-C. Birget. Design and evaluation of a shoulder-surfing resistant graphical password scheme. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '06. ACM, New York, NY, USA, 2006.

[80] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 2011.

[81] B. Xin, G. Chen, Y. Wang, G. Bai, X. Gao, J. Chu, J. Xiao, and T. Liu. The efficacy of immersive virtual reality surgical simulator training for pedicle screw placement: a randomized double-blind controlled trial. *World neurosurgery*, 2019.