2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR) | 979-8-3503-2838-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISMAR59233.2023.00096

Towards a Framework for Validating XR Prototyping for Performance Evaluations of Simulated User Experiences

Jan Hendrik Plümer*
Salzburg University of Applied Sciences

Markus Tatzgern†
Salzburg University of Applied Sciences

Abstract

Extended Reality (XR) technology has matured in recent years, leading to increased use of XR simulations for prototyping novel human-centered interfaces, approximating advanced display hardware, or exploring future user experiences, before realising them in real-world scenarios. However, the validity of utilizing XR prototyping (XRP) as a method for gathering performance data on novel user experiences is still underexplored, i.e, it is not clear if results gathered in simulations can be transferred to a real experience. To address this gap, we propose a validation framework that supports establishing equivalence of performance measures gathered with real products and simulated products and, thus, improve ecological validity of XRP. To demonstrate the utility of the framework, we conduct an exemplary validation study using a Varjo XR-3, a state-of-the-art XR head-mounted display (HMD). The study focuses on steering a small drone and comparing it to interactions with its real-world counterpart. We identify functional fidelity, i.e., functional similarity between real and simulated product, as well as simulation overhead from wearing an HMD as major confounding factors for XRP.

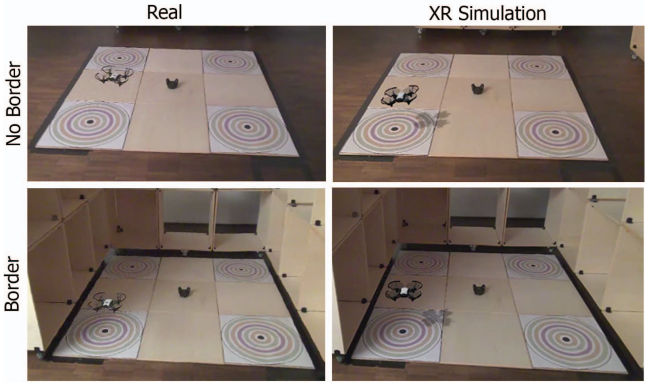

Figure 1: XR Prototyping Validation Study. The performance of a real drone was compared to a virtual drone in an XR simulation utilizing a VST HMD (Varjo XR-3) in order to explore the use of XR simulation as part of a prototyping methodology for evaluating user experiences. The validation study was designed based on a validation framework in order to guide the study design and identify confounding factors. As an additional confounding factor, shelves were introduced as bordering geometry to increase the perceived risk of users when steering the drone, which in turn should influence task performance.

1 Introduction

Extended Reality (XR) technology for creating Augmented Reality (AR), Mixed Reality (MR)[1] and Virtual Reality (VR) user experiences and applications has matured over the last years, leading to XR simulations of real-world scenarios being increasingly utilized in the research community for prototyping novel human-centered interfaces for MR [18, 25], approximating advanced display hardware that is not yet available [62], or imagining future use cases [67]. XR also has a long tradition of being utilized in various stages of the product design and product development process [5, 13]. Especially in early stages of design and development processes, XR simulations can serve as a valuable foundation to gather feedback from other designers and end users. However, only few attempts have been undertaken to explore the potential of Extended Reality Prototyping (XRP) for gathering quantitative performance data of users interacting with a product [39, 45, 53]. Providing reliable guidelines for practitioners regarding the suitability of XRP can further enhance tool sets for research and development and reduce costs for developing user experiences and products, as quantitative performance evaluations can be made earlier in the design process.

Previous efforts to validate XRP as a method for performing quantitative performance evaluations followed two general approaches. The first approach isolates specific factors that could potentially influence the outcome of performance evaluations in simulations, leading to differences between a simulation and the investigated real-world use case. Examples are the influence of latency [37], registration error [11], or visual realism [39]. The second approach follows a top-down evaluation procedure by recreating complete scenarios of a product or user experience as XR simulation and, ideally, comparing it against the corresponding real-world scenario. This top-down approach that compares against a real-world ground truth was utilized, for instance, to explore the use of VR to evaluate user behavior for public displays [41], flight deck interaction in a plane [3], or situated visualizations in MR [75]. However, while previous work demonstrated that useful qualitative feedback can be gathered with XRP [27], the ecological validity and, thus, transferability of the quantitative results from the simulation to the real-world scenario is still underexplored.

In this paper, we present a framework that aims at providing structure to the systematic validation of XRP in order to explore its potentials and limitations for evaluating products and user experiences, especially with respect to quantitative performance measurements. Our framework is applicable to various user experiences: physical products that users can directly interact with (e.g., car cockpits), completely virtual user interfaces (e.g., adaptive MR user interfaces), or a mix of both physical and virtual (e.g., MR visualizations for maintenance support). We focus on the top-down approach for validating XRP where interactions with an existing real-world artifact are compared against interactions with an XR simulation of this artifact. The optimal outcome of validation studies is that quantitative performance measures of both real-world and simulated use case are equivalent, thus, demonstrating absolute validity of the results [9,77]. Another useful result of a validation study is demonstrating relative validity, which means that statistically significant effects and order relationships between evaluated conditions in a simulation and in the real scenario are the same. Being able to demonstrate absolute or relative validity of performance measures would determine the value of XRP for early evaluations of new user experiences. After

*e-mail: jan.pluemer@fh-salzburg.ac.at
†e-mail: markus.tatzgern@fh-salzburg.ac.at

[1] We will use the term MR to describe both AR and MR.

DOI 10.1109/ISMAR59233.2023.00096

defining a framework for validation studies, we instantiate a study design from this framework and perform an exemplary validation study using state-of-the-art XR technology (Varjo XR-3). We base our use case on an indoor drone as an existing real-world product and designed a study that takes into account potential confounding factors. While steering a drone is a straightforward and apparently simple use case, the virtual XR simulation of drone behavior already presents a complex use case to explore the boundaries of XRP.

In summary, we make the following contributions:

- A first attempt at a framework for guiding the validation of XRP for user experience development that allows researchers to systematically design, analyze and discuss validation studies with the goal to explore the utilization of XRP for quantitative performance evaluations in early stages of the design process.
- An exemplary instantiation of a validation study to demonstrate the application of the framework utilizing a real-world drone as use case.
- We demonstrate the influence of confounding factors of simulation overhead and product behavior (i.e., functional fidelity), as well as a potential influence of controlling a virtual simulation compared to controlling the real product in terms of perceived risk that may influence user behavior.

2 Related Work

In the following, we describe previous work that utilized XR simulation to evaluate user experiences, but did not demonstrate ecologically validity, and previous work that performed validation studies by comparing the simulation against its real-world counterpart.

2.1 XR Simulation Studies

XR simulation has been utilized in various contexts to provide fast and cost-efficient design iterations, to study human behavior in a safe and controlled environment, or to evaluate and prototype novel systems without considering current technical limitations and before deploying them in real-world settings. For instance, product design research explored the use of XR to not only simulate the appearance of products, but also to allow usability evaluations in early design stages where detailed physical prototypes are not yet available. Barbieri et al. [5] utilized MR to overlay visual designs over a physical proxy of a household appliance so that users could interact with early designs. Bruno et al. [13, 14] simulated functional behavior of household appliance in combination with MR prototypes.

The use of XR simulation also has a long tradition in HCI research as novel concepts can be presented and evaluated without potential limitations of state-of-the-art technology [62], or the need of performing complex field studies. XR simulation also allows creating the same experimental conditions across participants. Hence, XR simulation has been utilized to prototype and evaluate advanced adaptive user interfaces for MR [18, 40], novel authentication procedures [44, 74], or MR guidance visualizations for industrial maintenance [16]. Furthermore, complex technological solutions that require extensive effort to implement, such as user-perspective rendering [6], and shared interactive spaces utilizing various technologies [31], have been efficiently prototyped and explored in simulations. Social contexts that influence user behavior, e.g., when placing virtual information in public [50, 56], have also been part of simulations, as well as the simulation of complex and potentially dangerous outdoor environments to explore MR visualizations in urban environments [70], or the use of augmentations to improve situational awareness of pedestrians [32, 43].

However, while previous work explored the use of simulation for prototyping novel interfaces and products [5, 31], or demonstrated general feasibility of novel solutions [25, 43], the question of ecological validity remains, i.e., to what degree the results gathered in XR simulations are transferrable to real-world scenarios.

2.2 XR Validation Studies

Previous work has attempted to identify the potentials and limitations of XRP to determine the degree of ecological validity of results gathered in simulations. Research has investigated the effect of various simulation parameters such as latency [11, 37, 38, 55] or registration error [11, 61], or attempted to replicate real-world scenarios in XR [20, 27, 57]. Ideally, the result of a simulation in the context of XRP is compared against the real-world scenario [71, 75] to provide conclusive evidence of the ecological validity of gathered results.

Bruno et al. [15] explored the use of XRP in a participatory product design process that involved usability testing of products. As part of their investigation, they compared interacting with a simulated product in VR against interacting with the real product and demonstrated the feasibility of XRP for usability testing. While general feedback was similar between virtual and real product, VR interaction performance was worse due to technical limitations. Recent work by Faust et al. [26] and Min et al. [53] utilized XRP to create prototypes for electronic devices and compared these prototypes against real product interactions. Similar to Bruno et al. [15], they determined that while qualitative feedback appeared to be the same, performance measures differed across real and virtual conditions. However, Faust et al. [26] demonstrated that for both virtual and real product, performance scaled with task difficulty, indicating the same relative effects between simulated and real scenarios. Furthermore, they noted that the technology appears to impact user feedback, as the MR condition was "fascinating" for users.

Voit et al. [71] explored the use of XRP for smart artifacts and compared XR against real conditions. While qualitative feedback was similar between conditions, the results of standardized questionnaires differed. Similar to Faust et al. [26], they speculated that the nature of the technology influenced results, e.g., due to the novelty effect, or technical limitations such as a small field of view of MR devices. Weiß et al. [75] performed a similar validation study in the context of situated visualization and also found similar qualitative feedback, but performance measures differed between conditions.

XRP validation studies were also performed in other contexts, and found similar results in terms of user feedback and performance differences between real and XR conditions. Examples include flight deck interaction in a plane [3], in-vehicle interaction in a car [59], human locomotion analysis [1], pedestrian crossing behavior [65], map navigation [64], interacting with public displays [41], visual search tasks in MR [39], and exploring novel authentication methods [45, 46]. Performance differences between real and XR conditions are often attributed to technical limitations such as low resolution displays leading to legibility issues when using a VR head-mounted display (HMD) [3, 59, 64], or issues with the utilized tracking methods that can lead to inaccurate user input, e.g., when using hand or eye tracking [45]. XRP validation studies may suffer from a misalignment between real and XR conditions, when user input differs between conditions (e.g., interacting with controllers instead of real hands [41]), the user's behavior is not translated correctly to the XR simulation (e.g., when a walking in place navigation is used for locomotion, but it is not adjusted to the user's actual speed in the real world [1]), or the behavior of the simulation does not correspond to the real-world counterpart (e.g., when the utilized vehicle behaves differently in both conditions [59]).

Hence, real and XR conditions were often not aligned, leading to potential confounding factors influencing the user's performance in XR. Therefore, we present a framework for validation studies for XRP that guides researchers to facilitate the identification of confounding factors to improve the ecological validity of XRP.

3 Framework for Designing XRP Validation Studies

To validate XRP, our framework for validation studies aims at creating as realistic as possible user experience simulations and scenarios and comparing them against the real-world counterpart. In the fol-

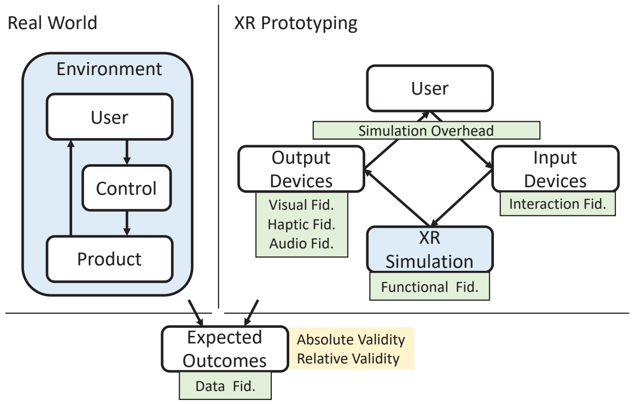

Figure 2: XRP Validation Studies. The framework structures comparing real-world user experiences against XR simulations. The simulation must represent all components of the user experience: environment, user, product control and the product itself. The framework is based on the Human-VE interaction loop [10,54] and indicates the various fidelity types and their location in the framework. Aside from fidelity differences, simulation overhead also introduces confounding factors influencing the user's performance. The goal is to determine absolute validity, where real and simulated experience lead to equivalent results, or relative validity, where effects and order relationships of results follow the same patterns.

lowing, we discuss requirements of the simulation based on various fidelity concepts in order to provide a framework for researchers and practitioners to validate and utilize XRP. We also discuss expected outcomes and their impact. Fig. 2 provides an overview of the framework. As the framework is generally applicable to various simulation scenarios, we provide an exemplary instantiation of the framework in a case study described in Section 4. Note that the presented XRP validation framework does not consider projector-based systems such as CAVEs [19] or spatial AR [8] and focuses on HMD hardware as widely available simulation technology.

3.1 Simulation Fidelity

We utilize the concept of fidelity that in various contexts describes the "realness" of a simulator component [2,48,49,54,63]. We follow Muender et al. [54] and utilize the Human-VE interaction loop concept proposed by Bowman and McMahan [10] to distinguish between necessary system and fidelity components for XRP validation studies. We focus on visual fidelity, audio fidelity, haptic fidelity as well as interaction fidelity as the basis of the discussion. We also add functional fidelity [2] to represent the simulated behavior of the product itself. We also integrate simulation overhead that potentially influences user performance of XR simulations [41,71,75].

We identify four basic components to set up a simulation environment for XRP validation studies: (1) the **product** representing the investigated artifact, (2) the product's **environment**, (3) the product's **controls**, and (4) a **user** representation. For a complete VR simulation, for all these components, digital representations must be created. Hence, ideally, the real-world product, controls and environment are digitized to create scenarios that are directly comparable to the real-world counterpart. However, due to technological limitations, instead of creating complete virtual representations, components may also be realized partially or fully utilizing real-world proxies. For instance, instead of putting efforts into recreating a highly realistic VR environment, MR technology could be utilized instead so that the virtual product can be overlaid over the available real-world environment representing its target scenario.

Visual Fidelity. We discuss visual fidelity based on the utilized XR simulation technology: using VR HMDs, a complete digital representation of the real-world scenario must be created, while with MR HMDs a mix of real and virtual content can be utilized. MR may be more feasible to achieve, as parts of the real world can be integrated into the simulation. However, while controls, environment and user may be (partially) real, the product itself is ideally purely virtual as it is not supposed to exist yet.

For a VR simulation, real-world products and target environments need to be digitized, using a combination of photogrammetric reconstruction and manual post-processing, or manual recreations. If working with companies, researchers can potentially utilize available digital 3D versions of products. In the future, NERFs [52] could be a feasible solution to recreate photorealistic digital versions of products and environments. In a VR simulation, also the user needs a digital representation as previous work has demonstrated that the user's perception in VR changes with the embodied virtual avatar [72]. Product controls also need a digital representation. Controls can take many shapes, such as physical buttons integrated into products [3,69], touch displays [41], or additional handheld devices such as tablets [4,31]. While it may be straightforward to reproduce controls in a simulation, the simulation hardware must ensure that users can read and understand the controls. Until recently, MR and VR HMDs did not offer a sufficiently high resolution to faithfully reproduce legible controls [3,57]. With the advent of retina resolution displays (e.g., Varjo XR-3), legibility of virtual content such as text on control panels significantly improved, avoiding potential perceptual issues that negatively impact XRP performance.

When relying on an MR simulation, the real target environment could be utilized for validation studies, i.e., no visual digital representation is necessary. However, when using MR, parts of the environment still need to be digitized to create a visually coherent integration of the virtual product simulation. For instance, the lighting situation should be considered for correct shading of virtual content, and shadow generation [42] as these are important depth and shape cues. Furthermore, in MR the real-world requires a digital representation so that occlusions between real and virtual parts of the scene can be modelled correctly to provide another important depth cue [12,33], and the product and the environment can physically interact. The MR simulation has to ensure that the user has a digital representation that creates correct occlusions with the virtual product for depth cues to work correctly [33]. Visual artifacts (e.g., incorrect borders of user's body) may negatively impact XRP performance.

Haptic Fidelity. To reduce confounding factors, the haptic feedback ideally resembles the anticipated real haptic feedback. The perfect haptic feedback system would encompass an exoskeleton to provide precise force-feedback for all body parts of the user so that all kinds of input methods can be simulated. A more feasible approach could model haptic feedback specifically for the input methods of a product, e.g., by recreating control boards resembling the real counterpart [47]. Furthermore, existing input methods (e.g., tablets) can be directly integrated into the simulation to provide the correct haptic feedback [4,45]. Haptic feedback can also be more unspecific and consist of vibrotactile feedback, or passive feedback from haptic proxies [28] that approximate the real shape of the product. For a more detailed discussion on haptic fidelity, we refer to Muender et al. [54]. For initiating research on validation studies, we argue that the haptic feedback should closely resemble the real world use case in order to avoid confounding factors.

Audio Fidelity. Product, environment, and controls make specific sounds and provide auditory feedback that can potentially influence user performance. For instance, auditory feedback supports the detection of machine faults [36], serves as feedback for control elements [3,36], or supports the estimation of performance characteristics, such as the speed of a vehicle [30,51]. The effect of sound also has to be considered with respect to VR simulations as virtual ambient sound has been shown to influence the sense of presence of users [21], and may be essential for the investigated use case, e.g., a system to increase situational awareness in urban environments [43]. Furthermore, environmental sounds, i.e., external influences to a VR simulation, have the potential to negatively impact the users' sense of presence [68] and, thus, the simulation experience. Hence, the influence of sound should not be underestimated in designing and analyzing XRP validation studies. While current 3D engines provide sophisticated spatial sound support, reproducing appropriate sounds of real-world products can be a time-consuming task.

Interaction Fidelity. Interaction fidelity describes the resemblance of the utilized input method to the real interaction performed by a user in terms of physiological and biomechanical parameters [48, 49]. Haptic feedback is an attribute of interaction fidelity as part of biomechanical response parameters. Therefore, haptic fidelity also influences interaction fidelity. Ideally, input methods for XRP resemble the ones of the real-world product to collect representative data in validation studies. However, simulated controls requiring fine motor manipulations can be challenging to reproduce [69], as hand tracking cannot yet capture small motions reliably. Furthermore, McMahan et al. [49] observed a potential uncanny valley effect, when input methods are utilized that closely resemble the real-world interaction, but do not match the real interaction, i.e., which have an objectively high interaction fidelity, but their performance is worse than an input method with lower fidelity. Therefore, the integration of existing input methods such as tablets or controllers [4, 45] appears to be the most feasible route for validation studies for XRP.

Functional Fidelity. Functional fidelity [2] refers to the degree to which the virtual product behaves like the real-world counterpart. This means that manipulating the controls of a product leads to an outcome representing its real-world operation. Functional fidelity not only includes the behavior of the product (e.g., speed, latency), but also its potential effects on the environment (e.g., collisions). Creating virtual products with high functional fidelity can be a challenging and time-consuming task. Validation studies could also rely on existing simulation frameworks, or seek partnerships with companies that often utilize digital twins of their products for testing and prototyping. As validation studies will be performed with specific products and are potentially costly to set up, researchers should consider how results generalize to other user experiences.

Data Fidelity. In terms of data collection, the measurements of real-world use case and XRP simulation must be directly comparable. For subjective data, measurement methods are typically the same, e.g., in the form of questionnaires. However, quantitative performance measurements may deviate, because measurement methods for error rate, accuracy or task completion time (TCT) may differ between a real product and the simulated product. For instance, when task completion is registered due to a product reaching a final position, the physical product may require positional tracking capabilities, while the virtual product is already tracked due to the utilized XR technology that shows the visual virtual representation.

Simulation overhead. The utilized simulation hardware itself introduces confounding factors that can skew the results. For instance, wearing HMDs for the XR simulation adds weight to the user's head impacting ergonomics and, thus, potentially performance [34, 58]. Current technology also induces perceptual issues due to the vergence-accomodation conflict [7], influences depth perception [60], or users may be prone to simulator sickness [73]. Validation studies should make the best effort to control for these confounding factors and should acknowledge these with respect to the gathered results. However, validation studies may also show that their adverse effects on performance may be negligible when considering the overall performance of a virtual product.

3.2 Expected Outcomes

Simulator research differentiates between two outcomes of validation studies [9,77]: **absolute validity** and **relative validity**. Absolute validity means that the measurements collected in the XR simulation are equivalent to the ones of the real-world experience, which is the ideal outcome of validation studies. Relative validity means that effects found in the simulation follow the same order and direction as the effects of the real-world scenario, even though their performance is not the same. For instance, two interaction methods perform relatively the same in the XR simulation and the real scenario.

Absolute Validity. To provide evidence for successful applications of XRP, validation studies must prove statistical equivalence [35, 39] between XR and real-world conditions. However, absolute validity for objective and subjective measurements is hard to achieve as performance differences between XR simulation and the real scenario are often influenced by confounding factors introduced by simulation overhead, such as inaccurate tracking [45], or issues with legibility [3]. Furthermore, study conditions between XR and reality often deviate in other fidelity parameters such as interaction fidelity, or haptic fidelity [41]. Hence, previous work on XRP mainly found similar user feedback for XR and real scenarios [26, 75], but not for objective performance criteria.

Relative Validity. While absolute validity is hard to achieve, relative validity may be a more likely outcome of validation studies. Interestingly, previous work on validation studies for XRP has rarely considered demonstrating relative validity. One example is Savino et al. [64] who compared phone and map navigation in a real and virtual environment. Another example is Mathis et al. [45] who compared different input modalities for an authentication task. Both could not demonstrate full relative validity, potentially due to differences in fidelity levels between XR simulation and real scenario.

4 Case Study: Drone Simulation

We designed an exemplary validation study with a real user experience based on the presented XRP validation framework. We utilized the framework to identify and control for external influences, e.g., by performing the study in a small room with controlled light setup. Increasing the environmental complexity would have introduced more confounding factors that would have made it harder to interpret the gathered data of the validation study. We decided on the use case of remote controlling a small indoor quadcopter drone as a real-world product that users can interact with. We chose a drone for several reasons. First, drones are readily available and, therefore, the validation study can be reproduced by related work [29]. Second, the drone simulation is also sufficiently complex to explore the confounding factor of functional fidelity in the experiment. Third, from a user's perspective, the controls of a drone as well as its behavior are straightforward to understand and learn. Fourth, as users may perceive a flying drone as fragile, the use case enables us to explore the potential confounding factor of perceived risk of damaging a controlled product, e.g., when flying into bordering walls in the environment (Fig. 5). In the following, we provide an overview of the study design, which allows us to discuss design decisions, fidelity levels and potential confounding factors of the study based on the XRP validation framework.

4.1 Study Design

We designed a within-subject user study to explore the ability of XR simulation to replicate a real-world scenario, in our case flying a drone. We had two independent variables: fidelity and border.

Fidelity had three conditions: a reference condition where participants perceive and interact with the real-world (REAL) drone; a mediated (MED) condition, where participants perceive the real-world scenario through a video see-through (VST) HMD device; an XR condition (XRP), where users control a simulated version of the product using the same XR technology as in the MED condition. We included the MED condition to isolate the potential confounding factor of the XR simulation technology influencing users' behavior and, thus, user performance. **Border** had two conditions: a scenario

Fidelity		REAL	MED	XRP
Visual	Product	real drone	real drone mediated through VST HMD	virtual MR drone; shading and shadows based on real env.
	Control	real gamepad	real gamepad mediated through VST HMD	
	Envir.	real environment	real environment mediated through VST HMD	
	User	real user	real user mediated through VST HMD	
Haptic		only haptic feedback from real gamepad required		
Audio		real drone audio		replicated drone audio adjusting to virtual drone behaviour
Interaction		real gamepad as input method		
Functional		real drone, restricted to 3DoF, collisions with environment		advanced drone simulation, restricted to 3DoF, collisions with environment
Data	Obj.	time measurement via real drone sensors; accuracy estimated by experimenter		time measurement via virtual drone position; accuracy estimated by system and experimenter
	Subj.	same measurement method over all conditions		
Sim. Overhead		none	HMD weight, perceptual issues due to VST	

Table 1: Fidelity analysis of the three studies conditions. Alignment with the real-world scenario (REAL) is indicated in blue, alignment with the mediated real-world scenario (MED) is indicated in yellow. Fidelity differences unique to the XR simulation are indicated in red.

that contained no bordering walls (NOB) that could endanger the drone while flying, or the flight path was enclosed by bordering walls (BO) that could potentially damage the drone on collision.

Task. The task resembled a Fitt's law task, where participants had to navigate the drone to a series of target locations (Fig. 1). Participants started from a location and navigated the drone to the next location to land it there. The locations were close to the border in the BO condition to increase the perceived risk of users.

Data Collection. To support generalization of the validation study results, the case study included dependent variables that are typically utilized in performance evaluations of products. Hence, we measured TCT for each start and landing of the drone, landing error distance, task load parameters with NASA TLX questions and task difficulty with the SEQ. We also measured risk perception with a custom questionnaire, which is the only parameter that is more specific to the scenario of controlling a drone. After finishing all fidelity conditions, participants ranked the conditions based on experienced realism and perceived risk, and the experimenter performed an unstructured interview. Error was measured by calculating the offset from the target position at the end of a repetition from the center of the drone. In the XRP condition, error was measured automatically, in the MED and REAL condition, the position was measured manually by the experimenter. To facilitate measurements, the accuracy measurements are discretized into a fixed set of distances. Hence, the target location consists of concentric rings spaced at 2 cm. As backup for measurements, the images through the HMD and a top-down view from an external camera were captured as well.

Apparatus. For XRP and MED condition, we utilized the VST mode of a Varjo XR-3 HMD. The drone simulation was implemented in Unity based on existing simulation software[2] and runs on a PC with an NVIDIA GeForce RTX 3080 Ti graphics card, AMD Ryzen 7 5700X CPU and 32 GB of RAM. To improve replicability, we provide the drone simulation as open source[3]. The physical drone was a Ryze Tech Tello drone. For the accuracy measurement in MED and REAL we use a Microsoft Azure Kinect DK mounted at a height of 2.5m. Target locations for the tasks are shown in Fig. 1. A target consisted of 11 rings, each ring having 2 cm thickness. The center had a 2 cm radius. The distance between targets was 97cm.

4.2 Simulation Fidelity

While study design already describes the study in a very standardized manner, important fidelity details are missing that need to be explicitly addressed when performing a XRP validation study. We utilize the presented framework to support the documentation and reasoning process. The fidelity analysis is summarized in Table 1.

Visual Fidelity. We relied on a VST MR HMD so that we could utilize the real-world environment in the XR simulation. The high resolution VST mode of the Varjo XR-3 ensured high visual fidelity of control, environment and user. As the user did not interact directly with the drone, no geometric user representation was required to model collisions or occlusions. For the environment, an invisible virtual representation of collision geometry was required in order to detect collisions between drone and environment, as well as for receiving virtual shadows. Modelling visual occlusions between drone and environment were not relevant, as the drone would never be occluded by the environment. The visual appearance of the virtual drone corresponded to the real drone (Fig. 1). Furthermore, the environment light influenced the virtual drone in terms of shading and shadow casting. Therefore, we controlled the light setup by using a LUPO Superpanel Dual Color 60 as the single light source in a darkened room. Controlling the light setup also avoided another confounding factor, because the VST eye cameras capturing the real-world scene perform white balancing and brightness adjustments when the lighting changes. By carefully considering visual aspects of product, control, environment and user, we ensured high visual fidelity of XRP, REAL and MED. However, due to the mediation of the real world via eye cameras in the MED and XRP conditions, the visual fidelity between those conditions was different to REAL.

Haptic Fidelity. In the study, users did not interact with the virtual drone directly. Therefore, we did not need to model haptic feedback. Users interact with the remotely controlled drone via the handheld gamepad, that is the same in each study condition. Therefore, haptic fidelity between conditions is equal.

Audio Fidelity. We recreated the drone sound for the virtual drone, which changes for starting, landing and flying when changing directions. We rely on spatialized sound provided by the Unity engine to realistically move the sound source of the drone through the environment. To ensure similar behavior to the real drone, we performed pilot tests to adjust the audio output. Therefore, we achieved a high fidelity for the audio. Audio fidelity is equal between REAL and MED conditions. However, due to the mediation of sound through head-phones worn by users, audio fidelity is not equal

[2] https://assetstore.unity.com/packages/tools/physics/yue-ultimate-fpv-drone-physics-231651

[3] https://github.com/DigitalRealitiesLab/DroneSimulation

between REAL and XR, respectively MED and XRP conditions.

Interaction Fidelity. As all conditions, i.e., REAL, MED and XRP utilized the same gamepad as input method for controlling the drone, interaction fidelity was equal across all conditions.

Functional Fidelity. We based the virtual drone behavior on an advanced drone physics simulation from a Unity asset and adjusted it for this type of drone in a series of pilot studies. Pilot studies showed that the real drone also created air turbulences when approaching real-world geometry. We approximated this behavior by adding randomized forces as turbulences to the flight directions of the drone. Part of aligning real and virtual drone behavior was also fine-tuning the drone's reaction to user input. To reduce confounding factors from an imprecise modelling of the drone flight behavior, we further restricted the controls of the drone to three degrees of freedom along the main axes, i.e., rotation and, thus, unrestricted movement was not allowed. We also animated the drone's propellers depending on the speed of the drone, as well as audio output as described before in the paragraph on audio fidelity. As the real drone would automatically shut down when colliding with an obstacle too hard, we also implemented this behavior into the virtual drone. Hence, when the drone collided with the borders, the drone was shut down and moved back to the start point using an animation. Overall, we designed functional fidelity to be high between REAL and XRP, and MED and XRP. Clearly, fidelity between REAL and MED is equal.

Data Fidelity. Subjective measurements between all conditions were the same. However, not all quantitative measurements were captured in the same way. TCT measurement was started with the button press that launches the drone, measurement was stopped when the drone touched down on the ground. For the real drone, we utilized the built-in ground sensor reading. For the virtual drone, we detected landing when the drone collided with the invisible virtual ground surface representation. Therefore, start and endpoints for TCT were measured equally. Landing accuracy was measured visually by the experimenter. Furthermore, the measurement of the XRP was supported by the simulation as the drone position in the simulation space was known. Based on the previous fidelity analysis, differences in performance measurements could occur due to differences in functional fidelity, e.g., because of potential speed differences between virtual and real drone, or due to simulation overhead, e.g., because the virtual ground surface is not precisely registered to the real ground surface. Subjective feedback from participants was collected in the same way, as users had to fill out questionnaires on a PC without wearing any XR hardware.

Simulation Overhead. The XR simulation clearly introduces simulation overhead due to the VST HMD. Hence, users perceived the real world mediated through cameras that did not have the same resolution as the human eye, had a smaller field of view, were positioned differently than the user's eyes and suffered from a small latency. Furthermore, the ergonomics of the use case changed when wearing an HMD. We attempted to reduce the impact of changed ergonomics by seating the participants in the study, so that, e.g., the pull of the HMD cable is reduced. We controlled for this confound by introducing the MED condition, where users perceived the real-world use case through the HMD cameras. We expected the overhead introduced by the HMD to be a main contributing factor to performance differences. Therefore, we did not expect equivalent performance between REAL and MED. Due to the study design, eventual performance differences between MED and XRP could be attributed to other fidelity parameters of the simulation.

5 Experiment

Here, we describe the experiment based on the study design presented in Section 4.

Procedure. The Institutional Ethics Committee of the Salzburg University of Applied Sciences approved this study. Participants were recruited via public mail to a university campus. At the time of the experiment, participants signed an informed consent and filled out a demographic questionnaire. Participants were instructed that they could remove the HMD or quit the experiment at any time. The border condition was counterbalanced, i.e., half of the participants started with bordering walls (BO), the other half without bordering walls (NOB). The fidelity condition was counterbalanced using a Latin Square table. Participants started each condition with training tasks, where they practiced flying and landing of the drone. The study started once participants felt familiar with the controls. After finishing all repetitions for a task, participants filled out SEQ, TLX and perceived risk questions. After finishing all fidelity conditions, they ranked the conditions. The process was then repeated for the second border condition and ended with an unstructured interview. Participants performed 8 repetitions of the task. Therefore, we collected 8 (repetitions) x 2 (border) x 3 (fidelity) = 48 data points per participant and, overall, 1152 data points for all participants.

Participants. 24 participants volunteered (age=28.7 (6.45), female=8). Participants covered the age range between 19 and 42 years. All participants had normal or corrected to normal vision. On a scale from one to seven (best), the mean self-rated drone flying experience was 1.96 (sd=1.68, median=1), the mean self-rated AR experience was 3.83 (sd=2.1, median=3.5), the mean self-rated HMD experience was 3.79 (sd=2.34, median=3.5).

Hypotheses. H1. We expected performance differences between the MED and REAL condition, as the simulation technology influences the participants' perception of the real world. **H2.** We expected equivalent performance between MED and XRP for the drone in the NOB condition as participants were equally impacted by the HMD during interaction and the functional fidelity was sufficiently high. **H3.** We expected performance differences between NOB and BO conditions for each fidelity condition due to the increased perceived risk of crashing the drone. **H4.** We expected performance differences between MED and XRP in the BO condition, as participants would potentially be more cautious when interacting with a real-world drone than with a simulated drone [66].

Results. The statistics software *R* was used, data was evaluated with a significance level of 0.05. The data residuals did not fulfill the normality requirement. Therefore, we utilized align-and-rank transform (ART) [76] tests and follow-up ART contrasts [22] for post-hoc analysis. The reported p-values are Bonferroni-Holm corrected. Equivalence testing was performed using TOST [35] with a conservative small effect size boundary of 0.3 due to a lack of comparable studies. For each fidelity and border condition, we calculated the mean over all task conditions for each participant. Descriptive statistics are summarized as box plots in Fig. 3 and Fig. 4, as well as the supplemental material. Statistically significant differences between border and fidelity conditions are presented in Table 2. All equivalence tests were not statistically significant. Fig. 5 shows error plots and statistically significant differences of interaction effects. In the following, the results of testing interaction effects are reported.

ART revealed a significant difference in **TCT** (F(2,115)=3.1,p=0.051), contrasts for NOB between MED and XRP (t(115)=3.8,p=0.002,d=0.5), for BO between REAL and XRP (t(115)=5.5,p<.001,d=0.4) and MED and XRP (t(115)=7.1,p<.001,d=0.4), and between XRP conditions themselves (t(115)=3.2,p=0.010,d=0.7). ART revealed a significant difference in task **success** (F(2,155)=4.1,p=0.019), contrasts for XRP conditions themselves (t(115)=3.5,p=0.006,d=0.6). ART revealed a significant difference in **commitment** (F(2,155)=4.2,p=0.018), contrasts for NOB between REAL and MED (t(115)=2.9,p=0.028,d=0.02) and between XRP and MED (t(115)=2.7,p=0.046,d=0.4), for BO between REAL and XRP (t(115)=3.4,p=0.008,d=0.5), and between the XRP themselves (t(115)=4.8,p<.001,d=0.7). ART revealed a significant difference in **stress** (F(2,155)=3.7,p=0.029), contrasts for NOB between REAL and MED (t(115)=2.9,p=0.031,d=0.08),

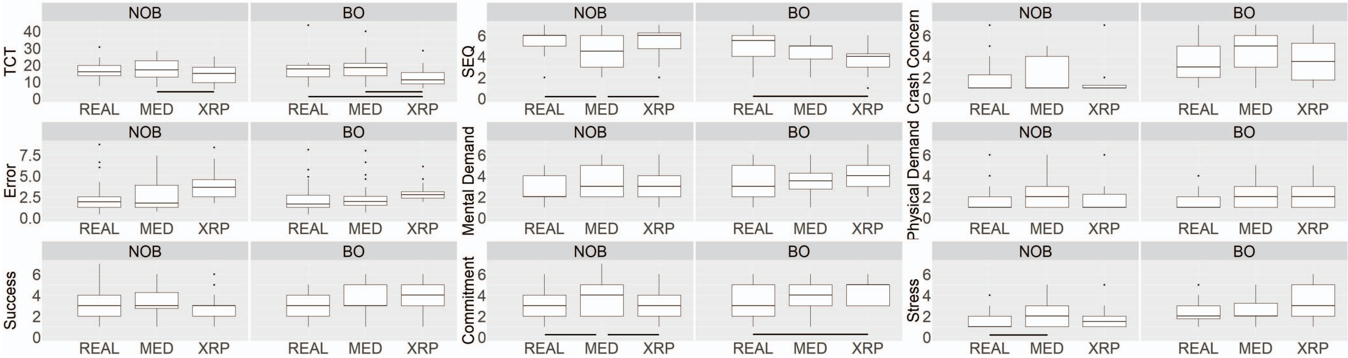

Figure 3: Box plots for each fidelity and border combination. Lines indicate statistically significant differences.

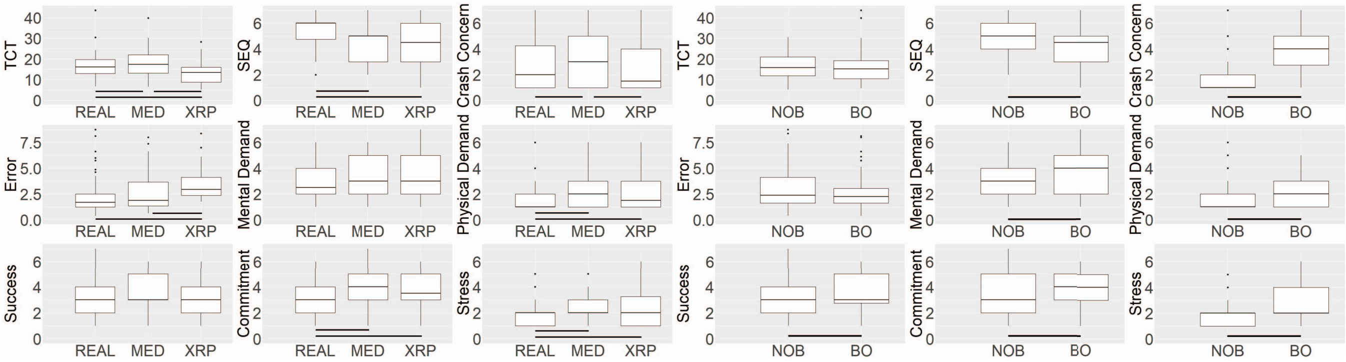

Figure 4: Box plots for fidelity and border conditions. Lines indicate statistically significant differences.

	Fidelity	REAL vs MED	REAL vs XRP	MED vs XRP	Border
TCT	F(2,115)=28.9,p<.001	t(115)=2.2,p=.030,d=0.5	t(115)=5.2,p<.001,d=1.1	t(115)=7.4,p<.001,d=1.5	-
Error	F(2,115)=22.9,p<.001	-	t(115)=6.5,p<.001,d=1.3	t(115)=5.0,p<.001,d=1.0	-
Mental Dem.	F(2,115)=3.4,p=.035	-	-	-	t(115)=3.1,p=.003,d=0.5
Physical Dem.	F(2,115)=6.3,p=.003	t(115)=3.1,p=.006,d=0.6	t(115)=3.1,p=.008,d=0.6	-	t(115)=2.5,p=.013,d=0.4
Task Pace	-	-	-	-	-
Success	-	-	-	-	t(115)=2.2,p=.033,d=0.4
Commitment	F(2,115)=5.7,p=.005	t(115)=3.2,p=.005,d=0.7	t(115)=2.5,p=.030,d=0.5	-	t(115)=4.8,p<.001,d=0.8
Stress	F(2,115)=5.8,p=.004	t(115)=3.1,p=.007,d=0.6	t(115)=2.8,p=.013,d=0.6	-	t(115)=4.7,p<.001,d=0.8
SEQ	F(2,115)=10.8,p<.001	t(115)=4.3,p<.001,d=0.9	t(115)=3.7,p<.001,d=0.7	-	t(115)=4.6,p<.001,d=0.8
Crash	F(2,155)=4.3,p=.016	t(155)=2.4,p=.032,d=0.5	-	t(155)=2.6,p=.031,d=0.5	t(115)=8.4,p<.001,d=1.4

Table 2: Inferential statistics for fidelity and border conditions. Results of ART and ART contrasts.

and overall between XRP conditions (t(115)=4.7,p<.001,d=0.6). ART revealed a significant difference in task difficulty (**SEQ**) (F(2,155)=8.1,p<.001), contrasts for NOB between REAL and MED (t(115)=3.9,p=0.001,d=0.3), and between MED and XRP (t(115)=3.3,p=0.007,d=0.7), for BO between REAL and XRP (t(115)=4.5,p<.001,d=0.9), and overall between XRP conditions (t(115)=5.8,p<.001,d=1.1).

6 Discussion

Our validation study design had the goal to identify confounding factors of the simulation overhead, in this case an XR HMD. Furthermore, the influence of controlling a real-world artifact compared to the simulated approximation was explored.

6.1 Hypotheses

H1. In terms of TCT, MED did perform worse than REAL. Furthermore, subjective feedback regarding task difficulty, physical demand, commitment, stress and crash concern were significantly higher for MED. 38% of participants mentioned issues of depth perception when using the VST HMD in MED, 21% general perceptual issues using VST HMD in MED. Hence, as all other parameters (Table 1) between MED and REAL are the same, wearing the HMD introduces a clear confounding factor. We accept hypothesis H1.

H2. MED and XRP did not show equivalent performance. Despite our best efforts, the simulated drone was still showing differences compared to the real drone. Participants mentioned that the virtual drone was more responsive (21%) and the simulated turbulences not realistic enough (21%). However, participants noted that the simulated drone was generally well implemented, and only small tweaks were required to achieve a better approximation. Participants rated the mean realism compared to the real drone high with 5.04 (sd=1, md=5) on a 7-point scale. While participants mainly commented on functional fidelity differences, note that visual and audio fidelity differences could also be confounding factors that influenced performance. Overall, we reject H2.

H3. While the borders did not significantly affect TCT and error,

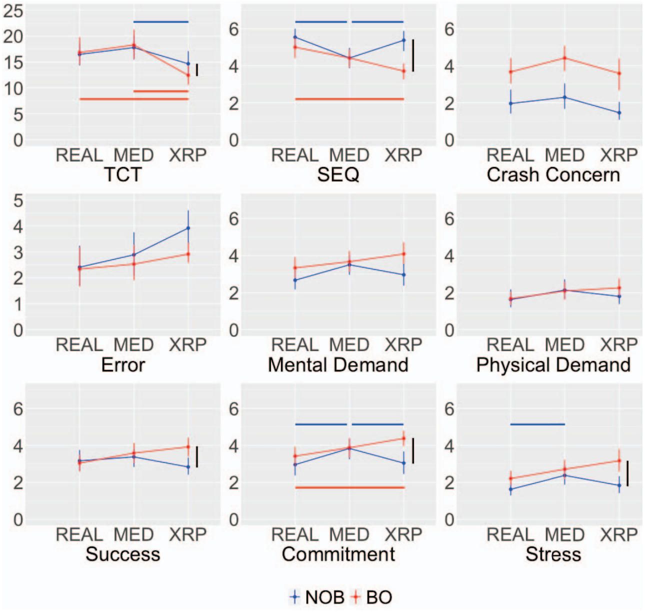

Figure 5: Error plots of interaction effects between fidelity and border. Lines indicate statistically significant differences.

users perceived the task with borders as being significantly more difficult, leading to higher mental demand, higher physical demand, a lower sense of success, higher required commitment, more stress, and a higher concern for crashing the drone. We partially accept H3, as BO lead to subjective performance differences in the conditions.

H4. MED was significantly slower than XRP in the border condition BO indicating that participants were more careful steering the real drone than the virtual drone. However, due to the virtual drone behaving differently than the real one in terms of speed and responsiveness, this result must be interpreted carefully. Furthermore, while the crash concern for the real drone was overall significantly higher for MED than for XRP, this effect was not present when looking only at the border condition BO (Fig. 5). This may indicate that participants also cared for the virtual drone in XRP and avoided crashes. However, this result is better explained with the conservative handling of crashes in the XRP condition, where small collisions more frequently lead to a crash of the drone, while the real drone did not crash as often. Hence, the concern of crashing the drone was potentially inflated in XRP due to this conservative behavior. A follow-up study should take care to improve functional fidelity of the drone behavior, i.e., avoiding overly conservative handling of crashes. However, based on the current data, we reject H4.

6.2 Learnings

With the drone use case, we explored a comparably simple user experience and attempted to recreate it in an XR simulation. While we could not demonstrate equivalence between the simulation and the real scenario, we gained valuable insights to inform the design of future validation studies due to the utilized validation framework.

Simulation Overhead. Unsurprisingly, wearing an HMD influences users and their performance in XRP validation studies. As depth perception was identified as a major issue by participants, it is likely better to utilize a VST HMD that support rendering the camera viewpoint from the user's eye perspective [17, 23, 24], which may improve depth perception. To navigate this issue, validation studies may also focus on evaluating user experiences that require the same or similar display hardware to make this confound part of the experience design, e.g., when evaluating novel adaptive MR interfaces [18, 25] that require wearing HMDs in any case. Alternatively, a validation study using HMDs should focus on proving relative validity instead of absolute validity of results.

Functional Fidelity. It is challenging to achieve a completely realistic approximation of a user experience, even though the anticipated simulation may appear straightforward to realize. For example, the drone use case appeared to be rather simple, but turned out more complex than anticipated. If recreation of the exact same behavior is too complex, and, thus, functional fidelity in a user experience hard to achieve, a validation study may focus on demonstrating relative validity instead of absolute validity.

Perceived Risk. The sensory perception of a virtual environment can influence the user's behavior in a simulation [66]. Our case study had the goal to determine if there is a difference in perceived risk of crashing a real or a virtual drone. While we found significant differences between MED and XRP, we could not conclusively relate the difference to the introduced borders and fidelity levels of the virtual and real drone. However, user feedback suggests that the perceived risk was higher for the real drone due to the mediation through the VST HMD, indicating insecurities when perceiving an environment mediated via VST technology. This confounding factor should be considered when designing validation studies. Furthermore, the rather high crash concern in XRP may mainly be an effect of issues of functional fidelity and not necessarily because participants were concerned about the safety of the virtual drone. This is backed up by additional user feedback, as users were frustrated due to the conservative crash handling and tried to avoid the borders in the BO condition when controlling the virtual drone. We will improve functional fidelity and investigate this issue further.

Path to Validation. As user studies generally require a lot of effort, performing exclusive XRP validation studies may often be infeasible. A more feasible validation approach may be to combine validation studies with the development of novel user experiences, i.e., when XR simulations are utilized for prototyping and evaluating a new user experience [18, 45], before realizing the final design as a real user experience. The created XR simulation can then be directly compared to the final real user experience, e.g., by demonstrating relative validity of variations of the experience. The validation framework presented in this paper can provide guidance in the design of the respective validation study.

7 Conclusion and Future Work

In this paper, we presented a framework for validation studies to systematically explore the use of XR simulations to prototype novel user experiences. The framework allows identifying and discussing potential confounding factors in a study design, as well as systematically aligning study conditions with each other. We provide a simple drone simulation as a first case study for applying the framework, and demonstrate the difficulty to achieve absolute validity of results gathered in a simulation compared to a real-world scenario. While functional fidelity of the simulated product was a main issue, another major confounding factor is the overhead of the XR simulation technology itself. In our case, participants remarked on issues of depth perception, likely due to the offset between the users' eyes and the VST cameras. A follow-up study may explore, if eye-perspective rendering provides a solution for this type of technology [17, 23, 24]. Overall, as a validation study can likely contain multiple confounding factors making absolute validity potentially infeasible to prove, studies may focus on demonstrating relative validity of results.

Acknowledgments

This work was supported by medPhoton GmbH and a grant from the Austrian Research Promotion Agency (grant no. 891105).

REFERENCES

[1] P. Agethen, V. S. Sekar, F. Gaisbauer, T. Pfeiffer, M. Otto, and E. Rukzio. Behavior analysis of human locomotion in the real world and virtual reality for the manufacturing industry. *ACM Trans. Appl. Percept.*, 15(3), jul 2018. doi: 10.1145/3230648

[2] A. L. Alexander, T. Brunyé, J. Sidman, S. A. Weil, et al. From gaming to training: A review of studies on fidelity, immersion, presence, and buy-in and their effects on transfer in pc-based simulations and games. *DARWARS Training Impact Group*, 5:1–14, 2005.

[3] S. Auer, J. Gerken, H. Reiterer, and H.-C. Jetter. Comparison between virtual reality and physical flight simulators for cockpit familiarization. In *Proceedings of Mensch Und Computer 2021*, p. 378–392, 2021. doi: 10.1145/3473856.3473860

[4] H. Bai, L. Zhang, J. Yang, and M. Billinghurst. Bringing full-featured mobile phone interaction into virtual reality. *Computers & Graphics*, 97:42–53, 2021.

[5] L. Barbieri, A. Angilica, F. Bruno, and M. Muzzupappa. Mixed prototyping with configurable physical archetype for usability evaluation of product interfaces. *Computers in Industry*, 64(3):310–323, 2013. doi: 10.1016/j.compind.2012.11.010

[6] D. Baričević, C. Lee, M. Turk, T. Höllerer, and D. A. Bowman. A hand-held ar magic lens with user-perspective rendering. In *IEEE ISMAR*, pp. 197–206, 2012. doi: 10.1109/ISMAR.2012.6402557

[7] A. U. Batmaz, M. D. Barrera Machuca, J. Sun, and W. Stuerzlinger. The effect of the vergence-accommodation conflict on virtual hand pointing in immersive displays. In *2022 CHI Conference on Human Factors in Computing Systems*, 2022. doi: 10.1145/3491102.3502067

[8] O. Bimber and R. Raskar. *Spatial augmented reality: merging real and virtual worlds*. CRC press, 2005.

[9] G. J. Blaauw. Driving experience and task demands in simulator and instrumented car: a validation study. *Human Factors*, 24(4):473–486, 1982.

[10] D. A. Bowman and R. P. McMahan. Virtual reality: How much immersion is enough? *Computer*, 40(7):36–43, 2007. doi: 10.1109/MC.2007.257

[11] D. A. Bowman, C. Stinson, E. D. Ragan, S. Scerbo, T. Höllerer, C. Lee, R. P. McMahan, and R. Kopper. Evaluating effectiveness in virtual environments with mr simulation. In *Interservice/Industry Training, Simulation, and Education Conference*, vol. 4, p. 44, 2012.

[12] D. E. Breen, R. T. Whitaker, E. Rose, and M. Tuceryan. Interactive occlusion and automatic object placement for augmented reality. In *Computer Graphics Forum*, vol. 15, pp. 11–22. Wiley Onl. Lib., 1996.

[13] F. Bruno, A. Angilica, F. Cosco, and M. Muzzupappa. Reliable behaviour simulation of product interface in mixed reality. *Engineering with Computers*, 29:375–387, 2013.

[14] F. Bruno, A. Angilica, F. Cosco, M. Muzzupappa, I. Horvath, F. Mandorli, and Z. Rusak. Functional behaviour simulation of industrial products in virtual reality. In *International symposium on Tools and Methods of competitive engineering*, vol. 2, 2010.

[15] F. Bruno and M. Muzzupappa. Product interface design: A participatory approach based on virtual reality. *International journal of human-computer studies*, 68(5):254–269, 2010.

[16] A. Burova, J. Mäkelä, J. Hakulinen, T. Keskinen, H. Heinonen, S. Siltanen, and M. Turunen. Utilizing vr and gaze tracking to develop ar solutions for industrial maintenance. In *CHI Conf. on Human Factors in Computing Systems*, p. 1–13, 2020. doi: 10.1145/3313831.3376405

[17] G. Chaurasia, A. Nieuwoudt, A.-E. Ichim, R. Szeliski, and A. Sorkine-Hornung. Passthrough+: Real-time stereoscopic view synthesis for mobile mixed reality. *Proc. ACM Comput. Graph. Interact. Tech.*, 3(1), may 2020. doi: 10.1145/3384540

[18] Y. Cheng, Y. Yan, X. Yi, Y. Shi, and D. Lindlbauer. Semanticadapt: Optimization-based adaptation of mixed reality layouts leveraging virtual-physical semantic connections. In *ACM Symp. on User Interf. Softw. and Techn.*, p. 282–297, 2021. doi: 10.1145/3472749.3474750

[19] C. Cruz-Neira, D. J. Sandin, and T. A. DeFanti. Surround-screen projection-based virtual reality: The design and implementation of the cave. SIGGRAPH '93, p. 135–142, 1993. doi: 10.1145/166117.166134

[20] S. Deb, D. W. Carruth, R. Sween, L. Strawderman, and T. M. Garrison. Efficacy of virtual reality in pedestrian safety research. *Applied ergonomics*, 65:449–460, 2017.

[21] H. Dinh, N. Walker, L. Hodges, C. Song, and A. Kobayashi. Evaluating the importance of multi-sensory input on memory and the sense of presence in virtual environments. In *IEEE VR*, pp. 222–228, 1999. doi: 10.1109/VR.1999.756955

[22] L. A. Elkin, M. Kay, J. J. Higgins, and J. O. Wobbrock. An aligned rank transform procedure for multifactor contrast tests. UIST '21, p. 754–768, 2021. doi: 10.1145/3472749.3474784

[23] G. Emsenhuber, M. Domhardt, T. Langlotz, D. Kalkofen, and M. Tatzgern. Towards eye-perspective rendering for optical see-through head-mounted displays. In *IEEE VR Abstracts and Workshops (VRW)*, pp. 640–641, 2022. doi: 10.1109/VRW55335.2022.00171

[24] G. Emsenhuber, T. Langlotz, D. Kalkofen, J. Sutton, and M. Tatzgern. Eye-perspective view management for optical see-through head-mounted displays. CHI '23, 2023. doi: 10.1145/3544548.3581059

[25] J. a. M. Evangelista Belo, M. N. Lystbæk, A. M. Feit, K. Pfeuffer, P. Kán, A. Oulasvirta, and K. Grønbæk. Auit – the adaptive user interfaces toolkit for designing xr applications. UIST '22, 2022. doi: 10.1145/3526113.3545651

[26] F. G. Faust, T. Catecati, I. de Souza Sierra, F. S. Araujo, A. R. G. Ramírez, E. M. Nickel, and M. G. Gomes Ferreira. Mixed prototypes for the evaluation of usability and user experience: simulating an interactive electronic device. *Virtual Reality*, 23:197–211, 2019.

[27] U. Gruenefeld, J. Auda, F. Mathis, S. Schneegass, M. Khamis, J. Gugenheimer, and S. Mayer. Vrception: Rapid prototyping of cross-reality systems in virtual reality. CHI '22, 2022. doi: 10.1145/3491102.3501821

[28] A. Hettiarachchi and D. Wigdor. Annexing reality: Enabling opportunistic use of everyday objects as tangible proxies in augmented reality. CHI '16, p. 1957–1967, 2016. doi: 10.1145/2858036.2858134

[29] K. Hornbæk, S. S. Sander, J. A. Bargas-Avila, and J. Grue Simonsen. Is once enough? on the extent and content of replications in human-computer interaction. CHI '14, pp. 3523–3532, 2014.

[30] M. S. Horswill and A. M. Plooy. Auditory feedback influences perceived driving speeds. *Perception*, 37(7):1037–1043, 2008.

[31] H.-C. Jetter, R. Rädle, T. Feuchtner, C. Anthes, J. Friedl, and C. N. Klokmose. "in vr, everything is possible!": Sketching and simulating spatially-aware interactive spaces in virtual reality. CHI '20, p. 1–16, 2020. doi: 10.1145/3313831.3376652

[32] J. Jung, H. Lee, J. Choi, A. Nanda, U. Gruenefeld, T. Stratmann, and W. Heuten. Ensuring safety in augmented reality from trade-off between immersion and situation awareness. ISMAR '18, pp. 70–79, 2018. doi: 10.1109/ISMAR.2018.00032

[33] D. Kalkofen, C. Sandor, S. White, and D. Schmalstieg. *Visualization Techniques for Augmented Reality*, pp. 65–98. Springer New York, New York, NY, 2011. doi: 10.1007/978-1-4614-0064-6_3

[34] E. Kim and G. Shin. Head rotation and muscle activity when conducting document editing tasks with a head-mounted display. *Hum. Fact. and Erg. Soc. Ann. Meeting*, 62(1):952–955, 2018. doi: 10.1177/1541931218621219

[35] D. Lakens, A. M. Scheel, and P. M. Isager. Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2):259–269, 2018.

[36] A. T. Lee. *Flight simulation: virtual environments in aviation*. Routledge, 2017.

[37] C. Lee, S. Bonebrake, D. A. Bowman, and T. Höllerer. The role of latency in the validity of ar simulation. In *2010 IEEE Virtual Reality Conference (VR)*, pp. 11–18, 2010. doi: 10.1109/VR.2010.5444820

[38] C. Lee, S. Bonebrake, T. Hollerer, and D. A. Bowman. A replication study testing the validity of ar simulation in vr for controlled experiments. ISMAR '09, pp. 203–204, 2009. doi: 10.1109/ISMAR.2009.5336464

[39] C. Lee, G. A. Rincon, G. Meyer, T. Höllerer, and D. A. Bowman. The effects of visual realism on search tasks in mixed reality simulation. *IEEE TVCG*, 19(4):547–556, 2013. doi: 10.1109/TVCG.2013.41

[40] F. Lu and Y. Xu. Exploring spatial ui transition mechanisms with head-worn augmented reality. CHI '22, 2022. doi: 10.1145/3491102.3517723

[41] V. Mäkelä, R. Radiah, S. Alsherif, M. Khamis, C. Xiao, L. Borchert, A. Schmidt, and F. Alt. Virtual field studies: Conducting studies on

public displays in virtual reality. CHI '20, p. 1–15, 2020. doi: 10.1145/3313831.3376796

[42] D. Mandl, K. M. Yi, P. Mohr, P. M. Roth, P. Fua, V. Lepetit, D. Schmalstieg, and D. Kalkofen. Learning lightprobes for mixed reality illumination. ISMAR'17, pp. 82–89, 2017. doi: 10.1109/ISMAR.2017.25

[43] A. Marquardt, C. Trepkowski, T. D. Eibich, J. Maiero, E. Kruijff, and J. Schöning. Comparing non-visual and visual guidance methods for narrow field of view augmented reality displays. *IEEE TVCG*, 26(12):3389–3401, 2020. doi: 10.1109/TVCG.2020.3023605

[44] F. Mathis, J. O'Hagan, K. Vaniea, and M. Khamis. Stay home! conducting remote usability evaluations of novel real-world authentication systems using virtual reality. AVI 2022, 2022. doi: 10.1145/3531073.3531087

[45] F. Mathis, K. Vaniea, and M. Khamis. Replicueauth: Validating the use of a lab-based virtual reality setup for evaluating authentication systems. CHI '21, 2021. doi: 10.1145/3411764.3445478

[46] F. Mathis, K. Vaniea, and M. Khamis. Can i borrow your atm? using virtual reality for (simulated) in situ authentication research. VR'22, pp. 301–310, 2022. doi: 10.1109/VR51125.2022.00049

[47] B. J. Matthews, C. Reichherzer, and R. T. Smith. Remapped interfaces: Building contextually adaptive user interfaces with haptic retargeting. CHI EA '23, 2023. doi: 10.1145/3544549.3583912

[48] R. P. McMahan. *Exploring the effects of higher-fidelity display and interaction for virtual reality games*. PhD thesis, Virginia Tech, 2011.

[49] R. P. McMahan, C. Lai, and S. K. Pal. Interaction fidelity: the uncanny valley of virtual reality interactions. VAMR'16, pp. 59–70. Springer, 2016.

[50] D. Medeiros, M. McGill, A. Ng, R. McDermid, N. Pantidi, J. Williamson, and S. Brewster. From shielding to avoidance: Passenger augmented reality and the layout of virtual displays for productivity in shared transit. *IEEE TVCG*, 28(11):3640–3650, 2022. doi: 10.1109/TVCG.2022.3203002

[51] N. Merat and H. Jamson. A driving simulator study to examine the role of vehicle acoustics on drivers' speed perception. In *Driving Assesment Conference*, vol. 6. University of Iowa, 2011.

[52] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1):99–106, dec 2021. doi: 10.1145/3503250

[53] X. Min, W. Zhang, S. Sun, N. Zhao, S. Tang, and Y. Zhuang. Vpmodel: High-fidelity product simulation in a virtual-physical environment. *IEEE TVCG*, 25(11):3083–3093, 2019. doi: 10.1109/TVCG.2019.2932276

[54] T. Muender, M. Bonfert, A. V. Reinschluessel, R. Malaka, and T. Döring. Haptic fidelity framework: Defining the factors of realistic haptic feedback for virtual reality. CHI '22, 2022. doi: 10.1145/3491102.3501953

[55] M. Nabiyouni, S. Scerbo, D. A. Bowman, and T. Höllerer. Relative effects of real-world and virtual-world latency on an augmented reality training task: An ar simulation experiment. *Frontiers in ICT*, 3, 2017. doi: 10.3389/fict.2016.00034

[56] A. Ng, D. Medeiros, M. McGill, J. Williamson, and S. Brewster. The passenger experience of mixed reality virtual display layouts in airplane environments. ISMAR'21, pp. 265–274, 2021. doi: 10.1109/ISMAR52148.2021.00042

[57] M. Oberhauser and D. Dreyer. A virtual reality flight simulator for human factors engineering. *Cognition, Technology & Work*, 19:263–277, 2017.

[58] S. A. Penumudi, V. A. Kuppam, J. H. Kim, and J. Hwang. The effects of target location on musculoskeletal load, task performance, and subjective discomfort during virtual reality interactions. *Applied Ergonomics*, 84:103010, 2020. doi: 10.1016/j.apergo.2019.103010

[59] I. Pettersson, M. Karlsson, and F. T. Ghiurau. Virtually the same experience? learning from user experience evaluation of in-vehicle systems in vr and in the field. DIS '19, p. 463–473, 2019. doi: 10.1145/3322276.3322288

[60] K. Pfeil, S. Masnadi, J. Belga, J.-V. T. Sera-Josef, and J. LaViola. Distance perception with a video see-through head-mounted display. CHI '21, 2021. doi: 10.1145/3411764.3445223

[61] E. Ragan, C. Wilkes, D. A. Bowman, and T. Hollerer. Simulation of augmented reality systems in purely virtual environments. VR'09, pp. 287–288, 2009. doi: 10.1109/VR.2009.4811058

[62] D. Ren, T. Goldschwendt, Y. Chang, and T. Höllerer. Evaluating wide-field-of-view augmented reality with mixed reality simulation. VR'16, pp. 93–102, 2016. doi: 10.1109/VR.2016.7504692

[63] K. Rogers, S. Karaosmanoglu, M. Altmeyer, A. Suarez, and L. E. Nacke. Much realistic, such wow! a systematic literature review of realism in digital games. CHI '22, 2022. doi: 10.1145/3491102.3501875

[64] 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. ICMI '19, p. 16–25, 2019. doi: 10.1145/3340555.3353741

[65] S. Schneider, P. Maruhn, N.-T. Dang, P. Pala, V. Cavallo, and K. Bengler. Pedestrian crossing decisions in virtual environments: behavioral validity in caves and head-mounted displays. *Human factors*, 64(7):1210–1226, 2022.

[66] E. Shaw, T. Roper, T. Nilsson, G. Lawson, S. V. Cobb, and D. Miller. The heat is on: Exploring user behaviour in a multisensory virtual environment for fire evacuation. CHI '19, p. 1–13, 2019. doi: 10.1145/3290605.3300856

[67] A. L. Simeone, R. Cools, S. Depuydt, J. a. M. Gomes, P. Goris, J. Grocott, A. Esteves, and K. Gerling. Immersive speculative enactments: Bringing future scenarios and technology to life using virtual reality. CHI '22, 2022. doi: 10.1145/3491102.3517492

[68] M. Slater and A. Steed. A virtual presence counter. *Presence*, 9(5):413–434, 2000. doi: 10.1162/105474600566925

[69] M. Tatzgern and C. Birgmann. Exploring input approximations for control panels in virtual reality. VR'21, pp. 1–9, 2021. doi: 10.1109/VR50410.2021.00092

[70] T. T. M. Tran, C. Parker, M. Hoggenmüller, L. Hespanhol, and M. Tomitsch. Simulating wearable urban augmented reality experiences in vr: Lessons learnt from designing two future urban interfaces. *Multim. Tech. and Intera.*, 7(2), 2023. doi: 10.3390/mti7020021

[71] A. Voit, S. Mayer, V. Schwind, and N. Henze. Online, vr, ar, lab, and in-situ: Comparison of research methods to evaluate smart artifacts. CHI '19, p. 1–12, 2019. doi: 10.1145/3290605.3300737

[72] T. Waltemate, D. Gall, D. Roth, M. Botsch, and M. E. Latoschik. The impact of avatar personalization and immersion on virtual body ownership, presence, and emotional response. *IEEE TVCG*, 24(4):1643–1652, 2018. doi: 10.1109/TVCG.2018.2794629

[73] J. Wang, R. Shi, W. Zheng, W. Xie, D. Kao, and H.-N. Liang. Effect of frame rate on user experience, performance, and simulator sickness in virtual reality. *IEEE TVCG*, 29(5):2478–2488, 2023. doi: 10.1109/TVCG.2023.3247057

[74] K. Watson, R. Bretin, M. Khamis, and F. Mathis. The feet in human-centred security: Investigating foot-based user authentication for public displays. CHI EA '22, 2022. doi: 10.1145/3491101.3519838

[75] 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 TVCG*, 27(2):1204–1213, 2021. doi: 10.1109/TVCG.2020.3030400

[76] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins. The aligned rank transform for nonparametric factorial analyses using only anova procedures. CHI '11, p. 143–146, 2011. doi: 10.1145/1978942.1978963

[77] R. A. Wynne, V. Beanland, and P. M. Salmon. Systematic review of driving simulator validation studies. *Safety science*, 117:138–151, 2019.