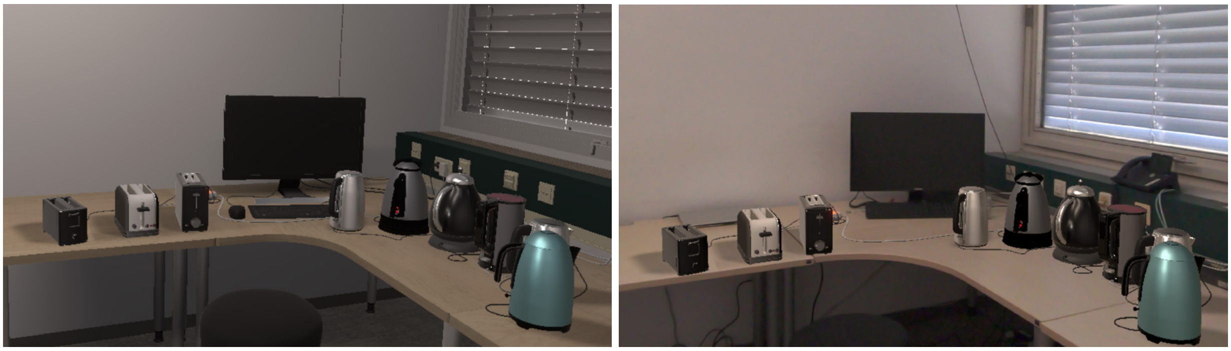
Exploring Plausibility and Presence in Mixed Reality Experiences

Franziska Westermeier, Larissa Brübach, Marc Erich Latoschik, and Carolin Wienrich



(a) VR scene view

(b) AR scene view

Fig. 1: VR and AR scene view with virtual interaction objects (kettles, toasters, and associated power cables and connectors). In the AR scene view, the background (including the monitor, the table, and the window) is content from the video see-through display. In the VR scene view, this background is completely remodeled. Background music was played from a virtual (VR) or physical audio source (AR).

Abstract—Mixed Reality (MR) applications along Milgram's Reality-Virtuality (RV) continuum motivated a number of recent theories on potential constructs and factors describing MR experiences. This paper investigates the impact of incongruencies that are processed on different information processing layers (i.e., sensation/perception and cognition layer) to provoke breaks in plausibility. It examines the effects on spatial and overall presence as prominent constructs of Virtual Reality (VR). We developed a simulated maintenance application to test virtual electrical devices. Participants performed test operations on these devices in a counterbalanced, randomized 2x2 between-subject design in either VR as congruent or Augmented Reality (AR) as incongruent on the sensation/perception layer. Cognitive incongruence was induced by the absence of traceable power outages, decoupling perceived cause and effect after activating potentially defective devices. Our results indicate that the effects of the power outages differ significantly in the perceived plausibility and spatial presence ratings between VR and AR. Both ratings decreased for the AR condition (incongruent sensation/perception) compared to VR (congruent sensation/perception) for the congruent cognitive case but increased for the incongruent cognitive case. The results are discussed and put into perspective in the scope of recent theories of MR experiences.

Index Terms—Plausibility, presence, congruence, mixed reality, virtual reality, augmented reality, spatial presence

1 Introduction

Milgram and Kishino's Reality-Virtuality (RV) continuum ranks Mixed Reality (MR) applications based on the proportion of virtual/physical content they exhibit [19]. For example, Augmented Reality (AR) is characterized by proportionally more physical than virtual content and, therefore, located closer to the continuum's physical reality endpoint. In contrast, Augmented Virtuality (AV) is located closer to the continuum's virtual reality endpoint since it is characterized by proportionally more virtual than real content. However, the RV continuum's scale is continuous. It provides a frame of reference for all proportions of virtual/physical environments. Recent see-through displays, such as the Varjo XR-3 device with its 115° field of view [39], enable almost seamless transitions across the RV continuum between fully immersive VR experiences and a purely physical environment with only minimal virtual AR content. These types of displays provide a vast design space for quite different MR applications.

This design space of the potential variety of mixed real/virtual interfaces poses interesting questions about the central effects and qualities of MR systems, as we are far beyond the point where these technologies are solely used for entertainment. Use cases nowadays include health and therapy [8, 24], training [40] and education [11], as well as co-working and creativity [14]. All these fields of application require high credibility of the system to achieve high efficacy. Traditional VR factors like presence [36], describing the feeling of "being there" [20] or place illusion [33] do not seem to accurately capture the essential user qualities of the RV continuum because there is no holistic perception of a whole new environment per se as in VR. Instead, depending on the virtual/physical proportion and the respective type of content, users are always more or less aware of the physical environment as it is part of the experience. Hence, the role of presence in MR remains largely unclear. Following the presence model as proposed by Slater [33] and recently updated by Slater et al. [34], presence is constituted by two orthogonal factors, which are the place illusion (PI) and the plausibility illusion (Psi). PI refers to the illusion of being in a place, whereas Psi refers to the illusion that something is actually happening. If we consider the PI to be at least questionable in its capability of capturing the essential qualities of the RV continuum, then what about Psi?

Accordingly, recent work by Skarbez et al. [30, 31] and Latoschik and Wienrich [15] focused more on plausibility as a central quality of

- *Franziska Westermeier and Larissa Brübach are with the Human-Computer Interaction (HCI) group and the Psychology of Intelligent Interactive Systems (PIIS) group from the University of Würzburg*
- *Marc Erich Latoschik is with the HCI group from the University of Würzburg*
- *Carolin Wienrich is with the PIIS group from the University of Würzburg E-mail: {franziska.westermeier | larissa.bruebach | marc.latoschik | carolin.wienrich} @uni-wuerzburg.de*

Manuscript received 14 October 2022; revised 13 January 2023; accepted 30 January 2023. Date of publication 22 February 2023; date of current version 29 March 2023. Digital Object Identifier no. 10.1109/TVCG.2023.3247046

MR systems. In empirical evaluations [2, 12], breaks in plausibility are utilized to study its role as a defining quality. However, both define slightly different objective system characteristics potentially leading to plausibility, in analogy to immersion potentially leading to spatial presence (or PI) [31, 33]. While Skarbez defines coherence "as the set of reasonable circumstances that can be demonstrated by the scenario without introducing unreasonable circumstances" [29, p. 42], Latoschik and Wienrich define congruence as "the objective match between processed and expected information on the sensory, perceptual, and cognitive layers" [15, p. 4]. While both approaches slightly differ concerning how they cope with the integration of fidelity and realism vs. fictitious worlds and behaviors into their theories (which both incorporate), both provide a suitable manipulation space to investigate the impact of breaks in plausibility. Overall, this leads to questions about the role of plausibility as a central quality of MR systems. The interplay between plausibility and presence is also of interest, and if and how perceived breaks in plausibility might impact spatial presence and presence in general.

In the first part of this paper, this work contextualizes different views on presence and plausibility, including the work of Slater [33], Slater et al. [34], Skarbez et al. [31], and Latoschik and Wienrich [15]. Besides this presence (and plausibility) domain, we reprocess important works concerning MR including the RV continuum of Milgram and Kishino [19] and its revisions of Skarbez et al. [32] and Wienrich et al. [41]. We combine both topics and search for a common ground on presence and plausibility in MR on which to build our experiment described in the second part.

In our experiment, we caused systematic incongruencies along the MR continuum and examined the effect on plausibility and presence ratings. We developed a simulated maintenance application to test virtual electrical devices. We caused incongruencies by the absence of traceable power outages, decoupling perceived cause and effect after activating potentially defective devices. Participants experienced these incongruencies either in VR or in AR.

We want to answer the question **RQ1:** *Do cognitive incongruencies lead to different plausibility judgments between VR and AR?* Hence, we aspire to learn more about the comparability of different MR experiences. Additionally, we intend to clarify more the role of spatial presence and presence in general, as there are currently only hypothetical assumptions of these concepts for MR [32, 34]. Specifically, we ask **RQ2:** *Do cognitive incongruencies lead to different presence ratings between VR and AR?* In addition, we investigate if spatial presence can be affected by our induced manipulations. Therefore, our last research question is **RQ3:** *Do cognitive and sensory/perceptual incongruencies lead to a break in spatial presence?*

Throughout this paper, we use the term MR synonymously with eXtended Reality (XR). In this paper, we prefer the term MR over XR because we introduce two conditions (VR and AR) that are placed at different points along the continuum, and we believe that the term MR better reflects that contrast.

2 Related Work

Milgram and Kishino [19] first introduced MR in the RV continuum, which makes the classification of MR applications on a technological basis possible. They define three axes: *Extent of World Knowledge* - the knowledge about the displayed world; *Reproduction Fidelity* - the display's ability to reproduce images of real or virtual objects; and *Extent to Presence Metaphor* - the intensity of the feeling of being in an unmediated environment. With the latter, the authors do not address a subjective feeling but rather a system characteristic comparable to immersion. Hence, the three axes objectively describe a system. The subjective reception of a system to a user is not reflected in this model at all. However, it is crucial when designing compelling user experiences for MR. In Skarbez et al.'s recent revision of this model [32], *coherence* is added to the set of axes defining an MR experience intending to move the focus from a collection of system characteristics to the actual interaction of these system characteristics and transmission to the user. According to the authors, in a VR experience, the judgment of coherence is internal, describing the interrelation of virtual objects or entities. Hence, users also adopt internal logic and laws that do not necessarily need to align with what they know from the real world. In science-fiction movies, for example, concepts might appear implausible compared to real-world logic. However, internal plausibility can be established by explaining the underlying internal mechanisms [3, 25]. As in AR, the physical world is omnipresent, the judgment is based more on real-world logic and knowledge, making it external. Wienrich et al. [41] also address the issue of how VR and AR experiences can be evaluated. They draw on a concept from media psychology, the so-called *reference frame*. So far, this concept has been used exclusively to distinguish between VR and real-world experiences. Thus, it has been associated with the feeling of "being there." Users feel spatially present in the virtual world if the virtual environment is set as a reference and virtual actions seem plausible [42]. Wienrich et al. [41] describe the reference frame as the result of weighted referential cues stemming from virtual and physical entities. In contrast to reference definitions in VR, these cues can not only be space-related (e.g., the environment in which the user interacts with objects) but also object-related (e.g., the actual objects the user interacts with). This makes it possible for MR experiences to be taken into account more and to make predictions about the plausibility evaluation when virtual and physical content is mixed.

These recent revisions [32, 41] of the RV continuum concentrate more on the user perception and judgment of an MR experience. They conclude that MR experiences are not to be evaluated according to the same criteria as pure VR experiences. However, both approaches lack empirical evidence.

In contrast, some theoretical models assume the general applicability of quality measures such as presence or plausibility, even for MR forms other than VR. Lombard and Ditton [16] do not directly refer to MR. However, they list the concept of presence as transportation, among others. Hence, they define three types: (1) the user being transported to a different location ("you are there"), (2) virtual objects are brought to the user ("it is here"), and (3) two users are transported to a different location together ("we are together"). Presence as transportation was mainly examined in the first type, which basically describes presence in VR. However, the second option can be linked more closely to MR. Slater [33] proposes a model with the two orthogonal factors PI and Psi leading to presence in virtual environments. PI is used in analogy to *spatial presence*, the feeling of "being there". In prior work [20] this feeling was described with the term *presence*. Slater [33], however, broadens this presence term to be the origin of the realistic response of users to virtual reality. Besides PI, he defines the other factor, Psi, as the "overall credibility of the scenario being depicted in comparison with expectations" [33, p. 3549]. Psi is evaluated on a cognitive level, whereas PI describes the sensory/perceptual reception of an experience. In their update to PI and Psi [34], the authors postulate that these two orthogonal factors apply not only to VR but to other forms of MR as well. In this case, PI becomes inverted, so there is no allocation of the user into the virtual world (as it would be in VR) but a placement of virtual content into the real world (similar to Lombard and Ditton's idea of transportation [16]). The quality of PI is then measurable by coverage of sensorimotor contingencies that allow the perception of virtual *objects*. Although they do not discuss Skarbez et al.'s [32] and Wienrich et al.'s [41] current thinking in this context, they, too, conclude that in MR, interaction objects may play a more significant role as anchors for PI than in pure VR experiences. The second factor, Psi, has the same requirements as in VR - responsiveness, interactability of virtual entities, and consistency between physical and virtual objects.

Where previous work limited the plausibility (or Psi) factor to be a pure cognitive construct, Latoschik and Wienrich [15] argue that plausibility is more deeply-rooted and requires sensory and perceptual processing in addition to cognitive processing. They propose a new model for plausibility called the *Congruence and Plausibility (CaP) model*. Similar to Skarbez et al. [31], they focus on plausibility and congruence to evaluate XR experiences. Thus, they make the concept of plausibility applicable to XR, that is, to the entire MR spectrum.

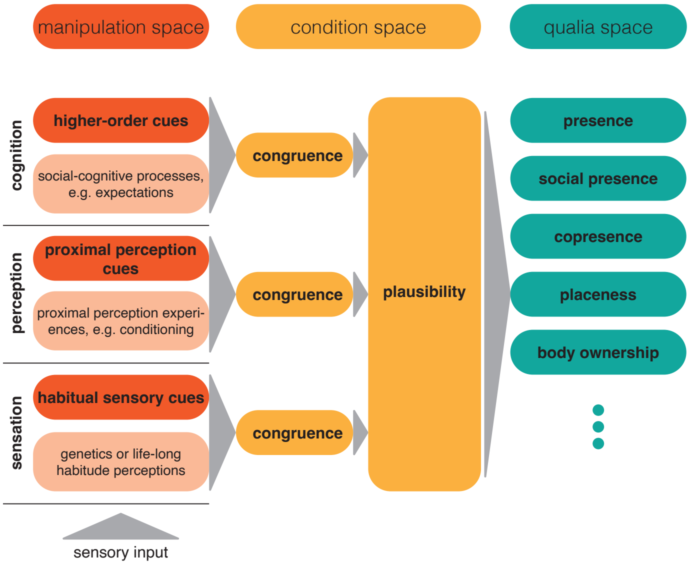


Fig. 2: The CaP model consists of a three-layered manipulation space on which every layer provides congruence. The different sources of congruencies are then added together to result in a weighted activation function of plausibility (redesigned by authors after the model in the original literature [15]).

This proposed model (see Fig. 2) consists of a three-layered manipulation space, including the sensory, the perceptual, and the cognitive layer. These layers define different information processing routes (i.e., bottom-up or top-down). Latoschik and Wienrich argue that the bottom-up layers represent the fundamental level of processing physical and physiological signals, while interpreting these signals happens on higher-order layers. Thus, bottom-up processing inheres processing on these higher-order layers as well. The higher the layer, the more plasticity is provided, and the more users can adapt to the given circumstances. The layers provide congruencies that are set in relation to the overall experience. On every single layer, the congruence can contribute differently to plausibility. Thus, (in)congruencies on the different layers should lead to different XR experiences. The authors assume that manipulation on a higher layer cannot influence plausibility on a lower layer. Thus, the model enables precise predictions about which manipulations lead to which effect in the evaluation of plausibility and presence of XR experiences. One way of manipulation is the introduction of incongruencies. Consequently, a break in plausibility shall be perceived by the user.

These models are said to be valid for MR experiences [15, 34]. However, up until now, empirical evaluations solely exist in VR. Systematic evaluations of plausibility, such as Hofer et al. [12] and Brübach et al. [2], make use of breaks in plausibility in VR and measure their outcome on different qualia, such as presence or spatial presence. Hofer et al. [12] manipulated the physical behavior of virtual objects. They hypothesized that a higher plausibility leads to a higher spatial presence. They understand plausibility as higher-order system processing while spatial presence is processed on the sensory/perceptual level. They successfully measured a break in plausibility. However, a break in spatial presence was not found. The work of Brübach et al. [2] links even more closely to current theories and poses a first evaluation of the CaP model in VR. To do so, they combined a manipulation of objects (gravity-defying behavior) with a framing manipulation (reading a story about a spaceship/container ship beforehand) in a VR scenario. Even though the authors define the manipulation of objects on the perception layer, we argue that this manipulation was cognitive as well. We would expect manipulations on, for example, motion or stereo parallax to affect the perception layer. However, we assume that there are gradations within each layer. These layers might also be continuous. Thus, the object manipulation was on a relatively low grade within the cognitive layer, while the framing manipulation was on a higher internal level. The plausibility was measured with thirteen questions. Specific questions target either the internal or the external plausibility. However, this set of questions has not been validated yet. The results revealed that the lower-layer incongruence significantly decreased the perceived plausibility while presence and spatial presence were unaffected. The higher-layer incongruence did not affect or counteract the perceived break in plausibility. Thus, the authors assumed that the higher-level cognitive manipulation was much weaker than the lower-level cognitive manipulation. In both experiments [2, 12], cognitive manipulations did not cause breaks in lower layers (such as spatial presence). In contrast, perceptual manipulations given by a priori device-specific measures, such as those examined in Wolf et al.'s work [44], influenced spatial presence (the Varjo XR-3 [39] vs. the Hololens 2 [18]). However, using the same device (the Varjo XR-3 [39]) for a video-see-through AR solution and an analog VR solution did not appear to affect spatial presence. In their experiment [44], the plausibility of a virtual human was assessed with a 3x1 within-subject design (VR, video-see-through AR, and optical-see-through AR). They used the *Virtual Human Plausibility Questionnaire* [17] with the subscales *appearance and behavior plausibility* and the *virtual humans' match to the virtual environment*. While they could not identify a difference in appearance and behavior plausibility along the different conditions, they measured a higher match to the virtual environment for the VR condition compared to both AR conditions. These results might be related to the idea of the distinction between internal (the appearance and behavior) and external (the match to the virtual environment) plausibility.

3 Summary and Present Study

To summarize, empirical studies investigating the impact of breaks in plausibility found a substantial impact on the subjective plausibility rating in VR. In Brübach et al. [2], two different cognitive manipulations resulted in different degrees of influence on plausibility. Because the manipulation of gravity (seen as the lower level) had a higher impact on plausibility than the manipulation of the framing (seen as the higher level), the authors found evidence for the strengths of bottom-up effects in contrast to higher-order cues. Thus, it might be interesting to investigate a different kind of operationalization, especially in MR. For example, the causal connection between events can be manipulated as causality is said to be an essential component to establishing a condition of plausibility [7]. Cavazza et al. [4] stated that their experiment manipulated the causality to create plausible and implausible co-occurring events. As a result, they could measure an effect of this causal perception on presence. Therefore, an operationalization of causality seems promising as it appears to have implications for both plausibility and presence. Following their model, we adapt the idea of manipulation of causality by modifying co-occurrences of events and thus, induce a semantic incongruence on the cognitive layer.

In addition to the first manipulation, we define the use of AR as an a priori incongruence on the sensory/perceptual layer and hence, our second manipulation. We assume an AR environment to convey incongruent information on the sensory and perception layers. This is due to slight reconstruction errors caused, for example, by inaccuracies or imprecisions of object tracking or unknown parameters of the current real-world light transport, given the used AR device, rendering engine, and sensory equipment. Accordingly, a seamless mix of virtual and physical content with undetectable borders between both is hard to achieve, even with the significant advantages that have been made to replicate real-world light-material interactions and to solve the registration problem of object placements within AR. Hence, the resulting incongruencies are continuously presented to the users, specifically concerning important cues on the sensory and perceptual layers [6].

We intend to cause breaks in plausibility with these incongruencies. With breaks in plausibility, we address users' perception of discrepancies in the experience in analogy to breaks in presence. Previous presence research defined breaks in presence as countable events when the attention is shifted from the virtual to the real environment [35]. However, these breaks in presence were evaluated based on the perception of users ("If and only whenever you experience a transition to Real, please say 'Now' very clearly and distinctively" [35, p. 425]). Chertoff et al. [5] defined different types of breaks in presence, such as contradictory mediation or inconsistent mediation. Our cognitive manipulation incorporates *contradictory mediation* (contradiction with expectations) and, thus, is capable of causing a break in presence and/or plausibility. This manipulation is also event-based (coupled with the power outage). Our sensory/perceptual manipulation incorporates *inconsistent mediation* (e.g., hardware limitations). We argue that participants rate the sum of perceived breaks. Every time the visual overlay is not aligned, the occlusion is wrong, or simply the different contrasts of virtual and real content are perceived by the participant, there are breaks. Thus, we see this sensory/perceptual break as continuous and do not expect a measurable spontaneous reaction from participants but a measurable effect on the overall experience. We define a break as a significant discrepancy between conditions.

We adopt Slater's [33] notion of *presence* as we see it as a broader construct than just the spatial aspect of being somewhere. To describe the latter, we use the term *spatial presence*.

3.1 Hypotheses

We induce cognitive and sensory/perceptual incongruencies. The work of Brübach et al. [2] and Hofer et al. [12] successfully measured a break in plausibility by inducing a cognitive manipulation. Similarly, we aspire to cause a perceived break in plausibility with cognitive incongruence. Our first hypothesis is as follows:

- **H1** Breaks in plausibility induced by cognitive incongruencies have different strengths between *VR* and *AR*.

To see if the concept of presence can be transferred unaltered from VR to AR, we want to determine if presence is rated differently between VR and AR. We check beforehand if breaks in presence were established in VR and AR to be able to compare both results:

- **H2** Breaks in presence induced by cognitive incongruencies have different strengths between *VR* and *AR*.

As the CaP model hypothesizes that top-down manipulations do not impact downwards, we do not expect that the cognitive incongruence affects the spatial presence as it is a low-layer quale. Instead, we expect the sensory/perceptual incongruence (given a priori through AR) to influence spatial presence. Therefore, the following hypotheses were formed:

- **H3** A break in spatial presence cannot be induced by an incongruence on the cognitive layer.
- **H4** A break in spatial presence can be induced by an incongruence on the sensory/perceptual layer.

3.2 Design

The study is based on a counterbalanced, randomized 2 x 2 between-subject design. The participants are distributed into the different conditions by a randomized list, including the four conditions independently of demographic data. The list also guarantees a counterbalanced number of participants per condition. The demographic distribution across the conditions showed no large deviation.

The first factor, the *sensory/perceptual congruence*, splits into a *VR* and an *AR* condition (see Fig. 1). We define VR as congruent and AR as incongruent. While the AR condition uses the video-see-through functionality to provide the physical environment, the VR condition uses an exact virtual replica of the physical room [23]. In addition to the visual sense, the auditory sense is addressed by an audio source emerging from a physical/virtual radio. We define physical audio as a sound which is reverberating in the physical room. In its most straightforward form virtual audio does not include spatial information. However, it can be synthetically spatialized in the virtual environment. In our experimental design, we include only the latter version of virtual sound and link it to the respective location of virtual objects. For example, if a device is switched on, a matching sound occurs from the direction of this device. Both the VR and AR conditions share virtual objects that the participants interact with (e.g., kettles and toasters placed on a table and a fuse box mounted on the wall; see Fig. 3). When interacting with these objects, virtual sound effects are triggered.

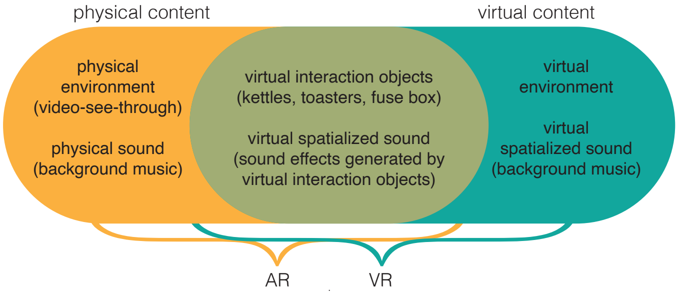

physical content
virtual content
physical environment (video-see-through)
physical sound (background music)
virtual interaction objects (kettles, toasters, fuse box)
virtual spatialized sound (sound effects generated by virtual interaction objects)
virtual environment
virtual spatialized sound (background music)
AR
VR

Fig. 3: Distribution of virtual and physical content in the VR and AR conditions. The outer components are not modified in the CI condition.

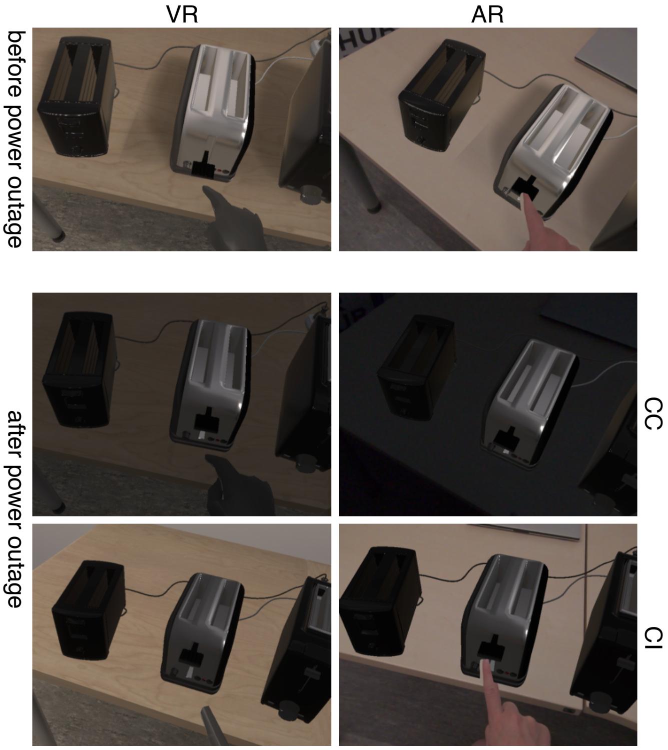


Fig. 4: Before the power outage, the CC and CI conditions do not differ in the respective VR and AR scenarios. After interaction with the virtual toaster, the power outage only affects the whole scene (CC) or solely the virtual interaction objects (CI). Besides visual cues (change of lighting and reflections), audio cues are deployed.

The second factor is the *cognitive congruence* of the scenario with the conditions *cognitive-congruent (CC)* and *cognitive-incongruent (CI)*. In the experiment, participants trigger a power outage. In the *CC* condition, this power outage affects the virtual interaction objects, the environment, and the background music. Semantically speaking, the power outage affects the whole electricity in the room. In the *CI* condition, solely the virtual interaction objects are affected. Hence, the lighting and reflections on the kettles, toasters, and fuse box are changed while the environment is still enlightened and the background music is still playing (see Fig. 4).

In the *AR* condition, the physical environment is manipulated using Philips Hue devices [37] that are controlled via a unity package [38].

4 METHOD

4.1 Participants

Eighty-three participants took part in the experiment. Due to technical issues, two participants were excluded, resulting in 81 remaining participants for data analyses. Twenty participants were in the VR-CC condition, 19 participants were in the VR-CI condition, while the AR-CC condition included 22, and the AR-CI condition included 20 participants. The demographic distribution and media usage of participants can be seen in Tab. 1. The study was approved by the institution's ethics committee.

Table 1: Demographic data and media usage over all four conditions and all participants.

demographics	# participants (from a total of 81 participants)
biological gender	49 female; 31 male; 1 diverse
social gender	50 female; 30 male; 1 transgender
age	$M = 27.83$ ($SD = 9.58$)
occupation	57 students, 21 employees, 2 job-seeking, 2 with other employment
duration of VR experience	22 <1h, 21 1-3h, 12 3-5h, 9 5-10h, 3 10-20h, 14 >20h
frequency of VR experience	10 never used, 32 used 1-3 times, 16 used 3-6 times, 5 used 6-10 times, 4 used 10-20 times, 14 used > 20 times
duration of AR experience	48 <1h, 14 1-3h, 5 3-5h, 4 5-10h, 2 10-20h, 8 >20h
frequency of AR experience	32 never used, 26 used 1-3 times, 9 used 3-6 times, 1 used 6-10 times, 3 used 10-20 times, 10 used > 20 times
Computer usage	$M = 5.88$ ($SD = 3.91$) h/day
Internet usage	$M = 6.49$ ($SD = 3.48$) h/day
Gaming usage	$M = 1.12$ ($SD = 1.95$) h/day
TV usage	$M = 0.60$ ($SD = 0.94$) h/day
Streaming service usage	$M = 1.12$ ($SD = 1.17$) h/day

4.2 Material

The application ran on a high-end computer with an Nvidia Geforce RTX 3080 GPU and an Intel i9-11900K CPU with 64 GB of RAM. We used a Varjo XR-3 [39] as the Head-Mounted Display (HMD) as it can be used both for VR and AR with the same visual parameters (field of view, brightness, resolution). In addition, the Varjo XR-3 offers hand tracking, which is used to interact with virtual objects. Masking is used to cut out the hand outline to ensure that virtual objects do not occlude the hands. In VR, rigged hand models are used to visualize the hands. In the AR condition, the video-see-through function of the HMD was used. The HMD provides two different types of display per eye, a focus area (1920 x 1920 px per eye) and a peripheral area (2880 x 2720 px per eye). The horizontal field of view reaches 115° with a refresh rate of 90 Hz. Physical and virtual content are displayed together on the same pixel display, ensuring equality of the physical display (unlike, e.g., the content depicted in the Hololens [18] in composition with the real unrasterized background). However, virtual and physical content differ in their light-material interaction. The application was developed in Unity Engine (v2020.3.21f1) with Varjo's Unity XR SDK (v3.0.0) and Varjo Base (v3.5.1.7). We used a Sennheiser SC 60 USB ML headset for virtual and a Tivoli Audio Model One for physical audio reproduction. We used R (v3.6.1) and Python (v3.8.15) for the statistical analysis.

4.2.1 Measures

To measure plausibility and to answer *H1*, we used the questions of Brübach et al. [2] as the basis (see Tab. 2). We adapted the questions according to our use case and changed the main focus from the object behavior to the cause-and-effect scenario. We adopt the separation of *internal* and *external plausibility*. For the remaining general questions, we assume that they cannot easily be classified as either internal or external. We measured a high reliability (with Cronbach's α) for the external plausibility of $\alpha = 0.83$ and a medium reliability for the internal plausibility of $\alpha = 0.58$. The plausibility questions are rated on a seven-point Likert Scale with the labels (1) do not agree at all, (2) disagree, (3) somewhat disagree, (4) agree partially, (5) somewhat agree, (6) agree, and (7) fully agree. Although we acknowledge that these questions have not yet been validated, previous results look promising, considering that there are no suitable validated alternatives to our knowledge.

Presence is the subject of *H2*. It is measured with two questionnaires: the Witmer and Singer presence questionnaire (WSPQ) [43] and the Igroup presence questionnaire (IPQ) [28]. Witmer and Singer define presence as "the subjective experience of being in one place or environment" [43, p. 225]. The WSPQ provides the four factors of *involvement*, *sensory fidelity*, *adaption/immersion*, and *interface quality*. Schubert et al. [28] define three subscales of presence: *spatial presence*, *involvement*, and *experienced realism*, which shows some similarities to Slater's model of presence [33]. Answers to both questionnaires are reported on a scale from 1 - 7. *H3* and *H4* shall be assessed with this IPQ subscale of spatial presence.

In contrast to the IPQ, the WSPQ provides a broader range of questions, including a higher variance for *control and predictability* compared to the IPQ [4], which could be beneficial regarding causality. However, we included the IPQ for a better comparison to previous plausibility studies [2] and presence models [33]. The term "virtual environment" was replaced by "shown environment" in both questionnaires to fit both the VR and AR conditions with minimal text change.

To investigate possible interpersonal differences in the perception of the experience, we include the tendency of immersion assessed by the Immersive Tendencies Questionnaire (ITQ, [43]), including an overall score and the subscales *focus*, *involvement*, and *games*. Participants rated the ITQ on a seven-point Likert Scale. Additionally, we evaluated the Simulator Sickness Questionnaire (SSQ, [13]) with the subscales of *nausea*, *oculomotor*, and *disorientation* on a scale from 1 - 4. The NASA-TLX was assessed on a discrete slider from 0 - 20 [9] with the subscales *mental demand*, *physical demand*, *temporal demand*, *own performance*, *effort*, and *frustration*. Lastly, we tracked the exposure time in seconds.

4.3 Procedure

The experiment takes approximately 40 minutes. The procedure can be seen in Fig. 5. The participants fill out consent forms and their demographic data, including VR and AR experience. They complete the SSQ and the ITQ.

After reading the instructions, the participants start the exposure counterbalanced either in VR or AR. Participants are told to test electronic devices (i.e., toasters and kettles). This narrative is consistent over all four conditions. Participants shall spot the devices and the fuse box in a short orientation phase. Afterwards, the participants test the residual current circuit breaker inside the fuse box. They have to switch it off and on again four times. Within this phase, the power outage is congruently affecting the whole room (virtual or physical) with the implication of accustoming the participants to this coherent effect regardless if they are in the CC or CI condition. Participants test the electronic devices in the next phase by turning on one after the other in a predefined order from left to right. Some predefined devices pass the test and turn off after a few seconds without causing a power outage. Other predefined devices fail and cause a power outage. There is a divergence in the effect of the power outage between the CC condition and the CI condition. In the CC condition, the whole room is affected by the power outage, and in the CI condition, solely the virtual interaction objects are affected. After a power outage, participants restore the power by switching on the residual current circuit breaker and continue to test the next device. When all nine devices are tested, the exposure is finished.

Participants fill out a free text area to qualitatively describe what

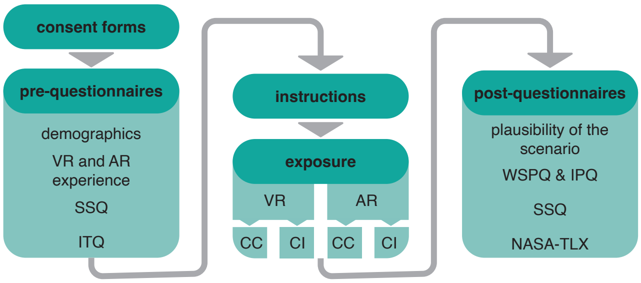


Fig. 5: The experiment procedure starts with consent forms and pre-questionnaires, followed by instructions. The exposure is either in VR or AR, and the scenario is either congruent (CC) or incongruent (CI). Post-questionnaires were assessed after the exposure.

happened to control for the efficacy of our manipulation. They fill out the post-questionnaires (scenario plausibility, WSPQ, IPQ, SSQ, and NASA-TLX). In the end, the participants are informed about the study's intention.

5 Results

We report descriptive statistics (i.e., mean and standard deviation (SD)) for each condition in Tab. 2. We calculated all tests against an alpha of <.05. We used non-parametric Mann-Whitney-U tests to identify significant main effects. To examine interaction effects, we calculate two-way ANOVAs even though we explicitly acknowledge that not all assumptions were met for all cases (i.e., normal distribution, homoscedasticity). Previous work [1, 10, 27] stated that ANOVAs are robust toward violations of normal distribution. We marked items that did not meet the assumption of homoscedasticity according to Levene's test in Tab. 2. In addition, the test statistics from the two-way ANOVA are listed in Tab. 2 as well. To cross-check the results of the ANOVA and to better compare the perception of the power outage manipulation between VR and AR, we calculate simple main effects *within* the VR and AR factor with Mann-Whitney-U tests (e.g., we calculate VR-CC versus VR-CI for the effect of the cognitive congruence factor within VR, i.e., the main effects of a factor level within one level of the respective other factor). We also calculated simple main effects within the CC and CI factors and marked significant simple main effects between the respective conditions in Tab. 2 with matching symbols.

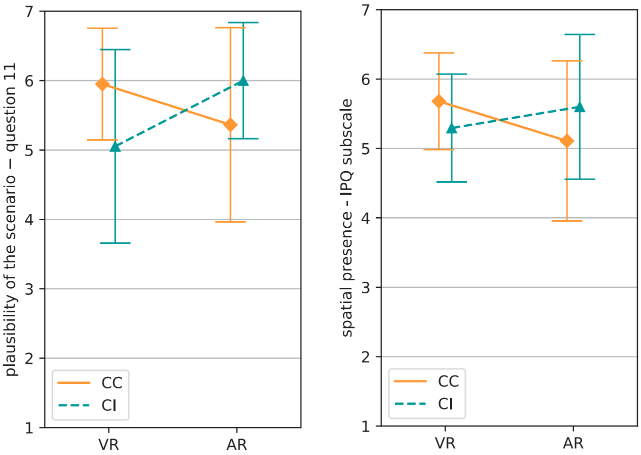
5.1 Plausibility

5.1.1 Main Effects with Mann-Whitney-U Tests

Based on the questionnaire of Brübach et al. [2], we evaluated the corresponding subscales and total score. In addition, we examined the single items to interpret our results further. Mean values were inverted for all negated questions so that a higher value indicates a higher plausibility. No main effect was found for the total plausibility score or the two subscales. However, we identified a significant main effect in question 8 between the mean values of $M = 4.8$ ($SD = 1.64$) for the CC and $M = 4.02$ ($SD = 1.70$) for the CI condition with $U = 604.0, p = .040$.

5.1.2 Interaction Effects with two-way ANOVAs

We detected three significant interaction effects in questions 11, 12, and 13, respectively (see Tab. 2 and Fig. 6 (a)). For these questions, the mean values of VR-CC and AR-CI are rated highest, while the mean values of the VR-CI and AR-CC provide the lower values for plausibility. The total score and subscales of internal and external plausibility did not show significances. However, the medium effect size $\eta_p^2 = 0.4$ of the total score underlines the assumption that a larger sample of participants potentially leads to significance. All results of the interaction can be seen in Tab. 2.



(a) Plausibility question 11

(b) Subscale *spatial presence* (IPQ)

Fig. 6: A similar interaction pattern can be observed for the significant plausibility questions (e.g., question 11 "The cause and effect behavior matched the scenario."), and for the IPQ subscale *spatial presence*. The interaction plots show the mean values ±*SD*.

5.1.3 Simple Main Effects with Mann-Whitney-U Tests

Examining the results of question 11, we found a significant simple main effect between the higher-rated VR-CC and the lower-rated VR-CI condition in *VR* ($U = 263.5, p = .031$). Within the *CI* factor, a significant simple main effect could be detected with $U = 112.5, p = .024$, whereas the plausibility in VR-CI was rated lower than in AR-CI. Question 12 revealed a significant effect between the VR-CC and VR-CI conditions in *VR* ($U = 257.0, p = .049$) with the higher value for the VR-CC condition. Additionally, a significance was discovered in question 13 within the *AR* factor($U = 146.0, p = .035$) with the higher-rated plausibility in the AR-CI condition. Within the *CI* factor, we found a significance between the VR-CI and the higher-rated AR-CI condition of $U = 119.0, p = .025$. Exact mean values and standard deviations can be obtained from Tab. 2.

All the found simple main effects strengthen the validity of interaction effects found with the two-way ANOVA.

5.2 Presence and Spatial Presence

5.2.1 Main Effects with Mann-Whitney-U Tests

To assess presence, we evaluated the WSPQ and the IPQ. In the WSPQ, the subscale *sensory fidelity* showed a significant main effect ($U = 586.5, p = .028$) with mean values of $M = 4.5$ ($SD = 0.16$) for the VR and $M = 5.0$ ($SD = 0.15$) for the AR condition. In the WSPQ subscale *involvement*, a significant main effect was detected ($U = 606.5, p = .045$) between the CC ($M = 4.92, SD = 0.13$) and the CI condition ($M = 5.18, SD = 0.13$). The IPQ and its subscales did not show significant main effects.

5.2.2 Interaction Effects with two-way ANOVAs

The results of the IPQ and the WSPQ of the four separate conditions can be seen in Tab. 2. The evaluation of the IPQ revealed a significant interaction effect in the subscale for *spatial presence*, with the highest rating for the VR-CC condition, followed by the AR-CI condition and the VR-CI and AR-CC conditions.

5.2.3 Simple Main Effects with Mann-Whitney-U Tests

A significant simple main effect was found in the WSPQ subscale *involvement* with $U = 141.0, p = .048$ within the AR factor, with a higher value for the AR-CI condition. Even though the IPQ subscale *spatial presence* showed a significant interaction effect, we could not identify significant simple main effects.

Table 2: Results from the calculation of two-way ANOVAs of the plausibility questions, the IPQ, and the WSPQ showing the four conditions (sensory/perceptual and cognitive (in)congruencies) and their interaction effects. Questions regarding the plausibility of the scenario are inspired by Brübach et al. [2].

no.	question	VR-CC	VR-CI	AR-CC	AR-CI	$F(1,77)$	p	η_p^2
	Plausibility - subgroup *external plausibility*							
1	I am used to such a scenario having this effect.	3.95 (1.63)	4.05 (1.70)	3.64 (1.99)	4.30 (1.82)	0.47	.496	.006
2	In everyday life, I expect the cause and effect of the scenario to proceed like this.	4.80 (1.33)	4.58 (1.63)	4.82 (1.85)	5.25 (1.45)	0.82	.369	.010
3	I have experienced this scenario in real life.	5.00 (2.17)	5.00 (1.72)	4.95 (2.18)	5.25 (1.70)	0.11	.743	.001
4	The cause and effect behavior is unfamiliar to me. [1] [2]	5.05 (1.77)	4.89 (1.97)	5.18 (1.67)	6.20 (0.75)	2.56	.114	.032
5	I do not know this scenario from real life. [1]	5.55 (1.77)	5.05 (1.91)	5.45 (2.95)	5.55 (1.99)	0.49	.487	.006
	mean value of answers in the *external plausibility* group	4.87 (1.40)	4.72 (1.39)	4.81 (1.55)	5.31 (1.05)	1.10	.296	.014
	Plausibility - subgroup *internal plausibility*							
6	I had a prior anticipation of how the scenario would proceed.	4.05 (1.80)	3.74 (2.07)	3.86 (2.03)	3.65 (1.85)	0.01	.911	<.001
7	I expected the effects in the scenario to be like this.	4.70 (1.45)	4.26 (1.77)	4.55 (1.92)	4.35 (1.71)	0.09	.760	.001
8	I could not anticipate what would happen next in the scenario. [1]	5.00 (1.18)	3.89 (1.92)	4.59 (1.95)	4.15 (1.46)	0.77	.384	.010
9	I was surprised by the scenario. [1]	5.20 (1.75)	4.16 (1.73)	3.41 (2.15)	4.60 (2.08)	1.93	.169	.024
10	I had no idea that the scenario would play out the way it did. [1]	5.35 (1.46)	5.05 (1.70)	4.95 (1.77)	5.60 (1.71)	1.53	.219	.020
11	The cause and effect behavior matched the scenario. [2]	5.95 (0.81) †	5.05 (1.39) †,*	5.36 (1.40)	6.00 (0.84) *	8.55	**.005**	.100
	mean value of answers in the *internal plausibility* group	5.04 (0.99)	4.36 (1.27)	4.62 (1.33)	4.72 (1.06)	2.16	.145	.027
	Plausibility - subgroup *general*							
12	The behavior of cause and effect made sense. [2]	6.05 (0.74) †	5.11 (1.48) †	5.55 (1.53)	5.85 (1.01)	4.86	**.030**	.059
13	I think this behavior of cause and effect is impossible. [1,2]	6.15 (1.46)	5.79 (1.58) *	5.68 (1.79) †	6.70 (0.56) †,*	4.43	**.039**	.054
	total mean value over all questions	5.14 (0.95)	4.66 (1.14)	4.85 (1.13)	5.19 (0.79)	3.12	.082	.039
	IPQ							
	spatial presence	5.68 (0.70)	5.29 (0.78)	5.11 (1.15)	5.6 (1.04)	4.13	**.045**	.051
	experienced realism	4.35 (0.85)	4.42 (0.85)	4.3 (1.07)	4.45 (0.97)	0.04	.847	<.001
	involvement	4.41 (1.11)	4.79 (1.14)	4.00 (1.21)	4.25 (1.53)	0.05	.825	<.001
	total mean value over all questions	4.95 (0.69)	4.95 (0.72)	4.59 (0.92)	4.9 (0.95)	0.69	.408	.009
	WSPQ							
	involvement	5.00 (0.84)	5.12 (0.78)	4.85 (0.78)	5.25 (0.83)	0.52	.472	.007
	sensory fidelity	4.61 (1.06)	4.39 (0.96)	5.12 (0.86)	4.88 (1.00)	<0.001	.956	<.001
	adaption	5.51 (0.46)	5.36 (0.69)	5.53 (0.63)	5.49 (0.77)	0.14	.708	.002
	interface	5.02 (0.99)	4.44 (1.25)	4.44 (0.87)	4.55 (1.06)	2.19	.143	.028
	total mean value over all questions	5.06 (0.59)	4.97 (0.68)	5.05 (0.61)	5.17 (0.70)	0.49	.485	.006

[1] Question is negated. Answers are already inverted, which means that a higher value corresponds to a higher plausibility. [2] Ratings did not pass Levene's test for homoscedasticity. †, * Significant simple main effects between two conditions (marked with respective matching symbols)

5.3 Other Control-Measures

We used gender, age, duration of VR and AR experience, gaming usage, and the total ITQ score as covariates to calculate main and interaction effects in an analysis of covariance (ANCOVA) for the items listed in Tab. 2. There was no deviation in significance of the main and interaction effects from the results of the ANOVAs.

The results of the SSQ showed high significances within the cognitive congruence dimension for the subscales *nausea* ($U = 1056.5, p = .015$; CC $M = 0.43$, $SD = 18.09$; CI $M = -12.16$, $SD = 19.07$), and *oculomotor* ($U = 1049.0, p = .024$; CC $M = 0.12$, $SD = 15.07$; CI $M = -11.41$, $SD = 19.26$). Analyzing the NASA-TLX did not lead to any significant findings, and also the time participants needed to fulfill the task did not differ significantly between the conditions. We asked participants for a qualitative description of what happened in the scenario to get insights if our manipulation was recognized. Forty-three participants explicitly mentioned the power outage. Hence, this can be considered a good indicator of a successful manipulation.

6 Discussion

6.1 Plausibility

We measured three interaction effects in the set of scenario plausibility questions. Additionally, the total plausibility score showed a medium effect size. Thus, we assume that our manipulations (cognitive and sensory/perceptual) influence plausibility to a certain extent.

Hypothesis *H1* holds true for our experiment. We could detect a break in plausibility in some plausibility questions in VR and AR. Contrary to our assumption, in AR the effect showed inverted plausibility ratings when inducing a cognitive incongruence.

Within the VR condition, we detected a break in two items. Question 11 (“The cause and effect behavior matched the scenario.”) does not ask for the plausibility of cause and effect itself but the plausibility in the context in which this effect appears. This aligns with the intention of the internal plausibility question group proposed by Brübach et al. [2] which rates plausibility based on the internal relation of entities. The question is indeed assigned to this internal plausibility question group. In question 12 (“The behavior of cause and effect made sense.”), a break was detected as well. Here, the context is not explicitly included in the wording, which might also be why Brübach et al. [2] did not define a distinct question group. However, our results could be a hint that question 12 describes internal plausibility as well. In AR, question 13 ("I think this behavior of cause and effect is impossible.") indicated a break in the perceived plausibility. Question 13 was not assigned to a subgroup. However, the question intends to assess the credibility of cause and effect. From its wording, no context frame is provided in which cause and effect have to cohere. Hence, participants might refer to their real-world knowledge when answering this question.

To summarize, one *internal* plausibility question in *VR* and two general unassociated plausibility questions in *AR* were significant. These results support Skarbez et al.'s hypothesis [32] that VR is judged internally and AR is judged externally. In VR, the participants perceived a more holistic view of the scenario and might have judged based on the internal interrelations of the virtual world. In AR, participants always perceive two different worlds because of a mixture of virtual and physical entities. Hence, the judgment in the AR scenario might have been external, given the portion of physical content that may be an anchor for this judgment. The inversion of the effect of congruence could be explained through the reference frame, Wienrich et al. [41] proposed. The reference frame is hypothesized to have different weights of space-related and object-related entities between different MR experiences. In AR, users might focus more on objects than on the environment. In contrast, in the holistic VR scenario, the environment (i.e., space-related entities) has a higher priority for the reference frame. Thus, in the AR-CC condition, participants could have had expectations concerning the virtual interaction objects and then could have been surprised or confused about the effect the virtual entities had on the physical world. These findings are the first empirical indication of a different reference setting in AR compared to VR.

Moreover, the measured effect might be cognate of the uncanny valley effect [21]. Nilsson et al. [22] proposed that there is an analogous concept for realism in general. They postulate that the higher the fidelity (i.e., the objective degree of realism) of a system, the higher the risk for users to find an experience less realistic. Similar findings could be substantiated in Pouke et al.'s work [26]. In our case, we argue that the more concrete previous expectations and experiences are (regarding an anthropomorphic human or the world), the higher the risk of being classified as uncanny in case of smaller deviations. In our AR condition, we have a very concrete idea of the physical world in which virtual content triggers a mismatch. In contrast, environments that are out of touch with reality cannot trigger such strong reactions because their expectation is not concrete but very plastic.

In Brübach et al. [2], there was a dominant effect on plausibility from the lower-layered manipulation, namely the manipulation of object behavior, compared to the relatively weak top-down manipulation of the framing, which could not counteract a break in plausibility. In our experiment, the missing significances in the results indicate that the induced cognitive manipulation might have been too weak to affect the participants. We controlled for noticeability of the power outage by letting the participant fill out a free text area to describe what happened. Forty-three participants explicitly mentioned the power outage, which strengthens our assumption that the power outage was noticed by most of the participants. Due to the qualitative nature of this assessment, participants were free to write anything and might have noticed the power outage without mentioning it afterwards. However, the resulting effect might have been too weak to be reflected in the plausibility questions. To achieve higher indicative power, we want to use closed questions in the future to control for noticeability. Furthermore, we conclude that the plausibility questions are solely cognitive-oriented. They do not cover the whole body of plausibility, including the perception and sensation layer. Finding the right measurement tools to assess XR experiences entirely is of utmost importance.

To conclude, further studies need to be conducted to get more insights into the reference frame as a baseline for plausibility evaluations of MR experiences. Furthermore, appropriate measurement tools need to be developed to assess plausibility.

6.2 Presence

Hypothesis *H2* cannot be answered as we did not find any breaks in presence. We evaluated two questionnaires, the WSPQ and the IPQ. Both total scores did not become significant. These findings align with Brübach et al. [2]. When we examined the simple main effects of the cognitive manipulation within VR and AR, respectively, we did not find a significant effect for the overall presence. From the results, we now assume that there is no difference between VR and AR concerning presence which could (mis)lead to the interpretation that the AR experience has the same quality as the VR experience. However, looking at the plausibility and spatial presence ratings, we found an indication that there is indeed an impact from the induced and a priori incongruencies. Fitting these results into the model and notion of Slater [33], we manipulated PI with AR and VR and Psi with the cognitive manipulation. The combination of both manipulations should inevitably lead to an effect on the overall presence, which was not the case. This result strengthens the assumption that presence is not the sole responsible factor to take into account for the quality of MR experiences and motivates to have a look at other constructs, such as plausibility [15].

In addition, the lack of comparability between presence questionnaires is problematic. Even if the total scores of the IPQ and the WSPQ were equally non-significant, that does not mean they ask for the same construct because their initial definition of presence differs. This is already made evident by the different subscales: while the IPQ has the subscales *spatial presence*, *experienced realism*, and *involvement*, the WSPQ has the subscales *involvement*, *sensory fidelity*, *adaption/immersion*, and *interface quality*. The WSPQ subscale *sensory fidelity* had a significant main effect between the VR and AR scenarios which can be attributed to the quality of the physical audio sources. However, in the IPQ, audio quality was not assessed with any items. Additionally, we detected a significant main effect in the WSPQ subscale for involvement between CC and CI. In contrast the same named subscale involvement of the IPQ did not reveal a significant main effect between CC and CI. The IPQ involvement subscale focuses on the attention to and awareness/captivation of the real and virtual environment (e.g., "I still paid attention to the real environment."). The WSPQ involvement subscale captures a wider variety of items such as naturalness, control, responsiveness, and compellingness.

The spatial presence subscale of the IPQ showed a significant interaction effect, leading to the next hypotheses.

6.3 Spatial Presence

We cannot reject *H3*. Even though the IPQ subscale of *spatial presence* revealed a significant interaction effect, the simple main effects in the respective VR and AR factors were not significant. Therefore, we assume that in VR and AR, the cognitive manipulation did not affect spatial presence measured with the IPQ subscale *spatial presence*. Even though we could not find an effect of a cognitive incongruence on presence in our experiment, our results cannot be generalized for other forms of cognitive manipulations.

With regard to our specific manipulations, we have to reject *H4* since - again - the simple main effects within the CC and CI conditions did not show significance.

However, we found a significant interaction effect (see Fig. 6 (b)), in which the VR-CC scenario elicited a higher spatial presence than the VR-CI scenario. In the AR scenarios, this effect is inverted, meaning that the AR-CI scenario achieved a higher presence than the AR-CC scenario. These results might have arisen from (missing) adaption effects. As Latoschik and Wienrich [15] argue, on lower layers, users cannot easily adapt to a manipulation when it is continuously present (e.g., incongruencies effectuated by the use of AR). On higher layers, more plasticity is provided, and users adapt more easily to these kinds of manipulation, even if they are unrealistic. For example, in Brübach et al. [2], participants could adapt to the gravity manipulation, which did not affect the spatial presence rating.

Given the measured interaction effect, we assume that the cognitive manipulation even supported this effect of adaption/plasticity in our experiment. When the participants triggered a congruent power outage, *both* groups of entities, the physical and the virtual have undergone a change (visually by change of lighting, auditory by stopping the background music), making it even harder to adapt. The incongruent power outage only affected virtual content and, thus, might have been less of a spatial disturbance for participants.

Moreover, the interaction effect looks similar to the one found for plausibility (see Fig. 6). Thus, we assume that a possible reference frame also influences spatial presence and sets the baseline for this measurement.

Our findings concerning spatial presence fit into the CaP model, which proposes that incongruencies on different layers have different influences. Hofer et al. [12], and Brübach et al. [2] could not identify an effect of plausibility on spatial presence. This could be explained by the nature of the manipulation used in the experiments. They used cognitive manipulations, which - given the ordering of layers - cannot affect the bottom-up spatial presence. However, in our experiment, we manipulated the sensory/perceptual layer with the outcome of an effect on spatial presence (with the cognitive manipulation as an amplifier). Thus, the more basal the manipulation is, the stronger the resulting effect. In addition, we assume that the sensory/perceptual incongruence impacts the effectiveness of the cognitive congruence. This supports Latoschik and Wienrich's [15] statement that the bottom-up layers determine higher-order processes to a certain extent as well.

6.4 Implications

Our results support the implications from the CaP model that lower layers influence higher-order layers. Thus, we assume that a cognitive sense-making connection between virtual and physical entities can only be established when a certain amount of congruence is provided on the lower layers. This is of relevance along the whole RV continuum, where real and physical entities are combined to result in one experience. It is necessary to consider these influencing bottom-up factors in the design of MR experiences and corresponding experimental setups.

6.5 Limitations and Future Work

A limitation of this study was the fact that we could not rely on standardized questionnaires. We slightly modified the presence questionnaires to be applicable in AR as well. However, this could have changed the intention of some questions or left the participant confused (e.g., "In the shown environment, I had a sense of 'being there").

Our plausibility questions were inspired by another study [2]. We altered them from addressing the plausibility of object behavior to the plausibility of the scenario as a whole. Unfortunately, there are currently no standardized plausibility questionnaires since plausibility has not been evaluated extensively. Hence, we based our questionnaire on previous studies that attempted to measure plausibility and adapted their questions accordingly. Similar to our results, Brübach et al. [2] reported a tendency towards an interaction effect in two plausibility questions. These two questions, in their modified forms, revealed a significance in our experiment as well. Thus, we assume that at least these questions assess plausibility in a certain way and, thus, potentially are a good starting point to find the right measurement for plausibility. We explicitly acknowledge that the plausibility questions are not validated yet. As reported, we did find significant effects for some of the questions. However, we cannot say with certainty whether the differences in plausibility were too small to be mapped by the remaining questions or whether the questionnaire cannot assess plausibility as suspected. In addition, some single items had no equal group variance. The task participants had to perform may not have been part of their real-world knowledge. Thus, the external plausibility questions might have caused a lower plausibility independent from the (in)congruence of the scenario. In contrast, participants with much experience with this scenario (e.g., electricians) might have rated the plausibility very low because they notice the discrepancies from the real world more intensely.

In general, there is ongoing progress in the development of MR devices. Different quality aspects might influence the sensory/perceptual incongruence more or less. Thus, our results cannot be generalized for all types of HMDs. The severeness of the sensory/perceptual incongruencies might decrease more and more in the future as the quality of HMDs increases.

The visibility of the own body could have been a confound in our study. In AR, the own body was visible (as is naturally the case with the video-see-through functionality), while in VR, only hand models which moved according to hand movement were visible. Further studies shall control for this factor.

For future research, the reference frame shall be examined in more detail. With the AR and VR conditions, we decided to choose the more extreme points in the RV continuum. In this context, it would be fruitful to examine Augmented Virtuality (AV), located between AR and VR in the RV continuum, in future work. In our experiment, we addressed the visual and auditory senses. This interrelation of different senses in MR shall be investigated more intensively. For example, the set of senses could be extended by olfactory stimuli opening up the room for interesting research questions regarding multi-sensory (in)congruencies.

7 Conclusion

Recent advances in the development of XR and MR displays significantly increase the design space of potential interfaces and experiences along Milgram's Reality-Virtuality continuum, supporting a variety of different proportions of virtual and physical content. Therefore, examining the underlying constructs of how users perceive such MR experiences is crucial. This article presented and discussed current theoretical approaches to model and describe the qualities and effects of MR experiences.

We conducted a study to find out possible differences in perceived plausibility, presence, and spatial presence between VR and AR. Our conditions included a manipulation of congruence on the sensory/perceptual layer with the factors AR and VR and a manipulation of the congruence on the cognitive layer with the factors CC and CI.

We could identify a difference between the plausibility ratings of AR and VR, which might be connected to the recent discussion of a reference frame. We did not find an effect of our incongruencies on the overall presence, which motivates the establishment of a new measurement tool to evaluate MR experiences. Through the introduction of a sensory/perceptual manipulation (through the use of AR) we could find evidence that the corresponding processing layer is capable to affect spatial presence. Thus, we can corroborate that the assumptions about (in)congruencies on the three layers introduced in the CaP model are correct. We argue that congruence in the sensory/perceptual layer is important for establishing a sense-making connection of virtual and physical entities. Furthermore, we assume that the spatial presence might also be affected by a possible reference frame. Uncovering this underlying structure appears crucial for future MR experiences and should be subject to further research.

Acknowledgments

This research has been funded by the Bavarian State Ministry For Digital Affairs in the project XR Hub (project number A5-3822-2-16).

References

[1] M. J. Blanca, R. Alarcón, and J. Arnau. Non-normal data: Is ANOVA still a valid option? *Psicothema*, 29:552–557, Nov. 2017. doi: 10.7334/psicothema2016.383 6

[2] L. Brübach, F. Westermeier, C. Wienrich, and M. E. Latoschik. Breaking plausibility without breaking presence - Evidence for the multi-layer nature of plausibility. *IEEE Transactions on Visualization and Computer Graphics*, 28(5):2267–2276, May 2022. doi: 10.1109/TVCG.2022.3150496 2, 3, 4, 5, 6, 7, 8, 9

[3] R. Busselle and H. Bilandzic. Fictionality and perceived realism in experiencing stories: A model of narrative comprehension and engagement. *Communication Theory*, 18(2):255–280, Apr. 2008. doi: 10.1111/j.1468-2885.2008.00322.x 2

[4] M. Cavazza, J.-L. Lugrin, and M. Buehner. Causal perception in virtual reality and its implications for presence factors. *Presence: Teleoperators*

and Virtual Environments, 16(6):623–642, Dec. 2007. doi: 10.1162/pres.16.6.623 3, 5

[5] D. B. Chertoff, S. L. Schatz, R. McDaniel, and C. A. Bowers. Improving presence theory through experiential design. *Presence: Teleoperators and Virtual Environments*, 17(4):405–413, Aug. 2008. doi: 10.1162/pres.17.4.405 4

[6] J. Collins, H. Regenbrecht, and T. Langlotz. Visual coherence in mixed reality: A systematic enquiry. *Presence: Teleoperators and Virtual Environments*, 26:16–41, Feb. 2017. doi: 10.1162/PRES_a_00284 3

[7] L. Connell and M. T. Keane. A model of plausibility. *Cognitive Science*, 30(1):95–120, Jan. 2006. doi: 10.1207/s15516709cog0000_53 3

[8] N. Hamzeheinejad, D. Roth, S. Monty, J. Breuer, A. Rodenberg, and M. E. Latoschik. The impact of implicit and explicit feedback on performance and experience during VR-supported motor rehabilitation. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 382–391. IEEE, Mar. 2021. doi: 10.1109/VR50410.2021.00061 1

[9] S. G. Hart and L. E. Staveland. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In *Advances in Psychology*, vol. 52, pp. 139–183. Elsevier, 1988. doi: 10.1016/S0166-4115(08)62386-9 5

[10] M. R. Harwell, E. N. Rubinstein, W. S. Hayes, and C. C. Olds. Summarizing monte carlo results in methodological research: The one- and two-factor fixed effects ANOVA cases. *Journal of Educational Statistics*, 17(4):315–339, Dec. 1992. doi: 10.3102/10769986017004315 6

[11] R. Hein, C. Wienrich, and M. Latoschik. A systematic review of foreign language learning with immersive technologies (2001-2020). *AIMS Electronics and Electrical Engineering*, 5:117–145, Apr. 2021. doi: 10.3934/electreng.2021007 1

[12] M. Hofer, T. Hartmann, R. Ratan, and L. Hahn. The role of plausibility in the experience of spatial presence in virtual environments. *Frontiers in Virtual Reality*, 1, Apr. 2020. doi: 10.3389/frvir.2020.00002 2, 3, 4, 9

[13] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal. Simulator Sickness Questionnaire: An enhanced method for quantifying simulator sickness. *The International Journal of Aviation Psychology*, 3(3):203–220, July 1993. doi: 10.1207/s15327108ijap0303_3 5

[14] F. Kern, P. Kullmann, E. Ganal, K. Korwisi, R. Stingl, F. Niebling, and M. E. Latoschik. Off-the-shelf stylus: Using XR devices for handwriting and sketching on physically aligned virtual surfaces. *Frontiers in Virtual Reality*, 2, June 2021. doi: 10.3389/frvir.2021.684498 1

[15] M. E. Latoschik and C. Wienrich. Congruence and plausibility, not presence: Pivotal conditions for XR experiences and effects, a novel approach. *Frontiers in Virtual Reality*, 3, June 2022. 1, 2, 3, 8, 9

[16] M. Lombard and T. Ditton. At the heart of it all: The concept of presence. *Journal of Computer-Mediated Communication*, 3(2), Sept. 1997. doi: 10.1111/j.1083-6101.1997.tb00072.x 2

[17] D. Mal, E. Wolf, N. Döllinger, M. Botsch, C. Wienrich, and M. E. Latoschik. Virtual human coherence and plausibility – Towards a validated scale. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 788–789, Mar. 2022. doi: 10.1109/VRW55335.2022.00245 3

[18] Microsoft HoloLens | Mixed Reality technology for business, https://www.microsoft.com/hololens. 3, 5

[19] P. Milgram and F. Kishino. A taxonomy of Mixed Reality visual displays. *IEICE Transactions on Information Systems*, vol. E77-D(12):1321–1329, Dec. 1994. 1, 2

[20] M. Minsky. Telepresence. *OMNI magazine*, OMNI magazine:45–52, 1980. 1, 2

[21] M. Mori, K. F. MacDorman, and N. Kageki. The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, 19(2):98–100, June 2012. doi: 10.1109/MRA.2012.2192811 8

[22] N. C. Nilsson, R. Nordahl, and S. Serafin. Waiting for the ultimate display: IEEE 3rd workshop on everyday virtual reality. *2017 IEEE 3rd Workshop on Everyday Virtual Reality (WEVR)*, Mar. 2017. doi: 10.1109/WEVR.2017.7957710 8

[23] S. Oberdörfer, D. Heidrich, S. Birnstiel, and M. Latoschik. Enchanted by your surrounding? Measuring the effects of immersion and design of virtual environments on decision-making. *Frontiers in Virtual Reality*, 2, Aug. 2021. doi: 10.3389/frvir.2021.679277 4

[24] L. Plabst, S. Oberdörfer, O. Happel, and F. Niebling. Visualisation methods for patient monitoring in anaesthetic procedures using augmented reality. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology*, VRST '21, pp. 1–3. Association for Computing Machinery, New York, NY, USA, Dec. 2021. doi: 10.1145/3489849.3489908 1

[25] L. N. Popova. Perceived reality of media messages: Concept explication and testing. 2010. 2

[26] M. Pouke, K. J. Mimnaugh, A. P. Chambers, T. Ojala, and S. M. LaValle. The plausibility paradox for resized users in virtual environments. *Frontiers in Virtual Reality*, 2, Mar. 2021. doi: 10.3389/frvir.2021.655744 8

[27] E. Schmider, M. Ziegler, E. Danay, L. Beyer, and M. Buehner. Is it really robust? *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 6:147–151, Jan. 2010. doi: 10.1027/1614-2241/a000016 6

[28] T. Schubert, F. Friedmann, and H. Regenbrecht. The experience of presence: Factor analytic insights. *Presence: Teleoperators and Virtual Environments*, 10(3):266–281, June 2001. doi: 10.1162/105474601300343603 5

[29] R. Skarbez. Plausibility illusion in virtual environments. The University of North Carolina at Chapel Hill, 2016. 2

[30] R. Skarbez, F. P. Brooks, and M. C. Whitton. Immersion and coherence: Research agenda and early results. *IEEE Transactions on Visualization and Computer Graphics*, 27(10):3839–3850, Apr. 2020. doi: 10.1109/TVCG.2020.2983701 1

[31] R. Skarbez, F. P. Brooks, Jr., and M. C. Whitton. A survey of presence and related concepts. *ACM Computing Surveys*, 50(6):1–39, Nov. 2017. doi: 10.1145/3134301 1, 2

[32] R. Skarbez, M. Smith, and M. Whitton. Revisiting Milgram and Kishino's Reality-Virtuality continuum. *Frontiers in Virtual Reality*, 2, Mar. 2021. doi: 10.3389/frvir.2021.647997 2, 8

[33] M. Slater. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3549–3557, Dec. 2009. doi: 10.1098/rstb.2009.0138 1, 2, 4, 5, 8

[34] M. Slater, D. Banakou, A. Beacco, J. Gallego, F. Macia-Varela, and R. Oliva. A separate reality: An update on place illusion and plausibility in virtual reality. *Frontiers in Virtual Reality*, 3, 2022. doi: 10.3389/frvir.2022.914392 1, 2, 3

[35] M. Slater and A. Steed. A virtual presence counter. *Presence: Teleoperators and Virtual Environments*, 9(5):413–434, Oct. 2000. doi: 10.1162/105474600566925 3, 4

[36] M. Slater and S. Wilbur. A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, 6(6):603–616, Dec. 1997. doi: 10.1162/pres.1997.6.6.603 1

[37] Smart lighting, https://www.philips-hue.com, 2022. 4

[38] M. Teyssier. unity-hue, https://github.com/marcteys/unity-hue, 2022. 4

[39] Varjo XR-3 - The industry's highest resolution XR headset | Varjo, https://varjo.com/products/xr-3/, 2022. 1, 3, 5

[40] C. Wienrich, N. Döllinger, and R. Hein. Behavioral framework of immersive technologies (BehaveFIT): How and why virtual reality can support behavioral change processes. *Frontiers in Virtual Reality*, 2, June 2021. doi: 10.3389/frvir.2021.627194 1

[41] C. Wienrich, P. Komma, S. Vogt, and M. Latoschik. Spatial presence in mixed realities – Considerations about the concept, measures, design, and experiments. *Frontiers in Virtual Reality*, 2, Oct. 2021. doi: 10.3389/frvir.2021.694315 2, 8

[42] W. Wirth, T. Hartmann, S. Böcking, P. Vorderer, C. Klimmt, H. Schramm, T. Saari, J. Laarni, N. Ravaja, F. R. Gouveia, F. Biocca, A. Sacau, L. Jäncke, T. Baumgartner, and P. Jäncke. A process model of the formation of spatial presence experiences. *Media Psychology*, 9(3):493–525, May 2007. doi: 10.1080/15213260701283079 2

[43] B. G. Witmer and M. J. Singer. Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and Virtual Environments*, 7(3):225–240, June 1998. doi: 10.1162/105474698565686 5

[44] E. Wolf, D. Mal, V. Frohnapfel, N. Döllinger, S. Wenninger, M. Botsch, M. E. Latoschik, and C. Wienrich. Plausibility and perception of personalized virtual humans between virtual and augmented reality. pp. 489–498. IEEE, Oct. 2022. doi: 10.1109/ISMAR55827.2022.00065 3