Virtual Reality (2020) 24:783–796
https://doi.org/10.1007/s10055-019-00411-y

S.I. : XR (VR, AR, MR) AND IMMERSIVE LEARNING ENVIRONMENTS

Level of immersion affects spatial learning in virtual environments: results of a three-condition within-subjects study with long intersession intervals

Kimberly A. Pollard[1] · Ashley H. Oiknine[2,3] · Benjamin T. Files[1] · Anne M. Sinatra[4] · Debbie Patton[5] · Mark Ericson[6] · Jerald Thomas[7] · Peter Khooshabeh[1,2]

Received: 1 March 2019 / Accepted: 13 November 2019 / Published online: 28 February 2020

Abstract

Virtual reality and immersive technologies are used in a variety of learning and training applications. However, higher levels of immersion do not always improve learning. The mixed results in the literature may partly arise from the use of between-subjects designs, insufficient time intervals between sessions in within-subjects designs, and/or overreliance on binary comparisons of immersion levels. Our study examined the influence of three levels of audiovisual immersive technology on spatial learning in virtual environments, using a within-subjects design with long intersession intervals. Performance on object recognition and discrimination was improved in the highest immersion condition, whereas performance on directional bearings showed a U-shaped relationship with level of immersion. Examination of our data suggests that these results likely would not have been found had we used a between-subjects design or a binary comparison, thus demonstrating the value of our approach. Results suggest that different levels of immersion may be better suited to more or less cognitively complex types of spatial learning. We discuss challenges and opportunities for future work.

Keywords Spatial learning · Virtual reality · Longitudinal design · Immersion · Head-mounted display · Spatial audio

✉ Kimberly A. Pollard
kimberly.a.pollard.civ@mail.mil

Ashley H. Oiknine
aoiknine@dcscorp.com

Benjamin T. Files
benjamin.t.files.civ@mail.mil

Anne M. Sinatra
anne.m.sinatra.civ@mail.mil

Debbie Patton
debra.j.patton4.civ@mail.mil

Mark Ericson
mark.a.ericson.civ@mail.mil

Jerald Thomas
thoma891@d.umn.edu

Peter Khooshabeh
peter.khooshabehadeh2.civ@mail.mil

[1] Combat Capabilities Development Command Army Research Laboratory, Los Angeles, USA

[2] Department of Psychological and Brain Sciences, University of California, Santa Barbara, USA

[3] DCS Corporation, Los Angeles, USA

[4] Combat Capabilities Development Command Soldier Center – Simulation and Training Technology Center, Orlando, USA

[5] Combat Capabilities Development Command Data and Analysis Center, Aberdeen Proving Ground, USA

[6] Combat Capabilities Development Command Army Research Laboratory, Aberdeen Proving Ground, USA

[7] Department of Computer Science, University of Minnesota, Minneapolis, USA

1 Introduction

Virtual reality (VR) has become increasingly popular in the context of education and training (Johnson et al. 2016). Virtual environments and highly immersive head-mounted displays (HMDs) can be found in a slew of learning and training domains including undergraduate classrooms (Moreno and Mayer 2004; Parong and Mayer 2018), special needs education (Lányi et al. 2006; Jeffs 2010), environmental education (Markowitz et al. 2018), mindfulness training (Chandrasiri et al. 2019), construction (Wang et al. 2004; Jeelani et al. 2017), first responders (Stansfield et al. 1998; Mossel et al. 2017; Carlson and Caporusso 2019), sports (Huang et al. 2015), military (Bhagat et al. 2016; Khooshabeh et al. 2017), and law enforcement (Wei et al. 2018; Carlson and Caporusso 2019).

The promise of such technology is well known: Immersive VR technologies can provide realistic experiences while avoiding dangers or expenses of real-world training scenarios. It is also commonly believed that more advanced immersive technology should improve the effectiveness of learning in simulated environments (Andre and Wickens 1995; Bowman and McMahan 2007; Summers 2012; Cummings and Bailenson 2016; Brown 2016). This promise, along with the inherent allure of cutting-edge technology, has led to expensive investments by military, civil services, and industry in advanced HMDs, CAVE systems, and simulated virtual environments (Lewis 2017; Sintia 2018; Horowitz 2018), without compelling evidence of positive outcomes (Government Accountability Office 2016). However, students and trainees do not always see a learning benefit from highly immersive technology (Alexander et al. 2005). Learning performance in immersive VR has a mixed track record (for reviews, see McMahan et al. 2012; Cummings and Bailenson 2016).

Immersion is defined as a function or attribute of the technology itself (Regan 1995; Slater and Wilbur 1997; Cummings and Bailenson 2016) and takes into account the collection of immersive features (such as visual occlusion, degrees of freedom of head rotation, spatialized sound, etc.) offered by that technology (Cummings and Bailenson 2016). Higher levels of immersion sometimes do not improve learning performance (e.g., Moreno and Mayer 2000; Craig et al. 2002; Picciano 2002; Baylor et al. 2003; Frechette and Moreno 2010; Stevens et al. 2015). Sometimes, they even decrease performance (Mania and Chalmers 2001; Walker et al. 2009; Taylor and Barnett 2013; Stevens et al. 2015). Occasionally, we find the intuitively paradoxical result of trainees preferring immersive technologies that yield worse performance (Andre and Wickens 1995). It is important to understand these effects so that programs can better maximize learning while avoiding unnecessary expenditures.

1.1 Why might immersive technology influence learning?

Implementing a learning program in immersive technology, such as a VR headset, is an example of using an enhanced training component. Enhanced training components are "extras" that can be added to training programs with the intent of improving learning by increasing motivation, attention, engagement, enjoyment, or feelings of presence (e.g., Moreno and Mayer 2004; Lee and Nass 2005; von der Pütten et al. 2009; Kasap and Magnenat-Thalmann 2012; Files et al. 2019b). This in turn may serve to keep learners focused, to hold their interest, to make them feel invested and present in the scenarios, and to motivate them to invest sufficient time-on-task (Alexander et al. 2005; Landers and Landers 2014).

Immersion fits into the taxonomy of enhanced, gamified learning interventions as defined by Landers (2014). As per Landers' model (2014), immersive technology does not replace instructional content. Rather, the add-on serves to influence the behavior and attitudes of the learner, which in turn can moderate the impact of the instructional content on learning outcomes.

The use of enhanced training components can improve performance and training transfer (e.g., Waller et al. 1998; Moreno and Mayer 2000, 2004; Wang et al. 2008). Such interventions can also backfire (e.g., Mania and Chalmers 2001; Walker et al. 2009; Taylor and Barnett 2013; Stevens et al. 2015). The overall track record in the literature is mixed (for review, see McMahan et al. 2012; Cummings and Bailenson 2016).

1.2 Could between-subjects designs help explain the mixed track record?

The lack of consistency in these results may be due to a variety of factors, some of which may have to do with study design. Studies of immersive VR in learning often utilize a between-subjects design (e.g., Moreno and Mayer 2004; Mania et al. 2006; Tse et al. 2017; Cho 2018; Parong and Mayer 2018), wherein each research subject experiences only one type of technology or device setting and the results are then compared across groups. Individual differences in performance or response add noise to the data, potentially obscuring results. A within-subjects design, wherein each research subject experiences multiple conditions, would largely control for these individual differences but is difficult to implement (but see Mizell et al. 2002; Sousa Santos et al. 2009; Patton 2014; Christou et al. 2016; Patton and Gamble 2016; Shu et al. 2018 for some examples).

Executing a within-subjects design comes with a variety of challenges, such as adaptation and other order effects, plus attrition of participants and additional logistical difficulties, especially if intersession intervals are long (e.g., increased participant tracking, scheduling). When within-subjects designs are used in VR research, particularly with HMDs, allowing subjects enough time to return fully to baseline requires them to return to the laboratory on a separate day, or better yet after several days or weeks (Sharples et al. 2008; Moss et al. 2011; Wilson 2016), but studies often use much shorter intervals. For example, participants are often offered "short breaks" between displays, about enough time to set up the next display (Swindells et al. 2004), offered no time at all (Sousa Santos et al. 2009), or given a 5-min "washout" period (Tong et al. 2016). Carryover effects may further obscure effects of the conditions tested. A long intersession interval not only presents a logistical and scheduling challenge, but it exacerbates attrition, as subjects who come to an experimental session may not return for all subsequent sessions. Attrition also frequently implies increased recruitment costs, compensation costs, and potentially more time needed to complete data collection. It is likely that these various difficulties explain why within-subject designs with sufficient inter-condition intervals are rarer in the VR learning literature.

1.3 Could number of conditions help explain the mixed track record?

Another potential reason for mixed results in immersion versus learning literature may be due to the number of conditions used. Many studies compare only two conditions, a "high" and a "low" immersive technology condition (e.g., Moreno and Mayer 2004; Sousa Santos et al. 2009; Hsieh et al. 2018; Krokos et al. 2018; Shu et al. 2018). This approach allows comparison between the two levels used but limits the ability to assess the shape of more complex response curves to levels of immersive technology. For example, if the true participant response to levels of immersive technology plateaus after a certain point, or if it follows a diminishing returns or U-shaped curve, a high versus low design would be unable to detect this shape. Furthermore, a high versus low design may fail to uncover effects at all if, for example, the low and high levels chosen happen to fall on either side of an undiscovered U. Hypothetically, if three different research teams performed binary comparison studies of points that happened to be sampled from an undiscovered U-shaped curve, these three teams could come to *three entirely different conclusions* about the relationship between immersion and learning. The highly contradictory results across the literature may partly stem from different teams choosing different pairs of points along what may be a complex curve function. Looking at three or more levels of immersive technology in the same study makes this occurrence less likely, but it also increases the logistical challenges of designing and running the study. Elucidating a curve response may help reveal why studies of immersion levels versus learning performance have shown such mixed results.

To better understand the relationship between immersive VR and learning performance, we conducted a study of three different levels of immersive technology in a training scenario for spatial learning, using a within-subjects design with long intersession intervals (≥ 14 days).

2 Methods

2.1 Equipment

We examined three different levels of immersive technology, including aspects of both visual and auditory immersion.

The Low level condition featured a Dell Ultra Sharp 24" Desktop Monitor and Dell OEM AX210 2.0 2-piece USB powered desktop speakers. The Medium level used a partially occlusive, mid-grade HMD (nVisor ST50, NVIS, Reston, VA, hereafter called "NVIS") fitted with an InterSense InertiaCube4 system, and supra-aural headphones (Sony MDR-G45). The High level used a fully occlusive HMD (Oculus Rift CV1, Facebook Technologies, Menlo Park, CA) and Audio-Technica ATH-M50x circumaural headphones (Fig. 1).

Additionally, audio playback algorithms used in the Low, Medium, and High conditions differed. The spatial audio cues used in the three auditory immersion levels were created using Oculus' 3D audio spatialization effects from their 3D audio spatialization Software Development Kit. The Low level was created by playing only distance-based acoustic intensity cues. The Medium level was rendered using distance cues as well as directional head-related transfer function cues with head tracking. The High level used distance cues, directional head-motion-tracked sounds, and room acoustic cues. The room acoustic cues were chosen to match

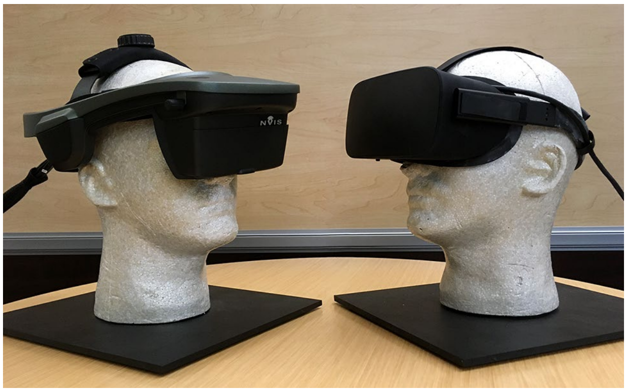


Fig. 1 Head-mounted displays (HMDs) used in this study: nVisor ST50 (NVIS, left) and Oculus Rift CV1 (right)

the smooth, hard surfaces of the virtual spaces. Low acoustic absorption values of ten percent and low scattering values of five percent were implemented for the virtual environments.

In all conditions, participants used an Xbox ONE game controller (wired connection) to traverse horizontal planar distance (i.e., "walk") in the virtual environments. The virtual study environments were large, containing multiple connected rooms. This necessitated a navigational method that could accommodate smooth real-time audiovisual traversal within, through, and between rooms in a large environment. Neither the desktop monitor, nor the NVIS, nor the Rift HMD has native functionality for significant distance traversal of this sort, so a peripheral (game controller) was used for this purpose. This setup is representative of what a typical classroom, military schoolhouse, or recreational gamer would utilize to explore large virtual environments in VR.

View rotation occurred via game controller in the Low condition, via game controller or head movement in the Medium condition, and via game controller or head movement in the High condition. All participants could, and did, physically rotate, tilt, swivel, crane, etc., their heads and body to change their gaze in the Medium and High conditions. The full 360° of motion and additional degrees of freedom movement affordances of the NVIS and Oculus Rift were enabled using appropriate software code and drivers to present the environments in these display devices and use their unique affordances.

An ASUSTek model G752VS gaming laptop running Windows 10 Pro, with NVIDIA Geforce GTX 1070 GPU, was used for all experiment runs. Questionnaires were implemented on a Windows Surface Pro 3 tablet running Windows 10, using Qualtrics software.

2.2 Virtual environments and scavenger hunt task

Virtual environments were created in Unity 3D, with additional support scripts in Python. Virtual environments used in this study included one simple controls-familiarization environment, three "mini" environments used for practice, and three larger "main" environments which served as the test environments. The controls-familiarization environment was a plain 3D space with a grid, populated with squares and spheres on which the participant could practice moving and targeting. The Mini environments consisted of small, four-room simulated indoor environments populated with theme-relevant objects. Themes included (1) History museum, (2) Recreation center, and (3) Holiday rooms. The mini environments were all the same size, had the same number of rooms (four), and had the same number of doorways (three doorways). The spatial orientation of rooms differed for each mini environment, as did the objects populating the environments. The Main environments consisted of larger simulated indoor environments populated by theme-relevant objects (Fig. 2). These themes were (1) Home, (2) Office, and (3) School. The main environments were all the same size and had the same number of doorways (14). The spatial orientation of rooms differed for each main environment, as did the objects populating the environments. Each main environment had an equal number of rooms (13), an equal number of scavenger hunt items (8), and a similar number of incidental (nonscavenger hunt) items. For realism, we included audio in the environments. Six objects in each main environment had an audio component (e.g., a ringing telephone). Three of these audio objects were included in the scavenger hunt; three were not.

The task scenario in our study contains a primary mission (connecting with a sequence of objects; i.e., an ordered scavenger hunt) followed by questions probing about observed objects and their spatial relationships, as recalled from the experience in the environment during the task. An ordered scavenger hunt with subsequent recall questions was selected as the main task for two reasons: (1) it forced participants to experience every room in the environment, and (2) this task is representative of spatial tasks involved in many real-world behaviors and training scenarios. Examples include search and rescue operations, bomb identification and defusing, building clearing, contraband discovery and securing, or simply a person exploring a new city or school to accomplish their goals. For example, a first responder must search a disaster environment for survivors, must address them in triage order, and must later be able to recall the layout of the environment and key objects within it to facilitate the ingress of subsequent personnel and to facilitate the process of care and evacuation. Similarly, a newly arrived freshman must identify the locations of key elements in her dorm (dining hall,

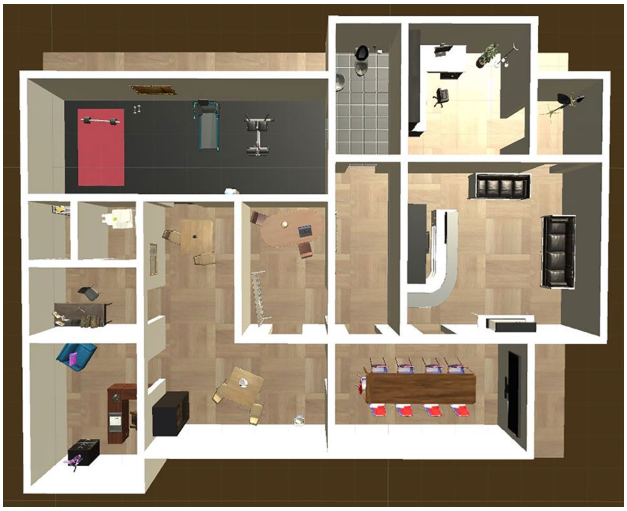


Fig. 2 Top-down view of the Office-themed main environment

mail boxes, laundry room, etc.) and must connect with them in sequential order to accomplish her time-sensitive goals (e.g., eating before laundry). She would later need to recall the layout and locations of objects to accomplish future goals (e.g., Where is the recycling bin?)

The main environments were the test environments, so great care was given to ensuring that the difficulty of these environments, and the scavenger hunts within them, was equivalent. Main environment floorplans were created by taking a single floorplan, cutting it into pieces, and rearranging the pieces to form three floorplans of identical size and identical individual room proportions. Doors were placed to provide as close as possible to equivalent connectivity between rooms as measured by graph theoretic metrics, and scavenger hunt items were ordered to allow as close as possible to equivalent minimum path lengths for scavenger hunt completion (Files et al. 2019a). These efforts were successful; time to complete the scavenger hunts did not significantly differ across the three main environments (Files et al. 2019a), suggesting that their difficulty was well balanced.

2.3 Participants

Participants were recruited through online platforms associated with the University of California, Santa Barbara community, including the Psychology Department's paid participant pool. Inclusion criteria included being at least 18 years of age and having normal hearing, normal color vision, and normal visual acuity (with or without contact lenses). Participants were screened for normal, symmetrical hearing using a MAICO MA 40 portable audiometer, for visual acuity using a traditional Snellen acuity chart, and for color vision using the 14-plate Ishihara's Test for Color Deficiency. Participants were screened for contraindicated conditions (such as motion sickness susceptibility, alcohol influence, illness, or pregnancy) via self-report and excluded if they reported such conditions. Participants were also excluded if they reported more than 11 h of experience with HMDs or VR, as we wished to ensure that no participants had any advantage of exposure to VR and/or retained potential adaptations to VR. Sixty-one participants (43 F, 18 M, ages 19–29) completed all three sessions of the experiment.

2.4 Research design

This study sought to examine the influence of three different levels of immersion (Low, Medium, and High) on spatial learning in virtual environments. Each participant experienced all three conditions. The independent variable was the level of immersive technology used. The outcome variables were performance on after-environment transfer tasks (yes/no object recognition, multiple-choice object discrimination vs. foil objects, and bearings estimation [i.e., angular distance estimation, headings estimation]). The level of immersive technology, the exact virtual environment, and the order of presentation were counterbalanced. Participants took a break of at least two weeks between each experimental session.

2.5 Experimental procedure

Testing took place at the University of California, Santa Barbara. Participants took part in three experimental sessions, each separated by at least two weeks to minimize any carryover effects (Moss et al. 2011). In each session, a participant provided written informed consent, was tested for normal hearing and vision (initial session only), and filled out self-report questionnaires on a tablet computer. The participant was then handed a game controller and was set up with one of three display technologies (desktop monitor with desktop speakers, NVIS HMD with supra-aural headphones, or Oculus Rift HMD with circumaural headphones). The order of display type was counterbalanced. This was followed by a brief controls-familiarization session in which the participant was asked to use the game controller buttons (and HMD head motions, if applicable) to view, move to, and select objects in a simplified 3D virtual environment. After this, participants practiced navigating in a "mini" environment to further familiarize them with the game controller, display technology, scavenger hunt concept, and types of assessment questions that would later be asked. Participants remained in the mini environment for 5 min and were instructed to complete a simple four-item scavenger hunt followed by free exploration. Scavenger hunt items were marked with numbered flags, and participants were instructed to find each number in serial order and to press a button when close to each object to mark each one as found. Numbers were used instead of object names to ensure that when participants engaged with the assessment questions, they would be using their memory of their experience in the virtual environment rather than any given semantic information. Participants were then removed from the display hardware and asked to answer a set of practice questions of the types they would later be asked after the main environment. Participants then entered one of the three "main" virtual environments (order counterbalanced) and were asked to complete a scavenger hunt for eight objects in order (marked by numbered flags), following instructions on the screen.

Scavenger hunt items were chosen to avoid duplication across environments. For example, if a towel was a scavenger hunt object in one environment, we ensured that (a) no other environment had the same towel and (b) no towels served as scavenger hunt items in the other environments.

In order to ensure that distinct questions could be asked about each environment, the designation of scavenger hunt items was tightly coupled with the creation of the post-environment assessment questions (Sect. 2.6).

Participants spent a total of 15 min in each main virtual environment. When they completed the scavenger hunt, participants were asked to freely explore the virtual environment for the remainder of their 15 min. After exiting the virtual environment and removing the display technology, the participants answered post-environment assessment questions (transfer tasks) to demonstrate spatial learning in the virtual environment. Participants returned after a break of at least 2 weeks (Moss et al. 2011) and then engaged in another session of the experiment with a different display technology and a different environment and scavenger hunt. The technology, environments, and order of presentation were counterbalanced.

Participants were compensated for each experimental session they completed. In addition, participants who completed all three sessions received an additional bonus payment.

2.6 Learning assessment and transfer

A series of learning assessments occurred after engaging with the environments. The learning assessments involved post-environment quiz questions (i.e., knowledge transfer tasks) aimed at assessing the participants' spatial learning from their immersive experience.

The questions were based on models of levels of spatial knowledge (Siegel and White 1975; Thorndyke and Hayes-Roth 1982; Stern and Leiser 1988; Darken and Peterson 2002), which consider object/landmark learning as among the cognitively simplest forms of spatial knowledge and consider tasks that require integration of more complicated survey-level spatial information as being the most cognitively complex. Our tests included yes/no object recognition questions, multiple-choice object recognition questions, and bearings estimation questions.

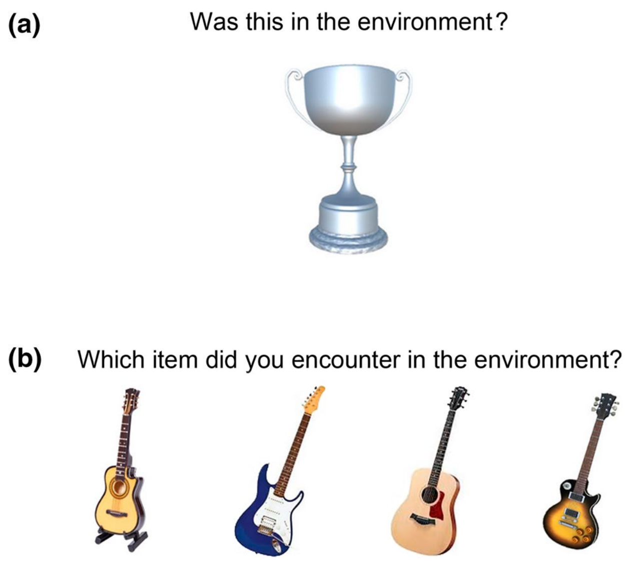
In a yes/no object recognition question, the participant was presented with an image of an object that could plausibly have been in the virtual environment and was asked to indicate whether they had observed that object in the environment (Fig. 3a). Yes/no object recognition questions have been used in a variety of spatial learning studies to assess landmark-level spatial knowledge (see Taillade et al. 2013; van der Ham et al. 2015; Zhong and Moffat 2016; Kraemer et al. 2017 for recent examples). In a multiple-choice object recognition question, an image of an object from the virtual environment was displayed along with images of three "foil" objects which could plausibly have been in the environment but were not (Fig. 3b). A similar test has been used previously (e.g., Nys et al. 2015; Boccia et al. 2017). For the bearings test, participants were presented with a circular diagram with clickable tick marks at six-degree intervals. Participants were instructed to imagine they were standing on one object and facing a different object. They were then asked to indicate the direction in which a specified third object could be found (Fig. 4). The format of the bearings questions was modeled after Ragan et al. (2017), who used a circle and tick marks with the position object in the center, the facing object above the circle, and the target object pictured within the instruction sentence. We made modest tweaks to the tick marks and the wording of the instruction sentence used by Ragan et al. (2017) and implemented the questions in an electronic survey platform (Oiknine et al. 2019). Similar bearings tests were used by Witmer et al. (1995), van der Ham et al. (2015), Zhong and Moffat (2016), and Kraemer et al. (2017) to assess survey-level spatial knowledge.

The types of questions were presented in an order that was intended to test increasing levels of cognitive processing without influencing performance by "giving away" what was in the environment. We avoided showing or mentioning information in quiz questions that could be used to help participants answer subsequent questions. For example, bearings questions included images of objects that were definitely in the environment. We therefore presented these questions after the yes/no and multiple-choice object recognition questions to ensure that participants answered the recognition questions based only on their remembered experiences from the environment, not based on what they may have seen in a previous quiz question. This approach ensured that participants were answering questions based on their direct experience in the virtual environment, not from what they saw in previous questions.

The after-environment questions serve as assessment of learning of the environment and as examples of knowledge transfer tasks. Near transfer may be defined as performing a task that is highly similar to the task on which a person was trained. Farther forms of transfer may be defined as tasks that require the use of knowledge gained during training to perform a largely new and different task. We see the range of near-to-far transfer as a continuum.

As very near transfer and low spatial knowledge complexity (landmark knowledge), yes/no recognition questions addressed participants' learning of objects in the environment. Identifying objects was fairly pertinent to and largely unavoidable in the environment exploration task. Identifying objects again in the quiz is therefore a rather near form of transfer.

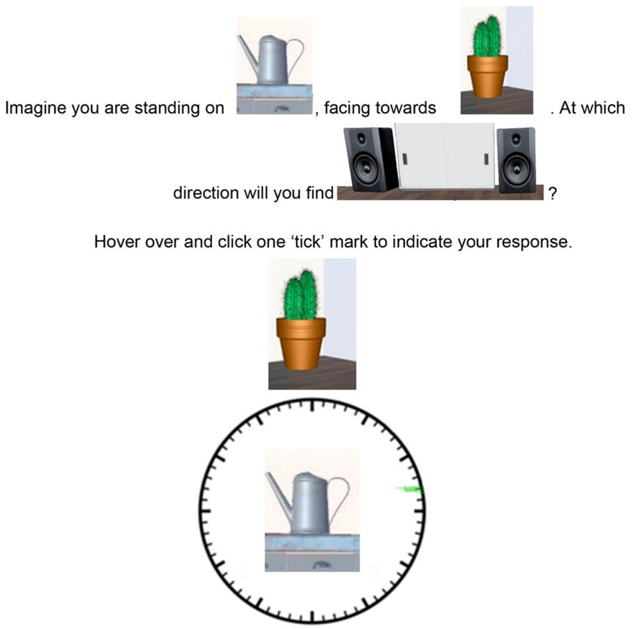
The multiple-choice object recognition/discrimination test we would consider to be a slightly farther example of transfer (albeit still fairly near). Here, the participant needs to identify which item they encountered and furthermore must recall sufficient detail to discriminate the object from



Was this in the environment?

Which item did you encounter in the environment?

Fig. 3 Example object recognition questions in **a** yes/no format and **b** multiple-choice format



Imagine you are standing on , facing towards . At which direction will you find ?

Hover over and click one ‘tick’ mark to indicate your response.

Fig. 4 An example of a bearings question, showing the participant’s selected tick mark

three foils. Discriminating among similar objects was not part of the original task. It is application of knowledge to a new task.

As significantly farther transfer, and requiring significantly more complex spatial processing (i.e., survey-level spatial knowledge), bearings questions addressed the participants’ understanding of the relative positions of objects in the environment. There was little need to calculate bearings during the actual environment exploration, and estimating a bearing direction (angular distance) required significant post-processing of the experience in the environment to perform. To answer a bearings question correctly, a participant had to: remember multiple objects, remember their locations, mentally reconstruct their relative spatial placement, hold all this in short-term memory, and then mentally construct a novel visual perspective. The bearings questions are thus a means to assess whether the participant learned enough from exploring the environments that they could apply that knowledge to a rather different context (i.e., farther transfer). The process of answering a bearing question involves cognitively intensive activities, considerably more so than simple object recognition and discrimination (Siegel and White 1975).

The difficulty of transfer questions was balanced across environments. For example, multiple-choice questions in each environment included 2 where the foils were schematically similar to the correct answer (e.g., all things related to exercise), 2 where the foils were functionally similar (e.g., all things that keep time), 2 where foils differed in stylistic details (e.g., phones of slightly different design), and 2 where foils differed only by color. Images used as foils in the recognition questions were carefully selected to be 3D models of similar visual/polygon quality as the actual objects from the environment, and all were displayed in an isolated manner with no environmental background present (Sinatra et al. 2019). Bearings questions for each environment were similarly difficulty balanced in terms of the distance between the objects and the number of items that had been targets versus nontargets in the scavenger hunt (Sinatra et al. 2019).

Participants were asked ten yes/no recognition questions per session, eight multiple-choice recognition questions per session, and four bearings questions per session. Due to a computer error, 15 sessions (out of 183 total sessions) included only seven multiple-choice recognition questions. Results on yes/no recognition and multiple-choice recognition were scored as proportions correct (the number correct out of the number presented). Results on the bearings questions were scored as the absolute angular error of the bearings (in degrees), summed across the bearings questions.

2.7 Analysis

The goal of this study was to compare learning performance in the main environments across different levels of immersive technology. The independent variable of interest was level of immersive technology (Low, Medium, High). Outcome variables included scores on yes/no object recognition questions (proportion correct out of questions given), scores on recognition multiple-choice questions (proportion correct out of questions given), and scores on bearings (sum of all absolute errors in directional bearings). Analyses were performed in SPSS (IBM, Armonk, NY) using repeated-measures design. As the outcome variables were not normally distributed, nonparametric approaches were used. We controlled for multiple comparisons using the Benjamini–Hochberg false discovery rate (FDR) method (Benjamini and Hochberg 1995), using an FDR of 0.05.

To compare the utility of a within-subjects versus a between-subjects design, we also performed a second analysis. In this second analysis, we examined our data as if it had been acquired under a between-subjects design. We did this by including only data from each participant's first session. This raised our sample size to 73 (50 F, 23 M, ages 19–33), because we were able to include participants that were lost to attrition in the within-subjects design. Independent and dependent variables were as above, and nonparametric independent-samples statistics were used to analyze these data.

3 Results

Performance on the simplest assessment question type (yes/no recognition), the more challenging question type (multiple-choice recognition/discrimination), and the most challenging type (bearings questions) was compared across levels of immersive technology to determine which levels of immersion provided greatest spatial learning of the main virtual environments. Level of immersive technology significantly influenced results on yes/no and multiple-choice object recognition tests (Friedman tests, $\chi^2(2) = 15.367$, $p < 0.001$ for yes/no; $\chi^2(2) = 6.612$, $p = 0.037$ for multiple choice). Post hoc analyses revealed that participants performed significantly better on yes/no object recognition questions in the High-immersion condition than they did in the Medium and Low conditions (Wilcoxon signed-rank tests, $Z = -0.588$, $p = 0.556$ for Low versus Medium; $Z = -2.547$, $p = 0.011$ for Low vs. High; $Z = -3.380$, $p = 0.001$ for Medium vs. High). Participants also performed significantly better on multiple-choice recognition questions in the High-immersion condition than they did in the Medium and Low conditions (Wilcoxon signed-rank tests, $Z = -0.248$, $p = 0.804$ for Low vs. Medium; $Z = -2.327$, $p = 0.020$ for Low vs. High; $Z = -2.775$, $p = 0.006$ for Medium vs. High). Figure 5 depicts the level of immersive technology versus learning performance in object recognition in yes/no questions and in multiple-choice questions.

For bearings questions, a U-shaped response curve was observed. Level of immersive technology significantly influenced the results on bearings questions (Friedman test, $\chi^2(2) = 6.426$, $p = 0.040$). Post hoc analyses revealed that participants performed significantly worse at bearings in the Medium condition than in the Low and High conditions (Wilcoxon signed-rank tests, Z=-2.155, $p = 0.031$ for Low vs. Medium; $Z = -0.273$, $p = 0.785$ for Low vs. High; $Z = -2.557$, $p = 0.011$ for Medium vs. High). Figure 6 shows the level of immersive technology versus errors on bearings questions (mean sum of all absolute errors on bearings questions). Because this outcome variable measures errors made, greater values of the variable indicate worse performance. Note the y-axis of Fig. 6 is flipped to more intuitively visually convey differences in performance.

All significant results remained significant after controlling for false discovery rate (Benjamini and Hochberg 1995).

In order to approximate the results we might have gotten had we used a between-subjects design, we created a dataset that included each participant's first visit only, including those participants we lost to attrition, and ran between-subjects (i.e., independent samples) analyses on these data.

Analyzing the data as if we had used a between-subjects design failed to uncover many of the above results. A Kruskal–Wallis test was conducted to examine between-subjects ($N = 73$; 24 in Low condition, 23 in Medium condition, 26 in High condition) differences in transfer test scores based on level of immersive technology used. No significant effects were found for bearings ($\chi^2(2) = 0.042$, $p = 0.979$) or for multiple-choice recognition scores ($\chi^2(2) = 0.759$, $p = 0.682$) using the between-subjects approach. Only the effects of level of immersive technology on yes/no object recognition remained significant under the approximated between-subjects design ($\chi^2(2) = 6.690$, $p = 0.035$). Post hoc Mann–Whitney U analyses indicated higher yes/no recognition scores for High ($Mdn = 0.80$) than for Medium ($Mdn = 0.60$) immersion levels ($U = 167.5$, $p = 0.007$). Scores for Low ($Mdn = 0.70$) and Medium conditions did not differ ($U = 223$, $p = 0.251$), nor did High vs. Low conditions ($U = 250.5$, $p = 0.223$) under the between-subjects approach. Significant values remained significant after controlling for false discovery rate (Benjamini and Hochberg 1995).

4 Discussion

A within-subjects approach with long intersession intervals was used to examine the effects of three levels of immersive audiovisual technology on spatial learning. Our goal was to examine spatial learning performance as a function of level of immersion considered in a holistic way, with focus on the overall level of immersive experience we expected the technology packages to impart. We therefore created Low, Medium, and High conditions which were each composites of multiple low-, medium-, and high-level immersive features, respectively. For example, the high condition brought together our highest fidelity audio processing algorithm, highest quality audio headphones, most occlusive audio headphones, and most occlusive HMD. Our design thus does not allow us to tease apart which specific individual aspects of immersion were driving the results. Although we speculate here, detailed isolation of relevant variables is left to other studies.

4.1 Benefit of high immersion for object recognition

Our results suggest that the highest immersion level (“High”) provided better learning performance on the simple yes/no recognition questions as well as the slightly more challenging multiple-choice recognition questions. The former type assessed whether participants remembered encountering the objects, whereas the latter assessed whether participants recalled objects in sufficient detail to distinguish them from similar “foil” objects. In both cases, using High immersion improved object recognition. Previous studies have found higher immersion levels to positively impact recall (Moreno and Mayer 2004; Bowman et al. 2009; Krokos et al. 2018). It may be tempting to hypothesize that these results are due to the increased visual affordances provided by the Oculus Rift HMD used in the High condition compared to the other conditions. In the High (Rift) condition, head tracking allowed participants to tilt and crane their heads, as they would in the real world, to view objects in the environment from multiple angles. While multiple-angle viewing was also possible in the desktop condition, it required joystick manipulation to do so. The increased naturalness of the movement may explain the Rift’s benefit over the desktop.

However, the NVIS used in the Medium condition also provided head tracking and natural perspective taking via head movement, yet was not associated with greater object recognition memory performance. This could be because natural viewing angle affordances are not important to object recognition memory, or because the Medium condition had other features that degraded performance despite natural viewing angles. One key parameter may be occlusion. While wearing the NVIS, the user still retains considerable peripheral vision of the external real world, and this may be sufficiently distracting to hamper object recognition encoding from the virtual world. Auditory occlusion also varied across our conditions. The supra-aural headphones of the Medium condition were only partially occlusive, compared to the more occlusive circumaural headphones used in the High condition. There may have been sufficient real-world audiovisual distraction in the Low and Medium conditions to interfere with performance; this suggests that full occlusion may be required to achieve object recognition benefit.

Another factor may be physical comfort. The NVIS is a significantly older HMD, with a bulkier, heavier design involving less soft cushioning and less precise fit adjustability. These factors may have been distracting to participants

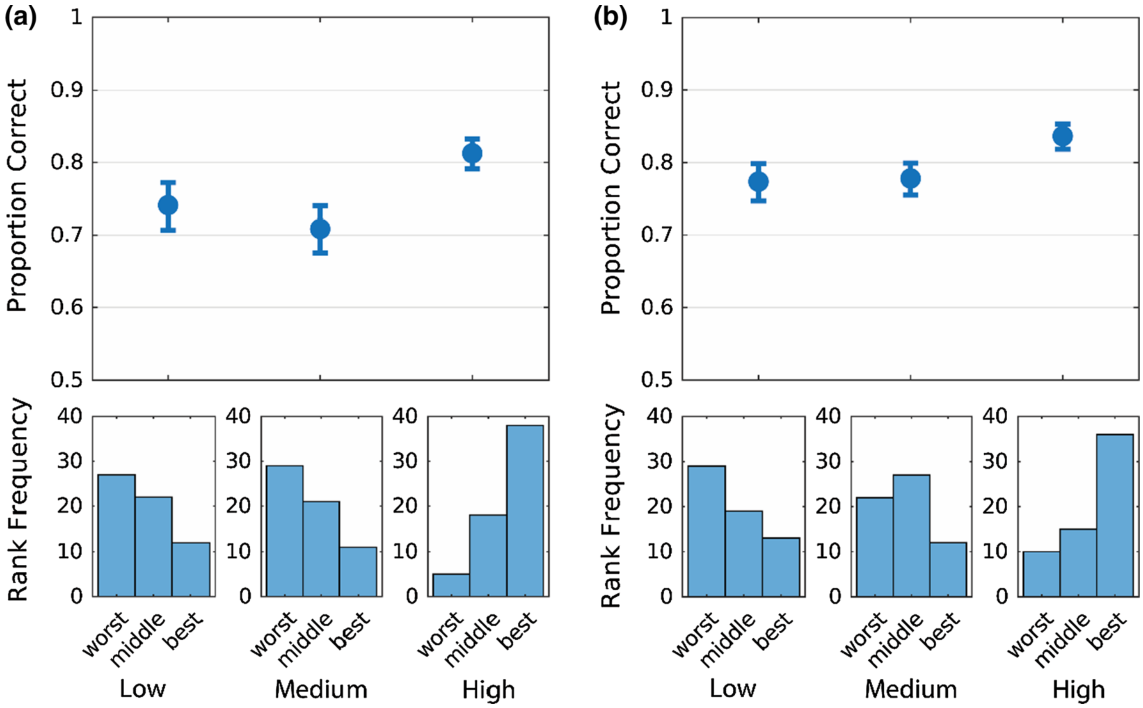


Fig. 5 Level of immersive technology versus learning performance in object recognition, **a** yes/no questions and **b** multiple-choice questions. Upper panels show mean accuracy for recognition questions. Error bars show 1 within-subject standard error of the mean (Morey 2008) computed in logit space (Warton and Hui 2011). Lower panels indicate how many participants for whom a given technology yielded the best, middle, or worst performance

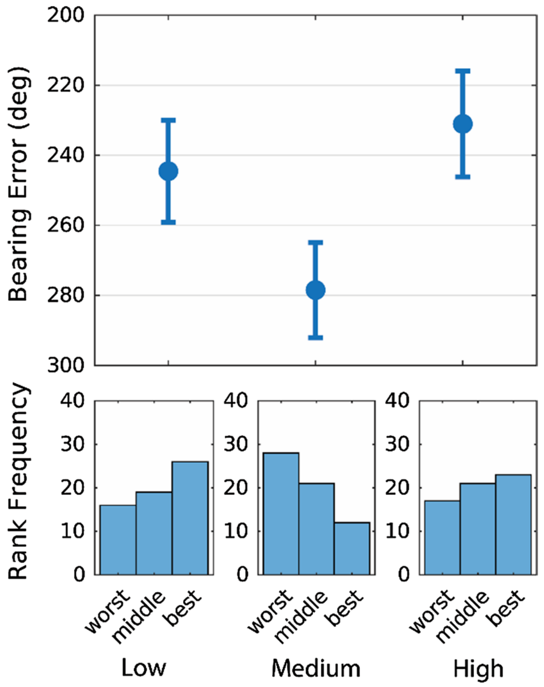


Fig. 6 Level of immersive technology versus errors on bearings questions. Upper panel shows mean sum of all absolute errors on bearings questions. Because this variable measures errors made, greater values of the variable indicate worse performance. Note the *y*-axis is therefore flipped to more intuitively visually convey differences in performance. Error bars show 1 within-subject standard error of the mean (Morey 2008). Lower panel indicates how many participants for whom a given technology yielded the best, middle, or worst performance

when engaging with the virtual environment and may have counteracted recognition benefits gained from head tracking.

Another potential factor that may help explain the object recognition results is the novelty that comes with experiencing a cutting-edge HMD. Our participants had little if any exposure to HMDs prior to the experiment. The novel exposure to an HMD could have put participants in a higher engagement state which may have heightened attention and exploration during the exposure to the virtual environment (Schomaker and Meeter 2015). While this may explain the improved performance in the High condition, it does not explain the lack of benefit in the Medium (NVIS) condition, which was also novel to the participants. Again, comfort or occlusion factors may have counteracted any novelty benefits.

Another candidate is the increased auditory affordance of the audio processing algorithm in the High condition, which included head-tracked spatial audio and simulated audio reflections from room surfaces. This may contribute to a more natural "viewing" of objects in the audio environment from multiple angles and may analogously assist in better object recognition. Other work has shown improved recall with enhanced audio fidelity (Davis et al. 1999). Spatial sound has previously been shown to improve object identification (Zhou et al. 2004) and navigation in augmented reality scenarios (Rumiński 2015). While the Medium-immersion condition also included head-tracked spatial audio, it lacked simulated reflections. The Low condition had neither. If audio reflections alone were the critical factor, we might have expected improved performance on yes/no object recognition but not on multiple-choice object recognition. The audio reflection algorithm was not sophisticated enough to convey the nuanced differences that were referenced in the multiple-choice questions. However, the existence of audio reflections may have enhanced performance simply via novelty, or perhaps as a result of enhanced ability to create spatial cognitive maps due to the affordance of the virtual sound reflections (Andreasen et al. 2019). Compared to effects of visual immersion on learning, audio immersion aspects receive less attention in the literature and may be a valuable avenue for future study.

4.2 An uncanny valley for immersion with difficult tasks?

The immersion versus performance response curve differed across the different types of learning assessments. While the cognitively simpler object recognition questions suggested no benefit until immersion level had passed a certain threshold, the cognitively complex, difficult bearings questions suggested a U-shaped curve: both Low and High immersion yielded better performance than the Medium level. McMahan et al. (2016) found similar patterns in studies of presentation fidelity. While recognition questions only required memory of objects encountered and visual details of those objects, bearings questions required a mental recreation of scenes and/or development of a mental map (Siegel and White 1975). The bearings transfer task requires participants to integrate a vast amount of knowledge gained from their virtual environment experience. To answer a bearings question correctly, the participant must remember multiple objects and their locations, must mentally reconstruct their relative spatial placement, and then must mentally construct a novel visual perspective integrating all of these. Learning the virtual environment in sufficient detail to accomplish this is a demanding task, and any property of the mediation that interferes with this could be expected to degrade performance.

There are several possibilities as to why the High level of immersion and the Low were similarly effective for the bearings test. It is possible that the novelty of the High and the familiarity of the Low were each similarly helpful when acquiring spatial knowledge to use for the complex bearings question processing. Or again, the physical comfort of the Medium level may have impeded the knowledge acquisition.

It may also be the case that spatial knowledge acquisition occurred via two different mechanisms, one used in the Low condition and one used in the High, both similarly effective. If the Medium condition led to a suboptimal mix of these methods, or to conscious or unconscious uncertainty over which mechanism to use, performance could decline as immersion increases, before then improving as immersion increases again.

The U-shaped curve we found calls to mind the uncanny valley curve (Mori et al. 2012; McMahan et al. 2016) reported for a variety of technological, immersive, and simulation contexts. Perhaps a related phenomenon is occurring in our study's Low, Medium, and High levels of immersion. The overall experience of the Low condition may be sufficiently abstract or fake to feel familiar and nonthreatening. After all, modern humans are used to interacting with flat desktop monitors or phone screens, button interfaces, and simple sound systems. It is just another day at work. The ease and familiarity of this interaction may free up attentional, memory, or other cognitive resources which may then be dedicated to the spatial learning task (see also McMahan et al. 2016). On the other end of the scale, the High condition may be sufficiently realistic and occlusive that the participants feel natural in their virtual interaction (head gaze, simulated audio reflections, no visual distractions, etc.), again freeing up cognitive resources which may be then be used for the spatial learning task. The Medium condition may fail to achieve either of these states, instead presenting an unsettling mix of both. The Medium condition in our study, like in classic interpretations of the uncanny valley, may represent a zone of tension between media that are clearly abstract and media that are sufficiently real. Or, it could represent a state where neither familiarity nor realism is sufficient to offer reduction in the cognitive load of the mediation experience, leading to poorer performance. This poorer performance might only be detectable on difficult tasks. These hypotheses, and the U-shaped response curve, warrant detailed investigation in future studies.

4.3 Better resolution when using within-subjects design

As a proof of concept, after we performed within-subjects analyses, we then "simulated" a between-subjects design by including only each participant's first session, adding back the participants we lost to attrition, and analyzing the resulting dataset without repeated measures. This was intended to approximate the results we may have gotten had we used a between-subjects design. It should be noted that an a priori between-subjects design likely would have taken a different approach to recruiting than we did, so our re-analysis of the data in a between-subjects style should be viewed as an approximation. Nonetheless, we found that most of our findings disappeared under the approximated between-subjects approach. Only a benefit for High immersion over Medium remained, and only in the yes/no object recognition questions. One of the strongest drawbacks of a between-subjects design is that latent individual differences are not controlled for. These can add noise to the data and obscure results. Because bearings questions are cognitively demanding and recruit multiple in-depth forms of cognitive processing, it is likely that performance on these questions is heavily influenced by a number of latent individual differences—exactly the sort that may obscure results in between-subjects designs. Thus, it is not surprising that only the within-subjects approach was able to detect any influence of immersive technology on bearings scores. Similarly, because the yes/no object recognition questions were the simplest (i.e., not requiring a cognitive map or configurational knowledge of landmarks), involving less cognitive complexity and potentially involving fewer individual difference factors, we may have expected this outcome variable to be the most likely one to show detectable differences in a between-subjects design. Individual differences are known to be of great importance in learning and may factor heavily into immersive technology effects as well. This is a fruitful avenue of study, but assessment in a between-subjects design requires identifying candidate individual differences. Because no study can identify and measure all potentially relevant individual differences, a within-subjects approach can still provide value in controlling for latent, unknown, unmeasured, or unnamed individual differences.

4.4 Complex curves require three or more conditions to uncover

Utilizing three levels of immersion also provided us with added precision over using a binary high versus low approach. We can "simulate" the results of such a study using our data as well. If we had only tested the Low and Medium conditions, we would have found no effect of immersion on yes/no object recognition or on multiple-choice object recognition. We may have erroneously concluded that level of immersion did not affect this type of learning, perhaps because object recognition is so simplistic that little benefit is gained from greater immersive technology. Based on our three-level study, we can instead suggest that perhaps the Medium level was not immersive *enough*, or that other factors are in play. Similarly, if we had only tested the Low and High conditions, we would have found no effect of immersion on bearings performance. We may have concluded that level of immersion did not affect this type of learning. If we had only tested the Low and Medium, we may have concluded that increasing immersion was a *detriment* to bearings learning, and vice versa if we had only tested the Medium and High. That these conclusions differ so dramatically highlights the downside of using a two-condition design and may help explain some of the highly contradictory results found in the literature.

5 Conclusions and future directions

Our three-condition, within-subjects study with long intersession intervals revealed effects of levels of immersion on spatial learning in virtual environments. We found a benefit to the highest level of immersion in object recognition questions, and a U-shaped response curve on bearings questions. Using our data to approximate a two-condition study resulted in a series of mutually contradicting conclusions, while using our data to approximate a between-subjects study yielded more null results. These approaches may partially explain the mixed findings in the broader literature regarding the effects of immersive VR technology on learning.

We therefore encourage VR researchers, when feasible, to consider within-subjects designs and to consider using sufficiently long intersession intervals in those designs. This reduces noise from individual differences, mitigates adaptation, and provides more powerful data to analyze effects. We also encourage the use of multiple levels of test conditions (at least three) to help elucidate the potentially nonmonotonic shapes of the underlying response curves. Of course, these recommendations become more difficult to follow when combined (e.g., a within-subjects design with ten conditions and 2-week intervals would be quite difficult to implement), so trade-offs must always be considered. Between-subjects studies and two-condition designs certainly do provide valuable data, but these data should be interpreted with appropriate caveats.

Acknowledgements We thank Jason Moss and Antony D. Passaro who contributed to the conceptualization and development of the study, and Bianca Dalangin who helped conduct the study. This work was funded by the United States Army Research Laboratory's Human Sciences Campaign. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval All procedures involving human participants were in accordance with the ethical standards of the Institutional Review Board of the U.S. Army Research Laboratory and with the 1964 Helsinki declaration and its later amendments.

Informed consent Informed consent was obtained from all individual participants included in the study, in accordance with Title 32, Part 219 of the CFR and Army Regulation 70-25.

References

Alexander AL, Brunyé T, Sidman J, Weil SA (2005) 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

Andre AD, Wickens CD (1995) When users want what's not best for them. Ergon Des 3:10–14

Andreasen A, Geronazzo M, Nilsson NC et al (2019) Auditory feedback for navigation with echoes in virtual environments: training procedure and orientation strategies. IEEE Trans Vis Comput Gr. https://doi.org/10.1109/TVCG.2019.2898787

Baylor A, Ryu J, Shen E (2003) The effects of pedagogical agent voice and animation on learning, motivation and perceived persona. Association for the Advancement of Computing in Education (AACE), pp 452–458

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B (Methodol) 57:289–300

Bhagat KK, Liou W-K, Chang C-Y (2016) A cost-effective interactive 3D virtual reality system applied to military live firing training. Virtual Real 20:127–140. https://doi.org/10.1007/s10055-016-0284-x

Boccia M, Rosella M, Vecchione F, Tanzilli A, Palermo L, D'Amico S, Piccardi L (2017) Enhancing allocentric spatial recall in preschoolers through navigational training programme. Front Neurosci 11:574

Bowman DA, McMahan RP (2007) Virtual reality: how much immersion is enough? Computer 40:36–43

Bowman DA, Sowndararajan A, Ragan ED, Kopper R (2009) Higher levels of immersion improve procedure memorization performance. In: Joint virtual reality conference of EGVE-ICAT-EuroVR

Brown RB (2016) Enhancing realistic training: delivering training capabilities in a complex world. Combined Arms Centers, Ft Leavenworth

Carlson G, Caporusso N (2019) A physically immersive platform for training emergency responders and law enforcement officers. In: Nazir S, Teperi A-M, Polak-Sopińska A (eds) Advances in human factors in training, education, and learning sciences. Springer, Berlin, pp 108–116

Chandrasiri A, Collett J, Fassbender E, Foe AD (2019) A virtual reality approach to mindfulness skills training. Virtual Real. https://doi.org/10.1007/s10055-019-00380-2

Cho Y (2018) How spatial presence in VR affects memory retention and motivation on second language learning: a comparison of desktop and immersive VR-based learning. Thesis, Syracuse University

Christou C, Tzanavari A, Herakleous K, Poullis C (2016) Navigation in virtual reality: comparison of gaze-directed and pointing motion control. In: 2016 18th mediterranean electrotechnical conference (MELECON), pp 1–6

Craig SD, Gholson B, Driscoll DM (2002) Animated pedagogical agents in multimedia educational environments: effects of agent properties, picture features and redundancy. J Educ Psychol 94:428–434. https://doi.org/10.1037/0022-0663.94.2.428

Cummings JJ, Bailenson JN (2016) How immersive is enough? a meta-analysis of the effect of immersive technology on user presence. Med Psychol 19:272–309. https://doi.org/10.1080/15213269.2015.1015740

Darken RP, Peterson B (2002) Spatial orientation, wayfinding, and representation. In: Stanney KM (ed) Human factors and ergonomics. Handbook of virtual environments: Design, implementation, and applications. Lawrence Erlbaum Associates Publishers, p. 493–518

Davis ET, Scott K, Pair J et al (1999) Can audio enhance visual perception and performance in a virtual environment? Proc Hum Factors Ergon Soc Annu Meeting 43:1197–1201. https://doi.org/10.1177/154193129904302206

Files BT, Oiknine AH, Thomas J et al (2019a) Same task, different place: Developing novel simulation environments with equivalent task difficulties. In: Proceedings of applied human factors and ergonomics

Files BT, Pollard KA, Oiknine AH, Passaro AD, Khooshabeh P (2019b) Prevention focus relates to performance on a loss-framed inhibitory control task. Front Psychol 10:726

Frechette C, Moreno R (2010) The roles of animated pedagogical agents' presence and nonverbal communication in multimedia learning environments. J Media Psychol Theor Methods Appl 22(2):61–72

Government Accountability Office USGAO (2016) Army training: efforts to adjust training requirements should consider the use of virtual training devices

Horowitz J (2018) Walmart buys 17,000 Oculus Go VR headsets to train a million employees. https://venturebeat.com/2018/09/20/walmart-buys-17000-oculus-go-vr-headsets-to-train-a-million-employees/

Hsieh T-J (Tracy), Kuo Y-H, Niu C-K (2018) Utilizing HMD VR to improve the spatial learning and wayfinding effects in the virtual maze. HCI International 2018 – Posters' Extended Abstracts 38–42

Huang Y, Churches L, Reilly B (2015) A case study on virtual reality American football training. In: Proceedings of the 2015 virtual reality international conference. ACM, New York, NY, USA, pp 6:1–6:5

Jeelani I, Han K, Albert A (2017) Development of immersive personalized training environment for construction workers. Comput Civil Eng 2017:407–415. https://doi.org/10.1061/9780784480830.050

Jeffs TL (2010) Virtual Reality and Special Needs. Themes Sci Technol Educ 2:253–268

Johnson L, Becker SA, Cummins M et al (2016) NMC horizon report: 2016 higher education edition. The New Media Consortium

Kasap Z, Magnenat-Thalmann N (2012) Building long-term relationships with virtual and robotic characters: the role of remembering. Vis Comput 28:87–97

Khooshabeh P, Choromanski I, Neubauer C et al (2017) Mixed reality training for tank platoon leader communication skills. In: 2017 IEEE virtual reality (VR), pp 333–334

Kraemer DJ, Schinazi VR, Cawkwell PB, Tekriwal A, Epstein RA, Thompson-Schill SL (2017) Verbalizing visualizing and navigating: the effect of strategies on encoding a large-scale virtual environment. J Exp Psychol Learn Memory Cognit 43:611

Krokos E, Plaisant C, Varshney A (2018) Virtual memory palaces: immersion aids recall. Virtual Real. https://doi.org/10.1007/s10055-018-0346-3

Landers RN (2014) Developing a theory of gamified learning: linking serious games and gamification of learning. Simul Gaming 45:752–768

Landers RN, Landers AK (2014) An empirical test of the theory of gamified learning: the effect of leaderboards on time-on-task and academic performance. Simul Gaming 45:769–785

Lányi CS, Geiszt Z, Károlyi P et al (2006) Virtual reality in special needs early education. Int J Virtual Real 5:55–68

Lee KM, Nass C (2005) Social-psychological origins of feelings of presence: creating social presence with machine-generated voices. Media Psychol 7:31–45

Lewis T (2017) Virtual reality helps reinvent law enforcement training. https://www.cbsnews.com/news/virtual-reality-law-enforcement-training/

Mania K, Chalmers A (2001) The effects of levels of immersion on memory and presence in virtual environments: a reality centered approach. Cyber Psychol Behav. https://doi.org/10.1089/109493101300117938

Mania K, Troscianko T, Hawkes R, Chalmers A (2006) Fidelity metrics for virtual environment simulations based on spatial memory awareness states. In: http://dx.doi.org.proxy.library.ucsb.edu:2048/10.1162/105474603765879549. http://www.mitpressjournals.org/doix/abs/10.1162/105474603765879549. Accessed 22 Feb 2019

Markowitz DM, Laha R, Perone BP et al (2018) Immersive virtual reality field trips facilitate learning about climate change. Front Psychol. https://doi.org/10.3389/fpsyg.2018.02364

McMahan RP, Bowman DA, Zielinski DJ, Brady RB (2012) Evaluating display fidelity and interaction fidelity in a virtual reality game. IEEE Trans Vis Comput Gr 18:626–633

McMahan RP, Lai C, Pal SK (2016) Interaction fidelity: the uncanny valley of virtual reality interactions. In: Virtual, augmented and mix reality (VAMR 2016) Lecture notes in computer science 9740:59–70

Mizell DW, Jones SP, Slater M, Spanlang B (2002) Comparing immersive virtual reality with other display modes for visualizing complex 3D geometry. University College London, technical report

Moreno R, Mayer RE (2000) Engaging students in active learning: the case for personalized multimedia messages. J Educ Psychol 92:724–733. https://doi.org/10.1037/0022-0663.92.4.724

Moreno R, Mayer RE (2004) Personalized messages that promote science learning in virtual environments. J Educ Psychol 96:165–173. https://doi.org/10.1037/0022-0663.96.1.165

Morey RD (2008) Confidence intervals from normalized data: a correction to Cousineau (2005). Tutorial Quant Methods Psychol 4:61–64

Mori M, MacDorman KF, Kageki N. (2012) The uncanny valley. IEEE Robotics & Automation Magazine June 2012:98-100

Moss JD, Austin J, Salley J et al (2011) The effects of display delay on simulator sickness. Displays 32:159–168. https://doi.org/10.1016/j.displa.2011.05.010

Mossel A, Froeschl M, Schoenauer C, et al (2017) VROnSite: towards immersive training of first responder squad leaders in untethered virtual reality. In: 2017 IEEE virtual reality (VR), pp 357–358

Nys M, Gyselinck V, Orriols E, Hickmann M (2015) Landmark and route knowledge in children's spatial representation of a virtual environment. Front Psychol 5:1522

Oiknine A, Files B, Pollard KA (2019) Web-based measurement of directional bearings (angular distance). CCDC-Army Research Laboratory, Aberdeen Proving Ground, MD, pp 1–11

Parong J, Mayer RE (2018) Learning science in immersive virtual reality. J Educ Psychol 110:785–797. https://doi.org/10.1037/edu0000241

Patton D (2014) How real is good enough? Assessing realism of presence in simulations and its effects on decision making. In: Schmorrow DD, Fidopiastis CM (eds) Foundations of augmented cognition. Advancing human performance and decision-making through adaptive systems. Springer International Publishing, Berlin, pp 245–256

Patton D, Gamble K (2016) Physiological measures of arousal during soldier-relevant tasks performed in a simulated environment. Foundations of augmented cognition: neuroergonomics and operational neuroscience. Springer, Cham, pp 372–382

Picciano AG (2002) Beyond student perceptions: issues of interaction, presence, and performance in an online course. J Asynchronous Learn Netw 6:20

Ragan ED, Scerbo S, Bacim F, Bowman DA (2017) Amplified head rotation in virtual reality and the effects on 3D search, training transfer, and spatial orientation. IEEE Trans Vis Comput Gr 23(8):1880–1895. https://doi.org/10.1109/TVCG.2016.2601607

Regan C (1995) An investigation into nausea and other side-effects of head-coupled immersive virtual reality. Virtual Real 1:17–31

Rumiński D (2015) An experimental study of spatial sound usefulness in searching and navigating through AR environments. Virtual Real 19:223–233. https://doi.org/10.1007/s10055-015-0274-4

Schomaker J, Meeter M (2015) Short- and long-lasting consequences of novelty, deviance and surprise on brain and cognition. Neurosci Biobehav Rev 55:268–279. https://doi.org/10.1016/j.neubiorev.2015.05.002

Sharples S, Cobb S, Moody A, Wilson JR (2008) Virtual reality induced symptoms and effects (VRISE): comparison of head mounted display (HMD), desktop and projection display systems. Displays 29:58–69. https://doi.org/10.1016/j.displa.2007.09.005

Shu Y, Huang Y-Z, Chang S-H, Chen M-Y (2018) Do virtual reality head-mounted displays make a difference? A comparison of presence and self-efficacy between head-mounted displays and desktop computer-facilitated virtual environments. Virtual Real. https://doi.org/10.1007/s10055-018-0376-x

Siegel AW, White SH (1975) The development of spatial representations of large-scale environments. In: Reese HW (ed) Advances in child development and behavior. JAI, pp 9–55

Sinatra AM, Oiknine AH, Patton D et al (2019) Development of cognitive transfer tasks for virtual environments and applications for adaptive instructional systems. In: Lecture notes in computer science. Springer, Orlando

Sintia R (2018) The U.S. military wants to lead the innovation game in VR. US News https://www.usnews.com/news/best-countries/articles/2018-03-20/the-us-military-wants-to-lead-the-innovation-game-in-vr

Slater M, Wilbur S (1997) A framework for immersive virtual environments (FIVE): speculations on the role of presence in virtual environments. Presence Teleoperators Virtual Environ 6:603–616

Sousa Santos B, Dias P, Pimentel A et al (2009) Head-mounted display versus desktop for 3D navigation in virtual reality: a user study. Multimed Tools Appl 41:161. https://doi.org/10.1007/s11042-008-0223-2

Stansfield S, Shawver D, Sobel A (1998) MediSim: a prototype VR system for training medical first responders. In: Proceedings of IEEE 1998 virtual reality annual international symposium (Cat. No. 98CB36180), pp 198–205

Stern E, Leiser D (1988) Levels of spatial knowledge and urban travel modeling. Geogr Anal 20:140–155

Stevens J, Kincaid P, Sottilare R (2015) Visual modality research in virtual and mixed reality simulation. J Def Model Simul. https://doi.org/10.1177/1548512915569742

Summers JE (2012) Simulation-based military training: an engineering approach to better addressing competing environmental, fiscal, and security concerns. J Wash Acad Sci 98:9–29

Swindells C, Po BA, Hajshirmohammadi I et al (2004) Comparing CAVE, wall, and desktop displays for navigation and wayfinding in complex 3D models. In: Proceedings computer graphics international, 2004. pp 420–427

Taillade M, Sauzéon H, Pala PA, Déjos M, Larrue F, Gross C, N'Kaoua B (2013) Age-related wayfinding differences in real large-scale environments: detrimental motor control effects during spatial learning are mediated by executive decline? PLoS ONE 2013(8):e67193

Taylor GS, Barnett JS (2013) Evaluation of wearable simulation interface for military training. Hum Factors 55:672–690. https://doi.org/10.1177/0018720812466892

Thorndyke PW, Hayes-Roth B (1982) Differences in spatial knowledge acquired from maps and navigation. Cogn Psychol 14:560–589

Tong X, Gromala D, Gupta D, Squire P (2016) Usability comparisons of head-mounted vs. stereoscopic desktop displays in a virtual reality environment with pain patients. Stud Health Technol Inform 220:424–431

Tse A, Jennett C, Moore J, et al (2017) Was I there: impact of platform and headphones on 360 video immersion. ACM, pp 2967–2974

van der Ham IJM, Faber AME, Venselaar M, van Kreveld MJ, Löffler M (2015) Ecological validity of virtual environments to assess human navigation ability. Front Psychol 6:637

von der Pütten AM, Krämer NC, Gratch J (2009) Who's there? Can a virtual agent really elicit social presence? Paper presented at the 12th Annual International Workshop on Presence Los Angeles CA

Walker AD, Carpenter TL, Moss JD et al (2009) The evaluation of virtual environment training for a building clearing task. In: Proceedings of the human factors and ergonomics society annual meeting. https://doi.org/10.1177/154193120905301809

Waller D, Hunt E, Knapp D (1998) The transfer of spatial knowledge in virtual environment training. Presence Teleoperators Virtual Environ 7:129–143

Wang X, Dunston PS, Skibniewski M (2004) Mixed reality technology applications in construction equipment operator training. In: Proceedings of the 21st international symposium on automation and robotics in construction (ISARC 2004), September 2125, Jeju, Korea, pp 393–400

Wang N, Johnson WL, Mayer RE, Rizzo P, Shaw E, Collins H (2008) The politeness effect: pedagogical agents and learning outcomes. Int J Hum Comput Stud 66:98–112

Warton DI, Hui FKC (2011) The arcsine is asinine: the analysis of proportions in ecology. Ecology 92:3–10. https://doi.org/10.1890/10-0340.1

Wei L, Zhou H, Nahavandi S (2018) Haptically enabled simulation system for firearm shooting training. Virtual Real. https://doi.org/10.1007/s10055-018-0349-0

Wilson M (2016) The effect of varying latency in a head-mounted display on task performance and motion sickness. Dissertation, Clemson University

Witmer BG, Bailey JH, Knerr BW (1995) Training dismounted soldiers in virtual environments: route learning and transfer (Technical Report 1022): U.S. Army Research Institute for the Behavioral and Social Sciences

Zhong JY, Moffat SD (2016) Age-related differences in associative learning of landmarks and heading directions in a virtual navigation task. Front Aging Neurosci 8(122):1–11. https://doi.org/10.3389/fnagi.2016.00122

Zhou Z, Cheok AD, Yang X, Qiu Y (2004) An experimental study on the role of 3D sound in augmented reality environment. Interact Comput 16:1043–1068. https://doi.org/10.1016/j.intcom.2004.06.016

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.