NOTICE: This is a paper that was accepted for publication in the **IEEE Transactions on Visualization and Computer Graphics**. It is not the final published version. The DOI of the definitive version available in IEEE Explore is

https://doi.org/10.1109/TVCG.2017.2653117

Effects of Different Types of Virtual Reality Display on Presence and Learning in a Safety Training Scenario

Fabio Buttussi and Luca Chittaro

Abstract—The increasing availability of head-mounted displays (HMDs) for home use motivates the study of the possible effects that adopting this new hardware might have on users. Moreover, while the impact of display type has been studied for different kinds of tasks, it has been scarcely explored in procedural training. Our study considered three different types of displays used by participants for training in aviation safety procedures with a serious game. The three displays were respectively representative of: (i) desktop VR (a standard desktop monitor), (ii) many setups for immersive VR used in the literature (an HMD with narrow field of view and a 3-DOF tracker), and (iii) new setups for immersive home VR (an HMD with wide field of view and 6-DOF tracker). We assessed effects on knowledge gain, and different self-reported measures (self-efficacy, engagement, presence). Unlike previous studies of display type that measured effects only immediately after the VR experience, we considered also a longer time span (2 weeks). Results indicated that the display type played a significant role in engagement and presence. The training benefits (increased knowledge and self-efficacy) were instead obtained, and maintained at two weeks, regardless of the display used. The paper discusses the implications of these results.

Index Terms— Virtual reality, displays, fidelity, training, user study, aviation, safety

1 Introduction

VIRTUAL Reality (VR) has been used to train professionals in many different domains (e.g., medicine [1], military [2], and firefighting [3]) since its early years. More recently, gaming elements have been added to VR training experiences, turning them into serious games that could be more engaging, especially for the general public [4].

Several studies showed the benefits of VR applications and serious games in training and education (e.g., see [5] and [6] for two recent meta-analyses). However, more research is needed to understand which characteristics of these applications, in particular VR fidelity, impact on their effectiveness [7]. A framework that categorizes the different aspects of VR fidelity is proposed in [7], and identifies three types of VR fidelity. *Interaction fidelity* is about the realism of the input devices, and the reproduction of real-world interactions in VR. *Display fidelity* is about the realism of the output devices, and the reproduction of sensory stimuli ("display fidelity" is sometimes referred to as "immersion", but the former term is preferred to avoid ambiguity [8]). *Scenario fidelity* is about the realism of the simulated scenario, and the reproduction of behaviors, rules, and properties in the simulation.

The increasing availability of head-mounted displays (HMDs) for home use motivates the study of the possible effects that adopting this new hardware might have on users. In addition, the effects of different types of display have been scarcely explored in the domain of procedural training, i.e. training that deals with the execution of procedures. For these reasons, the study we describe in this paper compares a new HMD for home VR with other types of displays, using them for training in cabin safety procedures with a serious game that simulates a runway overrun accident in VR. We chose cabin safety as a representative case study, since it requires mastering several skills that are relevant also to other procedural training applications. For example, cabin safety requires subjects: i) to memorize complex procedures that consist of several steps; ii) to analyze the environment, people, and other contextual factors that can affect the execution of the procedure; iii) to react to these factors, and update the execution of the procedure accordingly. Cabin safety is also particularly interesting, because passengers tend not to pay attention to current safety briefings (pre-flight safety briefings and safety card) provided by airlines, and even the few passengers who pay attention show an unacceptable level of safety knowledge [9]. This calls for other, more engaging, solutions [10]. A study of an immersive serious game that simulated a water landing scenario showed that VR was more engaging than safety cards, and more effective in helping users retain acquired knowledge a week after training [11]. However, that game was played with an HMD with narrow FOV and a 3-DOF tracker, a typical VR setup often used in the literature [12], [13], [14], [15], [16], [17], [18]. The increasing availability of new HMDs with wide FOV and 6-DOF tracking raises the question of whether such higher fidelity displays could be even more effective. Moreover, we wonder whether the immersive

• *Fabio Buttussi and Luca Chittaro are with the Human-Computer Interaction Lab, Department of Mathematics, Computer Science, and Physics, University of Udine, Italy. E-mail: {fabio.buttussi, luca.chittaro}@uniud.it.*

display played a role in the positive effects on engagement and knowledge gain found in [11] or a lower-fidelity desktop monitor could achieve the same results. For these reasons, the study in this paper compares: (i) a standard desktop monitor (representative of desktop VR), (ii) an HMD (Sony HMZ-T3W) with narrow field of view (FOV) and a 3-DOF tracker (representative of many setups for immersive VR used in the literature [11], [12], [13], [14], [15], [16], [17], [18]), and (iii) an HMD (Oculus Rift DK2) with wide FOV and 6-DOF tracker (representative of new devices for immersive home VR).

The paper aims at advancing knowledge in different directions. First, to the best of our knowledge, our study is the first to explore the possible effects of different types of display on safety knowledge acquisition.

Second, the study explores the effects of the different types of display on additional variables, namely self-efficacy (i.e., the person's belief in his/her ability to perform a specific behavior [19]), self-reported engagement, and self-reported presence (i.e., the sense of being in the virtual environment [20]). All these variables can play an important role in the acquisition and application of procedural knowledge. Studies in Social Cognitive Theory [21], [22] show that higher self-efficacy has a positive effect on performance. Presence is instead linked with the arousal of emotions, in particular with negative ones [23]. In turn, emotional intensity aroused by an experience, and negative emotions in particular, can increase the retention of memory allowing for a better recall [24], [25].

Third, unlike previous studies of the effects of different types of display, our study tests knowledge gain and self-efficacy not only immediately after using the serious game, but also two weeks later. This aspect is fundamental since an emergency can occur a long time after training, so effective training should support recall of procedures over time. Since different studies show connections between type of display and presence (see Section 2 and Table 1), between presence and emotions, and between emotions and retention, we hypothesize that the different types of display might play a role in the retention of procedural knowledge over time, and we assess this aspect in the study.

The paper is organized as follows. Section 2 introduces previous studies of the effects of displays in VR. Section 3 illustrates the training game we employed in the user study. The study is described in Section 4, while its results are reported and discussed in Section 5 and 6, respectively. Finally, Section 7 concludes the paper outlining future work.

2 Related Work

Several studies addressed the effects of using different displays in VR (see Table 1). Some of them contrasted different types of display (e.g., desktop monitor vs. HMD, HMD vs CAVE), others considered a single type of display and manipulated its features (e.g., with or without stereoscopy, with original or reduced FOV).

In many studies, change in display type included a change in interaction too. For example, head rotation is sensed by the system when using a head-tracked HMD, while it has to be controlled by the user (e.g., with a joystick) when using a desktop monitor. Therefore, in terms of the framework proposed in [7], all the studies in Table 1 concern changes in display fidelity, and many of them include also changes in interaction fidelity. The table describes in detail the aspects of VR fidelity assessed in each study as well as the considered independent variables, tasks, dependent variables, and main findings. In the following, we briefly summarize the results that are more relevant to our study.

Effects of different displays were studied especially with visual search tasks, i.e., tasks that require users to search for one or more targets in a virtual environment. For example, Pausch et al. [26] found that users wearing an HMD did not find targets faster than stationary monitor users, but were better at determining if a target was present or not. Ragan et al. [7] found that wider FOV led to better target detection while users trained in performing a given scanning strategy, but not in an assessment phase performed five minutes after training. Kim et al. [27] showed that a CAVE and an HMD elicited more emotional arousal than a monitor in terms of both self-reported and physiological (skin conductance) measures. Moreover, presence experienced with the CAVE was higher than with the HMD, and presence with the HMD was higher than with the monitor. However, the HMD caused more simulator sickness than the other conditions, and the CAVE caused more simulator sickness than the monitor.

The fidelity of the display plays an important role also on tasks involving data visualization. For example, Arns et al. [28] showed that an immersive application for statistical visualization in a CAVE led to better identification of data structures than a traditional desktop statistical tool, but also that users were more comfortable with the latter. However, although the immersive application was developed to offer the same functionality of the desktop tool, the two applications were different, and this could have contributed to the result. Bacim et al. [29] studied how the combination of display size, stereoscopy, and head tracking in a CAVE affected user performance in understanding mathematical graphs. The study found better overall task performance with the higher fidelity condition. The effects of different components of fidelity, namely head tracking, stereoscopy, and field of regard (FOR) on different tasks concerning volume data visualization were instead studied in [30] and [31]. FOR is the total size of the visual field surrounding the user [32], which depends on FOV and on the possibility to rotate the head. The studies found that increased fidelity generally led to better performance in analyzing volume data, and that the influence of the different components varied with the kind of task. The effects of head tracking, stereoscopy, and FOR were studied also on small-scale spatial judgement tasks [33]. The study found that increased FOR or the addition of head tracking reduced the number of participants' errors, and the combination of stereoscopy and head tracking allowed participants to perform faster.

In summary, the different studies on visual search tasks, data visualization tasks, and small-scale spatial judgement tasks indicate a positive overall effect of higher fidelity dis-

TABLE 1
STUDIES ABOUT THE EFFECTS OF DIFFERENT DISPLAYS

Reference	Aspects of VR fidelity	Independent variables	Tasks	Dependent variables	Main findings
Pausch et al. (1997) [26]	Display, Interaction	Display (6-DOF head-tracked HMD, stationary monitor)	Search for camouflaged targets in a virtual room	Time to find targets, time to determine lack of target	No difference in time to find targets, time to determine lack of target lower for HMD
Waller et al. (1998) [34]	Display, Interaction	Type of training (no training, 1' in real maze, 1' with map, 2' with 21" desktop monitor VR, 2' with 6-DOF head-tracked HMD VR, 5' with 6-DOF head-tracked HMD VR)	Reach 4 locations in a real maze by navigating blindfolded after training in one of the experimental conditions	Time to complete task, number of times participants touched the walls of the maze	2' with 6-DOF head-tracked HMD no more effective than 1'with map in reducing time to complete task, 5' with 6-DOF head-tracked HMD more effective than 1' in real maze, 2' with 6-DOF head-tracked HMD no more effective than 2' with desktop monitor, strong correlation between time to complete task and number of times participants touched the walls
Arns et al. (1999) [28]	Display	Display (4-screen CAVE, desktop monitor)	Identify structures in statistical data, select data points	Correct identification of structures, time to complete selection task	More correct identification of structures for CAVE, lower time to complete selection task for desktop monitor
Mania & Chalmers (2001) [43]	Display	Seminar presentation (real classroom, 3D virtual classroom on 21" desktop monitor, 3D virtual classroom on HMD without head-tracking, audio only)	Attend a 15-minutes seminar	Presence, knowledge gain	Higher presence only in real classroom, presence not associated with knowledge gain
Winn et al. (2002) [44]	Display	Display (HMD, desktop monitor)	Interact with a simulation about water movement and salinity	Presence, knowledge gain	Presence predictor for knowledge gain, higher knowledge gain about water movement for HMD, no effect on knowledge gain about salinity
Zanbaka et al. (2005) [39]	Display, Interaction	Setup (HMD with real walking, HMD with virtual walking and 6-DOF tracking, HMD with virtual walking and 3-DOF tracking, 17" desktop monitor with joystick)	Navigate in a virtual room	Scores of cognition questionnaire and map drawings, presence	Only HMD with real walking better than monitor for scores of cognition questionnaire and map drawings; all HMD conditions better than monitor for presence
Aoki et al. (2008) [17]	Display, Interaction	Setup (3-DOF head-tracked HMD with gamepad, 17" desktop monitor with gamepad, 17" desktop monitor with keyboard and integrated touchpad)	Point at targets and navigate in a simulated space station	Time to point from destination to start and from start to different destinations, egress time	Lower time to point with HMD, no effect on egress time
Elmqvist et al. (2008) [37]	Display, Interaction	Display (3-screen CAVE, 17" desktop monitor), navigation method (free, fully constrained, constrained but local deviations), virtual environment (outdoor, indoor, infoscape, conetree)	Recall landmarks and navigate to landmarks after navigating in the 4 virtual environments	Distance between recalled and actual landmark positions, missed landmarks, time to find landmarks	Without motion constraints: better performance using CAVE, with motion constraints: better performance using monitor
Johnsen & Lok (2008) [45]	Display	Display (6-DOF head-tracked HMD, large-screen projection display)	Interact with two virtual patients to train in communication skills	Simulated patient quality, presence, self-evaluation ratings, performance grades given by an expert	Higher self-ratings about use of empathy with HMD; more accurate reflection about use of empathy with large-screen projected display
Limniou et al. (2008) [42]	Display	Display (4-screen CAVE, single projected screen)	Study molecule structure and chemical reactions	Knowledge gain	Higher knowledge gain with CAVE
Sowndara-rajan et al. (2008) [47]	Display	Display (laptop display, 2-screen projection display)	Train in a simple and a complex procedure consisting in moving virtual objects from a location to another	Time to complete procedure, number of errors	Less time and less errors with 2-screen projection display on complex procedure, no effect on simple procedure
Ragan et al. (2010) [48]	Display	Software FOV (matched, unmatched), FOV (60°, 180°), FOR (1 screen, 3 screens), assessment environment (virtual, physical)	Train in a procedure consisting in moving virtual objects from a location to another, recall procedure in the assessment environment	Time to complete procedure in assessment environment, number of errors in assessment environment	No difference between assessment environments, less time and errors for matched software FOV, high FOV, and high FOR

Reference	Aspects of VR fidelity	Independent variables	Tasks	Dependent variables	Main findings
Litwiller & LaViola (2011) [40]	Display	Display (50” 3D stereo monitor, same monitor in 2D mode)	Play five commercial video games of different genres	User preference, various game performance metrics (e.g., scores and times)	User preference for 3D stereo monitor, no effect on performance
McNamara et al. (2011) [35]	Display	Display (continuous curve screen, tiled off-the-shelf LCD screens)	Navigate from a start to an end location in a virtual building	Time to complete task	No differences between the two displays
Fassbender et al. (2012) [46]	Display	Display (3-monitor setting, larger curved screen), background music (music, no-music)	Attend a virtual history lesson	Knowledge gain	Overall higher knowledge gain with 3-monitor setting, interaction between display and background music, higher knowledge gain with larger curved screen if there was music in the second half of the lesson
Kober et al. (2012) [38]	Display	Display (single stereo projected screen, 21” desktop monitor)	Navigate through a virtual maze from a start to an end location	Presence (user ratings, EEG data)	Higher presence with single stereo projected screen
Laha et al. (2012) [30]	Display, Interaction	Head tracking (6-DOF on, off), stereoscopy (on, off), FOR (4 screens, 1 screen)	Analyze volume data visualizations of tomography datasets (search and description tasks of different complexity)	Performance (task correctness, time to complete tasks), subjective measures (perceived difficulty, confidence in task correctness)	Better performance with higher fidelity, different influence of the variables in the different tasks
McMahan et al. (2012) [8]	Display, Interaction	Display (6-screen CAVE, single screen), input (6-DOF wand, mouse and keyboard)	Play a FPS video game	Game performance metrics (task completion time, damage taken, accuracy, headshot count), subjective scores (presence, engagement, usability)	Better performance in high display with high input and low display with low input conditions with respect to mixed conditions, higher subjective scores for high display with high input condition
Bacim et al. (2013) [29]	Display, Interaction	Display (4-screen CAVE with stereoscopy and 6-DOF head tracking, one screen with no stereoscopy and no head tracking)	Search for intersections, follow paths, identify connections, compare lengths in mathematical graphs	Performance (task correctness, time to complete task)	Better performance using 4-screen CAVE with stereoscopy and 6-DOF head tracking
Li and Giudice (2013) [36]	Display, Interaction	Setup (HMD with physical rotation, 15.6” desktop monitor with physical rotation, 15.6” desktop monitor with joystick-based rotation)	Point at and navigate to targets in a multi-level virtual building	Pointing time, pointing error, accuracy in reaching destination, navigation efficiency in following the shortest path, navigation error (distance between user and target location)	No difference between the setups
Lugrin et al. (2013) [41]	Display, Interaction	Setup (4-screen CAVE with wand tracker, 19” desktop monitor with keyboard and mouse)	Play a FPS video game	User preference, various game performance metrics (e.g., miss and hit shots, number of kills and deaths)	User preference for CAVE, better performance with monitor
Ragan et al. (2013) [33]	Display, Interaction	Head tracking (6-DOF on, off), stereoscopy (on, off), and FOR (1 screen, 4 screens)	Count interconnecting tubes in virtual underground cave structures	Number of errors, time to complete task	Less errors with higher FOR or head tracking on, less time with stereoscopy and head tracking on
Kim et al. (2014) [27]	Display, Interaction	Setup (17” monitor with standard mouse, 6-DOF head-tracked HMD with 3D mouse, 6-screen CAVE with 3D mouse)	Search for cards with specified color names in a virtual kitchen	Emotional arousal (self-report, skin conductance), performance (time to complete task, accuracy), presence, simulator sickness	Higher emotional arousal with CAVE and HMD than monitor, higher presence with CAVE than HMD, higher presence with HMD than monitor, more simulator sickness with HMD than other displays, more sickness with CAVE than monitor
Laha et al. (2014) [31]	Display, Interaction	Head tracking (6-DOF on, off), stereoscopy (on, off), FOR (1 screen, 4 screens)	Search for features, recognize patterns, judge positions and orientations, estimate properties, describe shapes in volume data visualizations of tomography scans	Performance (task correctness, time to complete tasks), subjective measures (perceived difficulty, confidence in task correctness)	Different tasks benefit from different combinations of independent variables values
Ragan et al. (2015) [7]	Display, Scenario	FOV (35.81°, 80.44°, 120.41°), scenario complexity (three levels with different number of objects and different detail of textures)	Search for targets in virtual city streets following a given scanning strategy	Detection performance (percentage of correct identifications, percentage of false-positive identifications) in training and in assessment, adherence to given	Better detection performance in training for higher FOV, no difference in assessment

plays on performance, with an influence of the different components that varies based on the particular task.

Results in the literature are instead less clear for navigation tasks, which combine spatial and visual skills with the cognitive process of building a mental map. For example, Waller et al. [34] tested participants' ability to navigate in a real-world maze blindfolded after training in different conditions. They found that a two-minute VR training in a virtual replica of the maze (either using an HMD or a monitor) was no more effective than maps, while a five-minute VR training with an HMD eventually surpassed training in the real world. However, no difference was found between an HMD and a monitor when used for the same amount of time. McNamara et al. [35] found no significant differences in navigation performance between a continuous curve screen and a cheaper alternative of the same size made by tiling off-the-shelf screens. Li and Giudice [36] found that display and rotation method did not affect pointing and navigation performance in a multi-level virtual building. Aoki et al. [17] found that users wearing an HMD were significantly faster than those using a monitor when pointing from destination to start location and from start toward a different destination. However, it is important to note that the input device was different, and no significant differences were found for other navigation performance measures. In [37], CAVE users performed better than monitor users in navigating without motion constraints, but monitor users outperformed CAVE users when motion constraints were enabled. Kober et al. [38] focused instead on sense of presence: a stereo projected screen produced more sense of presence than a monitor in users navigating a virtual maze. Zanbaka et al. [39] studied presence, cognition, and ability to draw a map of a virtual room. While their focus was on navigation methods using different input devices, they also compared three HMD conditions with a monitor condition. All HMD conditions led to higher presence than the monitor, while only the HMD condition with the highest interaction fidelity led to better cognition and drawing. In summary, in navigation tasks, displays with higher fidelity led to an increase in presence, while an increase in performance was found only in some cases (e.g., in absence of motion constrains).

In recent years, studies concerning displays with different fidelity have been extended to video games. In [40], users preferred playing games on a 3D stereo monitor rather than the same monitor in 2D mode, but performance in the games was similar. Lugrin et al. [41] contrasted a FPS game on a CAVE and on a monitor. Users' subjective preferences were clearly in favor of the CAVE, but performance in the game was better with the monitor, possibly because of a different input mechanism. Display and interaction fidelity in a FPS game were studied independently in [8]: users tried a CAVE (high display) with a 6-DOF wand (high input), a single screen (low display) with mouse and keyboard (low input), a low display and high input condition, a low input and high display condition. Results showed that performance in the game was better in low display with low input and high display with high input conditions rather than in the two mixed conditions. Subjective scores for presence, engagement, and usability were higher for the high display with high input condition. In summary, the studies concerning games do not show performance gains due to higher fidelity displays, but indicate that users tend to prefer higher fidelity displays to play games.

Finally, focusing on studies that assessed knowledge gain using different types of display, Limniou et al. [42] found that 3D animations in a CAVE are better than 2D animations on a desktop PC with a projector to understand molecule structure and chemical reactions. In [43], a 15-minute seminar was provided to different groups of users under four conditions: real classroom, 3D desktop, 3D HMD, and audio only. Authors found that only the real classroom was able to obtain presence results that were significantly higher than the other conditions. Moreover, reported presence was not associated with memory recall. Conversely, greater presence was reported as a predictor of better learning in a study that compared an HMD and a desktop monitor to understand water movement and salinity [44]. The higher fidelity display was more effective, but the result was found only for water movement. In [45], an HMD was compared with a large-screen projection display for training in communication skills with two virtual patients. Authors found that users' self-ratings of empathy were higher with the HMD, but users of the large-screen projection display were able to reflect more accurately on their use of empathy. Fassbender et al. [46] studied the effects of two types of display combined with background music for studying a virtual history lesson. The authors found that users remembered a higher number of facts using a 3-monitor setting rather than using a larger curved screen, but they also found an interaction between display and background music.

Only a few studies considered the effects of different types of displays on procedural knowledge acquisition. Sowndararajan et al. [47] tested the effects of using a typical laptop display and a large L-shaped two-screen projection display to memorize a simple and a complex procedure that consisted in moving objects among different spatial locations (e.g., "pick up the purple bottle" in a particular place, use it on a virtual character, and put it back). They found that users who tried the highest fidelity display performed the complex procedure faster and with less mistakes, while no difference was found for the simple procedure. A follow up study by Ragan et al. [48] used a CAVE to test independently the effects of different display fidelity components on the memorization of spatial procedures concerning the movement of objects with different forms and colors. They found that matched software FOV (i.e., the FOV set by the software was the same of the hardware), high FOV, and high FOR all contributed to memorization. They also showed that the training could be transferred from the virtual environment to the real world, by asking users to reproduce the procedure on physical objects.

In summary, no clear effect emerged from the different studies that concerned knowledge: further research is needed to understand the conditions under which different types of display could have an effect on presence and knowledge gain. In particular, the few studies concerning procedural knowledge indicated that using higher fidelity displays improves learning of spatial procedures, but, to the best of our knowledge, no study tested the effects of different displays on procedural training scenarios different from the movement of objects among different spatial locations. Another aspect that was not investigated in the literature concerns display effects on knowledge retention. The reported studies assessed knowledge gain immediately after using the displays, but did not assess if such knowledge was maintained over a longer time span.

Overall, current studies show that higher fidelity displays could increase presence, be preferred by users, and improve performance, but such effects greatly vary with the nature of the task. Our study aims to advance research on display effects in VR by assessing if, and which of, these effects apply to the largely unexplored domain of procedural training, considering procedures that are different from moving objects among spatial locations. Moreover, unlike previous studies, we extend our attention to a period of two weeks after the training to assess knowledge retention.

3 The Considered Serious Game

The serious game we used to test the effects of the different types of display allows players to experience a full emergency evacuation of a commercial twin-aisle, narrow-body aircraft after a runway overrun. The scenario is partially inspired by the real accident that occurred to Air France Flight 358 [49], which crashed into a field near Toronto International Airport after overshooting the runway because of inclement weather. The evacuation of the aircraft was made more complex by fire, which also made some exits unusable. The serious game vividly displays the consequences of players' errors in such scenario. In the following, we present in detail the plot of the game and the effects of players' actions.

3.1 Game Plot and Players' Actions

The game plot begins on-board the aircraft approaching the destination airport, twenty seconds before the captain announces very bad weather conditions, and ends when the player succeeds in reaching a safe place, after the aircraft overruns the runway and crashes in a nearby field. If players choose correct actions, they progress in the evacuation of the aircraft; if they choose wrong actions or omit right ones, they trigger negative feedback and recommendations about proper behavior. In particular, if the error is irreversible in the real world (e.g., jumping from wings instead of using the slides, or not keeping the brace position during impact), the game shows the negative consequences of the error, and pauses while a brief textual recommendation is displayed for 7 seconds. Then, it restarts from where the player took the wrong decision. On the contrary, if the error is reversible in the real world (e.g., taking luggage or trying to go towards an exit that is not the closest one), then the game does not stop, and nearby characters (passengers or flight attendants) give verbally the recommendation to the player. However, if the player ignores the verbal recommendation, and persists in the error (e.g., keeps luggage or keeps going in the wrong direction), then the game treats the error in the same way as irreversible ones.

In detail, the game plot develops following this sequence of steps:

1. The player is on a seat near the left aisle, two seat rows from the closest exit. The aircraft is flying normally, and passengers look calm. The player can hear normal engine sounds, and people chattering.
2. The captain announces that the aircraft is approaching the destination airport. He warns passengers that the weather is bad and can cause turbulence, asking them to fasten their seat belts. If players do not fasten seat belts within a few seconds, turbulence throws their avatar against the forward seat. The avatar hits the seat with the head, blood spatters on the view, and the player has to repeat step 2.
3. The aircraft is on the runway, but does not touch down. The captain realizes that the aircraft will overrun the runway, and asks passengers to prepare for an emergency by assuming the brace position (Figure 1A). Flight attendants keep shouting "Brace!" until the aircraft crashes on a field. If the player does not assume and keep the brace position until the aircraft comes to a complete stop, his/her avatar gets injured (as in step 2), and (s)he has to repeat step 3.
4. After the impact, some passengers are lightly injured, and have bloodstains on their faces (Figure 1B). The crew orders evacuation, and the player can choose among different actions: unfasten seat belts, take the life vest, and stand up. If the player tries to take the life vest, the nearby passenger tells him/her not to waste time with it, because the aircraft is not on water. If the player persists in wasting time on the seat, then (s)he hears an explosion, is reached by fire, and has to repeat step 4.
5. After the player stands up, (s)he can reach the aisle and possibly take his/her luggage. If the player takes the luggage, avatar movement becomes slow, and other passengers complain about the slowdown of the evacuation, telling the player to drop luggage. If the player does not drop it within a few seconds, then the error becomes irreversible, and the game restarts from the instant before luggage was taken.
6. The player can move towards an exit. If (s)he goes in the direction of the farther exits, a passenger blocks the way (Figure 1B) telling the player to go towards the closest exit. If the player persists in the wrong way, (s)he loses time, is reached by fire, and has to repeat step 6.
7. When the player approaches the closest exit, an explosion causes fire and smoke to propagate rapidly from the rear of the aircraft (Figure 1C). The flight attendant who assists passengers at that exit orders everyone to go towards the front of the aircraft because fire is coming. The player has to reach an exit

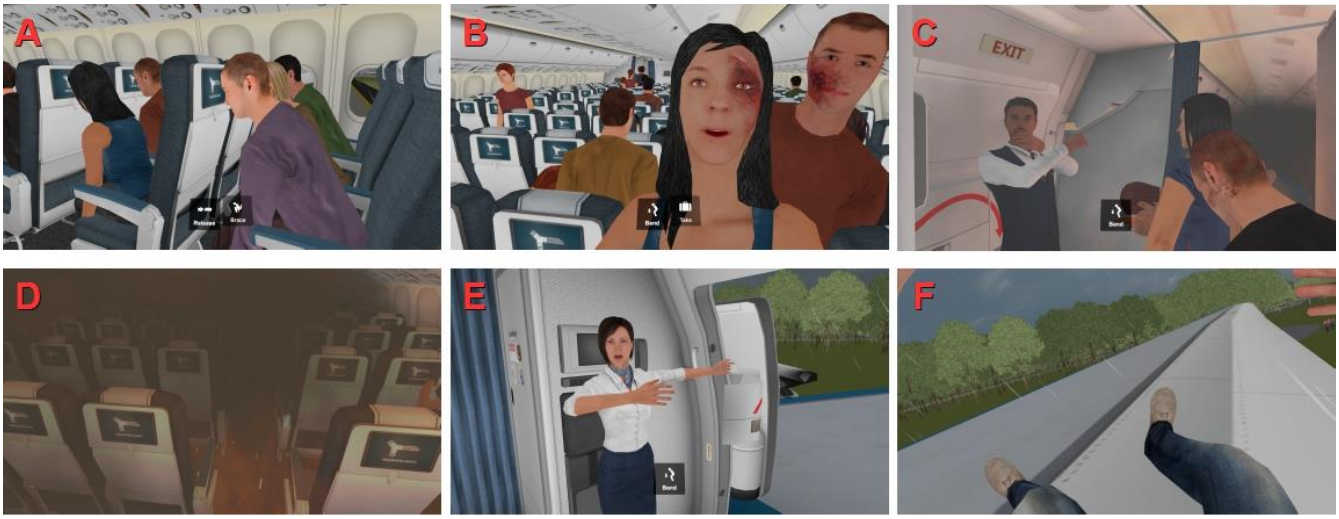


Fig. 1. Some screenshots from the serious game: a) passengers assuming the brace position, b) an injured passenger tells the player that the closest exit is in the opposite direction, c) fire and smoke begin to reach the closest exit, d) fire begins to reach the player's location, e) a flight assistant controls the evacuation at wing exit, f) player's avatar falls from the wing.

on the wings (the closest usable one), while smoke continues to propagate towards the front. If the player does not choose the "bend down" action, and keeps standing in smoke for more than 5 seconds, the avatar starts coughing, and loses consciousness. Then, the game restarts from the instant before coughing. If the player moves too slowly, fire reaches his/her avatar (Figure 1D), and the game restarts from 7 seconds before fire reached the player.

8. When the player reaches a wing exit, the flight attendant near the exit tells him/her to exit on the wing, reach the slide, and jump down (Figure 1E). If the player hesitates or tries to go towards the front, the flight attendant shouts to order him/her to use that exit. If the player does not comply within a few seconds, the avatar is reached by fire and the player has to repeat step 8.
9. When the player leaves the cabin and is on a wing, (s)he can move towards the wing slide and jump down. If (s)he tries instead to jump down directly from the wing without using the slide, the avatar gets injured in the fall (Figure 1F). Moreover, if the player wastes time on the wing before using the slide, an explosion injures the avatar. In both cases, the player has to repeat step 9.
10. When the player reaches the ground through the slide, (s)he can freely move around the aircraft. However, if (s)he remains close to the burning aircraft, an explosion eventually injures the avatar (as in step 9), and the player has to repeat step 10. The evacuation completes successfully when the player reaches a safe distance from the aircraft.

3.2 Interaction with the Game

Players interact with the game using an Xbox 360 controller (Figure 2). To move in the direction they are currently facing (resp. in the opposite direction), players push the left joystick forward (resp. backward). Left and right movement of the left joystick changes one's place when one is in a seat row. To rotate the avatar, players move the right joystick in the left or right direction. When the display is a desktop monitor, the right joystick can be moved backward/forward to rotate the head of the avatar up and down (in the HMD conditions, this rotation is instead controlled by head tracking). Actions described in the previous section (e.g., fastening seat belts, assuming brace position, and so on) are carried out by selecting semitransparent icons superimposed on the scene (see examples in Figure 1A, B, C, and E). Each time the player has to take an action, between one and three icons appear. Such icons refer to actions that can be taken in the real world at the current location. However, as seen in the previous section, not all actions one can take are appropriate for the emergency scenario, e.g. taking luggage after standing up from the seat. Players select icons using left and right arrows on the D-pad control. The currently selected icon is highlighted on the display by changing its: (i) distance from user's eyes (it becomes slightly closer), (ii) its background (fully opaque), and (iii) its brightness (higher). Players can perform the selected action by pressing the "A" button on the controller.

Fig. 2. The Xbox 360 controller.

4 User Evaluation

To evaluate the effects of the type of display, we carried out a between-groups study. In the following, we will refer to the group of participants who played the game on the desktop monitor as Low Fidelity (LF) group, those who used the HMD with narrow FOV and 3-DOF tracker as Medium Fidelity (MF) group, and those who used the HMD with wide FOV and 6-DOF tracker as High Fidelity (HF) group. It is important to remark that, in the group names, "fidelity" refers to display fidelity, but also to interaction fidelity, since the MF and HF conditions used head tracking (with 3DOF and 6 DOF, respectively) and LF used the joystick for head rotation. The other aspects of interaction fidelity as well as scenario fidelity were the same for all groups.

We formulated four hypothesis for the study. First, we hypothesized that all groups should show a knowledge gain immediately after use, since all participants were exposed to exactly the same scenario, and in [11] we found that playing a cabin safety scenario produced a knowledge gain. Second, since gaining experience in performing a given behavior is a major factor that contributes to increase self-efficacy [21], we hypothesized that being able to complete the serious game by successfully evacuating from the aircraft should increase user's self-efficacy immediately after use in all groups. Third, the display should instead play a role in how the scenario is experienced, affecting self-reported engagement (as suggested by previous studies concerning games [8], [40], [41]), and self-reported presence (as suggested by [8], [27], [38], [39]). In particular, these measures should return higher values as fidelity increases. Fourth, if the previous hypothesis holds, higher fidelity could also result in more knowledge retention, since presence is linked with emotions [23], and emotions have a positive effect on memory retention [24], [25].

4.1 Materials

The serious game was implemented using the Unity 4.6 game engine, and run on a PC equipped with a 3.60 GHz Intel i7-4790 processor, 16 GB RAM, and an NVidia GTX 970 graphic card. The monitor used by the LF group was an Asus VX279H 27" display with 1920x1080 resolution (Figure 3A). The Sony HMD used by the MF group had two OLED displays with 1280x720 resolution each and 45° FOV), and the 3-DOF sensor was an InterSense InertiaCube3 (Figure 3B). The Oculus HMD used by the HF group had an OLED display with 1920x1080 resolution, 100° FOV, and its own 6-DOF tracker (Figure 3C). In all conditions, participants listened to audio through Sennheiser HD 215 closed earphones. The Xbox 360 controller was connected wirelessly to the PC.

Fig. 3. Displays employed in the a) Low Fidelity group, b) Medium Fidelity group, and c) High Fidelity group.

4.2 Participants

The evaluation involved a sample of 96 participants (55M, 41F). Participants were volunteers who received no compensation and were recruited through personal contact. Age ranged from 18 to 36 (M=23.81, SD=3.58).

We asked participants to rate their frequency of use of video games on a 7-point scale (1=never, 2=less than once a month, 3=about once a month, 4=several times a month, 5=several times a week, 6=every day for less than an hour, 7=every day for more than one hour). Answers ranged from 1 to 7 (median=4; 15 users never played video games, 15 played less than once a month, 11 played about once a month, 26 played several times a month, 16 played several times a week, and 13 played every day).

We also assessed individual differences in frequency of air travel by asking participants to count their number of flights in the last two years, as in [9]. Each flight had to be counted individually (e.g., a round trip from airport A to airport C via a connection through airport B results in four flights). Answers ranged from 0 to 15 (M=3.03, SD=2.93).

Finally, we used the 32-items Flight Anxiety Situations questionnaire (FAS) developed by [50] to assess participants' anxiety in flight-related situations, and control for it in the analysis of engagement and presence, in case individual sensitivity to the considered situations could affect emotional response. Each FAS item is rated on a 5-point scale, ranging from 1 (no anxiety) to 5 (overwhelming anxiety). The total FAS score is obtained by summing all item scores, and can thus range from 32 to 160. In our sample, the FAS score ranged from 32 to 117 (M=57.66, SD=20.11).

Participants were assigned to the three groups in such a way that: (i) each group had 32 participants (18M, 14F in the LF and HF groups; 19M, 13F in the MF group); (ii) the three groups were similar in terms of age (LF: M=24.53, SD=3.82; MF: M=23.84, SD=4.07; HF: M=23.06, SD=2.65), frequency of video game use (LF, HF: median=4; MF: median=3.5), number of flights (LF: M=2.84, SD=3.10; MF: M=3.19, SD=2.81; HF: M=3.06, SD=2.96), and FAS score (LF: M=59.75, SD=20.99; MF: M=55.91, SD=21.02; HF: M=57.31, SD=18.66). Lack of significant differences among the three groups was confirmed by one-way ANOVA for age, frequency of air travel, and flight-related anxiety, and by Kruskal-Wallis test (used because the variable was ordinal) for frequency of video game use.

4.3 Measures

4.3.1 *Knowledge*

To measure participants' knowledge about cabin safety, we used a test with nine questions: 1) what to do in case of turbulence; 2) what to do in preparation for impact; 3) which exit should be the first choice for evacuation; 4) when it is not possible to use an exit; 5) what to do if the chosen exit cannot be used; 6) what to do if there is smoke in the cabin during evacuation; 7) what to do after using a wing exit; 8) what to do after leaving the aircraft; 9) what to do with luggage. Participants were asked to answer the questions orally to avoid suggesting possible answers (e.g., as in multiple-choice questionnaires). Answers were audio recorded and later rated by the experimenter as correct or wrong, following a codebook that listed the possible answers and their rating (right/wrong). Knowledge was measured as the number of correctly answered questions, and thus ranged between 0 and 9.

To measure knowledge acquisition as well as retention, we administered the knowledge test three times: before trying the serious game (pre-test), immediately after trying it (post-test), and two weeks later (retention-test). Mean pre-test knowledge score showed that participants were initially able to answer correctly only about half of the questions (M=4.56, SD=1.94). One-way ANOVA showed no significant differences in initial knowledge between the three groups.

4.3.2 Self-Efficacy

To measure self-efficacy, we used a questionnaire with six items: 1) I feel able to deal with an emergency evacuation of an aircraft; 2) I would be able to deal with an emergency evacuation even if the aircraft is on fire; 3) I would be able to deal with an emergency evacuation even if one or more exits are blocked; 4) I would be able to deal with an emergency evacuation even if most of the passengers scream or cry; 5) I feel confident of my ability to exit from the aircraft in time; 6) I would be able to help passengers in need. The questionnaire was designed by adapting items from well-known self-efficacy questionnaires (General Self-Efficacy scale [51]) to our domain, and following the recommendations on rigorous theory-based semantic structure for specific behaviors proposed by [52]. Each item was rated by participants on a 7-point scale (1=not at all, 7=very). We administered the self-efficacy questionnaire three times (pre-test, post-test, and retention-test). Answers were averaged to form a reliable scale (Cronbach's alpha pre-test=0.92, post-test=0.94, retention-test=0.92). Mean pre-test self-efficacy was only 2.91 (SD=1.23). One-way ANOVA showed no significant differences in initial self-efficacy between the three groups.

4.3.3 Self-Reported Engagement

To measure the level of engagement experienced by participants, we administered a questionnaire that asked them to rate their level of agreement about six statements on a 7-point scale (1=not at all, 7=very). The six items were: "It was boring", "It was engaging", "It aroused emotions in me", "The depicted situation looked real", "I forgot the passing of time", "I felt immersed in the depicted situation". After inverting the scale of the first item, the six ratings were averaged to form a reliable scale (Cronbach's alpha=0.87).

4.3.4 Self-Reported Presence

To measure the sense of presence experienced by participants while playing the game, we administered the Igroup Presence Questionnaire (IPQ) [53]. The IPQ is a 14-item self-report scale (available at http://www.igroup.org/pq/ipq/index.php), comprising a general item related to the sense of "being there", and three subscales that evaluate three independent dimensions of the VR experience: spatial presence (5 items), involvement (4 items) and experienced realism (4 items). Participants have to rate their degree of agreement with IPQ statements on a 7-point Likert scale, ranging from 0 to 6.

4.4 Procedure

Participants were told that we were testing a software system for learning safety procedures concerning emergency landing and evacuation of an aircraft. Written consent for participation and for recording verbal answers to the knowledge test was obtained from participants. They were also informed that they were going to be contacted again two weeks later for an additional questionnaire (without specifying what those further questions were going to concern), and that they could refrain from continuing the experiment at any time, without providing a reason to the experimenter. Then, participants filled an initial demographic questionnaire (gender, age, frequency of game use, and frequency of air travel), verbally answered the knowledge questions, and filled the FAS and the self-efficacy questionnaires.

The experimenter gave participants the game controller, and invited them to memorize the position of the controls highlighted in Figure 2. This was particularly important for the MF and HF groups, since those participants were not going to be able to see the controller while wearing the HMD. The experimenter helped participants of the MF and HF groups to adjust the HMD until they could see well and feel comfortable with it. Then, for all three groups, the experimenter explained the controls while participants were playing an initial tutorial level, in which they boarded the aircraft and moved in the cabin to reach their seat. When they were close to the seat, they had to perform the following sequence of actions: put their luggage in the overhead bin, sidestep in their seat row to reach the assigned seat, sit down, and fasten seat belts. In this way, participants could familiarize with all game controls needed to play the runway overrun scenario. Then, participants wore the closed earphones, and tried the runway overrun scenario.

After the experimental condition, participants filled the presence, engagement, and self-efficacy questionnaires. Then, they verbally answered the knowledge questions. Finally, the experimenter briefly interviewed participants, asking them what were their impressions about the serious game, what they liked or disliked, what they found difficult, and what they would change.

Two weeks later, the experimenter contacted participants to assess again knowledge and self-efficacy.

5 Results

5.1 Knowledge

Knowledge scores (Figure 4) were submitted to a 3 x 3 mixed design ANOVA, in which group served as the between-subjects variable, and time of measurement (pre-test, post-test, and retention-test) served as the within-subjects variable. Mauchly's test indicated that the assumption of sphericity had been violated ($\chi^2(2)=26.79$, $p<0.001$), therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon=0.80$). Statistically significant results revealed a main effect of time of measurement, $F(1.60, 148.49)=230.60$, $p<0.001$, $\eta_p^2=0.71$, while the main effect of group, and the group by time of measurement interaction were not statistically significant.

We thus explored the significant main effect, using Bonferroni post-hoc comparison to test each pair of time of measurement levels for significance. The differences between pre-test (M=4.56, SD=1.94) and post-test (M=7.63, SD=1.28) knowledge, as well as between pre-test and retention-test (M=7.63, SD=1.16) knowledge, were statistically significant ($p<0.001$), while the difference between post-test and retention-test knowledge was not statistically significant.

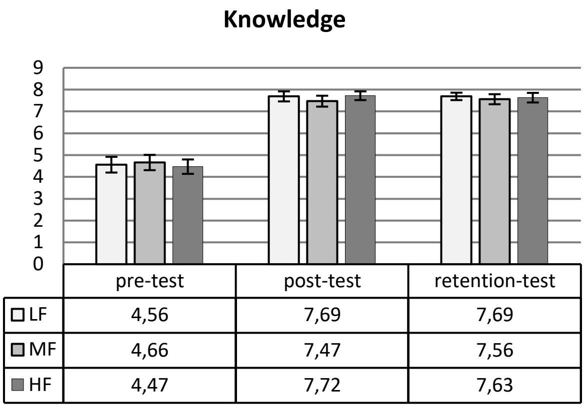


Fig. 4. Means of the knowledge score at pre-test, post-test, and retention-test. Capped vertical bars indicate ± SE.

5.2 Self-Efficacy

Self-efficacy scores (Figure 5) were submitted to a 3 x 3 mixed design ANOVA, in which group served as the between-subjects variable, and time of measurement (pre-test, post-test, and retention-test) served as the within-subjects variable. Mauchly's test indicated that the assumption of sphericity had been violated ($\chi^2(2)=11.81$, $p<0.005$), therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon=0.89$). Statistically significant results revealed a main effect of time of measurement, $F(1.79, 166.00)=89.71$, $p<0.001$, $\eta_p^2=0.49$, while the main effect of group and the group by time of measurement interaction were not significant.

We thus explored the significant main effect, using Bonferroni post hoc comparison to test each pair of time of measurement levels for significance. The differences between pre-test (M=2.91, SD=1.23) and post-test (M=4.08, SD=1.36) scores, as well as between pre-test and retention-test (M=4.39, SD=1.12) scores, were statistically significant ($p<0.001$). The difference between post-test and retention-test scores was also statistically significant ($p<0.005$).

Fig. 5. Means of the self-efficacy score at pre-test, post-test, and retention-test. Capped vertical bars indicate ± SE.

5.3 Self-Reported Engagement

Differences in self-reported engagement (Figure 6) were analyzed with a between-subjects ANCOVA, controlling for participant's flight-related anxiety (FAS score). The analysis revealed a statistically significant difference, $F(2, 92)=4.64$, $p<0.05$, $\eta_p^2=0.09$. Pairwise comparisons using Bonferroni test revealed that the difference between the LF group (M=4.48, SD=1.16) and the HF group (M=5.28, SD=1.14) was statistically significant ($p<0.05$). The difference between the MF group (M=4.58, SD=1.15) and the HF group as well as the difference between LF and MF did not reach significance.

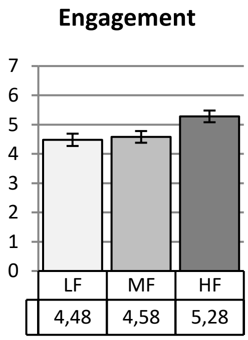


Fig. 6. Means of self-reported engagement. Capped vertical bars indicate ± SE.

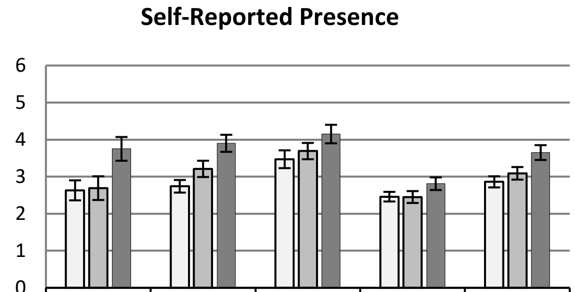
5.4 Self-Reported Presence

Differences in self-reported presence (Figure 7) were analyzed with a between-subjects ANCOVA, controlling for participant's flight-related anxiety (FAS score). Considering the IPQ total score, the analysis revealed a statistically significant difference, $F(2, 92)=5.80$, $p<0.005$, $\eta_p^2=0.11$. Pairwise comparisons using Bonferroni test revealed that the difference between the LF group (M=2.86, SD=0.82) and the HF group (M=3.65, SD=1.13) was statistically significant ($p<0.005$). The difference between the MF group (M=3.09, SD=0.94) and the HF group as well as the difference between LF and MF were not statistically significant.

Considering the general item about the sense of "being there", the analysis revealed a statistically significant difference, $F(2, 92)=4.47$, $p<0.05$, $\eta_p^2=0.09$. Pairwise comparisons using Bonferroni test revealed that the difference between the LF group (M=2.63, SD=1.54) and the HF group (M=3.75, SD=1.80) was statistically significant ($p<0.05$). The difference between the MF group (M=2.69, SD=1.82) and the HF group was also statistically significant ($p<0.05$). Finally, the difference between LF and MF was not statistically significant.

Considering the subscale about spatial presence, the analysis revealed a statistically significant difference, F(2, 92)=8.08, $p<0.005$, $\eta_p^2=0.15$. Pairwise comparisons using Bonferroni test revealed that the difference between the LF group (M=2.74, SD=0.98) and the HF group (M=3.90, SD=1.32) was statistically significant ($p<0.001$). The difference between the MF group (M=3.21, SD=1.26) and the HF group as well as the difference between LF and MF were not statistically significant.

No statistically significant differences were found for the involvement and realism subscales of the IPQ.



	Being there	Spatial	Involve-ment	Realism	Total Score
LF	2,63	2,74	3,47	2,46	2,86
MF	2,69	3,21	3,69	2,45	3,09
HF	3,75	3,90	4,15	2,81	3,65

Fig. 7. Means of self-reported presence. Capped vertical bars indicate ± SE.

6 Discussion

The results confirmed most of our hypotheses. First, the statistical analysis highlighted a significant gain in knowledge between pre-test and post-test. This is consistent with evidence provided by other studies (e.g., [11], [54], [55]) about the effectiveness of serious games for learning safety procedures. However, our study considered three displays with different fidelity, and showed that the serious game significantly increased participants' safety knowledge regardless of the display used to play it. It is worth to add that the type of display had no significant effect on playing time: one-way ANOVA showed no significant differences in the amount of time participants spent to complete the scenario (LF: M=258.9 s, SD=104.5; MF: M=266.3 s, SD=106.6; HF: M=237.0 s, SD=66.8), so all participants were exposed to the same safety materials for a similar amount of time. The fact that the HF condition was no better than the two lower fidelity conditions in terms of knowledge gain may seem in contrast with the findings in [47] and [48], which instead showed a positive effect of higher display fidelity on memorization of complex procedures. However, it is important to note that those previous studies concerned memorization of procedures that consisted in moving objects among different spatial locations. Therefore, as discussed in both [47] and [48], the displays with higher fidelity could better present spatial cues, allowing users to exploit a spatial memory strategy (similar to the "method of loci" [56]) for memorizing the procedure. Our study, instead, explored the effects of display type on safety procedures that are different from moving objects among different spatial locations. The questions asked in the knowledge test were not focused on assessing spatial abilities, but on the learning of cabin safety procedures. The identification of relations between events (e.g., the presence of smoke) and actions to perform (e.g., bending down), as well as relations between performed or omitted actions (e.g., not bending down) and avatar damage (e.g., coughing and losing consciousness), had thus a more prominent role than the acquisition of accurate spatial information. Therefore, scenario fidelity could have played a more important role than display fidelity in learning the procedures.

The results confirmed our hypothesis on self-efficacy, which significantly increased between pre-test and post-test in all three groups. After using the game, participants were more confident in their ability to deal with an emergency evacuation of an aircraft. As expected, there was no effect of display type on post-test self-efficacy, since a major factor that contributes to increase self-efficacy is gaining experience in performing the given behavior [21] and participants in all the three groups successfully tried the same scenario.

Display type played instead a role in self-reported engagement and presence. As hypothesized, higher fidelity resulted in higher engagement and higher presence, but the result did not reach statistical significance for all comparisons between displays. More precisely, we found that playing the serious game using the wider FOV HMD with the 6-DOF tracker was significantly more engaging than playing it on the monitor. Mean engagement was also higher using the higher fidelity HMD rather than the narrower FOV HMD with the 3-DOF tracker, but the difference between them did not reach significance. Interestingly, the narrower FOV HMD was not able to produce a significant difference with respect to the monitor. These results advance knowledge about the effects of display type on games: while previous studies compared only two displays, and showed that players' preference went to the higher fidelity display [8], [40], [41], our study compared three types of display, and indicated that only the highest fidelity HMD (wider FOV and 6-DOF head tracking) led to a significant increase in self-reported engagement with respect to the monitor. Participants' feedback elicited by means of short interviews helped us understand this finding: 13 participants in the HF group mentioned immersion aspects (which could be due to both wider FOV and 6-DOF tracking) as one of the features of the experience they liked the most, while four participants in the MF group complained about the display they used. In particular, two of them mentioned that they felt as if a "movie theater" screen was in front of them: the narrow FOV of the HMD might have reduced immersion and engagement, making it similar to a desktop monitor.

Participants' feedback was consistent with the results we found for presence. There was a statistically significant difference only between the HF group (highest presence scores) and the LF group (lowest presence scores) for the total score and the spatial subscale, and also a difference between HF and MF for the general item about the sense of "being there". The means of the other two subscales (involvement and realism) differed in the same way, but no statistical significance was found. Overall, our results confirm previous studies about the positive effects of higher fidelity displays on presence [8], [27], [38], [39], [44], and extend the findings to the domain of procedural training. Moreover, while those studies compared six-screen CAVE vs. single screen [8], desktop vs.

HMD vs. CAVE [27], desktop vs. single stereo projected screen [38], or desktop vs. HMD [39], [44], we compared desktop vs. two different types of HMDs and found that the two types of HMDs led to different presence scores. Moreover, we found that the wider FOV HMD with the 6-DOF tracker significantly increased presence with respect to the monitor, while the narrower FOV HMD with the 3-DOF tracker did not. This can be due to both the wider FOV and the 6-DOF tracker. Indeed, the wider FOV HMD allows expanding the peripheral view and likely reduces the possibility that users notice the case that holds the LCD panel. On the contrary, users of the narrower FOV HMD can easily notice the black borders around the LCD panel and feel as if a "movie theater" screen was in front of them. This has likely an impact on the sense of "being there" as well as on the ratings of some items in the spatial subscale of IPQ, such as "I felt like I was just perceiving pictures". At the same time, head tracking can also affect the sense of "being there" and spatial presence. The 6-DOF tracker can reproduce the effect of fully moving the head in the virtual environment. On the contrary, the translation of the head with the 3-DOF tracker can be perceived as "moving the environment" instead of moving inside the environment, because the camera position is not updated and thus the environment has no relative movement with respect to user's head. Further studies can separately address FOV and head tracking to understand how much each of the two affects presence. Nevertheless, our findings suggest that the new devices for home VR, such as the Oculus Rift, can enable virtual experiences where users can feel a sense of presence that is significantly higher than the past. This fosters future research work to investigate the exploitation of the new devices for procedural training not limited to the safety domain. Our findings also suggest the need to extend previous work that studied the effects of HMDs with narrow FOV and/or a 3-DOF tracker (e.g., [14], [17], [27], [43]). In particular, the outcomes of studies that compared presence with such HMDs and other displays (e.g., [27], [43]) might change using new, wider FOV HMDs with a 6-DOF tracker. Finally, our study extended previous work by considering the effect of display type on the different aspects of presence. On one hand, we found a clear positive effect of high fidelity on the sense of "being there", spatial presence, and overall sense of presence. On the other hand, such positive effect was not so evident for the involvement and realism subscales, suggesting that these measures might be more influenced by other aspects of fidelity that were kept constant for all groups in our study. The notion of involvement assessed by the IPQ was defined in [57] as "a psychological state experienced as a consequence of focusing one's energy and attention on a coherent set of stimuli or meaningfully related activities and events". Therefore, the generally positive scores for involvement could have been affected by the fidelity of the scenario, which presented a coherent sequence of events and required users to perform meaningfully related actions to deal with them.

The hypothesis about the effect of display type on knowledge retention was not met. The knowledge gained using the serious game was retained two weeks later regardless of display type. This was not expected, since presence was affected by the display used, and presence is linked with emotions [23], which have an impact on memory retention [24], [25]. Other aspects of the VR experience might have had a larger impact than presence on emotion and retention. For example, the surprising events in the game plot and the vivid depiction of the effects of player errors could have elicited negative emotions (which contribute to memory consolidation [24] and retention [25]) in all participants, including those who used lower fidelity displays and reported lower levels of presence. Unlike our previous work on knowledge retention [11], which assessed retention of safety procedures after 1 week, the present study considered a longer time span, showing that knowledge was retained even after 2 weeks.

Similarly, there was no effect of type of display on self-efficacy assessed after 2 weeks, and self-efficacy significantly increased also between post-test and retention-test in all groups. The fact that all participants succeeded in the evacuation and were successful at remembering over time how they did it might have contributed to the further increase in self-efficacy.

In summary, we found that changing the type of display affected users' engagement and sense of presence, while it did not significantly affect the increase in knowledge and self-efficacy obtained by playing the serious game. Although this result suggests that using desktop VR setups could be sufficient for procedural safety training, it must be noted that our participants spent a similar amount of time in the game, and none of them had the possibility to try it again between post-test and retention-test. On the contrary, in non-experimental settings, the availability of high-fidelity VR displays that – as we have seen – produce more engagement and presence could make users more likely to play the game again, refreshing and reinforcing the knowledge acquired as well as acquiring more knowledge.

7 Conclusion

In this paper, we studied the effects of different types of display in a VR-based procedural training scenario. To the best of our knowledge, this is the first study of the topic on safety procedures, and we found that changing the type of display affected users' engagement and sense of presence, while it did not significantly affect the increase in knowledge and self-efficacy. Analysis of the different aspects of presence found that overall presence, spatial presence, and the sense of "being there" were highest using the highest fidelity display, while no significant differences were found on the involvement and realism aspects of presence. It is worth noting that only the highest fidelity HMD led to significantly higher engagement and presence with respect to the monitor. Finally, our study was the first to explore the effects of different types of display at two weeks after the experimental condition, and showed that the acquired benefits in terms of knowledge and self-efficacy were retained, regardless of the display used.

Further research is needed to investigate the role of interaction fidelity. While this study investigated the overall effects of three typical VR display setups, we plan to carry out a study that will focus on the highest fidelity display, and assesses its effects with 3-DOF vs. 6-DOF head tracking enabled. This will allow us to explore how much of the presence and engagement results could possibly be due to the head tracking component.

We should also consider that different input devices for home VR are going to reach the market (e.g., Oculus Touch, Virtuix Omni,…), so it would be interesting to test the effects of further increases in interaction fidelity. In particular, using the same high fidelity display, we plan to compare the use of a joystick vs. hand tracking for activating actions, assessing if higher interaction fidelity will result in higher presence and engagement. Finally, it will be interesting to study if higher interaction fidelity could have a positive effect on procedural knowledge gain and retention, especially for procedures that require to perform physical actions.

Acknowledgment

Our research is supported by a grant of the Federal Aviation Administration (FAA).

We are grateful to Cynthia Corbett and Mac McLean for their precious feedback and encouragement.

Nicola Zangrando (HCI Lab, University of Udine) carried out 3D modeling activities for the development of the game.

References

[1] J.C. Silverstein, F. Dech, M. Edison, P. Jurek, W. S. Helton, and N.J. Espat, "Virtual reality: Immersive hepatic surgery educational environment," *Surgery*, vol. 132, no. 2, pp. 274–277, Aug. 2002.

[2] R. Hill, J. Morie, J. Rickel, M. Thiébaux, L. Tuch, R. Whitney, J. Douglas, W. Swartout, J. Gratch, W.L. Johnson, C. Kyriakakis, C. LaBore, R. Lindheim, S. Marsella, D. Miraglia, and B. Moore, "Toward the Holodeck: Integrating Graphics, Sound, Character and Story," *Proc. fifth int'l conf. on Autonomous agents - AGENTS '01,* 2001, pp. 409–416.

[3] M. Cha, S. Han, J. Lee, and B. Choi, "A virtual reality based fire training simulator integrated with fire dynamics data," *Fire Saf. J.,* vol. 50, pp. 12–24, May 2012.

[4] M. Zyda, "From visual simulation to virtual reality to games," *IEEE Comput.*, vol. 38, no. 9, pp. 25–32, Sep. 2005.

[5] Z. Merchant, E.T. Goetz, L. Cifuentes, W. Keeney-Kennicutt, and T.J. Davis, "Effectiveness of virtual reality-based instruction on students' learning outcomes in K-12 and higher education: A meta-analysis," *Comput. Educ.*, vol. 70, pp. 29–40, Jan. 2014.

[6] C. Girard, J. Ecalle, and A. Magnan, "Serious games as new educational tools: how effective are they? A meta-analysis of recent studies," *J. Comput. Assist. Learn.*, vol. 29, no. 3, pp. 207–219, Jun. 2013.

[7] E.D. Ragan, D.A. Bowman, R. Kopper, C. Stinson, S. Scerbo, and R.P. McMahan, "Effects of Field of View and Visual Complexity on Virtual Reality Training Effectiveness for a Visual Scanning Task," *IEEE Trans. Vis. Comput. Graph.*, vol. 21, no. 7, pp. 794–807, 2015.

[8] R.P. McMahan, D.A. Bowman, D.J. Zielinski, and R.B. Brady, "Evaluating display fidelity and interaction fidelity in a virtual reality game.," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 4, pp. 626–33, Apr. 2012.

[9] C.L. Corbett, G.A. McLean, and D.K. Cosper, "Effective Presentation Media for Passenger Safety I: Comprehension of Briefing Card Pictorials and Pictograms," Final Report DOT/FAA/AM-08/20, Washington, DC, USA, 2008.

[10] D. Cosper and G. McLean, "Availability of passenger safety information for improved survival in aircraft accidents," Technical Report DOT/FAA/AM-04/19, Washington, DC, USA, 2004.

[11] L. Chittaro and F. Buttussi, "Assessing Knowledge Retention of an Immersive Serious Game vs. a Traditional Education Method in Aviation Safety," *IEEE Trans. Vis. Comput. Graph.,* vol. 21, no. 4, pp. 529–538, Apr. 2015.

[12] M. Billinghurst, H. Kato, and I. Poupyrev, "The MagicBook: a transitional AR interface," *Comput. Graph.*, vol. 25, no. 5, pp. 745–753, Oct. 2001.

[13] B.H. Cho, J.M. Lee, J.H. Ku, D.P. Jang, J.S. Kim, I.Y. Kim, J.H. Lee, and S.I. Kim, "Attention Enhancement System using virtual reality and EEG biofeedback," *Proc. IEEE Virtual Reality 2002,* 2002, pp. 156–163.

[14] R.A. Ruddle and P. Péruch, "Effects of proprioceptive feedback and environmental characteristics on spatial learning in virtual environments," *Int. J. Hum. Comput. Stud.*, vol. 60, no. 3, pp. 299–326, 2004.

[15] D. Rand, R. Kizony, U. Feintuch, N. Katz, N. Josman, A. "Skip" Rizzo, and P.L. (Tamar) Weiss, "Comparison of Two VR Platforms for Rehabilitation: Video Capture versus HMD," *Presence Teleoperators Virtual Environ.*, vol. 14, no. 2, pp. 147–160, Apr. 2005.

[16] A.V. Sonar, K.D. Burdick, R.R. Begin, E.M. Resch, E.M. Thompson, E. Thacher, J. Searleman, G. Fulk, and J.J. Carroll, "Development of a virtual reality-based power wheel chair simulator," *Proc. IEEE Int'l Conf. on Mechatronics & Automation,* 2005, pp. 222–229.

[17] H. Aoki, C. M. Oman, D.A. Buckland, and A. Natapoff, "Desktop-VR system for preflight 3D navigation training," *Acta Astronaut.,* vol. 63, no. 7–10, pp. 841–847, Oct. 2008.

[18] B. Sousa Santos, P. Dias, A. Pimentel, J.-W. Baggerman, C. Ferreira, S. Silva, and J. Madeira, "Head-mounted display versus desktop for 3D navigation in virtual reality: a user study," *Multimed. Tools Appl.,* vol. 41, no. 1, pp. 161–181, Jan. 2009.

[19] B.J. Fogg, *Persuasive Technology: Using Computers to Change What We Think and Do.* San Francisco, CA, USA: Morgan Kaufmann, 2003.

[20] M. Slater and S. Wilbur, "A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments," *Presence Teleoper. Virtual Environ.*, vol. 6, no. 6, pp. 603–616, 1997.

[21] A. Bandura, *Self-efficacy: the exercise of control.* New York, NY, USA: Freeman, 1997.

[22] A. Bandura, "Social cognitive theory: an agentic perspective," *Annu. Rev. Psychol.*, vol. 52, pp. 1–26, Jan. 2001.

[23] G. Riva, F. Mantovani, C.S. Capideville, A. Preziosa, F. Morganti, D. Villani, A. Gaggioli, C. Botella, and M. Alcañiz, "Affective interactions using virtual reality: the link between presence and emotions.," *CyberPsychology Behav.*, vol. 10, no. 1, pp. 45–56, Feb. 2007.

[24] E.A. Kensinger, "Remembering the Details: Effects of Emotion," *Emot. Rev.*, vol. 1, no. 2, pp. 99–113, Jan. 2009.

[25] B. Finn and H.L. Roediger, "Enhancing retention through reconsolidation: negative emotional arousal following retrieval enhances later recall," *Psychol. Sci.*, vol. 22, no. 6, pp. 781–6, Jun. 2011.

[26] R. Pausch, D. Proffitt, and G. Williams, "Quantifying immersion in virtual reality," *Proc. 24th ann. Conf. on Computer graphics and interactive techniques - SIGGRAPH '97,* 1997, pp. 13–18.

[27] K. Kim, M.Z. Rosenthal, D.J. Zielinski, and R. Brady, "Effects of virtual environment platforms on emotional responses," *Comput. Methods Programs Biomed.*, vol. 113, no. 3, pp. 882–93, Mar. 2014.

[28] L. Arns, D. Cook, and C. Cruz-Neira, "The benefits of statistical visualization in an immersive environment," *Proc. IEEE Virtual Reality,* 1999, pp. 88–95.

[29] F. Bacim, E. Ragan, S. Scerbo, N.F. Polys, M. Setareh, and B.D. Jones, "The effects of display fidelity, visual complexity, and task scope on spatial understanding of 3D graphs," *Graphics Interface Conf. 2013,* 2013, pp. 25–32.

[30] B. Laha, K. Sensharma, J.D. Schiffbauer, and D.A. Bowman, "Effects of immersion on visual analysis of volume data," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 4, pp. 597–606, Apr. 2012.

[31] B. Laha, D.A. Bowman, and J.J. Socha, "Effects of VR System Fidelity on Analyzing Isosurface Visualization of Volume Datasets," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 4, pp. 513–22, Apr. 2014.

[32] D.A. Bowman and R.P. McMahan, "Virtual Reality: How Much Immersion Is Enough?," *IEEE Comput.*, vol. 40, no. 7, pp. 36–43, Jul. 2007.

[33] E.D. Ragan, R. Kopper, P. Schuchardt, and D.A. Bowman, "Studying the effects of stereo, head tracking, and field of regard on a small-scale spatial judgment task.," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 5, pp. 886–96, May 2013.

[34] D. Waller, E. Hunt, and D. Knapp, "The Transfer of Spatial Knowledge in Virtual Environment Training," *Presence Teleoperators Virtual Environ.*, vol. 7, no. 2, pp. 129–143, Apr. 1998.

[35] A.M. McNamara, F. Parke, and M. Sanford, "Evaluating performance in tiled displays," *Proc. 10th Int'l Conf. on Virtual Reality Continuum and Its Applications in Industry - VRCAI '11*, 2011, p. 483–489.

[36] H. Li and N.A. Giudice, "The effects of immersion and body-based rotation on learning multi-level indoor virtual environments," *Proc. Fifth ACM SIGSPATIAL Int'l Workshop on Indoor Spatial Awareness - ISA '13*, 2013, pp. 8–15.

[37] N. Elmqvist, M.E. Tudoreanu, and P. Tsigas, "Evaluating motion constraints for 3D wayfinding in immersive and desktop virtual environments," *Proc. twenty-sixth ann. CHI conf. on Human factors in computing systems - CHI '08*, 2008, pp. 1769–1778.

[38] S.E. Kober, J. Kurzmann, and C. Neuper, "Cortical correlate of spatial presence in 2D and 3D interactive virtual reality: an EEG study," *Int. J. Psychophysiol.*, vol. 83, no. 3, pp. 365–74, Mar. 2012.

[39] C. A. Zanbaka, B. C. Lok, S. V. Babu, A. C. Ulinski, and L. F. Hodges, "Comparison of path visualizations and cognitive measures relative to travel technique in a virtual environment," *IEEE Trans. Vis. Comput. Graph.*, vol. 11, no. 6, pp. 694–705, 2005.

[40] T. Litwiller and J.J. LaViola, "Evaluating the benefits of 3D stereo in modern video games," *Proc. 2011 ann. Conf. on Human factors in computing systems - CHI '11*, 2011, pp. 2345–2354.

[41] J.-L. Lugrin, M. Cavazza, F. Charles, M. Le Renard, J. Freeman, and J. Lessiter, "Immersive FPS games," *Proc. 2013 ACM int'l workshop on Immersive media experiences - ImmersiveMe '13*, 2013, pp. 7–12.

[42] M. Limniou, D. Roberts, and N. Papadopoulos, "Full immersive virtual environment CAVETM in chemistry education," *Comput. Educ.*, vol. 51, no. 2, pp. 584–593, Sep. 2008.

[43] K. Mania and A. Chalmers, "The Effects of Levels of Immersion on Memory and Presence in Virtual Environments: A Reality Centered Approach," *CyberPsychology Behav.*, vol. 4, no. 2, pp. 247–264, 2001.

[44] W. Winn, M. Windschitl, R. Fruland, and Y. Lee, "When does immersion in a virtual environment help students construct understanding?," *Proc. Int'l conf. of the learning societies*, 2002, pp. 497–503.

[45] K. Johnsen and B. Lok, "An evaluation of immersive displays for virtual human experiences," *Proc. IEEE Virtual Reality 2008*, 2008, pp. 133–136.

[46] E. Fassbender, D. Richards, A. Bilgin, W.F. Thompson, and W. Heiden, "VirSchool: The effect of background music and immersive display systems on memory for facts learned in an educational virtual environment," *Comput. Educ.*, vol. 58, no. 1, pp. 490–500, Jan. 2012.

[47] A. Sowndararajan, R. Wang, and D.A. Bowman, "Quantifying the benefits of immersion for procedural training," *Proc. 2008 workshop on Immersive projection technologies/Emerging display technologiges - IPT/EDT '08*, 2008, no. 2.

[48] E.D. Ragan, A. Sowndararajan, R. Kopper, and D.A. Bowman, "The Effects of Higher Levels of Immersion on Procedure Memorization Performance and Implications for Educational Virtual Environments," *Presence Teleoper. Virtual Environ.*, vol. 19, no. 6, pp. 527–543, 2010.

[49] Transportation Safety Board of Canada, "Aviation Investigation Report A05H0002," 2005.

[50] L.J. Van Gerwen, P. Spinhoven, R. Van Dyck, and R.F.W. Diekstra, "Construction and psychometric characteristics of two self-report questionnaires for the assessment of fear of flying," *Psychol. Assess.*, vol. 11, no. 2, pp. 146–158, 1999.

[51] R. Schwarzer and M. Jerusalem, "Generalized Self-Efficacy scale," *Measures in health psychology: A user's portfolio. Causal and control beliefs*, J. Weinman, S. Wright, and M. Johnston, Eds. Windsor, UK: Nfer-Nelson, 1995, pp. 35–37.

[52] A. Luszczynska and R. Schwarzer, "Multidimensional health locus of control: comments on the construct and its measurement," *J. Health Psychol.*, vol. 10, no. 5, pp. 633–642, Sep. 2005.

[53] T. Schubert, F. Friedmann, and H. Regenbrecht, "The Experience of Presence: Factor Analytic Insights," *Presence: Teleoper. Virtual Environ.*, vol. 10, no. 3, pp. 266–281, 2001.

[54] P. Backlund, H. Engström, M. Johannesson, and M. Lebram, "Games and Traffic Safety - an Experimental Study in a Game-Based Simulation Environment," *Proc. 11th Int'l Conf. Information Visualization - IV 2007*, 2007, pp. 908–916.

[55] C.D. Coles, D.C. Strickland, L. Padgett, and L. Bellmoff, "Games that 'work': using computer games to teach alcohol-affected children about fire and street safety," *Res. Dev. Disabil.*, vol. 28, no. 5, pp. 518–530, 2007.

[56] K. Anders Ericsson, "Exceptional memorizers: made, not born," *Trends Cogn. Sci.*, vol. 7, no. 6, pp. 233–235, Jun. 2003.

[57] B.G. Witmer and M.J. Singer, "Measuring Presence in Virtual Environments: A Presence Questionnaire," *Presence: Teleoper. Virtual Environ.*, vol. 7, no. 3, pp. 225–240, 1998.

Fabio Buttussi received his PhD in computer science from the University of Udine. He is a postdoctoral research fellow at Human-Computer Interaction (HCI) Lab in the Department of Mathematics, Computer Science, and Physics of the University of Udine, Italy. His research interests are in virtual reality, HCI, serious games, and their application in health and safety.

Luca Chittaro is full professor of Human Computer Interaction (HCI) in the Department of Mathematics, Computer Science, and Physics of the University of Udine, Italy, where he heads the HCI Lab (http://hcilab.uniud.it). He has authored or co-authored over 180 international academic publications, and he is an ACM Distinguished Speaker. His major research interests are in virtual reality, mobile HCI, serious games, persuasive technology, and their applications in health and safety. He has received research grants from a wide range of organizations, including the US Federal Aviation Administration (FAA), the European Union (EU), the Italian Ministry of University and Research (MIUR), and companies such as the Benetton Group and the Intesa Sanpaolo Bank group.