Embedding Test Questions in Educational Mobile Virtual Reality: A Study on Hospital Hygiene Procedures

Fabio Buttussi and Luca Chittaro

***Abstract*—Educational virtual environments (EVEs) can enable effective learning experiences on various devices, including smartphones, using nonimmersive virtual reality (VR). To this purpose, researchers and educators should identify the most appropriate pedagogical techniques, not restarting from scratch but exploring which traditional e-learning and VR techniques can be effectively combined or adapted to EVEs. In this direction, this article explores if test questions, a typical e-learning technique, can be effectively employed in an EVE through a careful well-blended design. We also consider the active performance of procedures, a typical VR technique, to evaluate if test questions can be synergic with it or if they can instead break presence and be detrimental to learning. The between-subject study we describe involved 120 participants in four conditions: with/without test questions and active/passive procedure performance. The EVE was run on a smartphone, using nonimmersive VR, and taught hand hygiene procedures for infectious disease prevention. Results showed that introducing test questions did not break presence but surprisingly increased it, especially when combined with active procedure performance. Participants' self-efficacy increased after using the EVE regardless of condition, and the different conditions did not significantly change engagement. Moreover, participants who had answered test questions in the EVE showed a reduction in the number of omitted steps in an assessment of learning transfer. Finally, test questions increased participants' satisfaction. Overall, these greater-than-expected benefits support the adoption of the proposed test question design in EVEs based on nonimmersive VR.**

***Index Terms*—Educational virtual environment (EVE), mobile learning, procedural knowledge, quizzes, user evaluation, virtual reality (VR).**

I. Introduction and Background

VIRTUAL reality (VR) is a technology that presents a synthetically generated 3-D virtual environment (VE) to the user through visual, auditory, and possibly other stimuli [1]. When the VE is based on pedagogical models, incorporates or implies didactic objectives, and provides users with experiences to foster learning outcomes [2], it is called educational virtual environment (EVE).

Received 12 February 2024; revised 24 May 2024 and 10 September 2024; accepted 19 October 2024. Date of publication 29 October 2024; date of current version 21 November 2024. *(Corresponding author: Fabio Buttussi.)*

The authors are with the Human-Computer Interaction Lab, Department of Mathematics, Computer Science, and Physics, University of Udine, 33100 Udine, Italy (e-mail: fabio.buttussi@uniud.it; luca.chittaro@uniud.it).

Digital Object Identifier 10.1109/TLT.2024.3487898

Since its early years, VR technology has included immersive and nonimmersive experiences [3]. Immersive VR exploits specific hardware, such as head-mounted displays (HMDs) or multiple large-size projections, called cave automatic virtual environments (CAVEs), to surround users with the VE. Nonimmersive VR exploits conventional screens like PC monitors or smartphones and tablets. In addition, some researchers consider an intermediate level of immersion, called semi-immersive VR, which includes single large-size projections [4], multiple small-size projections [5], projections over a physical workbench [6], and monitors with 3-D stereoscopy [7]. In a review covering EVEs from 1999 to 2009 [2], most of the 53 reviewed studies used nonimmersive VR, displaying the EVEs on desktop PC monitors, while only 16 studies used immersive VR, displaying the EVEs on HMDs or surrounding the users with CAVEs. The recent availability of consumer HMDs, with a wide field of view and six degrees of freedom (6DOF) tracking of users' head and hand movements, facilitates the development of immersive EVEs. Moreover, the massive availability of smartphones that can support interactive 3-D graphics now allows the delivery of nonimmersive EVEs to millions of learners [8].

To enable EVEs to deliver effective learning experiences on various devices, researchers and educators should identify the most effective pedagogical techniques. In doing so, they should not restart from scratch but explore how the vast amount of knowledge from traditional e-learning and VR literature can be effectively combined or adapted to EVEs.

Indeed, a considerable number of studies showed positive effects of EVEs in education and training, as discussed by several reviews, spanning different learners and domains [2], [9], [10], [11], [12], [13], [14], [15], [16]. Some studies also compared EVEs with traditional learning methods (see, e.g., [17], [18], [19], [20], [21], and [22]) or compared the same EVE in immersive versus nonimmersive VR (see [21] for a quick review of these studies). However, to better inform the design of effective EVEs, there is a need for new studies that address their different features (beyond immersiveness) and how different designs of those features can contribute to learning and related outcomes (e.g., learners' motivation). In traditional e-learning, a considerable number of studies evaluated the effects of several specific features, such as personalization, different types of multimedia content, provision of different kinds of feedback, support of different learning strategies, use of different language

styles, and alternative ways to control learning sequence and pace (see, e.g., [23]).

Recent research focuses on exploring the efficacy of such traditional e-learning features in the context of EVEs. For example, Meyer et al. [24] focused on the possibility of offering pretraining to validate the pretraining principle, which states that people learn more deeply from a multimedia message when they know the names and characteristics of the main concepts [25]. They compared a video and an immersive EVE condition with and without pretraining. In a 2×2 study involving 118 participants, they showed that pretraining had a positive effect on knowledge, transfer, and self-efficacy in the immersive EVE condition but no effect in the video condition. Albus et al. [26] instead studied the effects of signaling, i.e., highlighting relevant information, in an EVE. The study involved 107 participants and showed that signaling improved learners' recall performance, extending the positive effect of signaling found in traditional e-learning to EVEs. Bohné et al. [27] considered 14 features for a web-based EVE, including features from traditional e-learning (e.g., providing feedback about the state) and features inspired by games (e.g., giving badges for learners' achievements). A study with 355 online participants contrasted three versions of the EVE with two, six, or all the features without finding statistically significant differences favoring the versions with more features.

In this article, we focus on test questions (TQs), a feature typical of traditional e-learning for both learning and assessment purposes, and we explore if they can be effectively employed in an EVE through careful design. To achieve the same learning and assessment purposes, EVEs for teaching procedural knowledge often ask users to actively perform the steps of the procedure in VR. Some previous studies analyzed some effects of active versus passive performance in EVEs. Chittaro and Sioni [28] contrasted two modes of using an EVE about safety risks in a study with 42 participants where half of the participants interactively progressed through the experience by moving and acting in the EVE, while the other half passively watched the experience progressing automatically. Results showed that, while the two conditions increased participants' knowledge and self-efficacy in a similar way, interactively experiencing the EVE heightened emotional response during the experience (measured in terms of skin conductance and heart rate) and perception of the depicted risks (in terms of severity and vulnerability) after completing the experience. Roussou and Slater [29] compared learning how to solve arithmetical fraction problems in three conditions: 1) an interactive VR condition, where participants actively performed tasks in a CAVE-based EVE; 2) a passive VR condition, where participants observed a robot doing the tasks in the same CAVE-based EVE; and 3) a non-VR condition, where participants performed the tasks using physical plastic bricks. The study with 50 participants showed that, given the same starting level, there is greater learning gain among participants in the VR conditions than among participants in the non-VR condition, while no statistically significant difference was found between interactive and passive VR conditions. Ferguson et al. [30] assessed the effects of two aspects: 1) active and passive exploration of the EVE and 2) structure (explicit versus implicit) of the story narrated within the EVE. The 2×2 study with 42 participants showed that actively exploring the EVE had positive effects on cognitive interest and feeling of presence and that an implicit story structure led to increased recall of spatial information, while a passive (guided) exploration was beneficial for optimal learning of factual knowledge.

Our study researches if TQs inside EVEs can be synergically combined with the active performance (AP) of the procedure or, instead, they can break the EVE experience and be detrimental to learning. The specific graphical and interaction design of TQs we employ is aimed at blending well in the EVE to minimize the possible feeling of a break in presence in the EVE. The EVE we used in the study is a nonimmersive VR application for smartphones that teaches hand hygiene procedures in the prevention of infectious diseases, displayed through an animated pedagogical agent (APA) [31]. Effectively educating the general public and health workers about this topic is of particular importance, as dramatically pointed out by the COVID-19 pandemic. Our study analyzed two different aspects of TQs in EVEs: the effects of whether including or not TQs inside an EVE and their possible interaction with an active or passive experience of the procedures in the EVE. More specifically, we considered the possibility for participants to either passively watch the APA demonstrating the hand hygiene procedure or actively perform the procedures on the APA.

Besides assessing effects on objective measures of learning in a final assessment of learning transfer to the real world (correct steps and errors in performing the hand hygiene procedure), we considered subjective measures of learning confidence (changes in self-efficacy between before and after using the application) and of learning experience (sense of presence in the EVE, engagement, and satisfaction). We formulated the following hypotheses for the study.

1) TQs could increase learning of the procedure because testing is not only a means for assessment but also a means to improve learning [32].
2) AP could increase learning of the procedure because it may support learning by doing [33].
3) TQs might break the sense of presence in the EVE because they can be perceived as an extraneous element that does not belong to the virtual experience, and some anomalies in the VE will induce a break in presence [34].
4) AP could increase participants' self-efficacy regarding hand hygiene because gaining experience in performing a given behavior is a major factor that contributes to increase self-efficacy [35].

The study was exploratory about engagement and satisfaction. Lessons learned in our study can be useful to inform the design of new EVEs based on nonimmersive VR.

The rest of this article is organized as follows. Section II provides advice on designing TQs in a way that blends them well in the EVE. Section III describes the materials and methods of the user study, including the proposed hand hygiene application with the APA. Section IV reports the results. Section V discusses the results and the limitations of the user study, also outlining future work. Finally, Section VI concludes this article.

II. TQ Design

The design of TQs for introduction into EVEs should blend them well into the virtual experience. The final design of TQs for our EVE results from an iterative process based on the literature analysis and the feedback from a professional creator of e-learning courses in health and safety, who shared his expert knowledge with us. Before describing our EVE in detail in the following section, we share some general considerations for the design of TQs aimed at blending them well in EVEs and special considerations for nonimmersive ones.

A. Structure and Media for TQs

TQs offering multiple choices include a stem (the stimulus for the response, typically in question format), the correct choice (one undeniably correct answer), and distractors (the unquestionably wrong answers) [36]. Stem, correct choice, and distractors are typically displayed using text, but in rare circumstances, correct choice and distractors can be drawings or photographs [36]. In our iterative design process, we tried both text and pictorials for correct choice and distractors. On the one hand, reading long text displayed as an overlay on the EVE might draw learners' attention away from the EVE and break the learners' sense of presence. Moreover, reading on the small screen of smartphones can affect comprehension [37] and cause eye-related symptoms [38]. On the other hand, pictorials can be ambiguous and lead to low comprehension (see, e.g., [39] about the comprehension of pictorials in safety cards). Moreover, pictorials might also break presence if their graphics differ from the graphics of the EVE. For example, when considering hand hygiene, using simplified drawings (e.g., those used in some safety posters) or photographs of real nurses' hands will not match the 3-D graphics of the EVE. Therefore, to keep consistency and avoid breaking presence, in our final design, we represent correct choice and distractors as pictorials that render 3-D models used in the EVE with the same textures and colors, maximizing visual consistency between the answers and the EVE. To prevent possible pictorial ambiguity, we added text below the pictorials, but we limited it to a few words, in most cases only one [see Fig. 1(c), (d), and (e)] to minimize reading time.

B. Space Usage and Answer Number

A fundamental design aspect concerns the use of screen space because TQs should not hide relevant parts of the EVE to prevent a break in presence. For example, the APA has an essential role in our EVE, so TQs should be displayed below or above the APA. The use of screen space is particularly important for nonimmersive EVEs displayed on the small screens of smartphones. The number of displayed answers in the TQ is a major factor in determining screen space. Too many displayed distractors can lead to excessive use of screen space and long reasoning times, while too few distractors may easily lead to successful answers, even by pure chance. In our iterative design process, we created prototypes with two to six answers displayed in one or two rows. Displaying three answers in a row (i.e., one correct choice and two distractors) appeared to be appropriate to meet both limited screen space and readability needs. This choice is consistent with traditional education literature, which, based on other considerations (e.g., long reasoning times versus successful answers by pure chance), suggests that the optimal number of answers is three [40], [41], [42], [43], [44].

C. TQ Feedback

Another essential aspect of TQs is feedback, which can be defined as information regarding one's performance or understanding [45]. In e-learning, three main types of feedback have been identified [46]: 1) knowledge of results (i.e., revealing if the learners' answer is correct or wrong); 2) knowledge of correct response (i.e., revealing the correct choice); and 3) elaborated feedback (i.e., explaining why the learners' answer was correct or not). In our design, we preferred avoiding long textual elaborated feedback because of sense of presence and reading issues mentioned before [37], [38]. Regarding the remaining two types, we favored knowledge of results over knowledge of correct response because the former can give learners the possibility to retry if the feedback is implemented as answer-until-correct [47]. This has the potential to blend well with an interactive EVE, especially if learners can also actively try other actions, e.g., in our case, perform the steps of the procedure. In addition to the type of feedback, its style could also contribute to blending TQs well in the EVE. For example, the feedback style about correct and wrong answers could be made more consistent with the EVE by using the same sounds and animations of the EVE [see the description of virus animation to provide feedback for both TQs and performance of procedure steps in the next section; Fig. 1(e) and (h)].

D. Timing of TQs

A final design choice concerns the timing of TQs, which can be administered together at the end of instruction or distributed during instruction. Considering video lectures, adding TQs during video playback was found to be engaging [48], so we decided to try this approach in nonimmersive VR. More precisely, considering the case of procedural knowledge, we preferred not to break the flow of the procedure the first time it is introduced to the learner but instead to present questions in between following procedure reviews. In this context, the natural way to split the procedure reviews was to introduce TQs before each step. This was also meant to blend TQs well with the AP of the procedure since the learners could alternate a TQ with a step performance until all steps were completed. If results in [48] extend from videos to nonimmersive VR, this design should support engagement, while the alternance question/performance should lead to a flow in procedure review without breaking the flow of procedure presentation.

III. Material and Methods

The study followed a 2×2 design. One independent variable concerned whether including or not TQs and had two levels (Yes, No) indicating if participants had to answer TQs or not.

(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k)

Fig. 1. Screenshots of the hand hygiene EVE. (a) and (b) APA teaching the procedure. (c) TQ. (d) Feedback for correct answer to TQ. (e) Feedback for wrong answer to TQ. (f) AP of a procedure step. (g) Feedback for too long performance of a step. (h) and (i) Feedback for too short performance of a step. (j) APA automatically performing a step. (k) Final debriefing screen. The language of the text in the screenshots is the language spoken by participants.

TABLE I
FOUR GROUPS OF PARTICIPANTS AND CORRESPONDING LEVELS OF THE INDEPENDENT VARIABLES

	AP = Yes	AP = No
TQ = Yes	Participants answered test questions in the EVE and actively performed procedure steps	Participants answered test questions in the EVE and passively watched procedure steps
TQ = No	The EVE did not include test questions, and participants actively performed procedure steps	The EVE did not include test questions, and participants passively watched procedure steps

The other independent variable concerned the AP of the steps in the procedure to be learned and had two levels (Yes, No) indicating if participants had to actively perform the steps of the procedure on the APA or, instead, they watched them performed automatically by the APA.

The study was between-participants to exclude possible learning effects that trying multiple conditions can produce. Each participant was thus assigned to only one of the four groups described in Table I.

A. Hand Hygiene Application

To evaluate the effects of TQs and AP inside an EVE for a realistic purpose, we used a nonimmersive EVE where hand hygiene in the prevention of infectious diseases is taught with the involvement of an APA. The learning goal of the EVE is to teach how to sanitize the hands safely and quickly, so the APA must explain how to sanitize all parts of the hands and how much is the appropriate amount of time for each part: not too short, to avoid insufficient sanitization, and not too long, to avoid wasting time, which is an important aspect in work contexts. The application was developed for smartphones using the Unity game engine, appearing to users as an educational game or "serious game" in Zyda's terminology [49].

More precisely, the application went through the following phases with participants.

1) *Login:* The application displays a login screen requesting a password. The given password is associated with an anonymous identifier given to the participant. The anonymous identifier is, in turn, associated with the group the participant belongs to, allowing the application to apply the group settings (inclusion or not of TQs and active or passive performance of the procedure).
2) *Procedure presentation:* The application displays the EVE where the APA, represented as a nurse, briefly introduces the hand hygiene procedure [see Fig. 1(a)]. Then, for each step of the procedure, the application shows a brief textual description, a voice-over verbally explains the step, and the APA demonstrates how to perform the step [see Fig. 1(b)]. The procedure consists of seven steps, devoted to the sanitization of the different parts of the hands, in this sequence: a) palms; b) backs; c) areas between fingers; d) back of fingers; e) thumbs; f) fingertips; and g) wrists.
3) *First review of the procedure:* The application asks the participant to interactively review the procedure step by step. This phase differs according to the group the participant belongs to, as follows.

a) *Presence of TQs:* If TQ = Yes, the review of each step is preceded by a TQ asking which step must be performed next among three possibilities [see Fig. 1(c)]. If the participant selects the correct step, the answer is highlighted in green [see Fig. 1(d)]; otherwise, red viruses are animated over the wrong choice, which is highlighted in red [see Fig. 1(e)]. If the participant belongs to a group where TQ = No, the TQ is not displayed.
b) *AP of procedure:* If AP = Yes, then the participant must perform a gesture on the touchscreen to control the performance of the step until the corresponding part of the hand or wrist is sanitized [see Fig. 1(f)]. If the participant performs the step for too long, a message informs the participant [see Fig. 1(g)], who is invited to press the button on the bottom of the screen to continue. If the participant does not perform the step or performs it for too little time, animated red viruses appear, and the current part of the hand or wrist is highlighted in red to suggest that it is not sanitized enough [see Fig. 1(h)]. Then, a message informs the participant to continue rubbing the part of the hand or wrist until it is fully covered with disinfectant [see Fig. 1(i)]. If AP = No, the APA automatically performs the steps of the procedure, stopping after each step, and the participant should only press the button on the bottom of the screen to continue with the next step until the APA has completed the procedure [see Fig. 1(j)].
c) *Debriefing:* At the end of the procedure, in all groups, the application displays a debriefing screen [see Fig. 1(k)] that marks each step of the procedure with either a green tick or a red cross. The red cross highlights the steps for which the user made errors, and a comment on a red background describes the error (i.e., wrong answer to a question, too short, or too long performance of a step).

4) *Second review of the procedure:* The application asks the participant to interactively review the procedure a second time. This phase differs among groups as described in the previous phase. It concludes with the second debriefing and a final screen, which tells the participant that the application can be closed.

B. Participants

The study involved a sample of 120 participants (101 males, 19 females). They were undergraduate students who had followed a course on human–computer interaction. To be involved in the study, participants had to have an Android or iOS smartphone available and be able to install and use mobile applications in it. Participants' age ranged from 20 to 28 (M = 21.75, SD = 1.39). We asked participants the number of hours per week they used mobile applications with 3-D graphics (e.g., games).

Their answers ranged from 1 to 7 (M = 4.65, SD = 1.93). Finally, we administered the self-efficacy questionnaire described in Section III-C to assess participants' self-efficacy in hand sanitization before using the application (pretest self-efficacy). The obtained values ranged from 2.6 to 6.8 (M = 4.88, SD = 0.88).

Participants were assigned to the four groups in such a way that: 1) each group had 30 participants (24 males and 6 females in the group with TQ = No, AP = No; 25 males and 5 females in the other groups) and 2) the four groups were similar in terms of age, number of hours per week using mobile applications with 3-D graphics, and pretest self-efficacy. Each of these variables was submitted to a one-way analysis of variance (ANOVA) that confirmed the lack of statistically significant differences between groups.

C. Measures

Since we were interested in both objective and subjective measures of learning and of the learning experience in the EVE, we measured the following dependent variables.

1) Learning Transfer: To objectively evaluate the learning of the taught procedure, we involved participants in an assessment of learning transfer after they used the application with the EVE. In this assessment, we asked participants to physically perform all the steps of the hand hygiene procedure in the correct sequence, also verbally describing each step before performing it. The only difference with respect to performing the procedure in a real disinfection context was that we told participants to imagine having already put the sanitizer gel on their hands. We decided to do so to guarantee that participation in the study did not cause discomfort to participants, such as common dermatologic reactions to hand sanitizers (skin dryness in most people, eczema breakout trigger in some people with eczema) and other types of possible harm such as damage to the outer layer of the eye (if the participant accidentally touches his/her eyes shortly afterward hand sanitizer use). The experimenter recorded audio and video of the physical performance of the procedure focused on the hands and the forearms of the participants. Later, he reviewed the videos and objectively assigned one of the following unambiguous and mutually exclusive codes to each of the seven steps of the taught procedure.

a) *Correct:* The participant performed the step completely and correctly, as taught in the procedure.
b) *Incompliant:* The participant performed the step, but incompletely or in a way that did not match the one taught in the procedure.
c) *Omitted:* The participant did not perform the step in the assessment.

By summing up all the steps with the same code, we obtained three measures, respectively called correct steps, incompliant steps, and omitted steps. Each of these measures can range from 0 to 7, i.e., the total number of steps in the procedure. In addition, the experimenter computed the following measures.

a) *Misplaced steps:* This is the number of misplaced steps in the procedure. To compute it, each correct or incompliant step in the procedure was considered individually and counted as a misplaced step if it was performed before (respectively, after) at least one step that should have preceded (respectively, followed) it in the correct sequence of the procedure.
b) *Extraneous steps:* It is the number of steps that should not have appeared in the procedure, e.g., the participant performed two steps to clean the same part of the hands in two ways, one of which was not present in the taught procedure.

2) Presence: To measure the sense of presence in the EVE experience, we administered the widely used Igroup Presence Questionnaire (IPQ)[1] [50] to participants after they used the EVE. The IPQ asks participants to rate 14 items on a seven-point scale, ranging from 0 to 6. While eight of the items have the extreme values labeled in terms of agreement and disagreement as in Likert scales, the remaining six items have the extreme values labeled in other terms (e.g., "about as real as an imagined world"—"indistinguishable from the real world"). The IPQ includes a general item related to the sense of "being there" and three subscales (see confirmatory factor analysis in [50]): 1) spatial presence (five items); 2) involvement (four items); and 3) experienced realism (four items). Subscales and total presence score are calculated by averaging the items.

3) Engagement: To measure engagement, we adapted the Game Engagement Questionnaire (GEQ) proposed in [51]. The GEQ includes 19 items, and participants are asked to rate each of them on a three-point scale (1 = No, 2 = Sort of, 3 = Yes). Seven of the GEQ items are specifically about games or playing, so we adapted them to make them suitable for rating an EVE by changing "game" into "tool" and "playing" into "using", as in [21]. The ratings of the 19 items were summed up to form a scale that could range from 19 to 57.

4) Satisfaction: To measure participants' satisfaction with the EVE, we used the satisfaction subscale (seven items) of the USE Questionnaire [52], asking participants to rate their level of agreement on a seven-point Likert scale (1 = strongly disagree, 7 = strongly agree). The ratings were summed up to form a scale that could range from 7 to 49.

5) Self-Efficacy: To assess participants' self-efficacy in hand sanitization, we designed a questionnaire by adapting items from well-known self-efficacy questionnaires [53]. The self-efficacy questionnaire contained five items: 1) I feel confident of my ability to sanitize my hands; 2) I would be able to sanitize every part of my hands and wrists; 3) I could sanitize my hands without making mistakes; 4) I would be able to understand when my hands are sanitized; and 5) I could sanitize my hands quickly, but correctly. The questionnaire asked participants to rate their level of agreement on a seven-value Likert scale (1 = strongly disagree, 7 = strongly agree). The questionnaire score was obtained by averaging all item ratings. The measure was taken before (pretest self-efficacy) and after (posttest self-efficacy) the use of the application.

D. Procedure

The study was conducted during a period in which access to the university laboratories was discouraged due to the COVID-19 pandemic, so we opted for a remote

[1][Online]. Available: http://www.igroup.org/pq/ipq

videoconferencing evaluation to protect participants' health. Candidate participants were invited to book a slot of 30 min using an online calendar system. The week before the booked time, the experimenter sent participants an email with preliminary instructions and a form for obtaining participants' written consent. The instructions informed participants that they would be involved in an evaluation of a mobile learning application without specifying the taught topic. They also instructed participants how to send back the signed informed consent form and download and install the application on their smartphone before the booked slot. The installed application could not be run without the password they would receive during the videoconference meeting. The instructions invited participants to connect to the videoconferencing platform (Microsoft Teams) at least 10 min before the booked slot and to wait for the experimenter's call. Finally, the instructions informed participants that they would also fill out some questionnaires using an anonymous identification code.

At the booked time, the experimenter called the participants, summarized information they had already received in the instructions and consent form, and informed participants that the webcam had to be turned on only upon the experimenter's request. Then, the experimenter told participants their anonymous identification code and wrote it in the Teams chat. On the same chat, the experimenter sent the link to the initial questionnaire about participants' gender, age, and number of hours per week they used mobile applications with 3-D graphics. The experimenter told participants that they could ask any questions about the questionnaire and invited them to fill it out. Then, the experimenter did the same for the pretest self-efficacy questionnaire.

After participants had completed the two questionnaires, the experimenter asked them to check that their smartphone was connected to the Internet and that the sound volume was set to a comfortable level that could allow them to hear the sound from the application. He informed participants that during the use of the application, they had to follow the instructions provided by it and that they could not communicate with the experimenter while they were using the application. After asking if they had any questions, the experimenter gave participants their individual password to log in to the application, as described in Section III-A.

After the use of the application, the experimenter reminded participants what their code was and sent the links to presence, engagement, satisfaction, and posttest self-efficacy questionnaires, inviting participants to fill them. Then, participants were asked to turn on their webcam and place it in a way that could focus only on their hands and forearms. After checking that the camera viewpoint clearly included hands and forearms, the experimenter started recording, and participants carried out the final assessment of learning transfer, as described in Section III-C.

IV. Results

Since our groups differed on two independent variables, as described in Table I, a between-subjects 2 × 2 ANOVA was used to analyze the dependent variables measured once (all the variables described in Section III-C, except self-efficacy). In case of interactions, we analyzed simple effects of TQ separately at the two levels of AP, and simple effects of AP separately at the two levels of TQ using Bonferroni correction. Effect sizes are reported as partial eta squared (η_p^2).

We did not analyze extraneous steps because their occurrence was negligible in all groups (only seven users included a single extraneous step each: 2 in TQ = Yes, AP = Yes; 3 in TQ = No, AP = Yes, 1 in TQ = Yes, AP = No, and 1 in TQ = No, AP = No).

Since self-efficacy was measured twice over time (pretest and posttest), it was analyzed using a mixed-design 2 × 2 × 2 ANOVA, in which TQ and AP served as the between-subjects variables, and time of measurement served as the within-subjects variable.

A. Learning Transfer

Considering learning transfer (see Fig. 2), the analysis found no main effect of TQ, no main effect of AP, and no interaction, in correct, incompliant, and misplaced steps, $p > 0.05$ for all. The analysis of omitted steps instead revealed a main effect of TQ, $F(1,116) = 4.39$, $p < 0.05$, $\eta_p^2 = 0.04$; a main effect of AP, $F(1,116) = 4.39$, $p < 0.05$, $\eta_p^2 = 0.04$; and no interaction, $p > 0.05$. The number of omitted steps was lower when TQs were included (M = 0.57, SD = 0.59) than when they were not (M = 0.83, SD = 0.81) and was higher with AP of the procedure (M = 0.83, SD = 0.81) than passive watching of the procedure (M = 0.57, SD = 0.59).

B. Presence

Considering presence (see Fig. 3), the analysis revealed a main effect of TQ, $F(1,116) = 5.07$, $p < 0.05$, $\eta_p^2 = 0.04$; no main effect of AP, $p > 0.05$; and an interaction between TQ and AP, $F(1,116) = 6.23$, $p < 0.05$, $\eta_p^2 = 0.05$, in the total score. The total score for presence was higher when TQs were included (M = 3.29, SD = 0.95) than when they were not (M = 2.92, SD = 0.91). Investigating interaction, in case of passive watching of performance, we found no statistically significant difference in total score between participants who answered TQs and those who did not, $p > 0.05$. On the contrary, with AP, the total score for presence was higher when TQs were included (M = 3.64, SD = 0.75) than when they were not (M = 2.85, SD = 0.93), $p < 0.005$. With no TQs, we found no statistically significant difference in total score between active and passive performance, $p > 0.05$. On the contrary, with TQs, the total score for presence was higher with AP (M = 3.64, SD = 0.75), rather than passive performance (M = 2.94, SD = 1.02), $p < 0.005$.

For the general item about the sense of "being there," the analysis revealed no main effect of TQ, $p > 0.05$; no main effect of AP, $p > 0.05$; and an interaction between TQ and AP, $F(1,116) = 4.92$, $p < 0.05$, $\eta_p^2 = 0.04$. With passive performance, we found no statistically significant difference in total score between participants who answered TQs and those who did not, $p > 0.05$. On the contrary, with AP, the score for the general item was higher when TQs were included (M = 3.83, SD = 1.32) than when they were not (M = 2.90, SD = 1.73), $p < 0.05$. With no TQs, we found no statistically significant

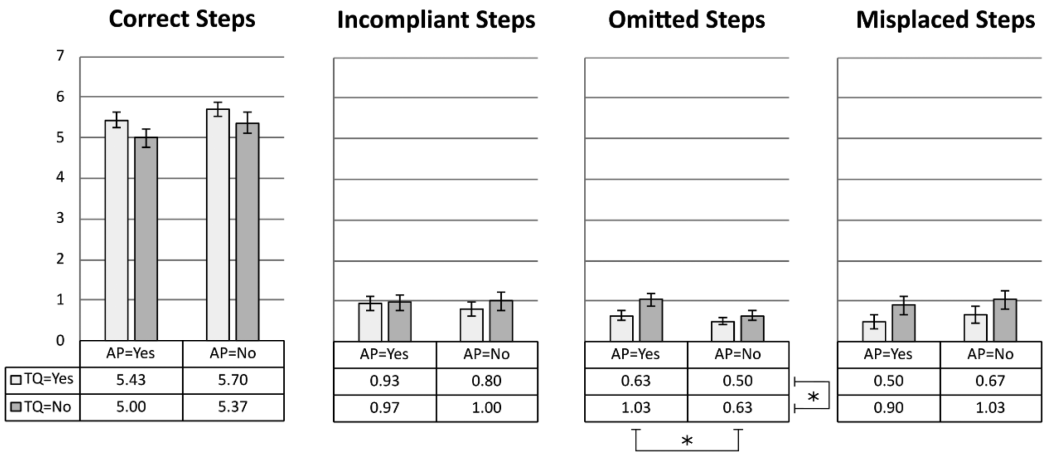

Fig. 2. Means of correct, incompliant, omitted, and misplaced steps in the four groups. Capped vertical bars indicate ± SE. The ∗ sign indicates differences with $p < 0.05$.

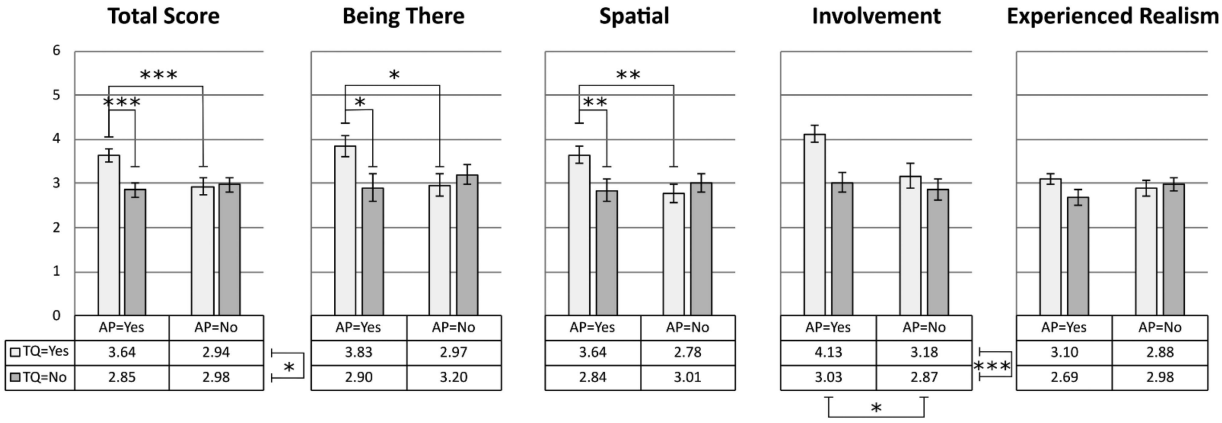

Fig. 3. Means of presence total score, general item about the sense of "being there," spatial subscale, involvement subscale, and experience realism subscale in the four groups. Capped vertical bars indicate ± SE. The ∗, ∗∗, and ∗∗∗ signs indicate differences with $p < 0.05$, $p < 0.01$, and $p < 0.005$, respectively.

difference in the score for the general item between active and passive performance, $p > 0.05$. On the contrary, with TQs, the score for the general item was higher with AP (M = 3.83, SD = 1.32), than passive performance (M = 2.97, SD = 1.45), $p < 0.05$.

For the spatial presence subscale, the analysis revealed no main effect of TQ, $p > 0.05$; no main effect of AP, $p > 0.05$; and an interaction between TQ and AP, $F(1,116) = 5.90$, $p < 0.05$, $\eta_p{}^2 = 0.05$. With passive performance, we found no statistically significant difference in spatial presence between participants who answered TQs and those who did not, $p > 0.05$. On the contrary, with AP, the score for spatial presence subscale was higher when TQs were included (M = 3.64, SD = 1.05) than when they were not (M = 2.84, SD = 1.36), $p < 0.01$. With no TQs, we found no statistically significant difference in spatial presence subscale between active and passive performance, $p > 0.05$. On the contrary, with TQs, the score for spatial presence was higher with AP (M = 3.64, SD = 1.05) than passive performance (M = 2.78, SD = 1.10), $p < 0.01$.

For the involvement subscale, the analysis revealed a main effect of TQ, $F(1,116) = 8.68$, $p < 0.005$, $\eta_p{}^2 = 0.07$; a main effect of AP, $F(1,116) = 5.37$, $p < 0.05$, $\eta_p{}^2 = 0.04$; and no interaction, $p > 0.05$. The score for the involvement subscale was higher when TQs were included (M = 3.65, SD = 1.42) than when they were not (M = 2.95, SD = 1.27), and it was higher with AP (M = 3.58, SD = 1.26) than passive performance (M = 3.02, SD = 1.45).

For the experienced realism subscale, the analysis found no main effect of TQ, no main effect of AP, and no interaction, $p > 0.05$ for all.

C. Engagement

Considering engagement (see Fig. 4), the analysis found no main effect of TQ, no main effect of AP, and no interaction, $p > 0.05$ for all.

D. Satisfaction

Considering satisfaction (see Fig. 4), the analysis found a main effect of TQ, $F(1,116) = 3.96$, $p < 0.05$, $\eta_p{}^2 = 0.03$; no main effect of AP, $p > 0.05$; and no interaction, $p > 0.05$. Participants' satisfaction was higher when TQs were included (M = 35.05, SD = 7.60) than when they were not (M = 32.07, SD = 8.69).

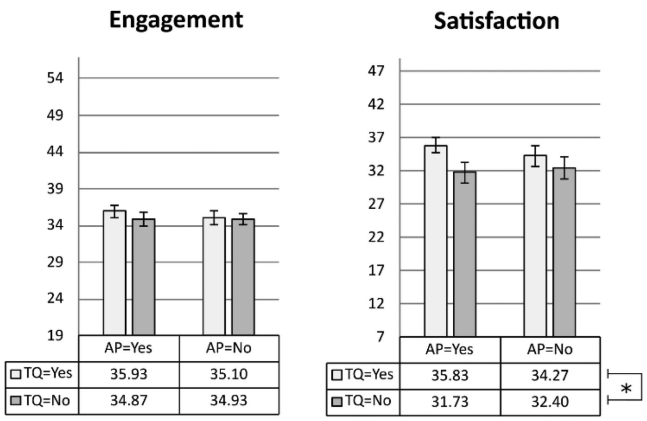

Fig. 4. Means of engagement and satisfaction in the four groups. Capped vertical bars indicate ± SE. The * sign indicates a difference with $p < 0.05$.

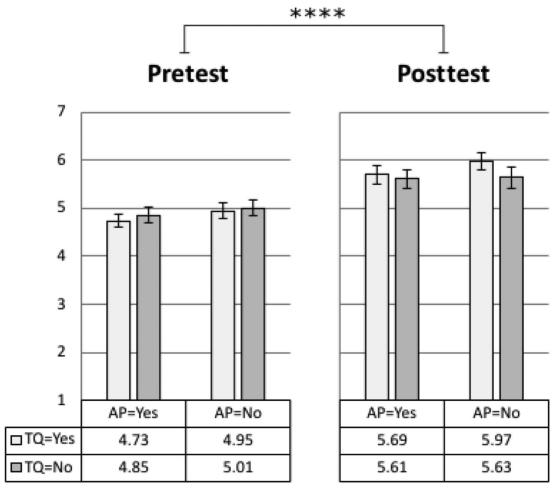

Fig. 5. Means of self-efficacy in the four groups before (pretest) and after (posttest) using the application. Capped vertical bars indicate ± SE. The **** sign indicates a difference with $p < 0.001$.

E. Self-Efficacy

Considering self-efficacy (see Fig. 5), the analysis found a main effect of time of measurement, $F(1,116) = 78.76$, $p < 0.001$, $\eta_p{}^2 = 0.40$, no main effect of TQ, $p > 0.05$, no main effect of AP, $p > 0.05$, and no two-way or three-way interactions, $p > 0.05$ for all. Overall, participants' self-efficacy increased between pretest (M = 4.88, SD = 0.88) and posttest (M = 5.73, SD = 1.06).

V. Discussion

Our hypothesis about the positive effect of TQs on learning was confirmed: the inclusion of TQs in the EVE resulted in a significantly smaller number of omitted steps in the final assessment of learning transfer. No significant difference was found in incompliant steps, possibly because TQs remarked which step to perform at what time, not how to perform them, so TQs did not make a difference in remembering such details. The inclusion of TQs could have made participants more successful by encouraging them to reason about which are the steps of the procedure, positively contributing to remembering more of them, as the significant reduction in omissions indicates. Although statistical significance was not reached for correct and misplaced steps, it is also interesting to note that the means are in favor of TQs: correct steps are higher, and misplaced steps are lower with TQs. The positive effect of TQs on learning is not new in education research, with studies pointing out their effectiveness in traditional assessment and teaching [32]. Our study advances knowledge in the literature by showing that the positive effect of TQs on learning also holds when they are blended inside an EVE experience.

Surprisingly, our hypothesis about a positive effect of AP of the procedure steps was contradicted by results: AP led to a higher number of omitted steps, while no statistically significant effect was found on other learning measures. Learning by doing theories point out that performing a task contributes to learning it, and EVEs are particularly suited to enable fail-safe task performance because they support interactivity [33]. However, the need to perform the gestures on the touchscreen could have increased participants' cognitive load, leading to a higher number of omitted steps. The application of cognitive load theory to educational technologies [54] can help articulate this hypothesis. The hand hygiene procedure taught by the EVE can be classified as domain-specific secondary information [55], [56]. Such kind of information is processed by a specific human cognitive architecture that can be described by five principles [57]. Two of these principles are particularly relevant to understanding our results: 1) the borrowing and reorganizing principle, because participants borrowed information from the APA and had to reorganize it considering their previous knowledge and 2) the narrow limits of change principle, because the borrowed information had to be processed by working memory with limited capacity and duration [58], [59]. AP might have been complex for working memory because participants had to deal with at least three interacting elements.

1) Which is the current step of the procedure?
2) How to perform the step?
3) How to perform the gesture on the touchscreen to perform the step?

When TQs were included, the identification of the current step could be processed and possibly sent to long-term memory in a separate moment before step performance, mitigating the load on working memory, and this could have contributed to the lower number of omitted steps in the condition with both TQs and AP with respect to the condition with AP and no TQs. With passive performance, the third element was not present because the step was performed by the APA automatically, resulting in a possible reduction of the cognitive load and leading to a lower number of omitted steps in the passive conditions.

Regarding the gestures, it is worth noting that, although they establish a direct mapping between the movement of the learner's finger and the movement of the APA's hands, the biomechanical fidelity of the mapping is inevitably lower compared to what can be obtained with immersive VR hardware. In the framework proposed by Ragan et al. [60], this kind of fidelity falls in the category of interaction fidelity, which concerns the

objective degree of exactness with which real-world interactions are reproduced in an interactive system. A previous study on procedural knowledge [21], which compared an EVE used on a smartphone and on an HMD, revealed a difference in omitted steps that favored the smartphone. However, in that study, the HMD was used with 6DOF tracked controllers, which have a higher interaction fidelity than gestures on a touch screen but possibly lower than haptic gloves or the optical bare hands tracking available in the latest models of HMDs. At the same time, 6DOF tracked controllers represent easy-to-use devices, and some studies showed that they could allow completing training tasks faster than haptic gloves [61], [62]. It would thus be interesting to repeat our study using 6DOF tracked controllers, haptic gloves, and hand tracking, also including measures of cognitive load, to explore both the hypothesis that the negative effect of AP on learning we found could be due to low interaction fidelity in controlling the hands of the APA and the hypothesis that it could be more generally due to an increased cognitive load associated to AP.

An important result of the study is that there is a way to blend TQs inside an EVE without breaking presence. Indeed, a main concern against the introduction of TQs inside EVEs is that they might be perceived as an extraneous element that does not belong to the EVE and could potentially break the sense of presence in it, as it happens with some anomalies in VEs [34]. Our results indicate that TQs did not break presence, and on the contrary, they increased it in terms of both total score and involvement subscale. The score for the involvement subscale was higher also with active rather than passive performance, showing that both TQs and AP contributed to get participants involved in the EVE.

The statistically significant interactions between TQ and AP for the total score, the general item about the sense of "being there," and spatial presence subscale provide additional insights showing that TQs and AP did not lead to a statistically significant increase when used alone, but synergically contributed to a statistically significant difference in total score, general item, and spatial presence subscale. The only subscale of presence for which we found no statistically significant result is experienced realism, but this was expected and in line with previous studies [63], [64], [65], where the VE did not change between the considered conditions. Since learners' presence is their sense of being in the EVE, our results about presence and its subscales can further encourage the integration of TQs inside EVEs, following the design considerations to blend them well we provided in Section II.

Considering self-efficacy, we did not find the increase hypothesized for AP. In [66], self-efficacy was higher for active players than passive observers of a serious game, while Chittaro and Sioni [28] did not find statistically significant differences in self-efficacy between active and passive conditions using a serious game. Interestingly, Peng [66] analyzed participants' identification with the character and found that it partially mediated the relationship between experience mode and self-efficacy, so in our EVE, the lack of differences in self-efficacy between active and passive performance could be due to limited identification with the APA. Notably, some EVEs—and particularly the Edu-Metaverse—allow learners to create their digital twins through avatars that can vividly represent themselves [67], [68], [69], [70]. It would thus be interesting to repeat our experiment including a customizable avatar with which learners can mimic the steps performed by the APA.

In addition, trying the experience with higher interaction fidelity technologies (e.g., in an immersive version of the EVE running on an HMD with optical bare hands tracking) could be interesting to assess if actively performing procedure steps in first person with the avatar can foster identification and consequently increases self-efficacy with respect to passive performance. Despite the lack of significant differences between active and passive performance, our study revealed that, in all experimental conditions, the use of the EVE led to an increase in self-efficacy between pretest and posttest. This is a positive result because, according to social cognitive theory [35], [71], self-efficacy significantly determines performance outcomes, and different people with similar skills may perform differently depending on variations in their self-efficacy. Therefore, the increase in self-efficacy displayed after using our EVE might contribute to a better performance of the taught procedure.

Finally, considering participants' engagement and satisfaction, we found no effect on engagement but a statistically significant effect on satisfaction, which was higher when TQs were included. Therefore, blending TQs in an EVE not only increases learning without breaking presence but also improves the learning experience.

A. Limitations

A limitation of our study is that the sample consisted of undergraduate students, who are not representative of health workers or the general population. Therefore, we cannot generalize the results to other categories, such as older users, with whom future studies are needed. Nevertheless, the sample is representative of young lay people who can benefit from learning hand hygiene procedures to prevent infectious diseases.

Another limitation concerning the sample is that it is male-dominated. Gender could play a role in some of the variables considered in our study. For example, in Britner and Pajares [72], girls reported stronger self-efficacy than boys, while Makransky et al. [73] found that girls learned better with a female APA and boys learned better with a male APA. Exploration of possible gender effects will require to perform the study on a gender-balanced sample.

Moreover, while in this article we considered an EVE concerning the topic of hand hygiene procedures in infection prevention, we cannot automatically extend the results to non-procedural types of knowledge, such as factual, conceptual, and metacognitive [74], because the positive learning effect found in this study was specific to recalling and performing the steps of the procedure in the correct order. TQs might have promoted recall instead of a deeper understanding that would be fundamental for other types of knowledge.

Interestingly, Finn et al. [75] found that providing elaborated feedback after learners' responses to TQs also supports conceptual understanding. We preferred not to provide written elaborated feedback to TQs to avoid breaking presence in the EVE, but a design option for future work could be using the APA to provide elaborated feedback orally.

Future studies are then needed to possibly extend the results we found to nonprocedural knowledge. Lessons learned in the study could instead likely be applied to topics different from hand hygiene but still involving procedural knowledge because the proposed design for TQs and the AP of steps can be suited also for other procedures.

Finally, while we showed the positive effects of TQs inside EVEs for nonimmersive VR on a mobile device, results cannot be generalized to EVEs displayed on immersive HMDs. In particular, since previous studies [21], [64], [76], [77] showed that presence is typically higher with higher fidelity displays, design changes could be needed to blend well the TQs also in immersive VR without breaking presence.

B. Future Research

Besides extending the sample of participants and the attention to nonprocedural knowledge, future research will include adapting the EVE for immersive VR and testing the effects of TQs and active procedures with HMDs. This would also allow us to better understand the effects of AP on learning and self-efficacy, as described earlier. To this purpose, in our future studies, we will also assess the cognitive load, including both subjective measures such as NASA TLX [78] and objective measures such as eye tracking [79]. Among subjective measures, it will also be interesting to consider the user experience in the EVE, e.g., using the questionnaire proposed and validated in [80].

Since human memory is subject to a natural decay, another measure that will be included in future work is knowledge retention after a period of nonuse. Previous studies showed that both immersive and nonimmersive VR are better than printed materials in terms of knowledge retention [17], [21], with limited or no difference between immersive and nonimmersive VR for the same measure [21], [64]. However, AP of procedures in a more natural way using immersive VR with tracking of users' hands might have an impact on retention that is worth testing. Moreover, TQs provide feedback to learners, and timely and relevant feedback is known to improve knowledge retention [81], also calling for future studies including this measure.

VI. Conclusion

In this article, we proposed a TQ design aimed at blending them well into a nonimmersive EVE. The results showed that introducing TQs did not break presence but surprisingly increased it, especially when combined with active procedure performance. Participants' self-efficacy increased after using the EVE regardless of condition, and the different conditions did not significantly change engagement. Moreover, participants who had answered TQs in the EVE showed a reduction in the number of omitted steps in an assessment of learning transfer. Finally, TQs increased participants' satisfaction. These greater-than-expected benefits support the adoption of the proposed TQ design in nonimmersive EVEs, and the results could likely be extended to nonimmersive versions of the Edu-Metaverse [67], a research topic gaining increasing attention in academia, especially in recent years [82], [83].

Acknowledgment

The authors would like to thank Nicola Zangrando for programming the different versions of the EVE used in the study and Hans Härting for sharing his expert knowledge on health and safety e-learning. This work involved human subjects in its research. It was an evaluation of a teaching method for which ethical approval was not required.

References

[1] R. Pausch, "Three views of virtual reality: An overview," *Computer*, vol. 26, no. 2, pp. 79–80, 1993, doi: 10.1109/2.192000.

[2] T. A. Mikropoulos and A. Natsis, "Educational virtual environments: A ten-year review of empirical research (1999–2009)," *Comput. Educ.*, vol. 56, no. 3, pp. 769–780, Apr. 2011, doi: 10.1016/j.compedu.2010.10.020.

[3] G. G. Robertson, S. K. Card, and J. D. Mackinlay, "Three views of virtual reality: Nonimmersive virtual reality," *Computer*, vol. 26, no. 2, pp. 81–83, 1993, doi: 10.1109/2.192002.

[4] A. F. Di Natale, C. Repetto, G. Riva, and D. Villani, "Immersive virtual reality in K-12 and higher education: A 10-year systematic review of empirical research," *Brit. J. Educ. Technol.*, vol. 51, no. 6, pp. 2006–2033, 2020, doi: 10.1111/bjet.13030.

[5] E. Pastorelli and H. Herrmann, "A small-scale, low-budget semi-immersive virtual environment for scientific visualization and research," *Procedia Comput. Sci.*, vol. 25, pp. 14–22, Jan. 2013, doi: 10.1016/j.procs.2013.11.003.

[6] L. Jacho, B. Sobota, Š. Korečko, and F. Hrozek, "Semi-immersive virtual reality system with support for educational and pedagogical activities," in *Proc. 12th IEEE Int. Conf. Emerg. eLearn. Technol. Appl.*, 2014, pp. 199–204, doi: 10.1109/ICETA.2014.7107584.

[7] E. T. Solovey, J. Okerlund, C. Hoef, J. Davis, and O. Shaer, "Augmenting spatial skills with semi-immersive interactive desktop displays: Do immersion cues matter?," in *Proc. 6th Augmented Hum. Int. Conf.*, 2015, pp. 53–60, doi: 10.1145/2735711.2735797.

[8] L. Chittaro and F. Buttussi, "Learning safety through public serious games: A study of 'prepare for impact' on a very large, international sample of players," *IEEE Trans. Vis. Comput. Graph.*, vol. 28, no. 3, pp. 1573–1584, Mar. 2022, doi: 10.1109/TVCG.2020.3022340.

[9] S. Haque and S. Srinivasan, "A meta-analysis of the training effectiveness of virtual reality surgical simulators," *IEEE Trans. Inf. Technol. Biomed.*, vol. 10, no. 1, pp. 51–58, Jan. 2006, doi: 10.1109/TITB.2005.855529.

[10] 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, doi: 10.1016/j.compedu.2013.07.033.

[11] E. Scott, A. Soria, and M. Campo, "Adaptive 3D virtual learning environments—A review of the literature," *IEEE Trans. Learn. Technol.*, vol. 10, no. 3, pp. 262–276, Jul.–Sep. 2017, doi: 10.1109/TLT.2016.2609910.

[12] Z. Feng, V. A. González, R. Amor, R. Lovreglio, and G. Cabrera, "Immersive virtual reality serious games for evacuation training and research: A systematic literature review," *Comput. Educ.*, vol. 127, pp. 252–266, Dec. 2018, doi: 10.1016/J.COMPEDU.2018.09.002.

[13] L. Jensen and F. Konradsen, "A review of the use of virtual reality head-mounted displays in education and training," *Educ. Inf. Technol.*, vol. 23, pp. 1515–1529, 2018, doi: 10.1007/s10639-017-9676-0.

[14] M. C. Howard and M. B. Gutworth, "A meta-analysis of virtual reality training programs for social skill development," *Comput. Educ.*, vol. 144, Jan. 2020, Art. no. 103707, doi: 10.1016/j.compedu.2019.103707.

[15] J. Radianti, T. A. Majchrzak, J. Fromm, and I. Wohlgenannt, "A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda," *Comput. Educ.*, vol. 147, Apr. 2020, Art. no. 103778, doi: 10.1016/j.compedu.2019.103778.

[16] N. Pellas, A. Dengel, and A. Christopoulos, "A scoping review of immersive virtual reality in STEM education," *IEEE Trans. Learn. Technol.*, vol. 13, no. 4, pp. 748–761, Oct.–Dec. 2020, doi: 10.1109/TLT.2020.3019405.

[17] 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, doi: 10.1109/TVCG.2015.2391853.

[18] S. J. Smith, S. Farra, D. L. Ulrich, E. Hodgson, S. Nicely, and W. Matcham, "Learning and retention using virtual reality in a decontamination simulation," *Nurs. Educ. Perspectives*, vol. 37, no. 4, pp. 210–214, Aug. 2016, doi: 10.1097/01.NEP.0000000000000035.

[19] E. D. Innocenti et al., "Mobile virtual reality for musical genre learning in primary education," *Comput. Educ.*, vol. 139, pp. 102–117, Oct. 2019, doi: 10.1016/j.compedu.2019.04.010.

[20] M. Alfadil, "Effectiveness of virtual reality game in foreign language vocabulary acquisition," *Comput. Educ.*, vol. 153, Aug. 2020, Art. no. 103893, doi: 10.1016/j.compedu.2020.103893.

[21] F. Buttussi and L. Chittaro, "A comparison of procedural safety training in three conditions: Virtual reality headset, smartphone, and printed materials," *IEEE Trans. Learn. Technol.*, vol. 14, no. 1, pp. 1–15, Feb. 2021, doi: 10.1109/tlt.2020.3033766.

[22] P. Araiza-Alba, T. Keane, W. S. Chen, and J. Kaufman, "Immersive virtual reality as a tool to learn problem-solving skills," *Comput. Educ.*, vol. 164, Apr. 2021, Art. no. 104121, doi: 10.1016/j.compedu.2020.104121.

[23] R. C. Clark and R. E. Mayer, *E-learning and the Science of Instruction: Proven Guidelines for Consumers and Designers of Multimedia Learning*. Hoboken, NJ, USA: Wiley, 2016.

[24] O. A. Meyer, M. K. Omdahl, and G. Makransky, "Investigating the effect of pre-training when learning through immersive virtual reality and video: A media and methods experiment," *Comput. Educ.*, vol. 140, Oct. 2019, Art. no. 103603, doi: 10.1016/j.compedu.2019.103603.

[25] R. E. Mayer and C. Pilegard, "Principles for managing essential processing in multimedia learning: Segmenting, pre-training and modality principles," in *The Cambridge Handbook of Multimedia Learning*. Cambridge, U.K.: Cambridge Univ. Press, 2014.

[26] P. Albus, A. Vogt, and T. Seufert, "Signaling in virtual reality influences learning outcome and cognitive load," *Comput. Educ.*, vol. 166, Jun. 2021, Art. no. 104154, doi: 10.1016/J.COMPEDU.2021.104154.

[27] T. Bohné, I. Heine, F. Mueller, P.-D. J. Zuercher, and V. M. Eger, "Gamification intensity in web-based virtual training environments and its effect on learning," *IEEE Trans. Learn. Technol.*, vol. 16, no. 5, pp. 603–618, Oct. 2023, doi: 10.1109/TLT.2022.3208936.

[28] L. Chittaro and R. Sioni, "Serious games for emergency preparedness: Evaluation of an interactive vs. a non-interactive simulation of a terror attack," *Comput. Hum. Behav.*, vol. 50, pp. 508–519, 2015, doi: 10.1016/j.chb.2015.03.074.

[29] M. Roussou and M. Slater, "Comparison of the effect of interactive versus passive virtual reality learning activities in evoking and sustaining conceptual change," *IEEE Trans. Emerg. Top. Comput.*, vol. 8, no. 1, pp. 233–244, Jan.–Mar. 2020, doi: 10.1109/TETC.2017.2737983.

[30] C. Ferguson, E. L. van den Broek, and H. van Oostendorp, "On the role of interaction mode and story structure in virtual reality serious games," *Comput. Educ.*, vol. 143, Jan. 2020, Art. no. 103671, doi: 10.1016/j.compedu.2019.103671.

[31] W. L. Johnson and J. C. Lester, "Face-to-face interaction with pedagogical agents, twenty years later," *Int. J. Artif. Intell. Educ.*, vol. 26, pp. 25–36, Mar. 2016, doi: 10.1007/s40593-015-0065-9.

[32] H. L. I. Roediger and J. D. Karpicke, "Test-enhanced learning: Taking memory tests improves long-term retention," *Psychol. Sci.*, vol. 17, no. 3, pp. 249–255, 2006, doi: 10.1111/j.1467-9280.2006.01693.x.

[33] M. Roussou, "Learning by doing and learning through play: An exploration of interactivity in virtual environments for children," *Comput. Entertainment*, vol. 2, no. 1, Jan. 2004, doi: 10.1145/973801.973818.

[34] M. Slater, "Presence and the sixth sense," *Presence Teleoperators Virtual Environ.*, vol. 11, no. 4, pp. 435–439, 2002, doi: 10.1162/105474602760204327.

[35] A. Bandura, *Self-Efficacy: The Exercise of Control*. New York, NY, USA: Freeman, 1997.

[36] T. M. Haladyna, *Developing and Validating Multiple-Choice Test Items*, 3rd ed. London, U.K.: Routledge, 2004.

[37] M. Honma et al., "Reading on a smartphone affects sigh generation, brain activity, and comprehension," *Sci. Rep.*, vol. 12, no. 1, Jan. 2022, Art. no. 1589, doi: 10.1038/s41598-022-05605-0.

[38] B. Antona, A. R. Barrio, A. Gascó, A. Pinar, M. González-Pérez, and M. C. Puell, "Symptoms associated with reading from a smartphone in conditions of light and dark," *Appl. Ergonom.*, vol. 68, pp. 12–17, Apr. 2018, doi: 10.1016/j.apergo.2017.10.014.

[39] J. K. Caird, B. Wheat, K. R. McIntosh, and R. E. Dewar, "The comprehensibility of airline safety card pictorials," in *Proc. Hum. Factors Ergon. Soc. Annu. Meeting*, 1997, vol. 41, no. 2, pp. 801–805, doi: 10.1177/107118139704100216.

[40] K. D. Crehan, T. M. Haladyna, and B. W. Brewer, "Use of an inclusive option and the optimal number of options for multiple-choice items," *Educ. Psychol. Meas.*, vol. 53, no. 1, pp. 241–247, Mar. 1993, doi: 10.1177/0013164493053001027.

[41] T. M. Haladyna and S. M. Downing, "How many options is enough for a multiple-choice test item?," *Educ. Psychol. Meas.*, vol. 53, no. 4, pp. 999–1010, Dec. 1993, doi: 10.1177/0013164493053004013.

[42] J. T. Sidick, G. V. Barrett, and D. Doverspike, "Three-alternative multiple choice tests: An attractive option," *Pers. Psychol.*, vol. 47, no. 4, pp. 829–835, 1994, doi: 10.1111/j.1744-6570.1994.tb01579.x.

[43] J. E. Bruno and A. Dirkzwager, "Determining the optimal number of alternatives to a multiple-choice test item: An information theoretic perspective," *Educ. Psychol. Meas.*, vol. 55, no. 6, pp. 959–966, Dec. 1995, doi: 10.1177/0013164495055006004.

[44] M. C. Rodriguez, "Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research," *Educ. Meas. Issues Pract.*, vol. 24, no. 2, pp. 3–13, 2005, doi: 10.1111/j.1745-3992.2005.00006.x.

[45] J. Hattie and H. Timperley, "The power of feedback," *Rev. Educ. Res.*, vol. 77, no. 1, pp. 81–112, 2007, doi: 10.3102/003465430298487.

[46] F. M. van der Kleij, R. C. W. Feskens, and T. J. H. M. Eggen, "Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis," *Rev. Educ. Res.*, vol. 85, no. 4, pp. 475–511, 2015, doi: 10.3102/0034654314564881.

[47] S. L. Pressey, "A simple apparatus which gives tests and scores and teaches," *Sch. Soc.*, vol. 23, pp. 373–376, 1926.

[48] S. Cummins, A. R. Beresford, and A. Rice, "Investigating engagement with in-video quiz questions in a programming course," *IEEE Trans. Learn. Technol.*, vol. 9, no. 1, pp. 57–66, Jan.–Mar. 2016, doi: 10.1109/TLT.2015.2444374.

[49] M. Zyda, "From visual simulation to virtual reality to games," *Computer*, vol. 38, no. 9, pp. 25–32, 2005, doi: 10.1109/MC.2005.297.

[50] T. Schubert, F. Friedmann, and H. Regenbrecht, "The experience of presence: Factor analytic insights," *Presence Teleoperators Virtual Environ.*, vol. 10, no. 3, pp. 266–281, 2001, doi: 10.1162/105474601300343603.

[51] J. H. Brockmyer, C. M. Fox, K. A. Curtiss, E. McBroom, K. M. Burkhart, and J. N. Pidruzny, "The development of the game engagement questionnaire: A measure of engagement in video gameplaying," *J. Exp. Soc. Psychol.*, vol. 45, pp. 624–634, 2009, doi: 10.1016/j.jesp.2009.02.016.

[52] A. M. Lund, "Measuring usability with the USE questionnaire," *Usability Interface*, vol. 8, no. 2, pp. 3–6, 2001.

[53] R. Schwarzer and M. Jerusalem, "Generalized self-efficacy scale," in *Measures in Health Psychology: A User's Portfolio. Causal and Control Beliefs*, J. Weinman, S. Wright, and M. Johnston, Eds., Windsor, U.K.: Nfer-Nelson, 1995, pp. 35–37.

[54] J. Sweller, "Cognitive load theory and educational technology," *Educ. Technol. Res. Develop.*, vol. 68, no. 1, pp. 1–16, Feb. 2020, doi: 10.1007/s11423-019-09701-3.

[55] A. Tricot and J. Sweller, "Domain-specific knowledge and why teaching generic skills does not work," *Educ. Psychol. Rev.*, vol. 26, no. 2, pp. 265–283, Jun. 2014, doi: 10.1007/s10648-013-9243-1.

[56] D. C. Geary, "Principles of evolutionary educational psychology," *Learn. Individual Differences*, vol. 12, no. 4, pp. 317–345, Jan. 2002, doi: 10.1016/S1041-6080(02)00046-8.

[57] J. Sweller and S. Sweller, "Natural information processing systems," *Evol. Psychol.*, vol. 4, no. 1, Jan. 2006, Art. no. 147470490600400135, doi: 10.1177/147470490600400135.

[58] N. Cowan, "The magical number 4 in short-term memory: A reconsideration of mental storage capacity," *Behav. Brain Sci.*, vol. 24, no. 1, pp. 87–114, Feb. 2001, doi: 10.1017/S0140525X01003922.

[59] L. Peterson and M. J. Peterson, "Short-term retention of individual verbal items," *J. Exp. Psychol.*, vol. 58, no. 3, pp. 193–198, 1959, doi: 10.1037/h0049234.

[60] 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, Jul. 2015, doi: 10.1109/TVCG.2015.2403312.

[61] M. Olbrich, H. Graf, J. Keil, R. Gad, S. Bamfaste, and F. Nicolini, "Virtual reality based space operations—A study of ESA's potential for VR based training and simulation," in *Virtual, Augmented and Mixed Reality: Interaction, Navigation, Visualization, Embodiment, and Simulation*, J. Y. C. Chen and G. Fragomeni, Eds., Berlin, Germany: Springer-Verlag, 2018, pp. 438–451.
[62] A. Forgiarini, F. Buttussi, and L. Chittaro, "Virtual reality for object location spatial memory: A comparison of handheld controllers and force feedback gloves," in *Proc. 15th Biannual Conf. Italian SIGCHI Ch.*, 2023, pp. 1–9, doi: 10.1145/3605390.3605403.
[63] Y. Ling, W.-P. Brinkman, H. T. Nefs, C. Qu, and I. Heynderickx, "Effects of stereoscopic viewing on presence, anxiety, and cybersickness in a virtual reality environment for public speaking," *Presence Teleoperators Virtual Environ.*, vol. 21, no. 3, pp. 254–267, Aug. 2012, doi: 10.1162/PRES_a_00111.
[64] F. Buttussi and L. Chittaro, "Effects of different types of virtual reality display on presence and learning in a safety training scenario," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 2, pp. 1063–1076, Feb. 2018, doi: 10.1109/TVCG.2017.2653117.
[65] G. Gonçalves, M. Melo, J. Vasconcelos-Raposo, and M. Bessa, "Impact of different sensory stimuli on presence in credible virtual environments," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 11, pp. 3231–3240, Nov. 2020, doi: 10.1109/TVCG.2019.2926978.
[66] W. Peng, "The mediational role of identification in the relationship between experience mode and self-efficacy: Enactive role-playing versus passive observation," *Cyber Psychol. Behav.*, vol. 11, no. 6, pp. 649–652, Dec. 2008, doi: 10.1089/cpb.2007.0229.
[67] M. Wang, H. Yu, Z. Bell, and X. Chu, "Constructing an Edu-Metaverse ecosystem: A new and innovative framework," *IEEE Trans. Learn. Technol.*, vol. 15, no. 6, pp. 685–696, Dec. 2022, doi: 10.1109/TLT.2022.3210828.
[68] X. Chen, Z. Zhong, and D. Wu, "Metaverse for education: Technical framework and design criteria," *IEEE Trans. Learn. Technol.*, vol. 16, no. 6, pp. 1034–1044, Dec. 2023, doi: 10.1109/TLT.2023.3276760.
[69] Z. Han, Y. Tu, and C. Huang, "A framework for constructing a technology-enhanced education metaverse: Learner engagement with human–machine collaboration," *IEEE Trans. Learn. Technol.*, vol. 16, no. 6, pp. 1179–1189, Dec. 2023, doi: 10.1109/TLT.2023.3257511.
[70] Y. Song, J. Cao, K. Wu, P. L. H. Yu, and J. C.-K. Lee, "Developing 'learningverse'—A 3-D metaverse platform to support teaching, social, and cognitive presences," *IEEE Trans. Learn. Technol.*, vol. 16, no. 6, pp. 1165–1178, Dec. 2023, doi: 10.1109/TLT.2023.3276574.
[71] A. Bandura, "Social cognitive theory: An agentic perspective," *Annu. Rev. Psychol.*, vol. 52, pp. 1–26, Jan. 2001, doi: 10.1146/annurev.psych.52.1.1.
[72] S. L. Britner and F. Pajares, "Sources of science self-efficacy beliefs of middle school students," *J. Res. Sci. Teach.*, vol. 43, no. 5, pp. 485–499, 2006, doi: 10.1002/tea.20131.
[73] G. Makransky, P. Wismer, and R. E. Mayer, "A gender matching effect in learning with pedagogical agents in an immersive virtual reality science simulation," *J. Comput. Assist. Learn.*, vol. 35, no. 3, pp. 349–358, 2019, doi: 10.1111/jcal.12335.
[74] D. R. Krathwohl, "A revision of Bloom's taxonomy: An overview," *Theory Pract.*, vol. 41, no. 4, pp. 212–218, 2002, doi: 10.1207/s15430421tip4104_2.
[75] B. Finn, R. Thomas, and K. A. Rawson, "Learning more from feedback: Elaborating feedback with examples enhances concept learning," *Learn. Instruct.*, vol. 54, pp. 104–113, Apr. 2018, doi: 10.1016/j.learninstruc.2017.08.007.
[76] 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, Nov./Dec. 2005, doi: 10.1109/TVCG.2005.92.
[77] K. Kim, M. Z. Rosenthal, D. J. Zielinski, and R. Brady, "Effects of virtual environment platforms on emotional responses," *Comput. Methods Program. Biomed.*, vol. 113, no. 3, pp. 882–893, Mar. 2014, doi: 10.1016/j.cmpb.2013.12.024.
[78] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (task load index): Results of empirical and theoretical research," *Adv. Psychol.*, vol. 52, pp. 139–183, Jan. 1988, doi: 10.1016/S0166-4115(08)62386-9.
[79] A. Korbach, R. Brünken, and B. Park, "Measurement of cognitive load in multimedia learning: A comparison of different objective measures," *Instructional Sci.*, vol. 45, no. 4, pp. 515–536, Aug. 2017, doi: 10.1007/s11251-017-9413-5.
[80] K. Tcha-Tokey, O. Christmann, E. Loup-Escande, and S. Richir, "Proposition and validation of a questionnaire to measure the user experience in immersive virtual environments," *Int. J. Virtual Reality*, vol. 16, no. 1, pp. 33–48, Jan. 2016, doi: 10.20870/IJVR.2016.16.1.2880.
[81] W. Arthur, Ed., *Individual and Team Skill Decay: The Science and Implications for Practice*, (Series in applied psychology). London, U.K.: Routledge, 2013.
[82] X. Chen, D. Zou, H. Xie, and F. L. Wang, "Metaverse in education: Contributors, cooperations, and research themes," *IEEE Trans. Learn. Technol.*, vol. 16, no. 6, pp. 1111–1129, Dec. 2023, doi: 10.1109/TLT.2023.3277952.
[83] C. Villalonga-Gómez, E. Ortega-Fernández, and E. Borau-Boira, "Fifteen years of metaverse in higher education: A systematic literature review," *IEEE Trans. Learn. Technol.*, vol. 16, no. 6, pp. 1057–1070, Dec. 2023, doi: 10.1109/TLT.2023.3302382.

Fabio Buttussi received the Ph.D. degree in computer science from the University of Udine, Udine, Italy, in 2009.

He is currently an Assistant Professor with the Department of Mathematics, Computer Science, and Physics, University of Udine. His major research interests include human–computer interaction, virtual reality, serious games, and their applications in health and safety education.

Luca Chittaro is a Full Professor of Human–Computer Interaction (HCI) with the Department of Mathematics, Computer Science, and Physics, University of Udine, Udine, Italy, where he heads the HCI Lab. He has authored or coauthored more than 200 international academic publications. His major research interests include virtual reality, serious games, persuasive technology, mobile HCI, and their applications in health and safety education.

Prof. Chittaro has received research grants from organizations, such as the US Federal Aviation Administration and the European Union, and companies, such as the Benetton Group and the Intesa Sanpaolo bank. He has been an ACM Distinguished Speaker.

Open Access provided by 'Università degli Studi di Udine' within the CRUI CARE Agreement