WP 6: Tools and guidelines for improving/designing a resilient BE assessed through case studies and virtual training

T6.1 - Virtual training development: identification of performance-based features; implementation protocols of innovative solutions within VR/AR environments, accessed on site or remotely (smartphone, tablet…) and targeted on different users' profiles (i.e. technicians, rescuers, users).

DELIVERABLE ID	D6.2.2
Deliverable Title	Report on users' feedback questionnaires
Delivery month	M23
Revision	1.0
Main partner	BA
Additional partners	AN
Authors of the contribution	Fabio Fatiguso (POLIBA); Mariella De Fino (POLIBA), Federica Cassano (POLIBA)
Deliverable type	report
Number of pages	20

Abstract

According to the workplan of the research project, the analysis of the results from the feedback questionnaires was carried out, based on the framework discussed in D.6.1.2, in order to validate the Bes2ecure Virtual Reality-Serious Game (VR-SG) prototype, as presented in D.6.1.1. In detail, the questionnaires were administered to 162 people, equally distributed per training mode (Video, PC, and Headset), age (18-35, 36-49 and 50-60 years old) and gender. The analysis of the questionnaires was focused on some key aspects, including (i) the efficacy of knowledge achievement depending on the interaction mode and the target groups; (ii) the engagement, perceived usefulness, perceived ease of use and simplicity/efficacy of recommendations depending on the immersivity level, with attention to the novel features of the prototype (crowd motion, multi-hazard approach, outdoor scenario); (iii) the overall realism of the typological scenes and the potential for knowledge transfer to real case settings and, (iv) the usefulness of the error/repetition approach in the game experience. The results are discussed as validation of the prototype and baseline for future demonstrators.

Keywords

Virtual Training, Serious Game, Multi-hazard scenarios, Prototype Validation, Feedback questionnaire, Real-word case studies.

Approvals

Role	Name	Partner	Date
Coordinator	Enrico Quagliarini	UNIVPM	-
Task leader	Fatiguso Fabio	POLIBA	25.10.2023

Revision versions

Revision	Date	Short summary of modifications	Name	Partner

Summary

1. Introduction

The present report is focused on the analysis of the results of the feedback questionnaires that were developed, according to the framework discussed in D.6.1.2, and administered to potential users of the Bes2ecure Virtual Reality-Serious Game (VR-SG) prototype, as presented in D.6.1.1. In detail, the prototype was exported in three interaction modes, i.e. non-immersive game through desktop, immersive game through VR headset and non-interactive recording of videogame exemplary sessions, in view of their pairwise comparison. Moreover, its testing was addressed to three target groups, 18-35 years old, 36-49 years old and 50-60 years old interviewees, corresponding to representative categories of the Italian population with comparable number of citizens (about 10 million people), based on the available statistical data. The analysis of the questionnaires was focused on some key aspects, including (i) the efficacy of knowledge achievement depending on the interaction mode, age and gender; (ii) the engagement, perceived usefulness, perceived ease of use and simplicity/efficacy of recommendations depending on the immersivity level, with attention to the novel features of the prototype (crowd motion, multi-hazard approach, outdoor scenario); (iii) the overall realism of the typological scenes and the potential for knowledge transfer to real case settings and, (iv) the usefulness of the error/repetition approach in the game experience. The results are discussed as validation of the prototype and baseline for future demonstrators.

2. Data analysis methodology

As extensively described in D.6.1.2., where the questions, assessment criteria and scales are detailed (see Appendix), the feedback questionnaire is composed of three parts, resulting in different data analysis methods.

- The first part, to be completed only before the training, is related to demographic information, such as age, gender, education level, previous experiences of training and virtual reality. Data were collected anonymously. Thus, an identification code was defined for each volunteer involved in the test, for demographic data, pre-training questionnaire, training SG-VR test and post-training questionnaire, without allowing the identification of the responder. The results are analyzed in terms of percentage distribution across the whole sample, in order to identify the users' profiling.

- The second part, to be completed before and after the training, is related to prior and acquired knowledge. It results in the number of right answers to both open and close ended questions, the latter ones following the visualization of the virtual tour of a real case study (i.e. Piazza dei Priori, Narni, Italy). The analysis, following the preliminary positive assessment of data internal consistency through the Cronbach's alpha coefficient, is carried out for each interaction mode and each target group, according to the following steps, resulting by well-established statistical methodologies (Gravetter, Frederick J; Wallnau, Larry B.; Forzano, Lori-Ann B.; Witnauer 2021):
  - Assessing that the knowledge before the training is not statistically different from zero for the three interaction modes, by suitable diversity testing, in order to verify consistent starting conditions. To this end, the Mann–Whitney U test[1] is used for pairwise comparison, since the samples are random and independent, the data is continuous and the scale of measurement is ordinal.

[1] The null hypothesis stating that there is no difference between the pre-training scores for each couple of interaction modes (Video/PC, Video/Headset, PC/Headset). Since the samples are larger than 10, it is assumed that the sampling distribution is approximately normal and, thus, the Z-ratio is used to calculate the p-value. If the p-value $< \alpha$ (0,05) the null hypothesis is rejected.

- Assessing that the knowledge before and after the training is statistically different from zero for each interaction mode, by suitable diversity testing, in order to verify meaningful improvements by the training. To this end, The Wilcoxon Signed-Ranks test[2] is used for pairwise comparison of knowledge (pre-training/post-training), since the matched samples are dependent and correlated (the same interviewees answered the same questions before and after the game), the data is continuous and the scale of measurement is ordinal. Furthermore, the content analysis of relevant keywords reported in the answers is carried out, in order to identify which training concepts feature both prior and acquired knowledge.
- Assessing whether or not the knowledge after the training is statistically different from zero when pairwise comparing different interaction modes (Video, PC, Headset) for all the interviewees and for different target groups, in terms of age (18-35ys, 36-49ys, 50-60ys) and gender (male, female). The Mann–Whitney U test[3] for ordinal, independent and continuous data is used. Moreover, the indexes of variability (mean value and standard deviation) are calculated, in order to detect relevant trends and correlations on the training efficacy.
- Assessing whether or not the Virtual Tour knowledge after the training is statistically different from zero when pairwise comparing different interaction modes (Video, PC, Headset) for all the interviewees and for different target groups, in terms of age (18-35ys, 36-49ys, 50-60ys) and gender (male, female). The Mann–Whitney U test[4] for ordinal, independent and continuous data is used. Moreover, the indexes of variability (mean values and standard deviation) are calculated, in order to observe relevant trends and correlations on the transferring of knowledge and applying of skills to new settings.

➢ The third part, to be completed only after the training, is related to the usability and user experience of the proposed tools, with reference to self-reported engagement, perceived ease of use, perceived usefulness, simplicity/efficacy of recommendations and realism. The analysis, following the preliminary positive assessment of data internal consistency through the Cronbach's alpha coefficient, is carried out by checking whether or not the assessment of the above-mentioned indicators is statistically different from zero when pairwise comparing the interactive modes (PC, Headset) for all the interviewees and for different target groups, in terms of age (18-35ys, 36-49ys, 50-60ys) and gender (male, female). The Mann–Whitney U test[5] for ordinal, independent and continuous data is used. Moreover, the indexes of variability (mean values and standard deviation) are assessed, in order to detect relevant trends and correlations on the training acceptability and involvement.

[2] The test is based on the null hypothesis stating that there is no difference between the pre-training and post-training scores for each interaction modes (Video, PC, Headset). Since the samples are larger than 10, it is assumed that the sampling distribution is approximately normal and, thus, the Z-ratio is used to calculate the p-value. If the p-value < α (0,05) the null hypothesis is rejected.
[3] The null hypothesis stating that there is no difference between each couple of compared samples. Since the samples are larger than 10, it is assumed that the sampling distribution is approximately normal and, thus, the Z-ratio is used to calculate the p-value. If the p-value < α (0,05) the null hypothesis is rejected.
[4] The null hypothesis stating that there is no difference between each couple of compared samples. Since the samples are larger than 10, it is assumed that the sampling distribution is approximately normal and, thus, the Z-ratio is used to calculate the p-value. If the p-value < α (0,05) the null hypothesis is rejected.
[5] The null hypothesis stating that there is no difference between each couple of compared samples. Since the samples are larger than 10, it is assumed that the sampling distribution is approximately normal and, thus, the Z-ratio is used to calculate the p-value. If the p-value < α (0,05) the null hypothesis is rejected.

Finally, since a log file was automatically generated after all the h PC and Headset game sessions, displaying the number of wrong answers (and consequent repetitions of questions) during the training, a direct evaluation against the knowledge gain was carried out, in order to detect potential correlations between errors and improvements.

3. Results

3.1 Demographic information

The feedback questionnaires were administered to 162 people, equally distributed per training mode (54 each) and age (54 each). According to national regulations and university rules, all the participants signed a waiver prior to testing and gave their consent to the exploitation of the anonymized data for research purposes. An independent-measure research design was carried out to compare conditions between the interaction modes, and the suitability of the sample was performed according to previous works methods (Latini et al. 2023). In particular, the definition of groups with 54 subjects ensures: (1) the normality of distribution, being the number of subjects > 30 subjects as normality threshold according to the central limit theory (Gravetter, Frederick J; Wallnau, Larry B.; Forzano, Lori-Ann B.; Witnauer 2021) the detection of significant effects with a statistical power equal to 0.80 according to the power analysis (effect size 0.50, α = 0.05) through the G*Power software (Franz et al. 2007).
Concerning the demographic profile (first part), a good gender balance was achieved (52% women and 48% male), with a predominance of medium-high levels of education (Figure 1) and low previous experiences in training (Figure 2), with high percentage of interviewees with no experiences in training for earthquake (92%) and heat wave (98%). Moreover, more than half of the sample declared no previous experiences in virtual reality (54,4% in Figure 3) and a great share declared low frequency in playing video-games (82,1% answering less than "at least once a week" in Figure 4), with some differences when plotting the data versus age, particularly for the latter indicator (Figure 5).

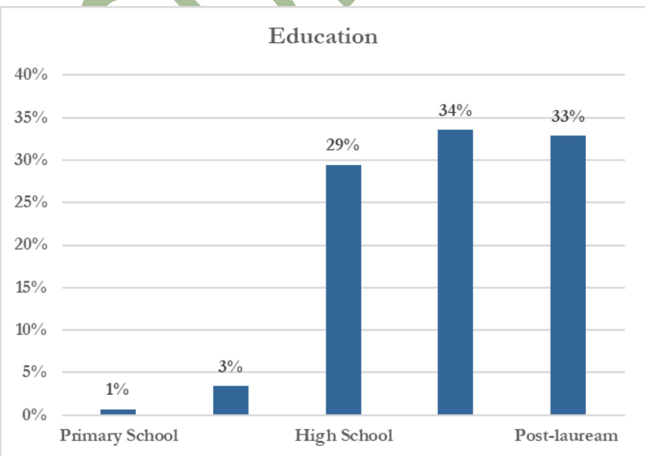


Figure 1. Percentage distribution of education

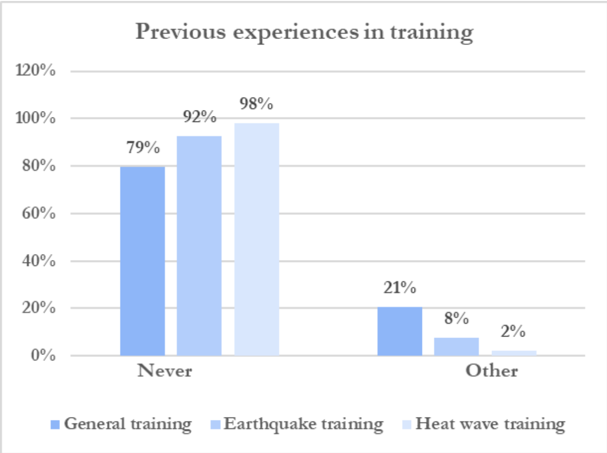


Figure 2. Percentage distribution of previous training experience

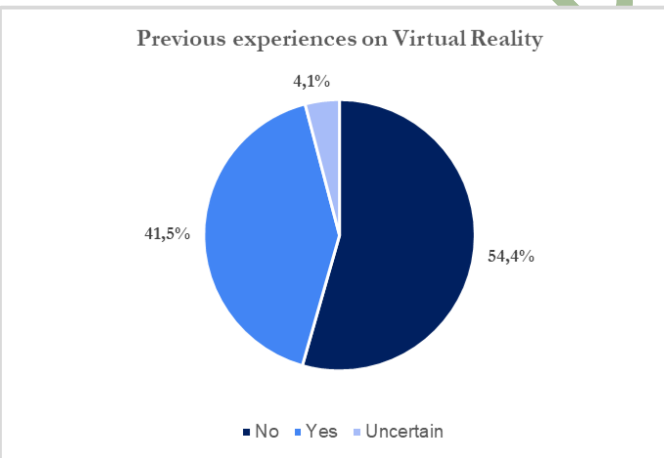


Figure 3. Percentage distribution of previous VR experience

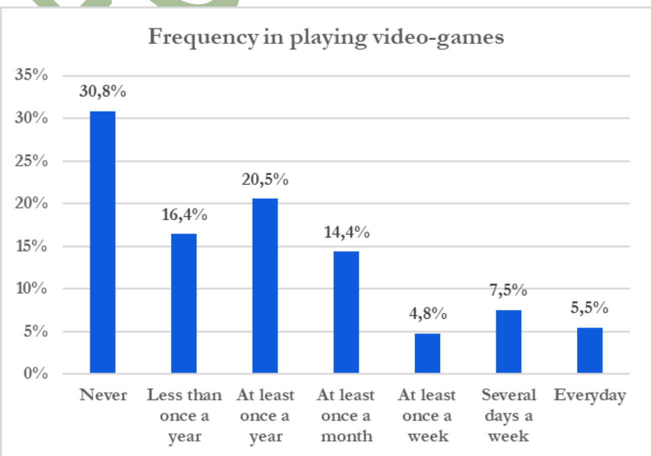


Figure 4. Percentage distribution of playing frequency

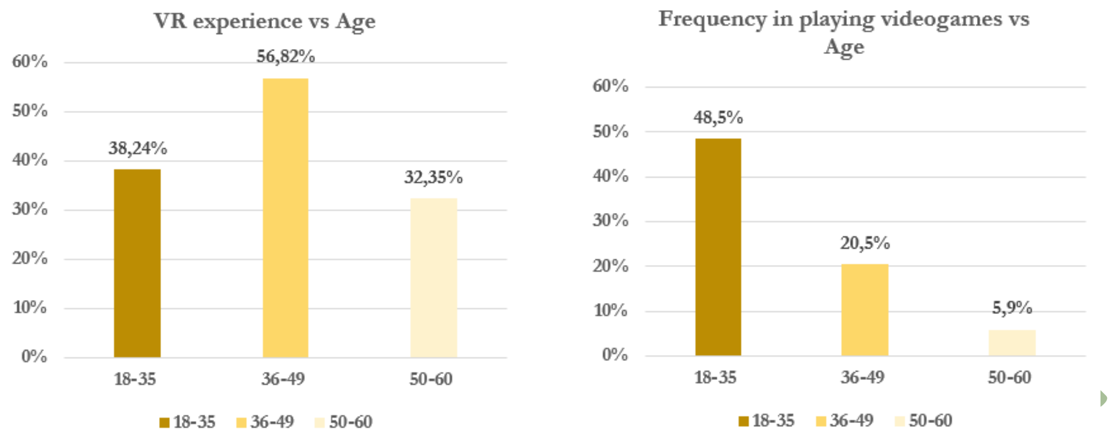


Figure 5. Percentage distribution of previous VR experience(left) and playing frequency (right) versus age

3.2 Knowledge

Concerning the knowledge (second part), no statistical differences were detected among the pre-training scores in the three interaction modes (p-value > 0,05 confirming the null hypothesis), while the pre- and post-training knowledge were found statistically different for each interaction mode (p-values < 0,05, rejecting the null hypotheses).

Consequently, it could be stated that the starting conditions were consistent in the three interaction modes and the training always resulted in meaningful knowledge improvements (Figure 6).

However, looking at the pairwise comparison of the knowledge after the training in the three interaction modes for all the interviewees, it was found that the PC mode was statistically different from Video, as well as Headset from Video, while PC and Video were not. Similar statistical results were observed for the target group 18-35 (Fig. 7 left). Differently, Headset mode was found statistically different from both PC and Video for the target group 36-49 (Fig. 7 middle) and no statistical differences were found between all the modes for the target group 50-60 (Fig.7 right).

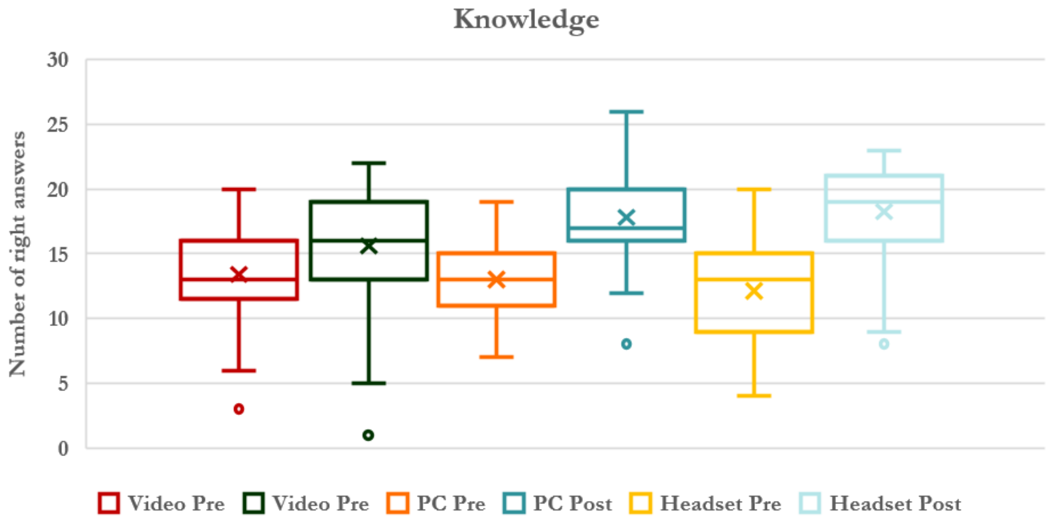


Figure 6. Box plot of pre- and post-training knowledge scores by interaction mode (all interviewees)

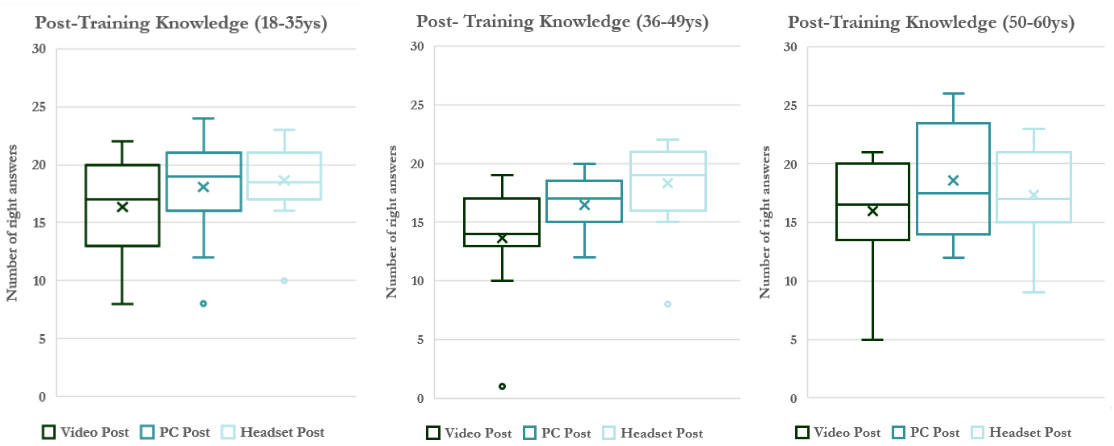


Figure 7. Box plots of post-training knowledge scores by interaction mode and age

The same analysis was carried out looking only at the post-training knowledge scores through close ended questions following the visualization of the virtual tour of a real case study.

In this case, for all the interviewees (Fig. 8), as well as for the target groups 18-35 (Fig. 9 left) and 36-49 (Fig. 9 middle), it was found that the PC mode was statistically different from Video, as well as Headset from Video. Differently, no statistical differences were found between all the modes for the target group 50-60 (Fig.9 right). Moreover, the indexes of variability of the knowledge assessment for all the above-mentioned indicators are listed in Table 1.

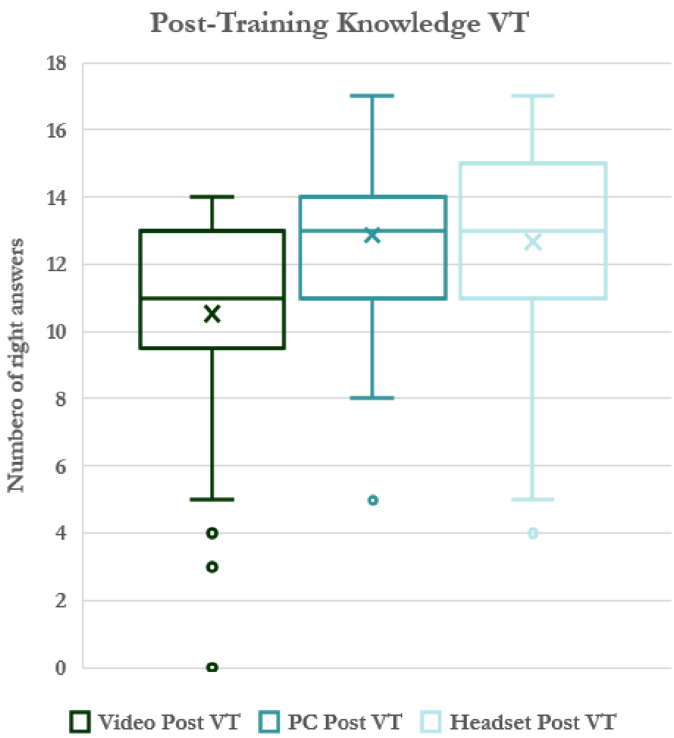


Figure 8. Box plots of post-training VT knowledge scores by interaction mode (all interviewees)

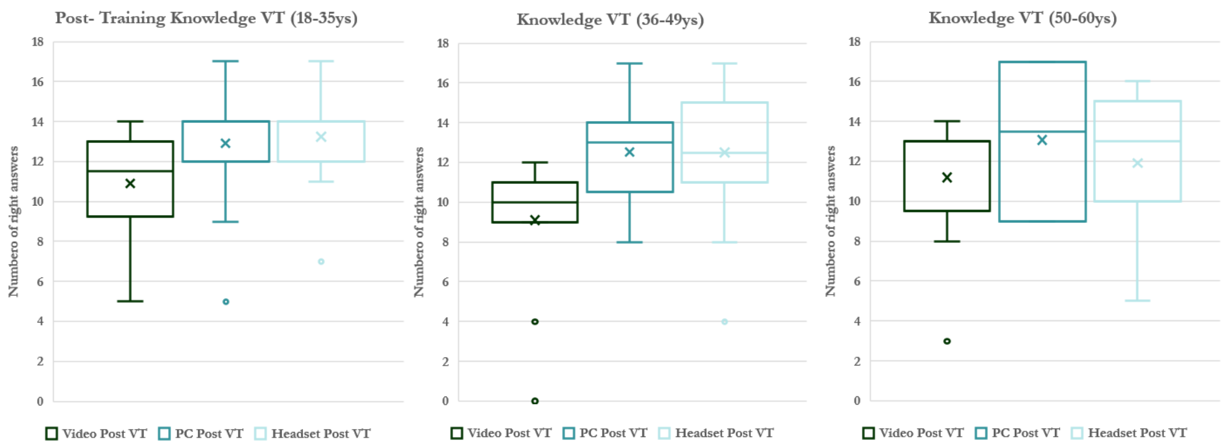


Figure 9. Box plots of post-training VT knowledge scores by interaction mode and age

Table 1. Indexes of variability for knowledge assessment

	Group	***Video (mean st.dev.)***	***PC (mean st.dev.)***	***Headset (mean st.dev.)***
Post-training number of right answers (out of 28)	*All*	15,6 ± 4,5	17,8 ± 3,8	18, 3 ± 3,4
	18-35	16,4 ± 4,1	18,1 ± 3,6	18,7 ± 3,0
	36-49	13,6 ± 4,9	16,5 ± 2,5	18,4 ± 3,5
	50-60	16,0 ± 4,8	18,6 ± 4,8	17,4 ± 4,2
Post-training VT number of right answers (out of 17)	*All*	10,5 ± 3,0	12,9 ± 2,8	12,7 ± 2,9
	18-35	10,9 ± 2,4	12,9 ± 2,6	13,3 ± 2,2
	36-49	9,1 ± 3,7	12,5 ± 2,4	12,5 ± 3,2
	50-60	11,2 ± 3,4	13,1 ± 3,4	11,9 ± 3,4

Concerning the gender based analysis, no statistical differences were found for overall post-training knowledge (Fig. 10) and post-training VT knowledge between men and women.

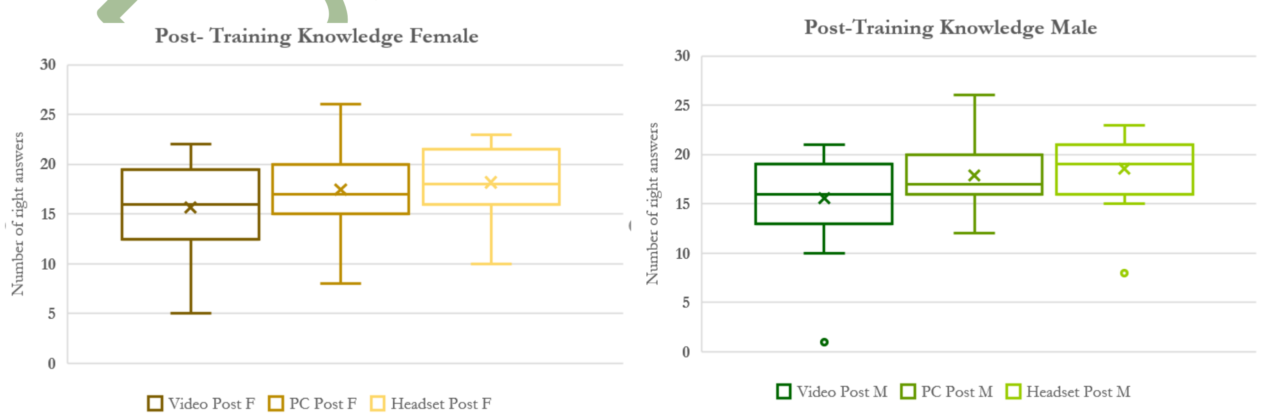


Figure 10. Box plots of post-training knowledge scores by interaction mode and gender

Finally, all the answers within the knowledge section of the feedback questionnaires were analyzed, with specific focus on the Heat Wave and Earthquake game sessions. It is worth recalling that the training is arranged in modules, corresponding to the training objectives/items suggested by national government documents. In detail, the recommendations by the Italian Ministry of Health (https://www.salute.gov.it/portale/caldo/homeCaldo.jsp) are related to the identification of outdoor mitigating elements, where the temperature peaks are supposed to be lower compared to the surroundings (i.e. TREES, FOUNTAINS, BUILDINGS), whereas, for the earthquake response in open spaces, the rules by the Italian Department of Civil Protection (https://www.protezionecivile.gov.it/it/approfondimento/in-caso-di-terremoto) are adopted on avoiding closeness to dangerous targets (i.e. BUILDINGS that could collapse, GLAZED SURFACES/OBJECTS that could break, VEHICLES that could hinder rescue operations, and ELECTRIC DEVICES that could catch fire or emit sparks). The same items are used as control answers in assessing the questionnaires.

In this regard, it was found that for the Heat Wave (Figure 11 left), the interviewees mostly referred to buildings before the training (about 57% of the questionnaires), with very low share related to the fountains/water sources (about 6%). After the training, the answers were more balanced and the knowledge gaps were filled by exceeding 50% for all the items. Similarly, for the Earthquake (Fig. 11 right), before the training the recommendation avoiding closeness to buildings was widely known (about 75% of the questionnaires), while the other targets were not. However, for the latter ones, a significant knowledge gain was assessed after the game testing.

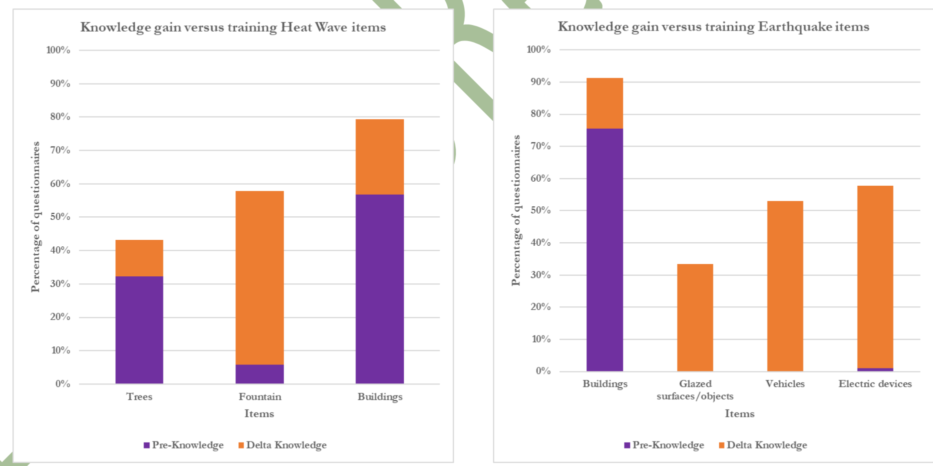


Figure 11. Knowledge gain versus training items

3.3 Usability and experience

Concerning the usability and user experience of the VR-SG tools, it should be first observed that there are no statistically relevant differences between the non-immersive and immersive modes, as resulted from the Mann–Whitney U test (p-value>0,05). Similarly, it should be noted that no statistical difference was found, by analyzing the data per age and gender (Figure 12). Nevertheless, some insights might come from the percentage of positive scores and the indexes of variability, as it follows.

In detail, concerning the engagement (Figure 13), most of the interviewees (94% PC, 98% Headset) has assessed positively (n>5) the criterion, resulting in mean scores of 5,9 (st.dev. 0,9) and 6,4 (st.dev.0,7) for PC and Headset, respectively.

Concerning the perceived usefulness (Figure 14), a share of 94% for PC and 100% for Headset has assessed positively (n>5) the criterion, resulting in mean scores of 5,8 (st.dev. 0,9) and 6,3 (st.dev. 0,6) for PC and Headset, respectively.

Concerning the ease of use (Figure 15), a share of 69% for PC and 67% Headset has assessed positively (n>5) the criterion, resulting in mean scores of 5,5 (st.dev. 1,3) and 5,5 (st.dev. 0,9) for PC and Headset, respectively.

Concerning the efficacy and simplicity of recommendations (Figure 16), a share of 92% for both PC and Headset has assessed positively (n>5) the criterion, resulting in mean scores of 6,2 (st.dev. 1,2) and 6,3 (st.dev. 0,7) for PC and Headset, respectively.

Concerning the realism (Figure 17), a share of 73% for PC and 90% for Headset has assessed positively (n>5) the criterion, resulting in mean scores of 5,5 (st.dev. 1,2) and 6,0 (st.dev. 0,7) for PC and Headset, respectively.

Furthermore, a summary of the mean values for each question are mapped in Table 2.

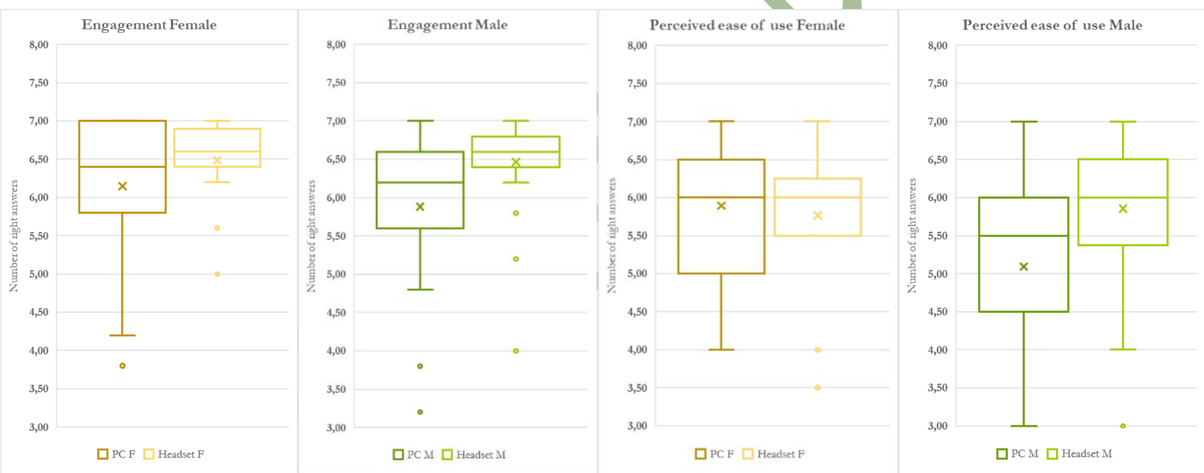
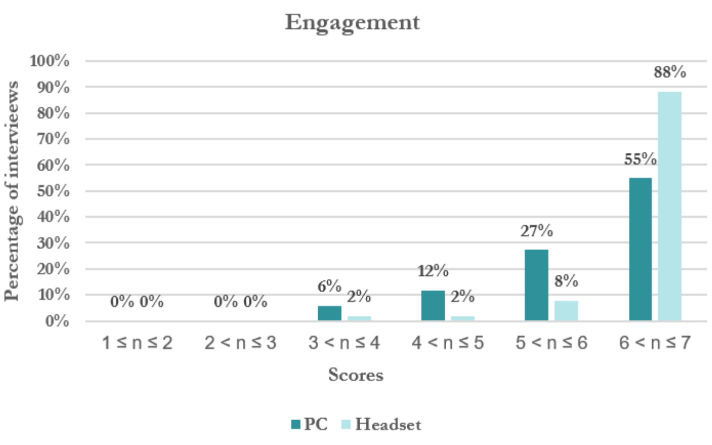


Figure 12. Box-plots of scores from representative criteria by gender



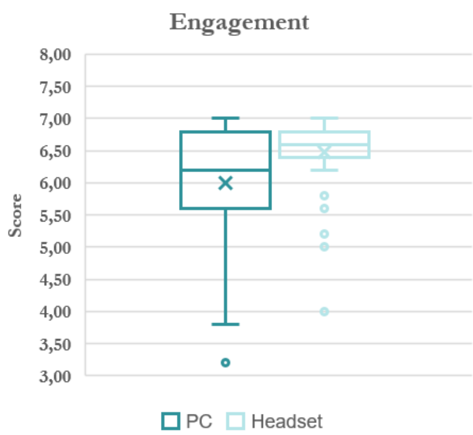


Figure 13. Distribution of scores (left) and box-plot (right) of data on engagement versus training mode

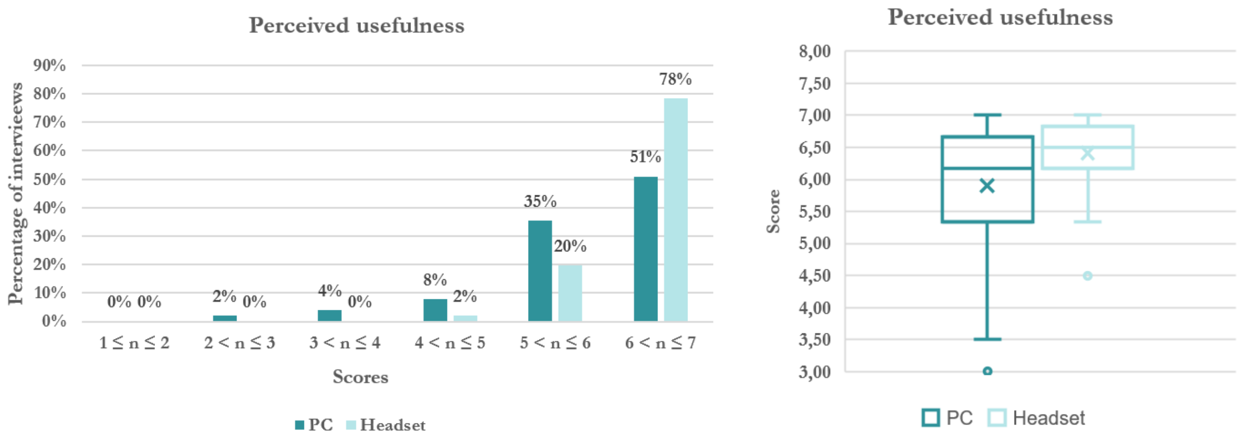


Figure 14. Distribution of scores (left) and box-plot (right) of data on perceived usefulness versus training mode

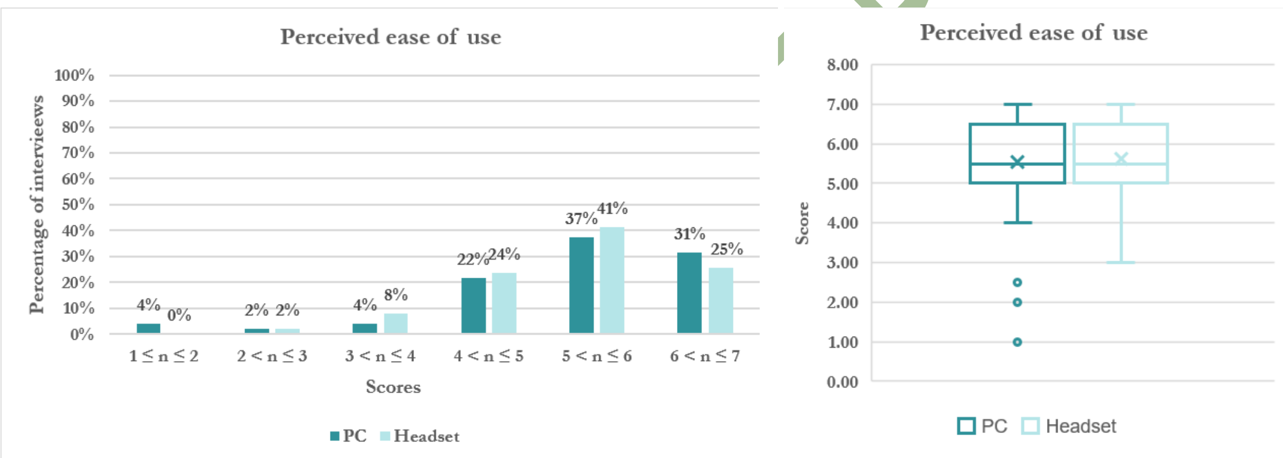


Figure 15. Distribution of scores (left) and box-plot (right) of data on perceived ease of use versus training mode

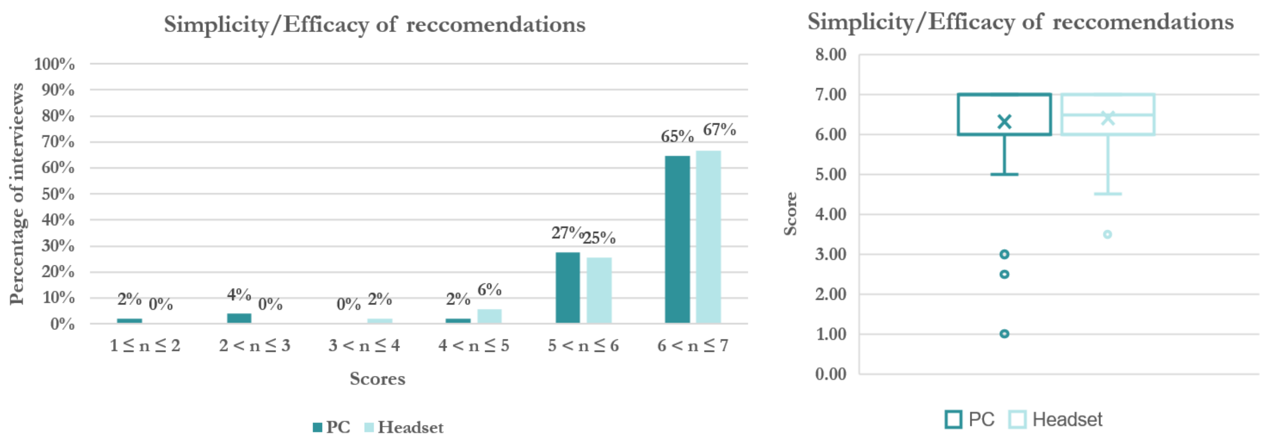


Figure 16. Distribution of scores (left) and box-plot (right) of data on efficacy/simplicity of recommendations versus training mode

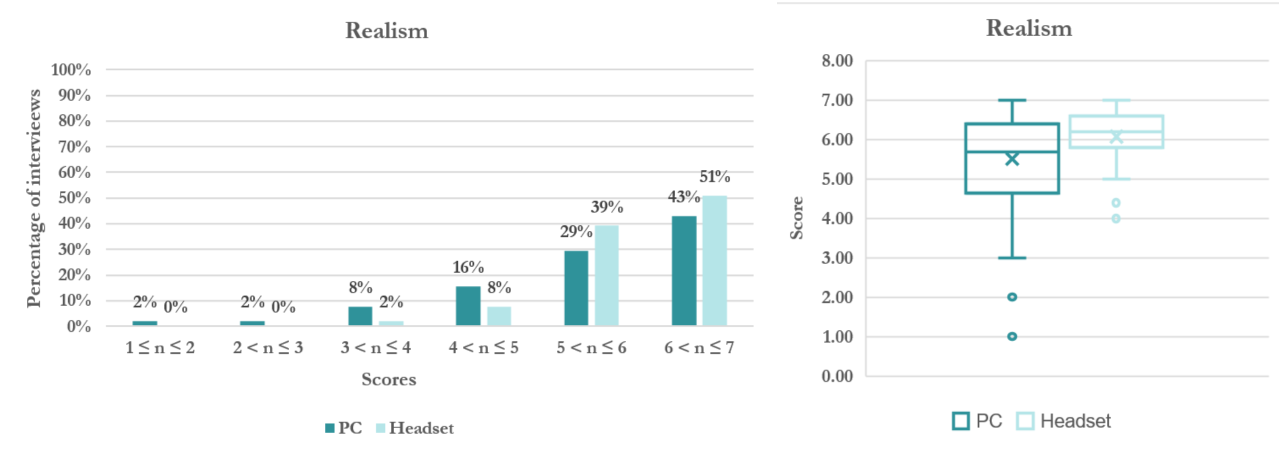


Figure 17. Distribution of scores (left) and box-plot (right) of data on realism versus training mode

Table 2. Maps of indexes of variability for all the questions of the third part (to be continued)
**the scores of the answers to the questions asked in negative form were mirrored*

Legend	
mean ≤ 5,5	
5,5 < mean ≤ 6	
6 < mean ≤ 6,5	
6,5 < mean ≤ 7	

Criterion	Questions	Group	PC	Headset
Self-reported engagement	The training experience was fun and enjoyable	18-35	6,1 ± 1,0	6,8 ± 0,5
		36-49	5,9 ± 1,3	6,6 ± 0,9
		50-60	5,7 ± 1,3	6,5 ± 0,5
	Safety training activities are boring*	18-35	6,0 ± 1,5	6,7 ± 0,7
		36-49	6,2 ± 1,4	6,8 ± 0,6
		50-60	6,0 ± 1,4	6,8 ± 0,4
	I would describe safety training as very interesting	18-35	5,9 ± 1,1	6,3 ± 1,0
		36-49	5,8 ± 1,4	6,5 ± 0,9
		50-60	5,9 ± 1,2	6,5 ± 0,5
	Safety training does not hold my attention at all*	18-35	6,6 ± 0,6	6,7 ± 0,5
		36-49	6,2 ± 1,5	6,4 ± 0,9
		50-60	6,2 ± 1,4	6,5 ± 1,3
	It was easy for me to concentrate on my learning	18-35	6,2 ± 1,4	5,7 ± 1,7
		36-49	5,1 ± 1,7	5,9 ± 1,2
		50-60	5,4 ± 1,2	6,5 ± 0,8
Perceived ease of use	This simulation tool is rigid and inflexible to interact with*	18-35	5,5 ± 1,3	5,9 ± 1,1
		36-49	4,9 ± 2,1	5,3 ± 1,6

		50-60	5,4 ± 1,9	5,1 ± 0,7
	I think this training tool is easy to use	18-35	6,3 ± 0,6	6,0 ± 0,7
		36-49	5,2 ± 1,7	5,9 ± 1,0
		50-60	5,5 ± 1,3	5,2 ± 0,8
Recommendation simplicity and efficacy	*I could easily remember the recommendations provided in the virtual experience*	18-35	6,8 ± 0,4	6,4 ± 0,9
		36-49	5,6 ± 1,9	6,3 ± 0,7
		50-60	6,1 ± 1,1	6,4 ± 0,9
	The recommendations provided in the training experience are useful for my safety	18-35	6,6 ± 0,6	6,4 ± 0,7
		36-49	6,1 ± 1,9	6,6 ± 0,7
		50-60	6,1 ± 1,2	6,5 ± 0,8

Table 2. Maps of mean value for all the questions of the third part (continued)

Question		***Group***	***PC***	***Headset***
Perceived usefulness	*Using this type of simulation as an educational tool will enhance my learning*	18-35	6,3 ± 1,0	6,6 ± 0,7
		36-49	5,8 ± 1,6	6,7± 0,5
		50-60	5,7 ±1,3	6,6 ± 0,67
	This type of simulation is as useful as indoor simulations	18-35	5,8 ± 1,6	5,9 ± 1,2
		36-49	5,8 ± 1,2	6,4 ± 1,1
		50-60	5,1 ± 2,0	6,5 ± 0,9
	This type of simulation is useful for behaving properly in real case, too	18-35	6,3 ± 0,8	6,4 ± 0,8
		36-49	5,9 ± 1,7	6,8 ± 0,4
		50-60	6,4 ± 0,7	6,5 ± 0,7
	This type of simulation is useful as a learning supplement	18-35	6,5 ± 0,8	6,7 ± 0,7
		36-49	5,9 ± 1,5	6,8 ± 0,6
		50-60	6,4 ± 0,8	6,7 ± 0,5
	The combination of two risks is effective because it simulates real conditions	18-35	5,5 ± 1,6	6,5 ± 0,7
		36-49	6,0 ± 1,2	6,2 ± 1,2
		50-60	6,3 ± 0,8	6,4 ± 0,8
	The simulation of the crowd helped me make the right decisions	18-35	5,4 ± 1,5	6,1 ± 1,1
		36-49	4,9 ± 1,9	6,1 ± 0,9
		50-60	5,1 ± 1,8	5,7 ± 1,5
Realism	*The built environment was realistic*	18-35	6,1 ± 1,0	5,7 ± 1,2
		36-49	5,8 ± 1,2	6,1 ± 1,0
		50-60	6,3 ± 0,6	6,0 ± 0,9
		18-35	5,2 ± 1,3	5,2 ± 1,0

	The built environment reminded me of a familiar place	*36-49*	*5,1 ± 1,8*	*5,7 ± 1,0*
		50-60	*5,5 ± 1,7*	*5,8 ± 0,8*
	The VR experience was realistic	*18-35*	*5,0 ± 2,2*	*5,8 ± 1,1*
		36-49	*5,2 ± 1,5*	*6,3 ± 1,0*
		50-60	*4,9 ± 2,3*	*6,1 ± 0,5*
	The realism of the virtual world motivates me to learn	*18-35*	*6,2 ± 1,0*	*5,9 ± 1,1*
		36-49	*5,1 ± 1,7*	*6,6 ± 0,6*
		50-60	*4,9 ± 2,6*	*6,4 ± 0,9*
	The virtual world makes learning more interesting	*18-35*	*6,6 ± 0,7*	*6,6 ± 0,7*
		36-49	*5,6 ± 1,7*	*6,8 ± 0,4*
		50-60	*5,5 ± 1,9*	*6,7 ± 0,5*

3.4 In-game reports analysis

The number of "wrong" answers throughout the game for each player was assessed against the questionnaires scores in order to detect relevant correlations.

In detail, Table 3 shows, for the Heat Wave and Earthquake separately, how the number of errors (and consequent repetitions), resulting from the log file automatically generated at the end of the game, is related to the average knowledge gain, in terms of difference of scores in open questions. The latter seems quite constant regardless the former one. Furthermore, Table 4 and Table 5 show how the average number of errors somewhat correspond to the frequency of playing videogames and the declared difficulty in being concentrated.

Table 3. In-game errors versus knowledge gain

	Number of errors	***Average improved scores in open questions***
Heat Wave	*0*	*0,685*
	1	*0,688*
	2	*0,688*
Earthquake	*0*	*1,492*
	1	*1,563*
	2	*1,563*
	3	*1,500*

Table 4. In-game errors versus frequency in playing video-games

Frequency	***Average number of errors by frequency groups***
Never *Less than once a year* *At least once a year*	1,45
At least once a month	1,05

At least once a week	
Several days a week *Everyday*	0,75

Table 5. In-game errors versus perceived concentration

Answer to the question "It was easy for me to concentrate on my learning"	***Average number of errors***
Disagree *Somewhat disagree*	*2,13*
Neutral *Somewhat agree*	*1,33*
Agree *Strongly agree*	*1,30*

4. Discussion

The overall results are herein analyzed for validating the Bes2ecure VR-SG prototype, as well as for drawing some recommendations on future development and testing of virtual training tools for risk communication within the safety management of the built environment.
In detail, the following key methodological and operational aspects are worth being pointed out.

Efficacy and versatility. The observed knowledge gain (Figure 6) in all the three interaction modes – Video, PC and Headset – proves the versatility of the tool that might potentially target users with different digital maturity and/or fitting demonstration venues with different technological equipment. However, looking at the results by age (Figure 7), it was found that the interactive modes – PC and Headset – are generally more effective in transferring information for young users (18-35ys) compared with Video, and that the Headset mode plays a distinctive role for adult users (36-49ys). For senior users (50-60ys), the outcome is less clear-cut, although the PC mode slightly exceeds on average the efficacy of both Video and Headset. As reported in Table 2, the explanation of these differences does not seem to lie in the low attractiveness of the immersive tool for senior users, who scored very high the Headset mode for all the questions related to the engagement, consistently with other studies (Tseng and Giau 2022), but rather in the ease of use that was scored lower for the Headset mode, both compared to the PC mode within the same target group and compared to the same mode for the other target groups. The lower ease of use is reasonably related to the frequency in playing video-games and previous VR experiences (Figure 5). Thus, specific attention should be paid to this target group in wide replication of the training, by ensuring that the users are aware of the options to lower the level of Immersivity by selecting the PC mode or they are fully aware on how the Headset mode, relative controllers and interaction functionalities work, even based on pre-training demonstration. This is paramount to prevent the player from being disoriented by the environment and distracted from the knowledge delivery. The analysis of results by gender did not show any evidence leading to further improvements in terms of equality and it is consistent with the outcome of other studies in the Architecture, Engineering and Construction training sector (Tastan and Tong 2023).

Realism and knowledge transfer. Although the prototype was developed by displaying a typological scene (repetitive patterns, schematic volumes, simple decorations, plain surfaces, neutral crowd figures), as representative of a wide set of case studies, the realism was scored positively, for both the built environment and the experience, even acknowledging the resemblance with familiar places and the added value of realism for training (Table 2). In this regard, the Headset mode systematically results as the most effective, reasonably due to the enhanced fidelity by immersivity. Moreover, the risk of desensitization to real-life disasters or the biases and stereotypes embedded in the virtual environment was not encountered. In fact, the post-training knowledge related to the close ended questions, following the visualization of the virtual tour of a real case study, resulted in very good scores for all the interviewees and modes (Table 1). Thus, it was confirmed the opportunity to integrate the virtual tour within the training itself in order to boost the overall consistency with real life scenarios though the skill transfer. This ability is acknowledged as highly beneficial in all fields of virtual training, because it makes it possible to transfer the acquired skills from an artificial environment to a real one that is similar or identical to the one in which the skills will be actually used in practice, and it was not previously investigated for risk training in buildings and built environments through virtual tools (Strojny and Dużmańska-Misiarczyk 2023). Finally, given that the perception of realism is not significantly affected by the representation fidelity, this aspect could be improved anyway by the employment of sensors, even embedded in wearable devices, to boost the immersiveness of the experience to reproduce sounds, smells, and heat to increase the credibility of the generated virtual world, in line with the insights of other studies (Gagliardi et al. 2023).

Simplicity and usefulness. All the target groups scored very highly the simplicity and efficacy of recommendations (Figure 16 and Table 2), confirming that the modularity of the prototype is beneficial in delivering the information. The results are not dependent from the immersivity level. Concerning the perceived usefulness, the scores are medium-high and high for PC and Headset modes respectively, including the aspects related to outdoor training and multi-risk approach, distinctively featuring the prototype compared to previous applications. The only question that is scored relatively lower is related to the role of the crowd, resulting from the agent-based simulation, which enabled to estimate the time and paths covered by Non Playable Characters (NPCs) to reach a safe area. The result might point out that the crowd motion does not plays a prominent role in leading/encouraging individual actions throughout the game, as informally reported by several interviewees, who found it . A solution could be to exploit the agent-based data for animating Playable Characters (PCs) with active roles in instructing the players, as tested in other studies (Tucker et al. 2018)(Feng et al. 2020a)(Feng et al. 2020b)(Serafini and Chittaro 2023).

Strengthening by repetitions. The possibility to repeat the question in case of errors is confirmed beneficial for the knowledge gain (Table 3) given that the players who failed several times to provide the right answer achieved a comparable level of knowledge gain compared to the players who failed no or a few times (Table 3). Moreover, the error/repetition approach is likely useful for compensating low familiarity with the tools, in terms of experience with videogames (Table 4) and consequent difficulties in concentrating during the training (Table 5).

Immersivity level. The differences between immersive and non-immersive modes were often found not statistically relevant, as in similar studies (Buttussi and Chittaro 2021), although the mean and standard deviation values generally showed a consistent increase in knowledge, engagement, usefulness and realism scores for the Headset mode compared to the PC one. Nonetheless, the efficacy and acceptability of non-immersive training solutions are confirmed in terms of successful learning and subject assessment, particularly compared with traditional approaches, as suggested by other authors (D'Amico et al. 2022)(Rahouti et al. 2021). Thus, the possibility to develop multiple modes for the same game is recommended, especially in case of great variability of the training receivers, including citizens and visitors of urban areas with different ages and experiences in VR and games.

Content analysis. Although applied to very simple keywords, as check indicators of the right answers, the analysis of the most recurring terms in open-ended questions, both from pre- and post- training questionnaire sections, was found quite interesting and useful, leading to the general assessment of those concepts that require strengthening in risk communication. This kind of content analysis, that could be empowered by customized design of questions and more complex methods of data mining for the assessment of the answers (e.g. by natural language processing algorithm) is successfully used in other fields combining VR and AI (Antel et al. 2022) and might be beneficial in risk training, too. Moreover, whenever the tool is arranged in subsequent modules and is based on teleporting, thus with a limited freedom of observable actions and movements of the player, it seems to be even more worthwhile than other approaches (e.g. behavioral analysis, sensor-based measurements) with different training layouts and motion modes.

Administration of individual and group questionnaires. The questionnaires were mainly administrated in targeted events where several users were involved within a demonstration initiative (e.g. building engineering university courses, post-degree specialization workshops). Only in a few cases, the testing was run individually, although with the assistance of the interviewer. It was observed that individual testing was less reliable than group testing, in terms of concentration of the user and reliability of the outcome, resulting in a few cases into the withdrawal of the results for evident inconsistencies (high Likert scale scores to both positive and negative questions, no answers to the post-training questions, out of topic answers). This is a shortcoming that requires great effort particularly for the non-immersive mode (and potentially the video recording as well), since the autonomy in developing the training and testing sessions should be guaranteed in view of large-scale dissemination of this kind of prototype, even by web-based smartphone and tablet applications (Chittaro and Buttussi 2022).

References

Antel R, Abbasgholizadeh-Rahimi S, Guadagno E, et al (2022) The use of artificial intelligence and virtual reality in doctor-patient risk communication: A scoping review. Patient Educ Couns 105:3038–3050. https://doi.org/https://doi.org/10.1016/j.pec.2022.06.006

Buttussi F, Chittaro L (2021) A Comparison of Procedural Safety Training in Three Conditions: Virtual Reality Headset, Smartphone, and Printed Materials. IEEE Trans Learn Technol 14:1–15. https://doi.org/10.1109/TLT.2020.3033766

Chittaro L, Buttussi F (2022) 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 28:1573–1584. https://doi.org/10.1109/TVCG.2020.3022340

D’Amico A, Bernardini G, Lovreglio R, Quagliarini E (2022) A Non-Immersive Virtual Reality Serious Game Application for Flood Safety Training. SSRN Electron J. https://doi.org/10.2139/ssrn.4110990

Feng Z, González VA, Amor R, et al (2020a) An immersive virtual reality serious game to enhance earthquake behavioral responses and post-earthquake evacuation preparedness in buildings. Adv Eng Informatics 45:101118. https://doi.org/10.1016/j.aei.2020.101118

Feng Z, González VA, Trotter M, et al (2020b) How people make decisions during earthquakes and post-earthquake evacuation: Using Verbal Protocol Analysis in Immersive Virtual Reality. Saf Sci 129:104837. https://doi.org/10.1016/j.ssci.2020.104837

Franz F, Edgar E, Albert-Georg L, Axel B (2007) G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods 39:175–191

Gagliardi E, Bernardini G, Quagliarini E, et al (2023) Characterization and future perspectives of Virtual Reality Evacuation Drills for safe built environments: A Systematic Literature Review. Saf Sci 163:106141. https://doi.org/10.1016/j.ssci.2023.106141

Gravetter, Frederick J; Wallnau, Larry B.; Forzano, Lori-Ann B.; Witnauer JE (2021) Essentials of statistics for the behavioral sciences. CENGAN

Latini A, Di Giuseppe E, D'Orazio M (2023) Immersive virtual vs real office environments: A validation study for productivity, comfort and behavioural research. Build Environ 230:109996. https://doi.org/10.1016/j.buildenv.2023.109996

Rahouti A, Lovreglio R, Datoussaïd S, Descamps T (2021) Prototyping and Validating a Non-immersive Virtual Reality Serious Game for Healthcare Fire Safety Training. Fire Technol 57:3041–3078. https://doi.org/10.1007/s10694-021-01098-x

Serafini M, Chittaro L (2023) Perception of Virtual Agents as Communicators in Virtual vs. Augmented Reality by a Male Sample

Strojny P, Dużmańska-Misiarczyk N (2023) Measuring the effectiveness of virtual training: A systematic review. Comput Educ X Real 2:100006. https://doi.org/10.1016/j.cexr.2022.100006

Tastan H, Tong T (2023) Immersive virtual reality in AECO/FM to enhance education: Systematic literature review and future directions. Comput Appl Eng Educ 31:1465–1491. https://doi.org/10.1002/cae.22631

Tseng KC, Giau DTN (2022) A feasibility study of using virtual reality as a pre-occupancy evaluation tool for the elderly. Autom Constr 134:104037. https://doi.org/https://doi.org/10.1016/j.autcon.2021.104037

Tucker A, Marsh KL, Gifford T, et al (2018) The effects of information and hazard on evacuee behavior in virtual reality. Fire Saf J 99:1–11. https://doi.org/https://doi.org/10.1016/j.firesaf.2018.04.011

Appendix - Feedback questionnaire

Questions	Answers
Gender	male / female
Age	(free text)
Educational level	primary school/ secondary school/ high school/ bachelor-msc degree/ post graduate specialization
Previous experience in training	never / once / twice / more than twice / unsure
Previous experience in earthquake training	never / once / twice / more than twice / unsure
Previous experience in heat wave training	never / once / twice / more than twice / unsure
Frequencies of playing videogames	never / less than once a year / at least once a year/ at least once a month / at least once a wee / several days a week / everyday
Experience with VR	no / yes / unsure

Questions and answers of the first section – PARTECIPANTS

Questions/Answers	Scores	Assessment
Where would you feel less exposed to high temperatures in an open space? (Free text)	3 points for knowing 3 out of 3 among the following items or similar: (i) area in shadow; (i) water sources; (iii) trees; 2 points for 2 out of 3 items; 1 point for 1 out of 3 items; 0 point for knowing nothing	Sum of scores
What would you avoid doing during an earthquake in an open space? (Free text)	4 points for knowing 4 out of 4 among the following items or similar: (i) stay close to buildings; (ii) stay close to glazed elements; (iii) stay close to electric devices; (iv) use vehicles; 3 points for 3 out of 4 items; 2 points for 2 out of 4 items 1 point for 1 out of 4 items; 0 point for knowing nothing	
Where would you go after an earthquake in an open space? (Free text)	3 points for knowing 3 out of 3 among the following items or similar: (i) reach the center of the square; (ii) reach pre-set designated areas; (iii) reach areas free of buildings and falling objects 2 points for knowing 2 out of 3 items 1 point for knowing 1 out of 3 items 0 points for knowing nothing	

Open-ended questions and scores of the second section – KNOWLEDGE

Questions/Answers	Scores	Assessment
How many times the wrong answer was selected during the HW training?	3 points for no wrong answers out of three; 2 points for one wrong answer; 1 point for two wrong answers; 0 points for three wrong answers	Sum of scores
How many time the wrong answer was selected during the E training?	4 points for no wrong answers out of four; 3 points for one wrong answer; 2 points for two wrong answers; 1 point for three	

	wrong answers; 0 points for four (or more) wrong answers	
How many time the wrong answer was selected during the PE training?	4 points for no wrong answers out of four; 3 points for one wrong answer; 2 points for two wrong answers; 1 point for three wrong answers; 0 points for four (or more) wrong answers	

Log file based scores of the second section - KNOWLEDGE

Questions	Scores	Assessment
Which hotspots of the virtual tour correspond to protective positions/items for heat waves?	1 point for each right item (A,B,C,E,H,I) up to 6 points	Sum of scores
Which hotspots of the virtual tour correspond to dangerous positions/items during an earthquake?	1 point for each right item (B,C,E,H,I,L,M) up to 7 points	
Which hotspots of the virtual tour correspond to safe area positions after an earthquake?	1 point for each right item (A,D,F,G) up to 4 points	

Closed-ended questions and scores of the second section – KNOWLEDGE

Topic	Questions	Scores	Assessment
Self-reported engagement	The training experience was fun and enjoyable (Shiradkar et al. 2021)	1 (strongly disagreed) – 7 (strongly agreed)	Mean value and standard deviation of the Likert scale scores
	Safety training activities are boring (Lovreglio et al. 2022)		
	I would describe safety training as very interesting (Lovreglio et al. 2022)		
	Safety training does not hold my attention at all (Lovreglio et al. 2022)		
	It was easy for me to concentrate on my learning (Feng et al. 2022a)		
Perceived usefulness	Using this type of virtual reality simulation as an educational tool will enhance my learning (Rahouti et al. 2021)	1 (strongly disagreed) – 7 (strongly agreed)	Mean value and standard deviation of the Likert scale scores
	This type of simulation is useful as a learning supplement (Davis 1989)		
	This type of simulation is as useful as simulation of indoor spaces		
	This type of simulation is useful for behaving properly in real case, too		
	The combination of two risks (heat wave + earthquake) is effective because it simulates real conditions		
	The simulation of the crowd helped me in taking the right decisions		
Perceived ease of use	This simulation tool is rigid and inflexible to interact with (Davis 1989)	1 (strongly disagreed) – 7 (strongly agreed)	Mean value and standard deviation of the Likert scale scores
	I think this training tool is easy to use (Davis 1989)(Rahouti et al. 2021)		
Recommendation simplicity and efficacy	I could easily remember the recommendations provided in the virtual experience (Chittaro and Sioni 2015)(Lovreglio et al. 2021)(Rahouti et al. 2021)	1 (strongly disagreed) – 7 (strongly agreed)	Mean value and standard deviation of the Likert scale scores
	The recommendations provided in the training experience are useful for my safety (Chittaro and Sioni 2015)(Lovreglio et al. 2021)(Rahouti et al. 2021)		
Realism	The built environment was realistic (Feng et al. 2022a)		

	The built environment reminded me of a familiar place	1 (strongly disagreed) – 7 (strongly agreed)	Mean value and standard deviation of the Likert scale scores
	The VR experience was realistic (Feng et al. 2022a)		
	The realism of the virtual world motivates me to learn (Rahouti et al. 2021)		
	The virtual world makes learning more interesting (Dalgarno et al. 2002)		

Likert scale questions and scores of the third section - TOOL