Multimodal Technologies and Interaction

MDPI

Article

Immersive Virtual Reality for Occupational Track Safety Training: A Quasi-Experimental Field Study Comparing Six Training Formats

Oliver Christ [1,*], Aaron Ettlin [2], Pascal Valentin Meier [1], Pascal Duchêne [3], Paul Hügli [3], Stefan Adam [3] and Stephan Gut [3]

[1] School of Applied Psychology, University of Applied Sciences and Arts Northwestern Switzerland (FHNW), 4600 Olten, Switzerland; pascal.meier@fhnw.ch
[2] Maxon Motor AG, 6072 Sachseln, Switzerland; aaron.ettlin@maxongroup.com
[3] Swiss Federal Railways SBB, 3014 Bern, Switzerland; pascal.duchene@sbb.ch (P.D.); paul.huegli2@sbb.ch (P.H.); stefan.adam@sbb.ch (S.A.); stephan.gut2@sbb.ch (S.G.)
* Correspondence: oliver.christ@fhnw.ch

Abstract

Conventional occupational safety training is often criticized for weak transfer of declarative knowledge into workplace behavior. Immersive virtual reality (iVR) has been proposed as a promising complement, yet comparisons of different iVR formats remain scarce in real occupational settings. This quasi-experimental field study (N = 127) compared six training formats—training-as-usual (control), desktop single-player, iVR single-player, iVR multi-player, iVR multi-player with retrospective perspective change (RPC), and iVR classroom—each embedded as an add-on to the Swiss Federal Railways (SBB) track safety course (SstA). The primary outcome was achieving the full score on a safety-critical checklist exam item; secondary outcomes were technology acceptance (TAM: perceived usefulness, ease of use, behavioral intention) and five practice-oriented items. Binomial logistic regression showed a significant overall model ($\chi^2(5) = 15.80$, $p = 0.007$): participants in all five digital conditions had significantly higher odds of achieving the full score than participants in training-as-usual (OR = 5.13–14.62). Kruskal–Wallis tests revealed significant group differences on all three TAM scales and three practice-oriented items, with the most favorable descriptive pattern observed for iVR classroom. The observed pattern suggests that instructional configuration may play an important role alongside technological immersion in occupational iVR training. However, given the quasi-experimental design and the presence of residual confounding, these findings should be interpreted as associative rather than causal.

Keywords: immersive virtual reality; occupational safety training; technology acceptance model (TAM); field experiment; collaborative learning; railway safety; head-mounted display; training transfer

Check for updates

Academic Editor: Arun K. Kulshreshth

Received: 15 July 2026
Revised: 19 August 2026
Accepted: 22 August 2026
Published: 28 August 2026

1. Introduction

Immersive virtual reality (iVR) is increasingly proposed as a way to close occupational safety training's persistent transfer problem—the well-documented gap between passing a course and reliably showing the trained behavior at the workplace [1–3]. Yet the existing evidence base for this claim rests almost entirely on comparisons between iVR and conventional, non-immersive training (e.g., [4,5]); it says comparatively little about how iVR training should be instructionally configured once the decision to use it has been made. Whether trainees learn individually or collaboratively and with or without direct instructor facilitation inside the virtual environment are design choices that occupational iVR research has rarely varied systematically within one authentic training pipeline—even though these choices may matter more for learning outcomes than the mere presence of immersive technology.

Conventional classroom instruction struggles to let trainees experience the relevance of safety content directly, particularly for situations rehearsed infrequently in daily work; without this felt relevance, training risks becoming a formal obligation whose content is quickly forgotten and decoupled from everyday practice [2,3]. One consequence is a distorted perception of risk and a weaker sense that safety outcomes can be influenced through one's own actions [2], with accurate risk perception being widely regarded as a precondition for safety motivation and appropriate safety behavior [6–9]. Because iVR affords realistic, situated rehearsal of hazardous scenarios without exposing trainees to actual risk [10], it is theoretically well positioned to strengthen exactly this link between training and workplace behavior. Risk perception itself was not measured in the present study, however, and is introduced here only as background motivation; the present study instead examines a direct behavioral proxy for this mechanism—performance on a safety-critical exam item requiring the same hazard judgment that risk perception is thought to support.

1.1. Related Work: Facilitating Factors and Immersive Virtual Reality

Beyond the medium itself, a training's benefit depends on learner factors such as self-efficacy and motivation [11,12] and on didactic design features such as interactivity, experience-based learning, and social interaction [2]. Simulation-based learning capitalizes on these features: its interactivity raises engagement [13], and iVR affords a level of spatial and sensory realism that further supports this engagement [14,15]. Reviews accordingly report that iVR is well suited to safety and workplace training because it enables risk-free, situated rehearsal of hazardous scenarios [4,5,16]. This benefit is conditional rather than automatic, however: perceived usefulness depends strongly on how well the iVR design fits the learning content, and poorly facilitated or overloaded designs can be counterproductive [15,17], which is precisely why the present study focuses on instructional configuration rather than on iVR technology as such. Unless stated otherwise, iVR in this study refers to fully immersive, head-mounted display-based VR with dedicated hand controllers, as distinct from non- to low-immersive desktop VR [18,19]; hardware specifications are reported in the Methods Section.

1.2. The Cognitive–Affective Model of Immersive Learning (CAMIL)

A generally accepted model explaining learning with immersive media has long been missing [20,21]. The Cognitive–Affective Model of Immersive Learning (CAMIL; [15]) fills this gap and serves as the conceptual backbone of the present study. The CAMIL proposes that immersion and interactivity elicit presence and agency, which in turn drive cognitive–affective mechanisms—interest, motivation, self-efficacy, embodiment, cognitive load, and self-regulation—that ultimately determine learning gains. Critically, presence is multidimensional: physical, social, and self-presence [22] can, in principle, arise from different sources within a training format, while agency describes the felt sense of being able to control one's own actions in the virtual environment. This distinction motivates our central theoretical expectation for the present comparison of six training formats: configurations that reduce individual agency and physical immersion (e.g., a shared classroom format observed via projection) may nonetheless sustain learning gains if they compensate with higher social presence through instructor facilitation and peer observation, whereas configurations that maximize individual agency but minimize social contact (e.g., single-player iVR) rely on a different presence pathway to reach the same cognitive–affective outcomes. Because medium and method must be considered jointly for iVR to be effective, the CAMIL predicts that these six formats should differ in learning outcomes not because they differ in how much VR they contain, but because they differ in which presence and agency pathways they engage.

1.3. Collaborative Learning

Multi-player and instructor-led iVR formats add a further layer to this presence pathway. Constructivist [23] and social–constructivist [24] learning theory, together with Kolb's [25] experiential learning model, predict that learners can accomplish more with the support of an instructor or peers—within their zone of proximal development [24]—than they can individually, provided this support facilitates reflection on the experience rather than merely adding coordination demands [26,27]. Concretely, instructor scaffolding, real-time feedback, and observational learning from peers' successes and errors are mechanisms through which collaborative or instructor-facilitated iVR could increase social presence and thereby sustain or exceed the learning gains of individual iVR, even at lower levels of individual immersion [28,29]. Whether this compensatory mechanism holds in practice, however, likely depends on the collaborative structure itself: purely peer-to-peer coordination may introduce a collaborative load that offsets its own benefit [26,27], whereas a shared, instructor-led format may retain the benefit without the added coordination cost. This distinction between collaboration with and without direct instructor facilitation is one of the specific instructional configurations compared in the present study.

1.4. Technology Acceptance

Whether a technology is merely available does not guarantee that trainees are willing to use it [30], and technology acceptance is therefore not simply an additional questionnaire outcome but a theoretically relevant one: within the CAMIL, the same presence and agency pathways that drive learning gains are also expected to shape perceived usefulness and enjoyment [15], and technology acceptance is therefore relevant to whether trainees intend to continue using a new training technology [31]. The Technology Acceptance Model (TAM; [31]), grounded in the Theory of Reasoned Action [32] and the Theory of Planned Behavior [33], operationalizes this readiness via perceived usefulness (PU) and perceived ease of use (PEOU), which predict behavioral intention to use (BI). TAM has already been applied to iVR training acceptance in vocational [34], medical [35], and academic contexts [36,37], with mixed results. Özgen et al. [36], for instance, found higher enjoyment and intention for iVR than paper-based training, but no difference in perceived usefulness or ease of use—underscoring that acceptance, like learning, may depend on instructional configuration rather than on iVR as such.

1.5. Research Gap and Hypotheses

Taken together, previous research suggests that immersive VR can enhance occupational safety training compared with conventional approaches and provides theoretical reasons, derived from the CAMIL and collaborative learning theory, to expect that different instructional configurations may be associated with different learning and technology acceptance outcomes. However, most occupational iVR studies have compared a single VR format with conventional training, leaving limited evidence on how different instructional configurations perform within the same authentic training context. The present quasi-experimental field study addresses this gap by comparing six training formats embedded within a single occupational railway safety course. Because this study was conducted under real-world field conditions without random assignment, it is intended to identify practically relevant patterns of association rather than to isolate the causal contribution of individual instructional or technological components.

We compared the six training formats in the context of Swiss railway track safety training—one instance of a broader class of high-risk occupational domains, alongside, e.g., construction or aviation ground operations, in which safety-critical judgment tasks are trained under similarly constrained conditions. Personnel entering the Swiss railway track area under self-protection must complete the two-day "Selbstschutz Arbeit" (SstA) course, which concludes with a theoretical exam; a specific, heavily weighted exam item on the safety checklist and approach distance calculation is a recurrent source of exam failure, giving SBB a concrete practical incentive—alongside the scientific motivation above—to evaluate whether digitally enhanced training add-ons could improve performance on this item. Following the identification of relevant safety elements and scenario design and the development of an interaction tutorial [38], the present study evaluated the complete set of training formats within real SstA courses. Because SBB's own priority was the practical, work-relevant usefulness of training rather than fine-grained theoretical constructs, five additional practice-oriented items on situational awareness and hazard perception were developed together with SBB safety experts to complement the TAM scales.

Building on the CAMIL and collaborative learning theory, the present quasi-experimental field study tested four hypotheses.

Hypothesis H1. *Participants in each digitally enhanced condition will have higher odds of achieving the full score on the safety-critical examination item than participants in training-as-usual [2,15].*

Hypothesis H2. *Examination outcomes in the instructor-facilitated iVR classroom condition will be descriptively similar to those in the iVR single-player condition.*

Hypothesis H3. *iVR conditions will show more favorable technology acceptance ratings (PU, PEOU, BI) than training-as-usual and desktop training.*

Hypothesis H4. *Collaborative and instructor-facilitated iVR conditions will show more favorable practice-oriented ratings than individual formats [24,25].*

2. Materials and Methods

2.1. Design

This study (see Table 1 for group statistics) followed a quasi-experimental, non-randomized field design with a control group, embedded as an add-on within the regular two-day SstA course, and compared participants' behavior across six between-subjects conditions (Table 2; [39]). A field design was chosen over a laboratory study because the research question concerned training effectiveness within the authentic SstA curriculum, trainers, and trainee population actually used by SBB; a laboratory analog would have sacrificed exactly the ecological validity—real trainers, real course pressure, real trainee motivation—that the practical and scientific questions required [40] at the cost of the tighter experimental control a laboratory setting affords [41]. Several features of the design were pragmatic constraints rather than deliberate methodological choices: because participation was voluntary and uncompensated, and because SBB required the regular course content to remain unaltered, randomization of individual trainees to conditions was not feasible. Condition was instead assigned at the level of the course run rather than the individual trainee: all trainees of a given SstA course run who opted to take part were assigned to the same condition, because running multiple conditions in parallel within one course was logistically impossible given room and equipment availability. Which condition a given course run received was pre-planned and only adjusted, when necessary, to keep the running tally of participants roughly balanced across the six conditions; this quota-based balancing targeted sample size only and did not take participant demographics into account (see Section 2.3). Because trainees could not be randomly allocated and rooms, trainers, and course dates varied across conditions, the design affords high external validity but a correspondingly reduced ability to rule out confounding variables such as differing rooms or trainers, which limits the causal interpretability of the results [40]; we return to this trade-off explicitly in Section 2.9 (Threats to Validity). Reporting follows, where applicable to a quasi-experimental field study, the spirit of the TREND recommendations for non-randomized behavioral and public health intervention evaluations [42].

2.2. Recruitment and Participant Allocation

Prospective participants were invited by the course trainer via e-mail approximately one week before the course began. Participation in the study was voluntary and uncompensated. On the first day of the course, the trainer recorded the number of trainees who agreed to participate and reported this number to the research team. Condition allocation occurred at the level of the entire course run rather than at the individual participant level, as described in Section 2.1. All participating trainees within a given course run therefore received the same training condition. The condition for each course run was selected with consideration of the number of participants already recruited to each condition and the technical capacity of the respective training format. No individual-level randomization or demographic balancing (e.g., by age, gender, or prior VR experience) was performed. Figure 1 summarizes the recruitment and course run-level allocation procedure. Sample characteristics by condition are reported in Section 2.3 and Table 1.

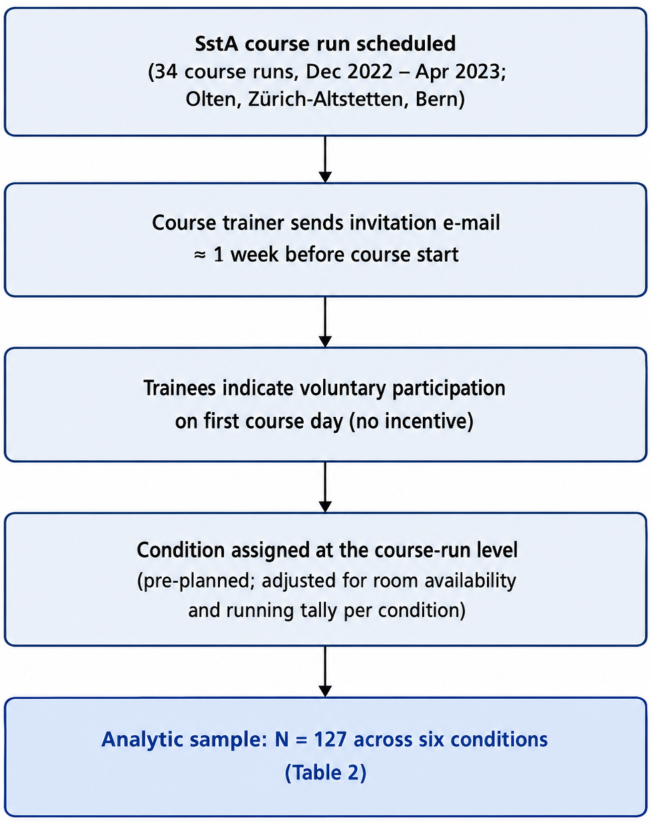

Figure 1. Recruitment and course run-level allocation procedure.

Table 1. Sample composition by condition.

Condition	N	M Age/Range	Prior VR Experience	Gender	Team (N)	Location (N)
Training-as-usual	23	19.43 years 15–41 years	13	19 male, 3 diverse, 1 not stated	Ost (23)	Zürich (23)
iVR single-player	20	37.45 years 16–53 years	10	1 female, 19 male	Mitte (13), Ost (7)	Olten (13), Zürich (7)
Desktop single-player	20	39.40 years 23–60 years	7	1 female, 18 male, 1 not stated	Mitte (20)	Bern (2), Olten (18)
iVR multi-player	23	35.04 years 21–60 years	5	4 female, 18 male, 1 not stated	Mitte (4), Ost (19)	Olten (4), Zürich (19)
iVR multi-player with RPC	21	38.33 years 20–62 years	8	1 female, 20 male	Mitte (9), Ost (12)	Olten (9), Zürich (12)
iVR classroom	20	36.45 years 23–59 years	6	1 female, 19 male	Mitte (15), Ost (5)	Olten (15), Zürich (5)
All groups	127	34.05 years 15–62 years	49 (38%)	8 female, 113 male, 3 diverse, 3 not stated	Mitte (61), Ost (66)	Bern (2), Olten (59), Zürich (66)

Note. N = sample size; M = mean age in years; Prior VR = number of participants with prior VR experience; iVR = immersive virtual reality; RPC = retrospective perspective change.

2.3. Sample

The required sample size was calculated a priori using G*Power 3.1.9.7 [43]. For the planned statistical procedures with six groups, a power of 0.95, and an estimated medium effect size of $f^2(V) = 0.15$ [44], a minimum sample of 90 participants was required. Data collection was conducted across 34 course runs and yielded a total analytic sample of N = 127 voluntary participants distributed across the six training conditions. Sample composition by condition is presented in Table 1.

Table 1 shows that the six conditions were reasonably balanced in sample size (n = 20–23) but not in demographic composition: mean age in training-as-usual (19.4 years) was descriptively far lower than in every experimental condition (35.0–39.4 years), reflecting that the control-condition course runs happened, by chance of the recruitment schedule, to draw more apprentices from the “Ost” team. Gender was heavily skewed towards male participants in all conditions (113 of 127 overall), and prior VR experience ranged descriptively from 5 to 13 participants per condition. Although participant-level demographic data were available, the present study was not designed or powered to evaluate demographic effects. Consequently, demographic variables were not included as covariates in the primary analyses, and the findings should be interpreted within the context of the quasi-experimental study design. We treat the resulting age imbalance, in particular, as a substantive threat to internal validity that is addressed explicitly in Section 2.9 (Threats to Validity) and revisited in Section 4.8.1 (Design Limitations).

2.4. Training Conditions

The six conditions are best understood along two largely independent dimensions (Figure 2): a technological dimension (non-immersive desktop vs. immersive HMD-based iVR) and an instructional dimension (individual, self-guided use vs. collaborative, instructor-facilitated use). Training-as-usual involved neither desktop nor iVR technology;

iVR single-player and desktop single-player differ only in technology while holding the individual, self-guided instructional format constant; and iVR multi-player, iVR multi-player with RPC, and iVR classroom all use identical HMD hardware while differing in exactly how collaboration and instructor facilitation are structured (Table 2). This conceptual distinction facilitates the comparison of technological and instructional characteristics across conditions. However, because several instructional and technological features varied simultaneously, the present design does not permit the independent causal attribution of observed differences to any single design component.

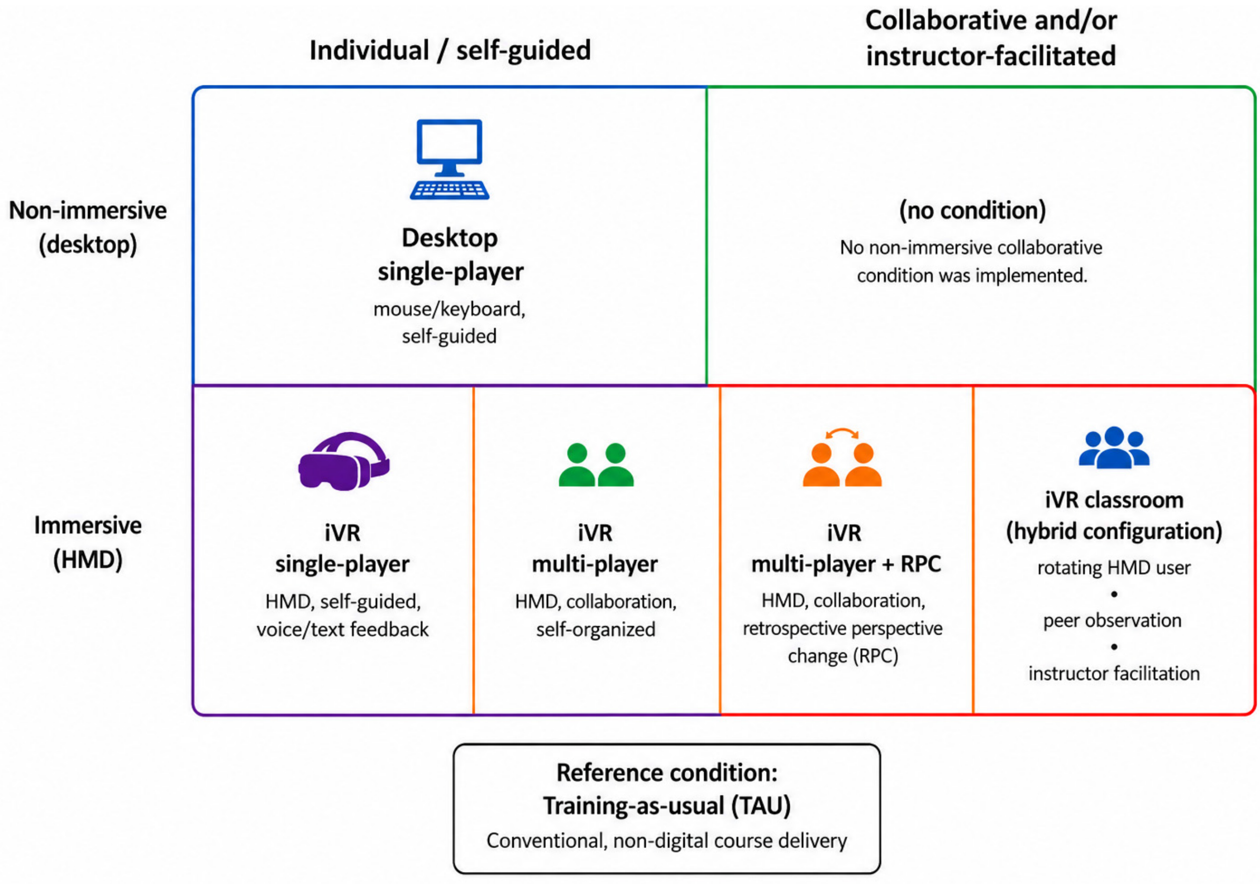

Figure 2. The instructional and technological dimensions of the digitally enhanced training conditions. The horizontal dimension distinguishes individual/self-guided from collaborative/instructor-facilitated training, whereas the vertical dimension distinguishes non-immersive desktop from immersive HMD-based training. Training-as-usual is shown separately as the conventional non-digital reference condition. Technical capacity differed between conditions. Depending on the available hardware, up to eight participants per course run could participate in the iVR single-player condition and up to six in the desktop single-player condition, whereas all trainees in a course run (maximum 12) could participate in training-as-usual.

Table 2. The six training conditions.

Condition	Group	Description
Training-as-usual (TAU)	Control	The SstA course took place as usual. Participants watched instructional films and worked with a wooden track model. A questionnaire was administered at the end of the day.
iVR single-player	Experimental	In addition to the regular course, participants individually experienced four weather conditions and train passages via HMD, using a T-wrench to tighten flashing screws, supported by voice/text feedback.

Table 2. *Cont.*

Condition	Group	Description
Desktop single-player	Experimental	As iVR single-player, but delivered via PC monitor instead of an HMD.
iVR multi-player	Experimental	Participants experienced the virtual world together with the trainer via HMD; the trainer assigned roles, triggered train passages, and changed the weather.
iVR multi-player with retrospective perspective change (RPC)	Experimental	As iVR multi-player, but at the end of each weather condition participants additionally relived their own trackside action from 30 s earlier from the train driver's viewpoint/cab view.
iVR classroom	Experimental	Participants experienced the virtual world in pairs via an HMD while the trainer assigned roles, triggered train passages and changed the weather from the real world; the remaining course observed the scene on a projection screen; After the pair was done, they went back to the observer role and the next two participants became immersed with the VR goggles until every participant was the observer and iVR-user

2.5. Materials

For the immersive conditions (iVR single-player, iVR multi-player, iVR multi-player with RPC, iVR classroom), a Meta Quest 2 HMD in standalone mode was used. Participants received one controller per hand, allowing them to teleport and to grip and use a virtual T-wrench, and built-in spatial audio conveyed sound, which was particularly important during the tutorial. Elite straps and lens spacers were used to increase wearing comfort, and a local wireless router maintained the network connection required by the HMDs throughout data collection. The desktop single-player condition used the same virtual learning environment delivered via a workstation, mouse, and headphones instead of an HMD, isolating the technological (immersive vs. non-immersive) contrast from the instructional format (Figure 2). The virtual learning environment (see Figure 3) itself included a virtual agent ("Emma") who guided participants through an introductory tutorial [38] before the main scenarios (the full training sequence is described in Section 2.7). Table 3 summarizes the principal design elements that were held constant across conditions and those that were intentionally varied as part of the comparative design.

Table 3. Comparison of experimental design elements across the six training conditions.

Design Feature	Held Constant Across Conditions	Intentionally Varied Between Conditions
Learning Scenario	✔	
Learning objectives	✔	
Safety task/examination	✔	
Tutorial	✔	
Scenario duration	✔	
Hardware (Desktop vs. HMD)		✔
Collaboration structure		✔
Instructor facilitation		✔
Perspective change (RPC)		✔
Feedback source (virtual agent vs. instructor)		✔

2.6. Measures

The study included one primary performance outcome and two groups of secondary self-report outcomes: (1) technology acceptance measures and (2) five practice-oriented items. The primary outcome was examination performance, operationalized as whether a participant achieved the full score on the safety-critical checklist/approach distance item of the standard SstA examination (vs. a score below 4). Scores were obtained directly from participants' completed examination checklists rather than from the follow-up questionnaire. This item was selected a priori because it captures the specific hazard judgment targeted by the training scenarios—evacuation timing and minimum visibility—and could be scored consistently across all 34 course runs without relying on self-report. The outcome should therefore be interpreted as performance on one predefined safety-critical examination task rather than as a comprehensive measure of occupational competence, training transfer, or subsequent workplace safety behavior. Secondary outcomes were assessed using a follow-up questionnaire created with the Enterprise Feedback Suite [45] and completed via QR code after the assigned training condition/study session. The full project questionnaire comprised 2 general questions, 4 demographic questions, and 49 items derived from the TAM and CAMIL frameworks together with the SBB-developed practice-oriented items. For the present study, only the TAM measures and the five practice-oriented items were analyzed and are therefore reported here; the remaining CAMIL-derived questionnaire items formed part of the broader project evaluation but were not analyzed for the present research questions. Technology acceptance was assessed using the iVR-adapted TAM measures reported by Özgen et al. [36], covering perceived usefulness (PU), perceived ease of use (PEOU), and behavioral intention (BI). Five additional practice-oriented items were developed in collaboration with SBB safety experts to assess participants' perceived preparedness for practical work, perceived safety preparedness, hazard awareness, and situational awareness. Their full wording is provided in Table 4. All five items were rated on a 5-point scale ranging from 1 ("strongly disagree") to 5 ("strongly agree"). In the present sample, the five items showed good internal consistency when combined (Cronbach's $\alpha = 0.88$). Given the available sample (N = 127), the project-specific status of these items, and the absence of an independently validated measurement model for this five-item set, no confirmatory factor analysis was undertaken. Although classical methodological guidance has treated substantially larger samples, including samples around 300, as preferable for stable factor-analytic solutions [46], more recent simulation research shows that CFA sample size requirements are model-dependent rather than governed by a universal numerical threshold [47]. Against this background, we considered N = 127 insufficient for a robust confirmatory validation of this project-specific five-item measurement model; scale scores were therefore calculated as unweighted item means rather than factor scores.

Table 4. Practice-oriented items developed for the present study in collaboration with SBB safety experts.

Practice-Oriented Items
Compared to training without the learning tool, I feel better prepared for practical work
Compared to training without the learning tool, I feel safer prepared for practical work
The hazards in the track area were made very clear to me by the learning tool
The learning tool conveyed the hazards in the track area vividly
The experience with the learning tool positively influenced my situational awareness

Each item was rated on a 5-point scale from 1 ("strongly disagree") to 5 ("strongly agree"); combined as a single scale, these five items showed good internal consistency computed on the present sample ($\alpha = 0.88$).

2.7. Procedure

Data collection took place on the first course day after completion of the regular SstA course (approximately 4:30 p.m.) so as not to alter the standard curriculum. The complete study session, including informed consent and questionnaire completion, was limited to a maximum of one hour; the target completion time for the questionnaire was ten minutes. Participants were explicitly informed that study participation was unrelated to passing or failing the SstA examination. At least two FHNW facilitators were present at every session to support participants and the course trainer and to address technical issues with the HMDs or PC equipment. Trainers followed a structured facilitation script ("Drehbuch") developed by the project team for each condition. The script specified the sequence of instructions, timing, and feedback to be provided and was intended to support consistent facilitation rather than prescribe a verbatim word-for-word delivery. Some variation in delivery across trainers therefore cannot be ruled out. Prior to data collection, an initial difference in HMD and script familiarity was identified between the "Ost" (Zürich) and "Mitte" (Olten, Bern) trainer teams. A dedicated train-the-trainer day was therefore conducted to familiarize all trainers with the HMDs and scripts, followed by a second refresher day. Both activities were intended to improve consistency across trainers and sites; the remaining risk of trainer- and site-related differences is discussed in Section 2.9 (Threats to Validity). For the digitally enhanced conditions using the virtual environment, the training sequence can be described in five consecutive phases: (1) interaction tutorial, (2) sensitization scenario, (3) safety checklist exercise, (4) four experimental training scenarios with condition-specific feedback, and (5) session ID assignment and post-training questionnaire completion. First, participants completed an approximately seven-minute tutorial individually, guided by the virtual agent Emma. The tutorial introduced locomotion, grasping and releasing the T-wrench, and finally using the tool to tighten three loose screws [38]. Second, participants entered a shared virtual environment (see Figure 3) for a sensitization phase. The trainer instructed them to stand between two tracks, after which a train passed at 160 km/h under poor visibility conditions. This scenario was intended to sensitize participants to sudden changes in visibility. Third, participants removed the HMD and completed the safety checklist under the trainer's guidance for a projected task. This exercise was used to calculate the safety-relevant evacuation time and minimum visibility distance applied in the subsequent scenarios. Fourth, participants re-entered the virtual environment and completed four scenarios presented in varying order: (a) rain with 250 m visibility and a passing train, (b) rain with 320 m visibility and a passing train, (c) fog with 250 m visibility and a passing train, and (d) fog with 320 m visibility and a passing train. Their task was to recognize independently, based on visibility or an approaching train, when to leave the track and retreat to the place of refuge while respecting the required 8 s evacuation time and the resulting minimum visibility of 320 m calculated during the checklist exercise. In the single-player scenarios, participants were self-guided via voice/text feedback, whereas the multi-player scenarios were guided by the trainer. Depending on participants' behavior, feedback was provided by the virtual agent in the single-player/desktop formats or by the trainer in the multi-player conditions. Five feedback messages were available: two negative-visibility variants, a negative refuge-without-cause variant, a negative too-slow variant, and a positive ("all correct") message. Finally, after completing all four scenarios, participants received an anonymous session ID within the VR environment to link their training data with the post-training questionnaire, which was completed via QR code as described in Section 2.6.

Figure 3. Examples of the virtual railway environment and training scenarios. (**Top**): Examples of the visibility conditions represented in the virtual environment, including clear weather and fog or rain at visibility distances of 320 m and 250 m. (**Bottom left**): Example of a track work scenario with an approaching train. (**Bottom right**): Cab view used in the iVR multi-player + RPC condition.

2.8. Statistical Analysis

The statistical analyses were organized according to the four study hypotheses. H1 was evaluated using binomial logistic regression comparing each digitally enhanced condition with training-as-usual, supplemented by Firth bias-reduced regression as a robustness analysis. H2 concerned the descriptive similarity of examination outcomes between iVR classroom and iVR single-player; no formal equivalence or non-inferiority test was performed. H3 was evaluated using Kruskal–Wallis tests of the TAM scales (PU, PEOU, and BI), followed by Dunn–Bonferroni pairwise comparisons. H4 was examined using the same non-parametric procedure for the five practice-oriented items, with these findings interpreted as exploratory because the items were developed specifically for the present project. Data exported from the questionnaire tool [45] were merged in Excel [48] with participants' exam scores; the analytic sample of N = 127 had no missing values on the primary exam score variable, as this was scored directly from every participant's completed paper checklist rather than relying on questionnaire response. For the secondary questionnaire outcomes, perceived usefulness had n = 127 observations, whereas perceived ease of use, behavioral intention, and the five practice-oriented items had n = 126 because one participant in the iVR multi-player condition had the relevant responses coded as zero and these values were treated as missing. The corresponding analyses used the available cases for each outcome. H1 was analyzed with a binomial logistic regression with a categorical predictor (condition, dummy-coded, with training-as-usual as reference category) predicting whether the full score on the checklist/approach distance exam item was achieved (vs. a score below 4). Each participant contributed one observation to one condition; however, observations were clustered within course runs and shared trainers/sites, so strict statistical independence cannot be assumed. This dependence could not be modeled reliably with the available cluster structure and is treated as an inferential limitation in Sections 2.9 and 4.8.1. Assumptions specific to continuous predictors, such as linearity of the logit, were not applicable because the model contained only the categorical condition predictor. Model fit (−2 log-likelihood, Cox & Snell R^2, Nagelkerke R^2), odds ratios with Wald-based 95% confidence intervals (relative to the control group), and the descriptive percentage of full vs. non-full scores per condition are reported in the Results Section. Because the five dummy-coded condition contrasts test a single a priori comparison structure (each experimental condition vs. the same training-as-usual reference) within one omnibus model rather than five independent pairwise tests, no additional multiple-testing correction was applied to these contrasts beyond the omnibus likelihood-ratio test of the full model; this differs from the H3/H4 pairwise KWT contrasts, where Bonferroni correction was applied because all 15 possible pairwise comparisons per scale were examined. H3 and H4 were analyzed with non-parametric Kruskal–Wallis tests (KWT), given visibly non-normal, ceiling-skewed distributions typical of 5- and 7-point Likert-type scale scores, followed by pairwise comparisons with Bonferroni correction; only overall significant KWTs are reported, with pairwise comparisons that remained significant after Bonferroni correction marked in the corresponding figures. All statistics reported in this manuscript—the binomial logistic regression (Table 5), the Firth [49] bias-reduced re-analysis (Table 6), and the Kruskal–Wallis tests with Dunn–Bonferroni pairwise comparisons (Technology Acceptance and Practice-Oriented Items Results)—were computed by the authors in Python 3.11 (statsmodels 0.14.6, pyreadstat 1.3.5, scipy 1.17.1, scikit-posthocs 0.14.0); figures were produced programmatically using matplotlib. We did not employ mixed-effects or multilevel models to account for the nesting of participants within 34 course runs, eight trainers, and three locations, because cluster sizes at the trainer and course run level (typically 1–4 participants per condition per run) were too small to support stable estimation of random effects or intraclass correlations; the resulting clustering and trainer/site confounding are addressed explicitly in Section 2.9 (Threats to Validity) and 4.8.1 (Design Limitations). For the same reason, and because condition assignment occurred at the course run rather than the individual level (see Section 2.1), covariate adjustment (e.g., for age) was not incorporated into the primary confirmatory model reported here; the plausibility of age as an alternative explanation is addressed explicitly in Sections 2.9 and 4.8.1.

Table 5. Binomial logistic regression coefficients and 95% confidence intervals by condition (reference category: training-as-usual).

Condition	B	SE	Wald	df	Sig.	OR (Exp[B])	95% CI
iVR classroom	2.682	1.109	5.851	1	0.016	14.62	[1.66, 128.46]
iVR multi-player with RPC	1.989	0.854	5.422	1	0.020	7.31	[1.37, 38.97]
iVR multi-player	1.635	0.749	4.770	1	0.029	5.13	[1.18, 22.27]
iVR single-player	2.682	1.109	5.851	1	0.016	14.62	[1.66, 128.46]

Table 5. *Cont.*

Condition	B	SE	Wald	df	Sig.	OR (Exp[B])	95% CI
Desktop PC	1.935	0.856	5.111	1	0.024	6.92	[1.29, 37.07]

Note. B = unstandardized logistic regression coefficient; SE = standard error; OR = odds ratio; CI = 95% confidence interval for OR, calculated using the Wald method (exp[B ± 1.96 × SE]). All values in this table were computed by the authors directly from the raw participant-level data using Python 3.11 (statsmodels 0.14.6); The identical coefficients for iVR classroom and iVR single-player are not a data-entry error: both conditions had exactly the same cell counts (19 of 20 participants achieving the full exam score), which mathematically yields identical log-odds, standard errors, and odds ratios relative to the same reference category in a single-categorical predictor logistic regression. The wide confidence intervals reflect the comparatively small per-condition sample sizes (n = 20–23) and should be considered when interpreting the precision of these estimates. As a robustness check for this small-sample bias, a Firth [49] bias-reduced re-analysis was additionally fitted (see Section 3.2 and Table 6).

Table 6. Comparison of standard maximum-likelihood (MLE) and Firth bias-reduced odds ratios for full exam score, by condition (reference category: training-as-usual).

Condition	OR (MLE)	95% CI (MLE)	OR (Firth)	95% CI (Firth)
iVR classroom	14.62	[1.66, 128.46]	10.11	[1.53, 66.98]
iVR multi-player with RPC	7.31	[1.37, 38.97]	6.07	[1.25, 29.44]
iVR multi-player	5.13	[1.18, 22.27]	4.56	[1.10, 18.87]
iVR single-player	14.62	[1.66, 128.46]	10.11	[1.53, 66.98]
Desktop PC	6.92	[1.29, 37.07]	5.76	[1.18, 28.07]

Note. MLE = maximum-likelihood estimate (as reported in Table 5); Firth = bias-reduced penalized-likelihood estimate [49], implemented directly following the algorithm of Heinze and Schemper [50] and cross-validated across three optimization methods. The Firth-corrected odds ratios are shrunk toward the null relative to the MLE estimates, and the corresponding confidence intervals are narrower (nearly halved for iVR classroom and iVR single-player, from a width of 126.8 to 65.45 odds-ratio units). In the Firth re-analysis, all five condition-versus-training-as-usual effects remained statistically significant (all $p < 0.04$), and the rank ordering of effect sizes was unchanged. This indicates that the H1 pattern is not solely an artifact of small-cell bias in standard maximum-likelihood estimation. The Firth analysis does not address the non-randomized allocation, clustering, age imbalance, unequal additional training exposure, or the absence of a direct equivalence test for H2; those limitations remain as described in Sections 2.9 and 4.8.1. Because the omnibus model and the condition-versus-training-as-usual coefficients do not provide a direct statistical comparison between iVR classroom and iVR single-player, H2 is evaluated descriptively rather than as statistically confirmed. Descriptively, H2 was supported: iVR classroom and iVR single-player showed identical full-score rates (19/20 participants, 95% in each condition). This descriptive similarity should not be interpreted as evidence of statistical equivalence, as no equivalence or non-inferiority test was performed.

2.9. Threats to Validity

Five interrelated threats to internal validity should be considered when interpreting the findings of this quasi-experimental field study. First, trainer and site effects: Eight different trainers delivered the six conditions across three locations, and condition was not crossed factorially with trainer or site. Consequently, trainer-specific facilitation style and site-specific contextual factors are partially confounded with condition. To minimize these differences, all trainers completed dedicated train-the-trainer sessions and refresher training before data collection, although some residual variation in delivery cannot be excluded.

Second, participation was voluntary and uncompensated. The analytic sample therefore reflects trainees who chose to participate rather than a random sample of the target population. Although this self-selection mechanism was identical across all six conditions, it may have increased general motivation and willingness to engage with the training and questionnaire, particularly for the technology acceptance measures.

Third, condition allocation occurred at the level of the course run rather than the individual participant and was not randomized. As a consequence, participant characteristics, trainer, site, and course run are partly confounded with condition. The most important example is the substantial descriptive age imbalance: participants in the training-as-usual condition were, on average, approximately 16–20 years younger than participants in the digitally enhanced conditions. Because age and condition are confounded in the present dataset, it was not possible to determine analytically to what extent the observed differences between conditions reflect the training formats themselves or pre-existing differences in participant characteristics. Accordingly, comparisons between training-as-usual and the digitally enhanced conditions should be interpreted with particular caution.

Fourth, participants were nested within 34 course runs and shared trainers and sites. This clustered dependence could not be modeled statistically because the number and size of clusters were insufficient to support stable multilevel estimation. Consequently, the participant-level standard errors, confidence intervals, and significance tests may be overly optimistic and should be interpreted cautiously.

Fifth, the digitally enhanced conditions received up to approximately one hour of supplementary study-related training after the regular course, whereas training-as-usual did not include an equivalent-duration active control activity. The comparison with training-as-usual therefore combines differences in instructional format with differences in additional training exposure and situated practice. Observed differences may consequently reflect additional practice time, the digital or immersive characteristics of the interventions, their instructional configuration, or a combination of these factors.

Taken together, these design characteristics limit the strength of the causal inferences that can be drawn from the present study. Alternative explanations related to participant characteristics, trainer effects, site differences, course run clustering, and unequal additional training exposure cannot be excluded. The findings should therefore be interpreted as associations observed under authentic field conditions that warrant confirmation in future randomized or otherwise more tightly controlled studies.

3. Results

This section reports the empirical findings for the four hypotheses specified in Section 1.5: H1 and H2 concern exam performance, and H3 and H4 concern technology acceptance and practice-oriented ratings. Descriptive statistics, inferential test results, and effect sizes are reported first; interpretation of these findings is reserved for the Discussion Section. In total, N = 127 individuals across 34 SstA course runs voluntarily took part in the study.

3.1. Participant Description

Sample composition by condition is summarized in Table 1 (see Section 2.3). Recruitment followed a single-stage flow without a separate randomization or attrition step: all trainees of a given course run who volunteered were assigned to the pre-determined condition for that run and completed the questionnaire once, immediately after training; so, no participant flow diagram in the CONSORT sense (with randomization and follow-up loss) applies to this quasi-experimental single-session design. No missing data were reported for the primary exam performance outcome; sample sizes for the secondary TAM and practice item outcomes are n = 127 for perceived usefulness and n = 126 for perceived ease of use, behavioral intention to use, and all five practice-oriented items (one participant in the iVR multi-player condition had entire practice item sections coded as zero and was recoded as missing). In summary, the achieved sample (N = 127, six conditions, 20–23 participants each) matched the a priori power calculation described in the Methods Section.

https://doi.org/10.3390/mti10090089

3.2. Primary Outcome: Exam Performance (H1, H2)

As specified in Section 2.8, H1 and H2 were evaluated using the examination-performance outcome. Figure 4 reports the absolute and relative frequencies of full versus non-full scores per condition: 13/23 (56.5%) in training-as-usual, 20/23 (87%) in iVR multi-player, 19/21 (90.5%) in iVR multi-player with RPC, 18/20 (90%) in desktop PC, and 19/20 (95%) in both iVR single-player and iVR classroom.

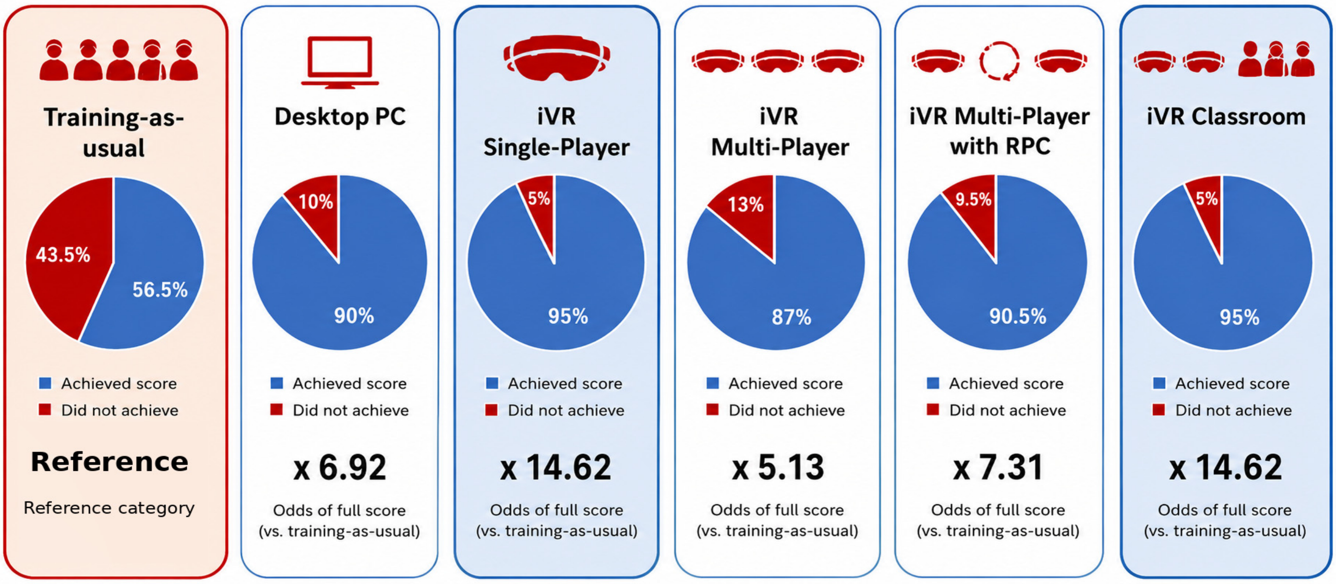

Figure 4. The percentage of participants with a full (blue) vs. non-full (red) score on the checklist/approach distance exam item across conditions.

As specified in Section 2.8, the primary outcome was analyzed using binomial logistic regression with training-as-usual as the reference condition. Because the model contained condition as its only categorical predictor, the model-predicted probabilities correspond to the observed proportions reported at the beginning of Section 3.2.

The overall model was significant, $\chi^2(5) = 15.797$, $p = 0.007$ (-2 Log-Likelihood = 91.397; Cox & Snell $R^2 = 0.117$; Nagelkerke $R^2 = 0.205$). Table 5 reports the regression coefficients, Wald statistics, odds ratios, and 95% confidence intervals (computed from the reported B and SE as $\exp[B \pm 1.96 \times SE]$) for each condition relative to training-as-usual. In practical terms, an odds ratio of 14.62 (iVR classroom, iVR single-player) means that the odds of achieving the full exam score were about 14.6 times higher in that condition than in training-as-usual, corresponding to the increase from 56% to 95% full scores reported above; the smallest, though still significant, increase was observed for iVR multi-player (OR = 5.13, corresponding to 56% vs. 87% full scores). The comparatively wide confidence intervals across conditions primarily reflect the modest per-condition sample sizes (n = 20–23) rather than instability of the observed effects.

In summary, participants in all five digital training conditions showed significantly higher odds of achieving the full score on the safety-critical examination item than participants in the training-as-usual condition. Within the present quasi-experimental field sample, the observed odds ratios ranged from approximately 5 to 15. Because the study was not randomized and condition assignment was partly confounded with participant and training characteristics, these findings should be interpreted as observed associations rather than isolated causal effects of the training formats.

Given the comparatively small per-condition sample sizes underlying Table 5 and the resulting wide confidence intervals—most notably for iVR classroom and iVR single-player (OR = 14.62, 95% CI [1.66, 128.46])—a Firth [49] bias-reduced (penalized-likelihood) logistic regression was fitted as a robustness check (see Table 6), following the iterative score modification algorithm described by Heinze and Schemper [50]. This method corrects for the small-sample bias that arises when cell counts are sparse; in this dataset, only 1 of 20 participants in each of the iVR classroom and iVR single-player conditions did not achieve the full exam score, and the resulting events-per-parameter ratio (19 non-full-score cases across 5 estimated condition effect parameters; EPP = 3.8) falls well below the conventional minimum of 10 recommended by Peduzzi et al. [51]. Firth's method does not alter the model specification; it adds a Jeffreys-invariant prior penalty to the log-likelihood, which shrinks estimates away from extreme values without discarding any data.

3.3. Technology Acceptance Results (H3–TAM Scales)

To evaluate H3, PU, PEOU, and BI were compared across the six training conditions using the non-parametric procedure specified in Section 2.8. Kruskal–Wallis tests (KWTs) showed significant overall differences between conditions for all three scales: perceived usefulness, $\chi^2(5) = 17.115$, $p = 0.004$, $\eta^2\_H = 0.10$; perceived ease of use, $\chi^2(5) = 22.137$, $p < 0.001$, $\eta^2\_H = 0.14$; and behavioral intention to use, $\chi^2(5) = 14.169$, $p = 0.015$, $\eta^2\_H = 0.08$ (effect sizes calculated by the authors as $\eta^2\_H = [H - k + 1]/[N - k]$, following [52], with k = 6 groups and N representing the number of observations included in the respective analysis). Figure 5A–C show the descriptive pattern underlying these tests.

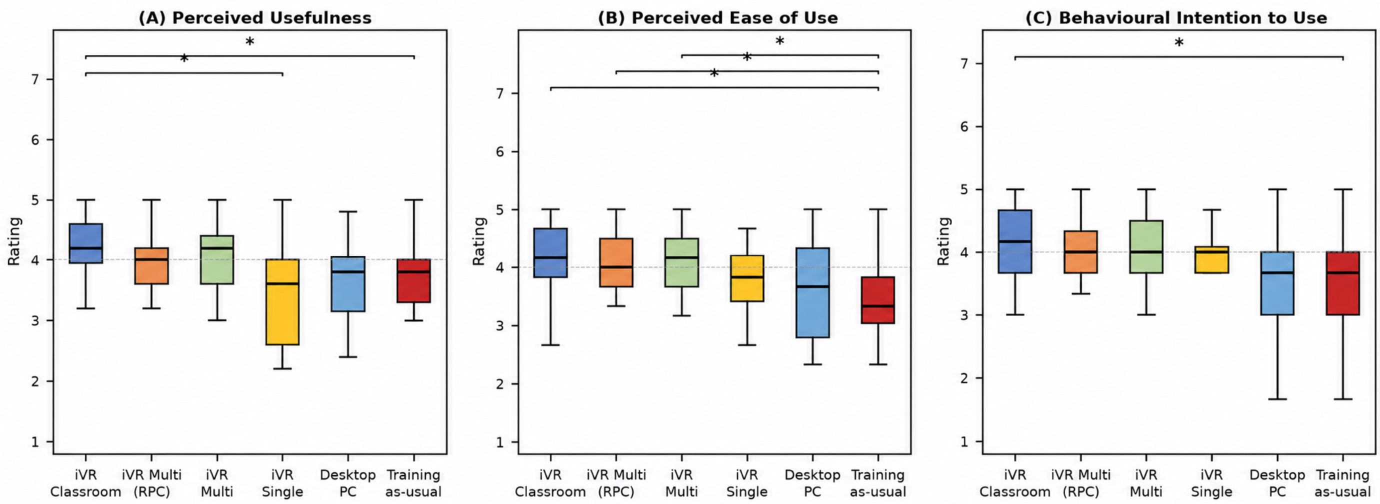

Figure 5. (**A**–**C**) Descriptive results of the significant Kruskal–Wallis tests for the TAM scales: (**A**) perceived usefulness, (**B**) perceived ease of use, and (**C**) behavioral intention to use. * $p < 0.05$ after Bonferroni correction.

Following Bonferroni correction for multiple pairwise comparisons, only a subset of the 15 possible pairwise contrasts per scale remained significant; all p-values reported for the post hoc pairwise comparisons are Bonferroni-adjusted and are evaluated against a family-wise $\alpha = 0.05$. Omnibus Kruskal–Wallis p-values are reported unadjusted. iVR classroom vs. training-as-usual was the only pairwise contrast that remained significant consistently across all three scales. Additional significant contrasts emerged on individual scales: for perceived ease of use, iVR multi-player with RPC vs. training-as-usual ($p = 0.015$) and iVR multi-player vs. training-as-usual ($p = 0.011$) also reached significance; for perceived usefulness, iVR classroom vs. iVR single-player ($p = 0.040$) also reached significance. Descriptively, training-as-usual and desktop PC scored below the scale midpoint of 4 on all three TAM scales, iVR single-player scored below the midpoint on two of three scales, and iVR multi-player, iVR multi-player with RPC, and iVR classroom scored above the midpoint on all three scales (see Table 7).

Table 7. Overview of primary and secondary outcomes by training condition.

Condition	N	Full Exam Score, n (%)	OR vs. TAU [95% CI]	TAM Scales (PU/PEOU/BI) vs. Scale Midpoint (4)
Training-as-usual (TAU)	23	13 (56.5%)	(reference)	Below midpoint on all three scales
iVR single-player	20	19 (95%)	14.62 [1.66, 128.46]	Below midpoint on two of three scales
Desktop PC (single-player)	20	18 (90%)	6.92 [1.29, 37.07]	Below midpoint on all three scales
iVR multi-player	23	20 (87%)	5.13 [1.18, 22.27]	Above midpoint on all three scales
iVR multi-player with RPC	21	19 (90.5%)	7.31 [1.37, 38.97]	Above midpoint on all three scales
iVR classroom	20	19 (95%)	14.62 [1.66, 128.46]	Above midpoint on all three scales

Note. TAU = training-as-usual (reference/control condition); RPC = retrospective perspective change; OR = odds ratio; CI = confidence interval; PU = perceived usefulness; PEOU = perceived ease of use; BI = behavioral intention to use. Full exam score counts (n) are based on direct calculation from the raw dataset. TAM-scale direction (below/above the scale midpoint of 4) is based on condition means from the raw dataset.

In summary, technology acceptance differed significantly between conditions on all three TAM scales, with small-to-medium effect sizes, and iVR classroom was the only condition that differed significantly from training-as-usual on every scale after correction for multiple comparisons. Accordingly, H3 was partially supported: technology acceptance ratings differed significantly between conditions on all three TAM scales, but the predicted general advantage of iVR conditions over both training-as-usual and desktop training was not observed consistently across all pairwise comparisons.

3.4. Practice-Oriented Items (H4, Exploratory)

To examine H4, the five practice-oriented items listed in Table 4 were compared across the six conditions using the non-parametric procedure specified in Section 2.8. Because these items were developed specifically for the present project and are not psychometrically validated instruments from the literature, the findings are interpreted as exploratory and project-specific rather than confirmatory. Kruskal–Wallis tests showed significant differences between conditions for three of the five items: "compared to training without the learning tool, I feel better prepared for practical work," $\chi^2(5) = 22.231$, $p < 0.001$, η^2_H = 0.14; "…I feel safer prepared for practical work," $\chi^2(5) = 17.799$, $p = 0.003$, η^2_H = 0.11; and "the hazards in the track area were made very clear to me by the learning tool," $\chi^2(5) = 18.488$, $p = 0.002$, η^2_H = 0.11 (effect sizes calculated by the authors using the η^2_H formula specified in Section 3.3). Figure 6A–C show the descriptive pattern underlying these tests.

As with the TAM scales, pairwise comparisons after Bonferroni correction identified iVR classroom as the condition contributing most often to the significant omnibus tests, most frequently in contrast to desktop PC and, less consistently, to training-as-usual. Because these items were purpose-built for this project and have not been validated in prior research, this pattern should be treated as a preliminary, hypothesis-generating finding rather than confirmatory evidence.

In summary, three of the five exploratory practice-oriented items showed significant differences between conditions, again most consistently favoring iVR classroom, but given the non-validated nature of these items this pattern warrants cautious interpretation and independent replication. Accordingly, H4 received partial exploratory support: significant condition differences were observed for three of the five practice-oriented items, but the predicted advantage of collaborative and instructor-facilitated formats was not consistently observed across all items and pairwise comparisons.

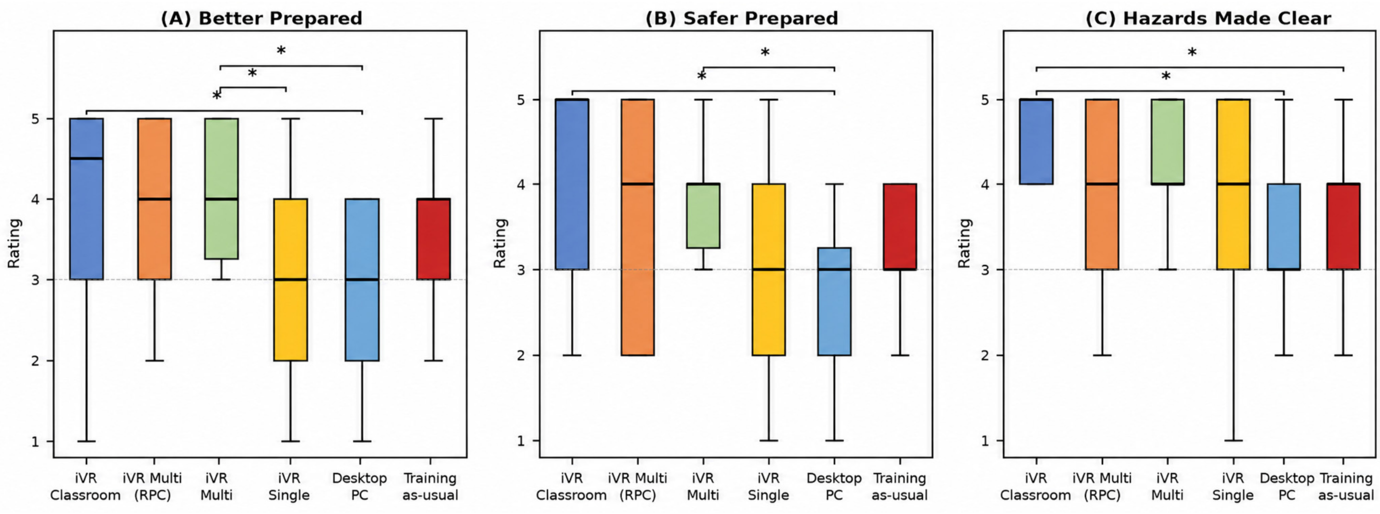

Figure 6. (**A**–**C**) Descriptive results of the significant Kruskal–Wallis tests for the SBB practice-oriented items: (**A**) "better prepared for practical work," (**B**) "safer prepared for practical work," and (**C**) "hazards made very clear.". * $p < 0.05$ after Bonferroni correction.

3.5. Overview of Exam Performance and Technology Acceptance Outcomes

Table 7 summarizes the primary exam performance outcome and the descriptive direction of the secondary technology acceptance outcomes across all six training conditions to allow for direct comparison of the relative standing of each format.

Taken together, the primary and secondary outcomes show a broadly consistent descriptive pattern across measures: iVR classroom and iVR single-player had the highest proportions of full examination scores, while iVR classroom, iVR multi-player, and iVR multi-player with RPC showed the most favorable technology acceptance profiles. Desktop PC and training-as-usual tended to show less favorable patterns across these outcomes. These comparisons are descriptive and should be interpreted in light of the non-randomized allocation, age imbalance, clustering, and unequal additional training exposure described in Section 2.9. The exploratory practice-oriented outcomes are reported separately in Section 3.4 because these project-specific items were not validated measures and are therefore not included in this summary table.

4. Discussion

This study examined whether different digitally enhanced training formats were associated with differences in examination performance and technology acceptance within an authentic occupational railway safety course. Across the present quasi-experimental field sample, participants assigned to the five digital training conditions had higher odds of achieving the full score on the safety-critical examination item than participants in the training-as-usual condition. The observed pattern further suggested differences between instructional configurations, with descriptively similar examination outcomes for the instructor-led iVR classroom and iVR single-player formats. Given the non-randomized allocation and remaining confounding factors, these findings should be interpreted as associations observed under authentic field conditions rather than as evidence of isolated causal effects. The observed pattern is consistent with the interpretation that the effectiveness of digital training add-ons in this occupational context may depend not only on immersive technology itself but also on the way the technology is pedagogically implemented. However, because several instructional and technological characteristics varied simultaneously, the present study cannot determine the independent contribution of individual design elements. To our knowledge, few occupational field studies have directly compared multiple instructional configurations of immersive VR within the same authentic training program. The present comparative design therefore extends the existing literature by enabling comparisons across several pedagogical configurations under otherwise similar field conditions.

4.1. Interpretation of Primary Findings

Within an already experience-based, simulated training format, the observed pattern suggests that pedagogical configuration may be an important contributor to learning outcomes. Given the quasi-experimental, non-randomized allocation, the marked age imbalance, and remaining confounding, however, the present data cannot isolate its causal contribution. In addition, dependence introduced by course run, trainer, and site clustering could not be modeled; the participant-level standard errors, confidence intervals, and significance tests may therefore be somewhat optimistic. The pattern is consistent with the possibility that instructional configuration matters within immersive training, but it should not be interpreted as direct evidence of an isolated causal mechanism.

4.2. The iVR Classroom Format

The iVR classroom condition is arguably the most novel and practically significant finding of this study. It combined the highest rate of full exam scores with consistently favorable technology acceptance ratings and required markedly fewer HMD units and less physical space than the multi-player formats, as only the instructor and, in pairs, the participants themselves needed a headset while the remainder of the course observed via projection. Several complementary mechanisms may explain why this comparatively low-resource format performed as well as, or better than, the fully individualized iVR single-player condition. First, instructor scaffolding: in iVR classroom the trainer remained continuously present and active in the scenario, able to adapt pacing, repeat explanations, and respond immediately to confusion—a facilitation role that is well documented to support reflection and debriefing in simulation-based learning generally [29] and that was structurally absent for participants working through the single-player scenario largely on their own initiative. Second, collaborative reflection and observational learning: because the wider course group observed the unfolding scenario on the projection screen, participants who were not currently wearing an HMD could still engage in vicarious, observational learning from peers' successes and mistakes, consistent with evidence that observing others' consequences shapes one's own risk perception and safety behavior [53] and with peer-mediated communication and feedback benefiting learning outcomes more broadly [28]. Third, social presence: the shared classroom moment—discussing the scenario together both inside and outside the headset—plausibly increased participants' sense of co-presence with peers and instructor [22], which in turn may explain the elevated behavioral intention observed for this condition, echoing our own prior observation that trainees who discuss content outside the iVR space also want to do so within it [54]. Fourth, cognitive load: by distributing the interaction burden between instructor and participants and reducing the technical self-management required of any single trainee, iVR classroom may have lowered extraneous cognitive load relative to the single-player conditions, in which participants had to manage locomotion, object interaction, and task execution simultaneously and without direct human guidance. Taken together, these mechanisms suggest that iVR classroom's advantage may lie less in the virtual environment itself and more in reproducing, inside iVR, the well-established ingredients of effective classroom facilitation. If this interpretation holds, the implications extend well beyond railway safety: any high-consequence occupational domain that currently relies on costly one-headset-per-trainee deployments—e.g., aviation, construction, or emergency response—may warrant testing instructor-led, shared-observation formats as a lower-hardware alternative to one-headset-per-trainee deployments. This remains, however, an interpretation of a descriptive pattern rather than a directly tested mechanism, and should be treated as a hypothesis for future confirmatory research rather than an established causal claim.

4.3. Connection to CAMIL

We use the Cognitive–Affective Model of Immersive Learning (CAMIL; [15]) as the primary theoretical framework for interpreting the observed findings while being explicit about which aspects are supported by the present data and which remain speculative. CAMIL predicts that immersion and interactivity elicit presence and agency, which in turn drive the cognitive–affective mechanisms (interest, motivation, self-efficacy, cognitive load) responsible for learning gains. Our finding that iVR single-player—high in individual immersion and agency but low in social presence—and iVR classroom—lower in individual immersion and agency but plausibly higher in social presence—showed descriptively similar learning outcomes is consistent with the CAMIL's premise that presence is multidimensional and that different presence sub-types (physical, social, self-presence) can compensate for one another in driving the model's downstream cognitive–affective constructs. This compensatory interpretation is plausible and consistent with the pattern of results, but was not directly tested, as the present questionnaire did not administer the full CAMIL construct battery to every condition (see Section 4.8.2, Measurement Limitations); it should therefore be read as a data-consistent hypothesis rather than a confirmed mechanism. What our data support directly is the observation that four iVR configurations using identical hardware and largely comparable scenario content were associated with different learning outcomes. This pattern is consistent with CAMIL's proposition that medium and instructional method should be considered jointly when designing immersive learning environments. However, because multiple instructional characteristics varied simultaneously, the present study cannot determine which specific design component, or combination of components, was responsible for the observed differences between conditions. Which specific CAMIL construct (e.g., agency, presence, cognitive load, or their interaction) may have contributed to the observed pattern cannot be determined from the present data and therefore remains an important target for future process-oriented research.

4.4. Collaborative Learning

The theoretical basis for collaborative iVR training draws on Vygotsky's [24] concept of the zone of proximal development (ZPD), which holds that learners can accomplish more with the guidance of a more competent instructor or in collaboration with peers than they can independently, and on Kolb's [25] experiential learning model, which identifies guided reflection as a necessary step for converting concrete experience into abstract, transferable knowledge. Both perspectives predict that multi-player and classroom formats, in which an instructor or peers are actively present during the experience, should outperform purely individual formats—particularly because expert feedback and debriefing are considered essential facilitators of the reflective phase of experiential learning [29] and because instructors can actively support task coordination and peer communication in multi-player settings [28]. Our results only partially align with this prediction: iVR classroom, which combines instructor presence with a collaborative, observational structure, performed among the best, consistent with ZPD-based expectations, whereas iVR multi-player—arguably the most directly collaborative condition, in which all participants share the same virtual space—showed the smallest improvement over training-as-usual. One plausible explanation, consistent with collaborative-load theory [26] and with findings that collaborative settings impose additional coordination and communication demands that can offset their pedagogical benefits, particularly for comparatively low-complexity tasks [27], is that the checklist/approach distance task used in this study was too procedurally simple to benefit from full multi-player collaboration, whose advantages may only emerge for more complex, multi-step tasks. This would suggest that collaborative interaction does not uniformly compensate for reduced individual immersion, but does so selectively—most clearly when collaboration is structured around a shared instructor-led narrative (as in iVR classroom) rather than distributed peer-to-peer coordination (as in iVR multi-player). This pattern is consistent with recent evidence that peer-to-peer VR-supported learning in higher education benefits from structured facilitation rather than unguided peer interaction [55]. For future collaborative VR training design, this implies that the specific social architecture of a multi-player scenario—who leads, who observes, and how feedback is distributed—may matter more than the mere presence of other participants in the same virtual space.

4.5. Technology Acceptance and Sustainable Implementation

Beyond documenting differences in acceptance ratings across conditions, it is important to consider why technology acceptance matters for implementation. Within the Technology Acceptance Model (TAM; [31]), perceived usefulness and perceived ease of use are associated with behavioral intention and can therefore inform whether a training technology is likely to be acceptable for continued organizational use. In the present sample, the iVR classroom condition combined favorable ratings of perceived usefulness, perceived ease of use, and behavioral intention with a high proportion of full scores on the safety-critical examination item. This pattern makes the format a relevant candidate for further implementation research, particularly because it requires fewer HMDs than fully individualized deployments. However, the present cross-sectional post-training assessment cannot establish whether acceptance predicts sustained use at SBB, whether it mediates longer-term training uptake, or whether the observed acceptance differences persist after the novelty of the training experience has diminished. These questions require longitudinal implementation studies.

4.6. Comparison with Previous Literature

Our single-player findings broadly align with the international literature on iVR safety training in construction and related high-risk trades [56], which reports gains in self-efficacy, risk perception, safety motivation and perceived usefulness relative to conventional instruction [56,57], higher declarative learning and motivation [58], increased safety behavior following iVR training [59], and gains in declarative and procedural learning and self-efficacy in confined-space rescue training [60] and in maritime deck safety training [61]. Recent work-at-heights research has also extended this literature through controlled protocols that explicitly assess practical skills, safety perception, and longer-term transfer, although those protocol reports do not yet provide outcome evidence [62]. Our technology acceptance results only partly replicate Özgen et al. [36], who found higher perceived enjoyment and behavioral intention for iVR relative to paper-based training but no differences in perceived usefulness or ease of use: in the present sample, perceived usefulness and ease of use did differ significantly between conditions, which may reflect the inclusion of multi-player and classroom formats absent from Özgen et al.'s single-player comparison. Where the present study departs most clearly from prior work is in its comparative scope: rather than confirming, once again, that iVR outperforms conventional training in a single-format comparison, it directly contrasts five distinct digital configurations against one another within the same real occupational curriculum and the same assessment framework—allowing patterns across instructional configurations to be compared without reducing the study to a simple VR-versus-conventional contrast. This comparative, format-contrasting design is, to our knowledge, rare in the occupational safety training literature and represents the primary novel contribution of the present study relative to prior single-format VR-versus-conventional comparisons.

4.7. Practical Implications

The practical implications of these findings extend beyond the specific SBB context in which they were generated. Railway track safety training is one instance of a broader category of high-risk occupational training—alongside domains such as construction, aviation ground operations, confined-space rescue, and energy sector maintenance—in which conventional training suffers from a comparable transfer problem [2] and in which the checklist-style, procedurally defined safety task studied here (a fixed evacuation-time and approach distance calculation) has direct analogs. Wherever such organizations face similar constraints—limited training time, a need to demonstrate concrete return on a technology investment, and a workforce with heterogeneous prior VR exposure—the iVR classroom format in particular may offer a resource-efficient option for further evaluation: it requires only one or two headsets per course rather than one per trainee, integrates directly into existing instructor-led curricula without redesigning the entire course around individual VR stations, and, in this study, showed examination performance and acceptance ratings that were descriptively similar to, or more favorable than, those observed for fully individualized iVR. We therefore frame the SstA checklist task studied here as one illustrative example of a wider class of high-risk occupational training problems for which this instructional configuration may generalize, pending confirmation in other domains.

4.8. Limitations

The limitations of this field experiment are organized into design limitations (Section 4.8.1), measurement limitations (Section 4.8.2), and generalizability (Section 4.8.3), each concluding with a corresponding recommendation for future research.

4.8.1. Design Limitations

Participants were not randomly allocated to conditions. Allocation was constrained by room availability (e.g., iVR was not possible in the Neptun room in Bern, and the multi-player condition in Olten required an additional room that was not always available) and by the voluntary, uncompensated nature of participation, which meant the research team did not know in advance how many participants would take part in a given course and filled remaining conditions based on the running tally already collected. This non-random allocation, together with the involvement of eight different trainers with varying experience, confidence, and facilitation styles across conditions and locations (“Ost”/Zürich vs. “Mitte”/Olten and Bern teams), means that condition and trainer/location effects are partially confounded, and that trainer effects cannot be fully ruled out as an alternative explanation for the observed differences between conditions—including the performance advantage of iVR classroom, which was delivered by a similarly mixed set of trainers as the other formats. This limits the causal interpretability of the present pattern of results, consistent with the generally lower control over confounding variables typical of field experiments relative to laboratory studies [40]. Future studies should employ randomized allocation of participants to conditions across multiple sites and trainers, ideally with trainers crossed factorially with conditions, to disentangle instructional-format effects from trainer and location effects. Relatedly, several odds ratios in Table 5 had wide confidence intervals due to small per-condition sample sizes and, in two conditions, only a single non-full-score case; a Firth bias-reduced re-analysis (Table 6) confirmed that all condition effects remained statistically significant and directionally consistent, but the precision of these estimates should still be interpreted with appropriate caution given the modest sample size. Because the dependence introduced by course run, trainer, and site clustering could not be modeled, the participant-level standard errors, confidence intervals, and significance tests may be overly optimistic and should be interpreted cautiously.

4.8.2. Measurement Limitations

The primary performance outcome was a single, binary exam item (full vs. non-full score on the checklist/approach distance question), which, while directly relevant to SBB's practical concern, does not capture broader declarative or procedural learning, nor any transfer of behavior to actual trackside work following the course. This restricts our conclusions to short-term exam performance and self-reported technology acceptance rather than to demonstrated on-the-job safety behavior. Future studies should directly measure the CAMIL-related constructs proposed as explanatory mechanisms (e.g., physical and social presence, agency, cognitive load, and self-efficacy) in order to test whether they statistically mediate condition differences, and should incorporate broader behavioral or supervisor-rated measures of safety conduct at the workplace, ideally with a delayed follow-up, to assess actual training transfer rather than immediate exam performance alone.

4.8.3. Generalizability

The sample was drawn exclusively from SBB SstA courses at three locations in the German-speaking part of Switzerland, was predominantly male (113 of 127 participants), and 38% of participants had prior VR experience, which may not represent the demographic or technological affinity profile of trainees in other occupational domains or countries. The findings are further specific to a single, procedurally simple safety task (the checklist/approach distance item); as discussed in Section 4.4, the relative advantage of collaborative formats may depend on task complexity [27], and it is unknown whether the present pattern—particularly the strong performance of iVR classroom relative to iVR multi-player—would replicate for more complex, multi-step procedural tasks. Future research should test whether the present pattern of results, and particularly the relative advantage of the iVR classroom format, replicates across multiple organizations, countries, and occupational safety domains, and across tasks varying systematically in procedural complexity.

4.9. Future Research

Future research should test the present descriptive pattern in randomized, multi-site designs in which condition assignment is independent of trainer and site and the number of course run clusters is sufficient for multilevel modeling. Studies should balance or model age and prior experience and include an active control matched for additional training time. Broader learning, transfer, and workplace safety outcomes should be assessed beyond the single examination item used here, ideally with delayed follow-up. Finally, direct measurement of CAMIL-related constructs such as physical and social presence, agency, cognitive load, and self-efficacy would allow the proposed instructional mechanisms to be tested rather than inferred from outcome patterns.

5. Conclusions

Rather than showing simply that immersive virtual reality improves occupational safety training, this study suggests that how iVR training is instructionally configured may be an important consideration alongside technological immersion. Across six training formats delivered within a real occupational curriculum, the collaborative, instructor-led, low-resource iVR classroom format showed examination performance and acceptance ratings that were descriptively similar to, or more favorable than, those observed for iVR single-player training, while the peer-to-peer iVR multi-player format showed a less favorable pattern on several outcomes. For occupational safety training, these findings motivate further evaluation of shared, instructor-led iVR formats that require fewer headsets than fully individualized deployments. Beyond railway safety, the results support shifting part of the research focus from whether VR works toward which instructional configurations are associated with favorable outcomes for a given task and workforce. These implications remain provisional because the present quasi-experimental design cannot isolate causal effects of individual technological or instructional components; future randomized, multi-site, and mechanism-focused research is needed to test them.

Author Contributions: Conceptualization, O.C., P.H. and S.G.; methodology, A.E., P.V.M., S.G., P.D. and O.C.; software, P.V.M., P.D., P.H. and O.C.; validation, A.E., S.G., P.D. and O.C.; formal analysis, A.E. and O.C.; investigation, A.E., P.V.M. and O.C.; resources, S.G., P.D. and S.A.; writing—original draft preparation, A.E. and O.C.; writing—review and editing, O.C., S.A., P.H. and S.G.; visualization, A.E.; supervision, O.C., P.H., S.A. and S.G.; project administration, O.C., P.H., S.A. and S.G.; funding acquisition, O.C. and S.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the SBB Research Fund on management in the transport sector represented by the Institute for Systemic Management and Public Governance (IMP-HSG) at the University of St. Gallen, Dufourstrasse 40a, 9000 St. Gallen.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The datasets analyzed for this study are available from the corresponding author upon reasonable request, subject to data protection agreements with SBB.

Acknowledgments: We thank the SUCAP steering board and advisory board, the SBB and Login AG trainers who supported data collection, and all trainees who voluntarily took part in the study.

Conflicts of Interest: A.E., P.V.M. and O.C., declare no conflicts of interest. S.G., S.A. and P.H. are employed by SBB, AG Bern Wankdorf, Switzerland. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Burke, M.J.; Sarpy, S.A.; Smith-Crowe, K.; Chan-Serafin, S.; Salvador, R.O.; Islam, G. Relative effectiveness of worker safety and health training methods. *Am. J. Public Health* **2006**, *96*, 315–324. [CrossRef] [PubMed]
2. Krauss, A.D.; Casey, T.W.; Chen, P.Y. Making Safety Training Stick. In *Contemporary Occupational Health Psychology: Global Perspectives on Research and Practice*; Leka, S., Sinclair, R.R., Eds.; Wiley: Chichester, UK, 2014; Volume 3, pp. 181–197. [CrossRef]
3. Ricci, F.; Chiesi, A.; Bisio, C.; Panari, C.; Pelosi, A. Effectiveness of occupational health and safety training. *J. Workplace Learn.* **2016**, *28*, 355–377. [CrossRef]
4. Grassini, S.; Laumann, K.; Skogstad, M.R. The Use of Virtual Reality Alone Does Not Promote Training Performance (but Sense of Presence Does). *Front. Psychol.* **2020**, *11*, 1743. [CrossRef] [PubMed]
5. Renganayagalu, S.K.; Mallam, S.C.; Nazir, S. Effectiveness of VR Head Mounted Displays in Professional Training: A Systematic Review. *Technol. Knowl. Learn.* **2021**, *26*, 999–1041. [CrossRef]
6. Laughery, K.R.; Wogalter, M.S. Designing Effective Warnings. *Rev. Hum. Factors Ergon.* **2006**, *2*, 241–271. [CrossRef]
7. Weinstein, N.D. Why it won't happen to me: Perceptions of risk factors and susceptibility. *Health Psychol.* **1984**, *3*, 431–457. [CrossRef] [PubMed]
8. Weinstein, N.D. Effects of personal experience on self-protective behavior. *Psychol. Bull.* **1989**, *105*, 31–50. [CrossRef] [PubMed]
9. Weinstein, N.D. Testing four competing theories of health-protective behavior. *Health Psychol.* **1993**, *12*, 324–333. [CrossRef] [PubMed]
10. Slater, M.; Sanchez-Vives, M.V. Enhancing Our Lives with Immersive Virtual Reality. *Front. Robot. AI* **2016**, *3*, 74. [CrossRef]
11. Colquitt, J.A.; LePine, J.A.; Noe, R.A. Toward an integrative theory of training motivation: A meta-analytic path analysis of 20 years of research. *J. Appl. Psychol.* **2000**, *85*, 678–707. [CrossRef] [PubMed]

12. Floyd, D.L.; Prentice-Dunn, S.; Rogers, R.W. A Meta-Analysis of Research on Protection Motivation Theory. *J. Appl. Soc. Psychol.* **2000**, *30*, 407–429. [CrossRef]
13. Felicia, P. (Ed.) *Handbook of Research on Improving Learning and Motivation Through Educational Games: Multidisciplinary Approaches (2 Volumes)*; IGI Global Scientific Publishing: Hershey, PA, USA, 2011. [CrossRef]
14. Warburton, S. Second Life in higher education: Assessing the potential for and the barriers to deploying virtual worlds in learning and teaching. *Br. J. Educ. Technol.* **2009**, *40*, 414–426. [CrossRef]
15. Makransky, G.; Petersen, G.B. The Cognitive Affective Model of Immersive Learning (CAMIL): A Theoretical Research-Based Model of Learning in Immersive Virtual Reality. *Educ. Psychol. Rev.* **2021**, *33*, 937–958. [CrossRef]
16. Naranjo, J.E.; Sanchez, D.G.; Robalino-Lopez, A.; Robalino-Lopez, P.; Alarcon-Ortiz, A.; Garcia, M.V. A Scoping Review on Virtual Reality-Based Industrial Training. *Appl. Sci.* **2020**, *10*, 8224. [CrossRef]
17. Pletz, C. Technologieakzeptanz von Virtuellen Lern- und Arbeitsumgebungen. Bachelor's Thesis, Universität Stuttgart, Stuttgart, Germany, 2022. Available online: https://nbn-resolving.de/urn:nbn:de:0009-5-56161 (accessed on 3 March 2026).
18. Di Natale, A.F.; Repetto, C.; Riva, G.; Villani, D. Immersive virtual reality in K-12 and higher education: A 10-year systematic review of empirical research. *Br. J. Educ. Technol.* **2020**, *51*, 2006–2033. [CrossRef]
19. Johnson, C.I.; Whitmer, D.E.; Entinger, K.; Peterson, E.K.; Sobel, B.M. Interacting with Virtual Reality with a Controller Instead of the Body Benefits Performance and Perceptions. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **2022**, *66*, 1294–1298. [CrossRef]
20. Jensen, L.; Konradsen, F. A review of the use of virtual reality head-mounted displays in education and training. *Educ. Inf. Technol.* **2018**, *23*, 1515–1529. [CrossRef]
21. Radianti, J.; Majchrzak, T.A.; Fromm, J.; Wohlgenannt, I. A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Comput. Educ.* **2020**, *147*, 103778. [CrossRef]
22. Makransky, G.; Lilleholt, L.; Aaby, A. Development and validation of the Multimodal Presence Scale for virtual reality environments: A confirmatory factor analysis and item response theory approach. *Comput. Hum. Behav.* **2017**, *72*, 276–285. [CrossRef]
23. Piaget, J. *Science of Education and the Psychology of the Child*; Orion Press: New York, NY, USA, 1970; 186p.
24. Vygotsky, L.S. *Mind in Society: The Development of Higher Psychological Processes*; Cole, M., John-Steiner, V., Scribner, S., Souberman, E., Eds.; Harvard University Press: Cambridge, MA, USA, 1978; Volume xi, 159p.
25. Kolb, D.A. *Experiential Learning: Experience as the Source of Learning and Development*, 2nd ed.; Pearson FT Press: Upper Saddle River, NJ, USA, 2014.
26. Dillenbourg, P.; Bétrancourt, M. Collaboration Load. In *Handling Complexity in Learning Environments: Theory and Research*; Elen, J., Clark, R.E., Eds.; Elsevier: Oxford, UK, 2006; pp. 142–163.
27. Sedlák, M.; Šašinka, Č.; Stachoň, Z.; Chmelík, J.; Doležal, M. Collaborative and individual learning of geography in immersive virtual reality: An effectiveness study. *PLoS ONE* **2022**, *17*, e0276267. [CrossRef] [PubMed]
28. Cui, P.; Zheng, L. A Meta-analysis of the Peer Evaluation Effects on Learning Achievements in Blended Learning Environment. In *Blended Learning. Enhancing Learning Success*; Cheung, S.K.S., Kwok, L., Kubota, K., Lee, L.-K., Tokito, J., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2018; Volume 10949, pp. 227–237. [CrossRef]
29. Decker, S.; Fey, M.; Sideras, S.; Caballero, S.; Rockstraw, L.; Boese, T.; Franklin, A.E.; Gloe, D.; Lioce, L.; Sando, C.R.; et al. Standards of Best Practice: Simulation Standard VI: The Debriefing Process. *Clin. Simul. Nurs.* **2013**, *9*, S26–S29. [CrossRef]
30. Davis, F.D.; Bagozzi, R.P.; Warshaw, P.R. User Acceptance of Computer Technology: A Comparison of Two Theoretical Models. *Manag. Sci.* **1989**, *35*, 982–1003. [CrossRef]
31. Davis, F.D. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Q.* **1989**, *13*, 319–340. [CrossRef] [PubMed]
32. Ajzen, I.; Fishbein, M. *Understanding Attitudes and Predicting Social Behavior*; Prentice-Hall: Englewood Cliffs, NJ, USA, 1980; 278p.
33. Ajzen, I. From Intentions to Actions: A Theory of Planned Behavior. In *Action Control: From Cognition to Behavior*; Kuhl, J., Beckmann, J., Eds.; Springer: Berlin/Heidelberg, Germany, 1985; pp. 11–39. [CrossRef]
34. Pletz, C.; Zinn, B. Technologieakzeptanz von virtuellen Lern- und Arbeitsumgebungen in technischen Domänen. *J. Tech. Educ. (JOTED)* **2018**, *6*, 86–105. [CrossRef]
35. Chen, F.-Q.; Leng, Y.-F.; Ge, J.-F.; Wang, D.-W.; Li, C.; Chen, B.; Sun, Z.-L. Effectiveness of Virtual Reality in Nursing Education: Meta-Analysis. *J. Med. Internet Res.* **2020**, *22*, e18290. [CrossRef] [PubMed]
36. Özgen, D.S.; Afacan, Y.; Sürer, E. Usability of virtual reality for basic design education: A comparative study with paper-based design. *Int. J. Technol. Des. Educ.* **2021**, *31*, 357–377. [CrossRef]
37. Travaglini, A.; Brand, E.; Meier, P.; Christ, O. Job relevance or perceived usefulness? What features of immersive virtual reality software predict intention to use in a future project-based-learning scenario: A mixed method approach. *Front. Virtual Real.* **2023**, *4*, 1286877. [CrossRef]
38. Urech, A.; Meier, P.V.; Gut, S.; Duchene, P.; Christ, O. Mapping or no Mapping: The Influence of Controller Interaction Design in an Immersive Virtual Reality Tutorial in Two Different Age Groups. *Multimodal Technol. Interact.* **2024**, *8*, 59. [CrossRef]

39. Charness, G.; Gneezy, U.; Kuhn, M.A. Experimental methods: Between-subject and within-subject design. *J. Econ. Behav. Organ.* **2012**, *81*, 1–8. [CrossRef]
40. Döring, N.; Bortz, J. *Forschungsmethoden Und Evaluation in Den Sozial- Und Humanwissenschaften*, 5th ed.; Springer: Berlin/Heidelberg, Germany, 2016; Volume XXVII, 1051p. [CrossRef]
41. Brosius, H.-B.; Haas, A.; Koschel, F. *Methoden Der Empirischen Kommunikationsforschung: Eine Einführung*, 7th ed.; VS Verlag für Sozialwissenschaften/Springer Fachmedien Wiesbaden: Wiesbaden, Germany, 2016; Volume XV, 265p. [CrossRef]
42. Des Jarlais, D.C.; Lyles, C.; Crepaz, N.; TREND Group. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: The TREND statement. *Am. J. Public Health* **2004**, *94*, 361–366. [CrossRef] [PubMed]
43. Faul, F.; Erdfelder, E.; Lang, A.-G.; Buchner, A. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **2007**, *39*, 175–191. [CrossRef] [PubMed]
44. Cohen, J. A power primer. *Psychol. Bull.* **1992**, *112*, 155–159. [CrossRef] [PubMed]
45. Tivian. *Enterprise Feedback Suite (EFS) [Computer Software]*; Tivian: Irvine, CA, USA; Available online: http://survey.fhnw.ch/ (accessed on 3 March 2026).
46. Comrey, A.L.; Lee, H.B. *A First Course in Factor Analysis*, 2nd ed.; Lawrence Erlbaum Associates: Hillsdale, NJ, USA, 1992.
47. Koran, J. Indicators per Factor in Confirmatory Factor Analysis: More is not Always Better. *Struct. Equ. Model.* **2020**, *27*, 765–772. [CrossRef]
48. Microsoft Corporation. *Microsoft Excel [Computer Software]*; Microsoft Corporation: Redmond, WA, USA.
49. Firth, D. Bias reduction of maximum likelihood estimates. *Biometrika* **1993**, *80*, 27–38. [CrossRef]
50. Heinze, G.; Schemper, M. A solution to the problem of separation in logistic regression. *Stat. Med.* **2002**, *21*, 2409–2419. [CrossRef] [PubMed]
51. Peduzzi, P.; Concato, J.; Kemper, E.; Holford, T.R.; Feinstein, A.R. A simulation study of the number of events per variable in logistic regression analysis. *J. Clin. Epidemiol.* **1996**, *49*, 1373–1379. [CrossRef] [PubMed]
52. Tomczak, M.; Tomczak, E. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends Sport Sci.* **2014**, *21*, 19–25.
53. Rosenstock, I.M.; Strecher, V.J.; Becker, M.H. Social Learning Theory and the Health Belief Model. *Health Educ. Q.* **1988**, *15*, 175–183. [CrossRef] [PubMed]
54. Christ, O.; Hirschi, M. Virtual Reality in der Berufsbildung. *Transf. Berufsbild. Forsch. Prax.* **2021**, *6*. [CrossRef] [PubMed]
55. Donald, C.; Rustamova, S.; Hirsu, L.; Rodolico, G.; Latkovskis, I.; McDonnell, N.; Veitch, N. Virtual reality peer-to-peer learning: A pilot study investigating peer-to-peer VR supported lessons in higher education. *Front. Educ.* **2026**, *11*, 1800038. [CrossRef]
56. Baikati, M.; Baikati, A.K.; Saravana Kumar, M.P.; Ramani, P.V. Integrating BIM and VR for enhanced safety communication and training in construction. *Front. Virtual Real.* **2026**, *7*, 1768058. [CrossRef]
57. Nykänen, M.; Puro, V.; Tiikkaja, M.; Kannisto, H.; Lantto, E.; Simpura, F.; Uusitalo, J.; Lukander, K.; Räsänen, T.; Heikkilä, T.; et al. Implementing and evaluating novel safety training methods for construction sector workers: Results of a randomized controlled trial. *J. Saf. Res.* **2020**, *75*, 205–221. [CrossRef] [PubMed]
58. Joshi, S.; Hamilton, M.; Warren, R.; Faucett, D.; Tian, W.; Wang, Y.; Ma, J. Implementing Virtual Reality technology for safety training in the precast/prestressed concrete industry. *Appl. Ergon.* **2021**, *90*, 103286. [CrossRef] [PubMed]
59. Adami, P.; Rodrigues, P.B.; Woods, P.J.; Becerik-Gerber, B.; Soibelman, L.; Copur-Gencturk, Y.; Lucas, G. Effectiveness of VR-based training on improving construction workers' knowledge, skills, and safety behavior in robotic teleoperation. *Adv. Eng. Inform.* **2021**, *50*, 101431. [CrossRef]
60. Lu, S.; Wang, F.; Li, X.; Shen, Q. Development and validation of a confined space rescue training prototype based on an immersive virtual reality serious game. *Adv. Eng. Inform.* **2022**, *51*, 101520. [CrossRef]
61. Makransky, G.; Klingenberg, S. Virtual reality enhances safety training in the maritime industry: An organizational training experiment with a non-WEIRD sample. *J. Comput. Assist. Learn.* **2022**, *38*, 1127–1140. [CrossRef]
62. Guerrero-Jaramillo, D.; de la Caridad Montero, R.; Campo, O. Work at heights training: Conventional approach with and without immersive virtual reality study protocol. *Methods Protoc.* **2026**, *9*, 55. [CrossRef] [PubMed]