Computers & Education: X Reality 8 (2026) 100140

Contents lists available at ScienceDirect

Computers & Education: X Reality

journal homepage: www.journals.elsevier.com/computers-and-education-x-reality

Effects of annotations and quizzes in a classroom-based immersive virtual field trip

Maximilian C. Fink [a], Carina Galler [a], Bianca Watzka [b], Bernhard Ertl [a,*]

[a] Learning and Teaching with Media, Institute of Education, Universität der Bundeswehr München, Neubiberg, Germany
[b] Didactics of Physics and Technology, RWTH Aachen University, Aachen, Germany

ARTICLE INFO

Keywords:
Virtual field trip
Science education
Annotations
Quizzes
Immersive virtual reality

ABSTRACT

Immersive virtual reality, and especially virtual field trips, can spark learners' interest and promote the learning of STEM content. However, immersive virtual reality can also increase visual and navigational demands, which may elevate cognitive load and complicate the use of conventional study strategies. To overcome these challenges, research is needed to determine which instructional support measures make learning in VR more effective. Among support methods that may be suitable for VR in STEM education are annotations (i.e., labeling of important terms) and quizzes (i.e., self-assessment through questions). We conducted an experiment with $N = 126$ 10th graders in a school setting. Participants engaged in a science learning environment with explanatory figures and auditory explanations. The experiment used a 2×2 factorial approach, with annotations and quizzes as the two instructional support conditions. We used a recall and a comprehension knowledge test, a cognitive load scale, and a simulation sickness survey as instruments. Annotations enhanced only recall knowledge, indicating that they can promote lower cognitive knowledge levels. Quizzes, on the other hand, did not impact the learning of either type of knowledge. We also investigated the extent to which annotations and quizzes affect the feeling of presence and different facets of cognitive load. Quizzes reduced the feeling of presence significantly while neither annotations nor quizzes impacted cognitive load. However, our results indicate strong effects of covariates respectively cyber sickness and prior knowledge. These suggest that the underlying learning processes and side effects of framing conditions have to be further analyzed in detail.

1. Introduction

Virtual reality (VR), including desktop-based VR and CAVE systems, has been a part of computer science research for several decades (Muhanna, 2015). However, it is only since the late 2010s, when immersive VR (IVR) with head-mounted displays (HMDs) became more accessible, that this topic has attracted increasing attention in the educational sector. One strand of educational VR research focuses on key factors, such as presence, influencing learning processes (e.g., Makransky & Lilleholt, 2018). Another strand of research conducts media comparisons, such as juxtaposing learning with HMDs with learning in desktop-based VR (e.g., Parong & Mayer, 2021). A third strand of research emerged a few years later, investigating support measures, such as signaling annotations embedded in IVR, to enhance learning (e.g., Albus et al., 2021). Many studies on these issues were conducted in highly standardized laboratory environments; therefore, there remains a need to examine IVR under authentic classroom conditions. Our study primarily aligns with the third-mentioned strand of research and examines the impact of support methods on learning in IVR within a virtual field trip (VFT) designed for an integrated Science, Technology, Engineering, and Mathematics (STEM) school unit focusing on content from physics, biology, and chemistry.

2. Theoretical background

Integrated STEM education refers to combining complex content in interdisciplinary units, establishing relationships between content from different subjects, and focusing on real-world problems (Thibaut et al., 2018). Technology is often utilized in integrated STEM education to teach content in a motivating and engaging manner. One didactic method commonly applied for this is the VFT, defined as "a journey taken without actually making a trip to the site" (Woerner, 1999, p. 5).

* Corresponding author. Universität der Bundeswehr München, Fakultät für Humanwissenschaften, Institut Bildungswissenschaft, Professur für Erziehungswissenschaft mit Schwerpunkt Lernen und Lehren mit Medien, Werner-Heisenberg-Weg 39, 85579, Neubiberg, Germany.
E-mail address: bernhard.ertl@unibw.de (B. Ertl).

https://doi.org/10.1016/j.cexr.2026.100140
Received 8 August 2025; Received in revised form 20 January 2026; Accepted 27 January 2026
Available online 17 February 2026

VFTs can convey complex processes and concepts using a digital guide or narrator, as well as didactic materials such as figures and exercises. VFTs make remote, inaccessible, or dangerous places and facilities accessible (Stainfield et al., 2000) and can authentically depict these places, for instance when enriched with 360-degree videos (Eisenlauer & Sosa, 2022). Whether presented on desktop-based, mobile, or HMD devices, participants can learn in VFTs in a situated, highly contextualized environment (Klippel et al., 2020). Along with these benefits, empirical studies also indicate the effectiveness of VFTs presented with HMDs (immersive virtual field trips; IVFTs). Such IVFTs can increase learners' interest and promote the learning of STEM content (Fink et al., 2023; Makransky, Petersen, et al., 2020).

2.1. IVFTs in the context of CTML

IVFTs may integrate verbal and pictorial information into a learning environment and thereby create conditions that reflect the assumptions of the Cognitive Theory of Multimedia Learning (CTML, Mayer, 2014). CTML assumes that learning with words and images involves two processing channels, verbal and pictorial, which operate in sensory memory. Relevant stimuli are selected and transferred to working memory, where organizational processes create coherent verbal and visual representations. Integration processes then combine these representations into an overall mental model, supported by the retrieval of related knowledge from long-term memory (Mayer, 2014).

In IVFTs, these processes may occur when verbal explanations and pictorial information are presented simultaneously, as described in CTML (Mayer, 2014). Learners select relevant auditory and visual elements, organize them into coherent representations, and integrate them into a unified mental model. IVFTs therefore reflect CTML's assumption that learning improves when verbal and pictorial input are processed together, facilitating the construction of multiple interconnected representations. CTML is particularly relevant for IVFTs because these environments rely on the simultaneous presentation of verbal and pictorial information, which directly corresponds to the dual-channel assumption of multimedia learning. However, the immersive presentation format introduces spatial and attentional characteristics that extend beyond the two-dimensional multimedia setting originally considered by Mayer (2014).

There is substantial empirical support for CTML and its underlying assumptions, including the *multimedia effect* (Alpizar et al., 2020; Hu et al., 2021; Xie et al., 2015). The multimedia effect states that people learn more effectively through a combination of spoken explanations and visuals than by reading comparable text-only materials (Mayer, 2003; Schweppe et al., 2015). Within IVFTs, this principle would be reflected when learners process verbal explanations together with visual information. Following CTML, such dual presentation promotes the creation of multiple representations, that can be more efficiently stored and integrated more efficiently in long-term memory (Mayer, 2014).

Explanatory figures and other visual elements used in IVFTs, such as images, videos, and animations, accompanied by audio or text explanations, represent direct applications of the multimedia principle. Such components have been successfully used in STEM subjects such as biology and chemistry, to illustrate complex content (Reid, 1990; Souza & Porto, 2012). Consequently, IVFTs provide immersive multimedia environments in which the central processes described by CTML are applied. Thus, explanatory figures are likely to enhance learners' outcomes in STEM IVFTs, consistent with CTML assumptions.

2.2. Cognitive load in VR learning environments

However, when explanatory figures are presented within a 360° IVFT, the extensive visual field and spatial complexity may affect the cognitive processing demands and thereby affect learners' cognitive load. Cognitive load theory posits that learning is driven by the demands of the learning material and environment on humans' limited working memory capacity (Sweller et al., 1998). Three types of cognitive load are distinguished. *Intrinsic load* is mainly determined by the number of items to be learned, their difficulty and interactivity, and the learner's amount of prior knowledge. *Extraneous load* arises from instructional design decisions irrelevant to learning, such as user navigation. Germane load is created when learning-related processes occur, like understanding material and generating representations (Sweller et al., 1998).

Several studies have investigated cognitive load in IVR with HMDs. A meta-analysis found that, in most comparisons with desktop-based VR, total and extraneous cognitive load were higher in IVR with HMDs (Poupard et al., 2025). Several studies, which sampled university and school students, confirmed that total cognitive load and extraneous load are negatively associated with learning outcome, including recall and comprehension (Albus et al., 2021; Andersen & Makransky, 2020; Schrader & Bastiaens, 2012).

2.3. Presence in VR learning environments

The Cognitive Affective Model of Immersive Learning (CAMIL) framework (Makransky & Petersen, 2021) attributes presence to influence cognitive load in IVR, especially extraneous cognitive load. Presence is "the subjective experience of being in one place or environment, even when one is physically situated in another" (Witmer & Singer, 1998, p. 225). It is a central concept in VR-based learning and is often used as a dependent or control variable in research (Coban et al., 2022; Cummings & Bailenson, 2016). Whitin the CAMIL framework (Makransky & Petersen, 2021), presence is conceptualized as an affordance that supports effective learning in immersive environments. Three main factors shape the level of presence: immersion, referring to objective system features such as the display resolution of HMDs (Cummings & Bailenson, 2016); control factors, which describe the possibilities for navigation, interaction, and immediate feedback (Witmer & Singer, 1998); and representational fidelity, which refers to the perceived realism and authenticity of digital environments, (Dalgarno & Lee, 2010). Makransky and Petersen (2021, p. 946) describe an effect of presence on extraneous cognitive load that stems from the visual field. According to their line of argumentation, a larger visual field that may constitute the perception of presence, e.g. when comparing IVR with desktop VR, may result in a higher extraneous cognitive load, because learners may need more cognitive resources to find relevant content. Thus, considering the negative effect of extraneous cognitive load on learning (Sweller et al., 1998), presence may also negatively affect learning outcomes through its impact on cognitive load.

Empirical findings on the relationship between presence and learning remain mixed. Krassmann et al. (2023) identified studies reporting positive associations, for instance between presence and psychomotor learning outcomes (e.g., Stevens & Kincaid, 2015), as well as studies reporting negative relations for cognitive outcomes (Huang et al., 2020). Presence can also be affected by contextual and procedural factors. For example, Frommel et al. (2015) showed that embedding survey questions implicitly in a virtual environment resulted in higher levels of presence than presenting them explicitly.

2.4. Cyber sickness in VR learning environments

Presence requires a certain level of physical well-being to be sustained during immersive experiences, as higher levels of cyber sickness have been found to reduce presence (Venkatakrishnan et al., 2020). Cyber sickness refers to a form of motion sickness induced by exposure to immersive technologies and includes symptoms such as nausea, disorientation, and headaches (Kennedy et al., 1993). It has been shown to impair cognitive task performance (Weech et al., 2019) and, more specifically, to reduce the recall of factual knowledge in VR-based learning (Polcar & Horejsi, 2015).

Research further suggests that higher levels of cyber sickness are

associated with increased cognitive load-related variables (Park, 2020; Sepich et al., 2022) and a decreased sense of presence (Venkatakrishnan et al., 2020). These findings suggest that cyber sickness not only affects learners' comfort but can also interfere with key cognitive and experiential processes in immersive learning environments.

Individual differences also contribute to the experience of cyber sickness. Several studies have found that female participants report higher symptom levels than male participants (Kelly et al., 2023; MacArthur et al., 2021), although a meta-analysis concluded that this difference is relatively small (Howard & Van Zandt, 2021). Possible explanations include inter-pupillary distance (IPD) mismatches between female users and head-mounted displays (Stanney et al., 2020), greater prior gaming or VR experience and usage among men, which may lower susceptibility (Kourtesis et al., 2024). Moreover, symptom severity has been shown to increase with longer exposure durations in VR (Stanney et al., 2020).

3. Facilitating learning in VR by different scaffolds

IVFTs may provide learners with unique and engaging experiences but also impose specific challenges related to cognitive processing. Therefore, appropriate instructional support is required to help learners manage cognitive load, direct attention, and integrate information effectively.

Media comparison studies suggest that traditional methods, such as direct instruction and computer-based learning, often perform better than IVR for conveying declarative knowledge (Makransky et al., 2019; Parong & Mayer, 2018). One reason for this finding may be that learning in IVR can be associated with increased cognitive load (Makransky et al., 2019). Another reason could be that IVR makes it challenging to apply conventional study strategies, such as self-testing (Hartwig & Dunlosky, 2012). Scaffolding, defined as embedding appropriate support methods within a (computer-based) learning environment (Belland, 2014), can make learning in VR more effective and overcome problems in conveying knowledge and complex skills (Chernikova et al., 2020; Schneider et al., 2018). In IVR, annotations and quizzes could be promising scaffolding methods to promote learning.

3.1. Signaling and annotations as scaffolds in IVR

The signaling effect states that people learn more effectively from materials when the relevant entities or connections between them are explicitly identified (Beege et al., 2021). This effect is also known as the cuing principle, and it encompasses several different signals and cues (Schneider et al., 2018). Prominent examples include color coding, spotlights, and text labels that can be added to enhance the learning of texts and figures (Schneider et al., 2018). Overall, such signals have shown medium positive effects on recall performance, with stronger effects observed when embedded in texts than in figures, in which signals were also effective (Schneider et al., 2018). A specific type of signal are annotations, which are cues similar to text labels that occur in a spatiotemporal contingency with the narration (Albus et al., 2021). The learning mechanisms behind the positive effect of annotations and other signals require further clarification. It is hypothesized that annotations reduce cognitive load, guide attention by signaling relevance, or help the integration and organization processes of content in the working memory (Mayer, 2014; Schneider et al., 2018).

Several studies have examined the effect of annotations and similar cues in IVR with HMDs. For example, Albus et al. (2021) compared a group learning with annotations to a control group without annotations in an IVR science learning environment, focusing on seawater desalination. They used a sample of 107 eighth-to tenth-grade students. Annotations improved recall but had no effect on comprehension knowledge. Moreover, annotations increased the germane cognitive load, while extraneous and intrinsic load remained unaffected. Vogt et al. (2021) investigated the effect of annotations on knowledge in a sample of 61 adults using the same IVR learning environment as Albus et al. (2021). Annotations had a small, non-significant but practically relevant effect on recall. No effects were observed in comprehension knowledge tests. In a study by Huang et al. (2024), 62 university students learned about cell biology in four IVR conditions. The control group received no cues, while the visual cues group learned with highlighting arrows and textual labels. The auditory cues only group had short, emphasized statements of relevant content presented by a voice narrator, and the audiovisual group combined the treatment of the visual cues and auditory cues groups. The study found that visual and auditory cues improved learners' attention on relevant stimuli and enhanced performance on a recall knowledge test. Moreover, visual and auditory-visual cues reduced extraneous cognitive load. Li et al. (2023) investigated the effect of scaffolding in an IVR lesson on the solar system. The experiment involved a sample of 152 elementary school students and employed a factorial design to investigate the effects of textual cues and summaries. Textual cues had a positive effect on learning, but did not impact intrinsic, extraneous, or germane cognitive load. In addition to these findings, several studies have demonstrated that the effects of annotations and other signals are more pronounced for students with lower prior knowledge and motivation (Han et al., 2023; Vogt et al., 2021).

3.2. Quizzes as scaffolds in IVR

Quizzes featuring multiple-choice questions and other assessment methods, such as sorting tasks, can be integrated into learning environments as a formative and engaging approach to assessment (Tsai et al., 2015). Integrating such self-assessment methods can potentially trigger several learning mechanisms: First, quizzes introduce activity into a learning environment and can therefore, according to Chi and Wylie (2014), transform otherwise passive settings into active learning environments that evoke higher cognitive processes according to their ICAP framework. Furthermore, quizzes allow learners to test their understanding and receive feedback (Dunlosky et al., 2013; Fiorella & Mayer, 2016). Plumed et al. (2021), e.g., implemented online exercises with immediate feedback in CAD training to strengthen students' understanding of geometric constraints. Although performance gains were not statistically significant, the approach promoted autonomous learning and was perceived as useful by students.

However, little empirical evidence is available for learning with quizzes in VR. Makransky et al. (2020) investigated the effects of integrating gamified quizzes into a desktop-based VR in a sample of 208 medical students. The study employed a one-sample design and embedded multiple quiz questions into the learning environment. Participants received feedback about the correct answer after answering each quiz question. Performance on a knowledge test significantly increased from pretest to posttest. However, due to the absence of a control group, the observed effects cannot be attributed exclusively to the quizzes but may also stem from the VR environment itself and other support measures included. Based on this result and other analyses, the authors argue that quizzes promote learning. Another study compared formative and summative assessment in VR with HMDs on emergency response knowledge among 52 university students (Seprum et al., 2025). Here, a two-group design was used. The formative assessment group received feedback on the correctness of their answers after each question, while the summative assessment group received this feedback for all questions at the end. The formative assessment group achieved higher learning outcomes than the summative assessment group. However, participants in the formative assessment group reported a descriptively lower sense of presence than the other group. The authors attribute this result to the more frequent interruptions caused by quizzes and feedback (Seprum et al., 2025).

Furthermore, Ahn and Noh (2025) investigated the effect of quizzes in a fire safety training experiment. The 176 participating university students were assigned to four groups: desktop-based VR without

quizzes, desktop-based VR with quizzes, VR with HMDs without quizzes, and VR with HMDs and quizzes. The HMD-based VR condition increased presence compared to videos; however, participants who encountered quizzes experienced lower presence than those who did not (Ahn & Noh, 2025).

3.3. Research questions and hypotheses

Generally, we have large evidence for effects of annotations and quizzes from e-learning and desktop VR. However, when it comes to IVR and IVFTs, evidence is still limited, particularly regarding their application in classroom settings. Therefore, we address three research questions. First, we examine the extent to which annotations and quizzes foster learning (RQ1). We hypothesized that annotations improve recall knowledge (H1.1) and that quizzes enhance comprehension knowledge (H1.2). Second, we investigate the extent to which annotations and quizzes influence presence (RQ2). We expected that annotations would boost presence (H2.1) and that quizzes would reduce presence (H2.2). Third, we explore to what extent annotations and quizzes affect cognitive load (RQ3). We assumed that annotations reduce extraneous load (H3.1) and that annotations increase germane load (H3.2). Unlike annotations, which are strongly grounded in CTML and signaling theory and have shown consisted effects on extraneous load, quizzes have not been theoretically linked to specific facets of cognitive load in IVR. Therefore, we formulated hypotheses for annotations but treated the effects of quizzes on cognitive load as exploratory.

4. Methods

The study was conducted in June 2023 and December 2023 with participants from two German grammar schools in the German state of Saxony-Anhalt.

4.1. Participants

In total, $N = 152$ 10th graders, aged 15 to 16 years, took part in the study. We excluded $n = 26$ participants who experienced technical issues, had insufficient language skills, or withdrew because to cyber sickness. The final sample therefore comprised $N = 126$ students. All of these participants completed the study in a classroom setting and spoke German fluently, the language of instruction. Of these participants, $n = 58$ (46%) identified as female, $n = 65$ (52%) identified as male, and $n = 3$ (2%) participants chose not to provide an answer.

4.2. Procedure

We integrated the study seamlessly into the physics lessons of the mentioned grammar schools. The participants first completed a 20-min pretest consisting of demographic questions, as well as recall and comprehension knowledge tests. Subsequently, the experimenters adjusted the VR headsets, and the participants completed a 5-min digital tutorial that explained the controls and the task to them. Then, the participants studied for 30 min in one of four VR environments structured according to four experimental conditions (see below). Allocation to the experimental groups was randomized using an experiment sheet on which one of the four conditions was indicated. The participants viewed the learning environment on standalone HTC Vive Focus III HMDs and used controllers for interacting with the user interface. After the intervention, all participants completed a 35-min posttest, that included cognitive load and cyber sickness scales, followed by recall and comprehension knowledge tests. Other surveys beyond the scope of this paper were also conducted.

Four experimental conditions make up the intervention of this study; see Fig. 1 for a visualization of all measurements and treatments. The participants in the control group had the learning environment without quizzes and annotations. The participants in the quiz group learned in the same environment that included quizzes but no annotations. The participants in the annotation group only received annotations, but no quiz questions. However, the annotations and quiz group participants were given annotations and quizzes. As time-on-task is an essential factor influencing performance in educational studies (Karweit & Slavin, 1982), we used digital timers in the learning environment to keep the duration equal across all conditions.

The study was incorporated into a single 90-min session within the physics curriculum of the two schools. The experimental groups consisted of 30 to 34 students (see Fig. 1). Participation was voluntary, and all students and their parents provided informed consent. A small number of students who chose not to participate in the study worked with an alternative text-based condition providing the same content and fall within the group of 26 students that were excluded from analysis.

4.3. Learning environment

All participants individually completed a VR science learning environment focused on wastewater treatment. This learning environment introduced participants to wastewater treatment through a virtual field trip that showed 360-degree videos of the different water basins.

Group (number of participants)	Pretest	Learning phase			Posttest
	Survey questions and knowledge tests	Tutorial	Annotations	Quizzes	Survey questions and knowledge tests
Control (n = 34)	●	●	/	/	●
Annotations (n = 32)	●	●	x	/	●
Quizzes (n = 30)	●	●	/	x	●
Annotations and Quizzes (n = 30)	●	●	x	x	●
Duration	20 minutes	35 minutes			35 minutes

Fig. 1. Procedure of the Experiment *Note.* Illustration of the experimental procedure, including approximate durations. Symbols: ● indicates a measurement or tutorial, x stands for a treatment, and / represents a missing treatment.

Moreover, the learning environment illustrated physical, chemical, and biological processes, with explanatory figures highlighting processes in facilities and under the water's surface. None of the participants had encountered this learning environment before. The topic of wastewater treatment was selected because environmental education is a key objective in the German school system. In addition, the topic fits well within an interdisciplinary unit that combines content from physics, biology, and chemistry, and is well suited for our study investigating explanatory figures. Knowledge about the water purification process was imparted at nine stations, which were presented as 360-degree videos of treatment facilities. There were stations for mechanical water purification (e.g., grit and grease removal), biological purification (e.g., the denitrification stage), and chemical purification (e.g., an activation tank with phosphorus addition). At each station, an interactive console prompted participants to view explanatory figures and select questions via the user interface, which were answered verbally by a recorded speaker. To generate annotations for the study, explanatory figures in the learning material were supplemented with additional cue words, colors, and arrows (see Fig. 2A). A multiple-choice question with three possible answers and one correct option (i.e., a single-choice question) was presented as a quiz at the end of each station (see Fig. 2B). Feedback on the correct answer was provided. The entire lesson was conducted in German; however, some parts of Fig. 2 were translated for the reader's convenience.

4.4. Instruments

First, we collected demographic data. Our questionnaire included questions about gender identity, vision, contact lens use, and computer and VR use.

Based on Bloom's taxonomy (1956), knowledge was assessed using recall and comprehension tests. Regarding subject matter, the knowledge tests focused on biological, chemical, and physical processes relevant to wastewater treatment.

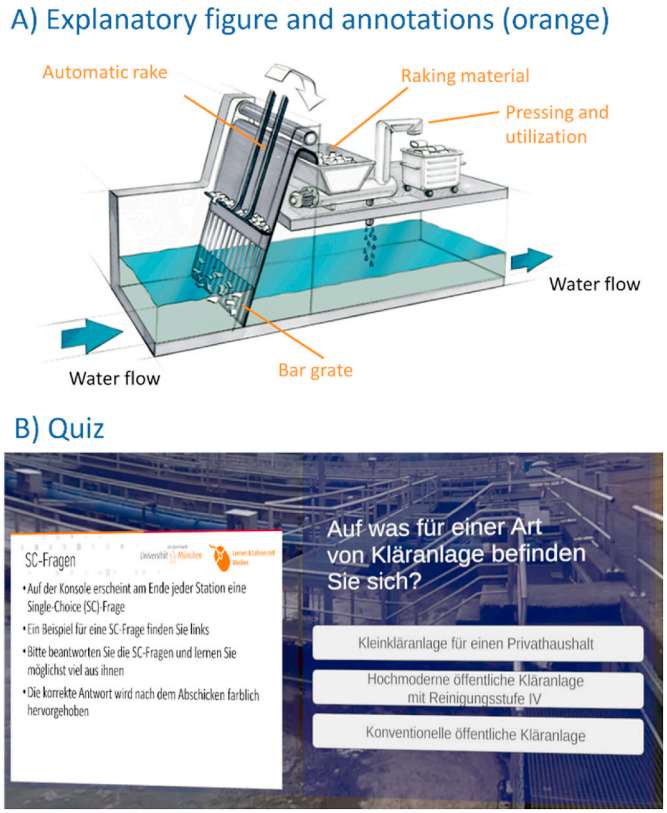

Fig. 2. Learning environment.

The recall test primarily measured the ability to recognize the information conveyed. It consisted of single-choice questions with three possible answers; only one was correct. An example item was as follows: "*What are the three types of treatment at wastewater treatment plants? a) Mechanical, biological, and chemical (correct) b) Mechanical, physical, and chemical, and c) Physical, chemical, and electrical.*" The pretest consisted of five items, while the posttest comprised the same five items plus five additional items. A percentage score ranging from 0 (no answer correct) to 100 (all answers correct) was calculated for the recall pretest and posttest. The recall posttest contained two rather easy items ($0.8 < p_i < 0.9$) while the other 8 items were in a medium range ($0.2 < p_i < 0.8$). The average item difficulty was $p_i = 0.656$.

The comprehension test assessed participants' ability to explain the mechanisms and describe processes and thus comprised three open-ended questions. One of the questions was: "*Where does wastewater come from, and what does it contain?*" The learners' answers were evaluated using a standardized scoring sheet. Like the recall test, a percentage score from 0 (no answers correct) to 100 (all answers correct) describes the participant's performance in the comprehension pre- and posttest. However, this score also allowed partial credit for each question. To ensure validity of the coding and the analysis, 10% of the comprehension tests were coded by two independent coders; a significant weighted Kappa of $\kappa = 0.862$ indicates a good inter-rater reliability. One of the comprehension posttest items was rather difficult ($0.1 < p_i < 0.2$) while the other 2 items were in a medium range ($0.2 < p_i < 0.8$). The average item difficulty was $p_i = 0.555$.

Cognitive load was assessed using an eight-item survey from Klepsch et al. (2017). This survey measures cognitive load with two items for intrinsic load (Cronbach's $\alpha = .782$), three for germane load (Cronbach's $\alpha = .505$), and three for extraneous load (Cronbach's $\alpha = .728$). This survey used a five-point Likert scale with scale anchors ranging from (1) "strongly disagree" to (5) "strongly agree". As this scale is quite frequently used, we kept it in the original composition although the α for the germane load was relatively low. The low α means that students' ratings notably distinguished between the extent to which they *had to highly engage themselves* and the extent to which they *had to think intensively about what things meant*, and that their evaluations of these two aspects of the germane cognitive load scale were less coherent. Therefore, results involving germane load should be interpreted cautiously and viewed as exploratory rather than as robust evidence for differences in germane load between conditions.

Presence was measured with a five-item scale from Makransky et al. (2017). This scale assesses the sense of physical presence experienced in VR learning environments (Cronbach's $\alpha = .837$). The scale was adapted to our virtual learning environment and featured items like "The virtual field trip to the water treatment plant seemed real to me". Participants rated their agreement to the items on a five-point Likert scale, with answer options from (1) "strongly disagree" to (5) "strongly agree".

Cyber sickness was assessed with a single question, describing the core symptoms of the condition, including nausea, dizziness, disorientation, and visual strain, based on Kourtesis et al. (2023). The item asked, "*How much cyber sickness did you experience while participating in the learning environment?*" Responses were given on a seven-point Likert scale ranging from (1) "no feeling" to (7) "extreme feeling". Single-item measures are commonly used in applied IVR field studies when time is severely limited (e.g., Weech et al., 2019), and short-form versions of the Simulator Sickness Questionnaire (SSQ) have been shown to capture the central variance of cyber sickness symptoms. Given the classroom time constraints of a 90-min physics lesson, the concise measurement represented a practical and commonly applied solution.

Students completed all tests and surveys on paper, to view color figures conveniently and easily provide handwritten responses. After the study, all tests and survey responses were entered as data and carefully reviewed. Scale means were calculated for all the described scales consisting of multiple items.

4.5. Statistical analyses (and power analysis)

We used R (Version 4.3.3; R Core Team, 2024) and the tidyverse package (Wickham et al., 2019) for data preparation, cleaning, and statistical analyses. For prior analyses, Pearson correlations and regression analyses were used. The rstatix package (Kassambara, 2023) and the function anova_test were used to compute ANCOVAs with a type III sum of squares. For RQ1, which addressed the effects of the annotations and quizzes on learning, ANCOVAs were performed for both knowledge measures. These ANCOVAs included cyber sickness and either the recall or comprehension knowledge pretest as covariates, with annotations and quizzes as fixed factors and the corresponding recall or comprehension posttests as dependent variables. RQ2 examined how annotations and quizzes influenced participants' sense of presence. We conducted ANCOVAs with presence as a criterion and annotations and quizzes as fixed factors. Cyber sickness and both knowledge pretests (recall and comprehension) served as covariates. RQ3 investigated the effect of annotations and quizzes on cognitive load. ANCOVAs were utilized here also, in which the two factors of annotations and quizzes were the independent variables and the different cognitive load scales were the outcomes. Cyber sickness, presence, and both recall and comprehension knowledge pretest scores served as covariates.

Given the final sample size of $N = 126$, we conducted a post-hoc power analysis using the pwr package (Champely, 2023). For a 2 × 2 ANCOVA with two covariates ($\alpha = .05$, power = 0.80), the analysis indicated that effects of $f = 0.256$ ($\eta_p^2 = 0.06$) can be detected.

4.6. Treatment check

As a treatment check we analyzed how well students engaged with the quizzes. On average, students completed 90 % of the quiz questions, and their first-attempt accuracy was 75 %.

5. Results

In this section we first report descriptive statistics and initial analyses regarding the framing conditions of the IVFT before addressing the research questions.

5.1. Descriptive statistics and prior analyses

We provide the descriptives for the full sample and the four different experimental groups in Table 1. Across all groups, students improved their recall and comprehension knowledge scores from pretest to posttest (see Table 1 and the electronic supplement, Tables 1–6 for the respective significance indicators). Paired-sample t-tests showed that improvements were statistically significant for recall knowledge, $t(125) = 9.37$, $d = 0.83$, $p < .001$, and comprehension knowledge, $t(125) = 9.20$, $d = 0.82$, $p < .001$.

Table 2 presents the intercorrelations among the knowledge tests, cognitive load scales, presence, and cyber sickness. The recall pre- and posttest knowledge scores were unrelated, whereas the comprehension pre- and posttests were significantly correlated. Cyber sickness correlated negatively with the recall posttest but was not associated with the comprehension posttest. Table 2 also shows the relationship between the different facets of cognitive load: Intrinsic load correlated positively with extraneous load, germane load correlated negatively with extraneous load, and germane load and intrinsic load were not related. Moreover, Table 2 reports on the associations between cyber sickness and cognitive load facets. Cyber sickness correlated positively with intrinsic and extraneous load, but not germane load while presence correlated positively with germane load and negatively with extraneous load.

We conducted prior analyses to examine gender differences in cyber sickness. Cyber sickness differed by gender, with female participants reporting higher cyber sickness scores ($M = 3.39$, $SD = 1.61$) than male participants ($M = 2.39$, $SD = 1.48$), $t(114) = 3.53$, $p < .001$.

Next, we examined the extent to which cyber sickness, gender, and posttest performance were related. We conducted a multiple regression analysis with the recall posttest as dependent variable, considering cyber sickness and gender identity as predictors. The global model was significant, and the model explained 9% of variance, $F(3, 117) = 3.88$, $p = .011$. The intercept was significant ($b = 0.69$, $p < .001$). Cyber sickness had a negative effect ($b = -0.02$, $\beta = -0.28$, $p = .008$), indicating that higher levels of cyber sickness were detrimental to recall posttest performance. Participants' pretest performance was not associated with posttest performance ($p = .261$) and gender identity did not have a significant effect ($p = .502$).

A second multiple regression used the comprehension posttest as a criterion, considering cyber sickness and gender identity as predictors. The global model was significant, explaining 19% of the variance, $F(3, 117) = 8.88$, $p < .001$. The intercept was significant ($b = 0.43$, $p < .001$). Performance in the comprehension pretest had a significant effect ($b = 0.45$, $p < .001$), supporting the finding that prior knowledge is an essential predictor for comprehension posttest performance. Gender identity ($p = .388$) and cyber sickness ($p = .139$) did not significantly affect the recall posttest outcome.

The analyses reported above indicate that cyber sickness and prior knowledge should be considered when analyzing the effects of annotations and quizzes.

5.2. To what extent do annotations and quizzes foster learning? (RQ1)

We examined whether annotations and quizzes affect the acquisition of recall knowledge. An ANCOVA was conducted with annotations and quizzes as fixed factors, while recall knowledge pretest and cyber sickness served as covariates. The overall model was significant,

Table 1
Descriptives in the full sample and across the experimental groups. Differences between the four experimental groups regarding the outcome measures and especially significance levels for the post-hoc comparisons can be found in the electronic supplement, Tables 1–6

Variable	Full sample	Groups			
		Control group ($n = 34$)	Quiz ($n = 30$)	Annotations ($n = 32$)	Annotations and quiz ($n = 30$)
Recall pretest (%)	44 (22)	49 (22)	44 (23)	42 (24)	42 (20)
Recall posttest (%)	66 (15)	63 (16)	61 (16)	67 (15)	72 (11)
Comprehension pretest (%)	36 (20)	34 (20)	38 (19)	34 (22)	37 (20)
Comprehension posttest (%)	47 (20)	42 (20)	51 (19)	46 (18)	49 (20)
Intrinsic load (SD)	2.58 (0.89)	2.81 (0.90)	2.38 (0.84)	2.52 (0.77)	2.57 (1.01)
Extraneous load (SD)	2.55 (0.78)	2.64 (0.79)	2.56 (0.82)	2.40 (0.78)	2.60 (0.77)
Germane load (SD)	3.73 (0.60)	3.64 (0.60)	3.64 (0.72)	3.83 (0.49)	3.80 (0.60)
Cyber sickness (SD)	2.85 (1.60)	3.16 (1.61)	2.40 (1.54)	2.94 (1.64)	2.90 (1.58)
Presence (SD)	2.99 (0.81)	3.07 (0.69)	2.63 (0.68)	3.26 (0.77)	2.95 (0.97)

Scale ranges: Knowledge 0-100 %, all cognitive load scales: 1 (low) – 5 (high).
Note: All significance indicators can be found in the electronic supplement, Tables 1–6

Table 2
Intercorrelations between knowledge tests, cognitive load, and cyber sickness.

Variable	1.	2.	3.	4.	5.	6.	7.	8.
1. Recall pretest	—							
2. Recall posttest	0.12	—						
3. Comprehension pretest	−0.07	0.12	—					
4. Comprehension posttest	0.08	0.33***	0.41***	—				
5. Intrinsic load	−0.19*	−0.10	−0.00	−0.16	—			
6. Extraneous load	−0.17	−0.03	−0.02	−0.11	0.38***	—		
7. Germane load	0.14	−0.00	−0.10	0.10	−0.02	−0.41***	—	
8. Cyber sickness	−0.07	−0.27**	0.04	−0.15	0.18*	0.28**	−0.12	—
9. Presence	−0.13	−0.02	−0.06	−0.12	0.05	−0.28**	0.25**	0.15

Note. Two-tailed Pearson correlations of the variables. ***$p < .001$, **$p < .01$, *$p < .05$.

$F(5, 118) = 4.58$, $R^2 = 0.14$, $p < .001$. A significant main effect for the factor of annotations was observed, $\eta_p^2 = 0.07$, $p = .003$. Cyber sickness also had a significant effect ($\eta_p^2 = 0.09$, $p < .001$), while the recall pretest ($\eta_p^2 = 0.01$, $p = .197$), quizzes ($\eta_p^2 = 0.00$, $p = .840$), and the interaction between annotations and quizzes ($\eta_p^2 = 0.02$, $p = .104$) were not significant. Fig. 3A illustrates the adjusted recall posttest scores among all experimental conditions. Post-hoc comparisons and contrasts in the scores between the four experimental groups could be found in the electronic supplement, Table 1. These results provide support for hypothesis H1.1, suggesting that annotations improve recall.

We also investigated whether annotations and quizzes affected the acquisition of comprehension knowledge. An ANCOVA was performed with annotations and quizzes as fixed factors; comprehension knowledge of the pretest and cyber sickness were included as covariates. The overall model was significant, $F(5, 118) = 5.80$, $R^2 = 0.20$, $p < .001$. However, neither quizzes ($\eta_p^2 = 0.02$, $p = .169$) nor annotations ($\eta_p^2 = 0.00$, $p = .916$) showed a significant main effect. Among the covariates, only the comprehension knowledge of the pretest was significant ($\eta_p^2 = 0.16$, $p < .001$), while cyber sickness ($\eta_p^2 = 0.02$, $p = .122$) and the interaction between annotations and quizzes ($\eta_p^2 = 0.00$, $p = .900$) were not. Fig. 3B displays the estimated marginal means of comprehension posttest scores across all conditions. Contrary to hypothesis H1.2, the inclusion of quizzes did not lead to a significant improvement in comprehension knowledge.

5.3. To what extent do annotations and quizzes influence presence? (RQ2)

For RQ2, we conducted an ANCOVA to determine how annotations and quizzes influence participants' presence perception. Presence was used as the outcome variable, quizzes and annotations as fixed factors. Cyber sickness and both pretest results were included in the model as covariates. The overall model was significant, $F(6, 117) = 2.31$, $R^2 = 0.12$, $p = .038$. Quizzes ($\eta_p^2 = 0.06$, $p = .009$) had a significant effect. Annotations ($\eta_p^2 = 0.03$, $p = .080$) and cyber sickness ($\eta_p^2 = 0.03$, $p = .088$), missed the significance level of $p < .05$; yet their p-values were $p < .1$, which means they show a marginal trend. Prior knowledge (recall: $p = .335$, $\eta_p^2 = 0.02$; comprehension: $p = .721$, $\eta_p^2 = 0.00$) and the interaction effect of annotations and quizzes ($\eta_p^2 = 0.00$, $p = .480$) were not significant. Fig. 4, illustrates the adjusted presence scores across all experimental conditions. Our post-hoc tests examined the main effects reported above in detail. While we couldn't find a convincing main effect of annotations that could support hypothesis 2.1, the reported main effect for quizzes gave support for hypothesis 2.2 that the inclusion of quizzes reduced the feeling of presence.

5.4. To what extent do annotations and quizzes affect cognitive load? (RQ3)

For each facet of cognitive load, we conducted a separate ANCOVA. In all analyses, annotations and quizzes were included as fixed factors, while recall and comprehension knowledge pretest scores, along with cyber sickness and presence, were included as covariates. All three ANCOVAs are visualized in Fig. 5.

Although we did not formulate a hypothesis for intrinsic load, we report its results first, as this construct is usually presented first. The overall model missed the significance threshold, $F(7, 116) = 1.85$, $R^2 = 0.08$, $p = .084$. Neither annotations ($\eta_p^2 = 0.01$, $p = .425$) nor

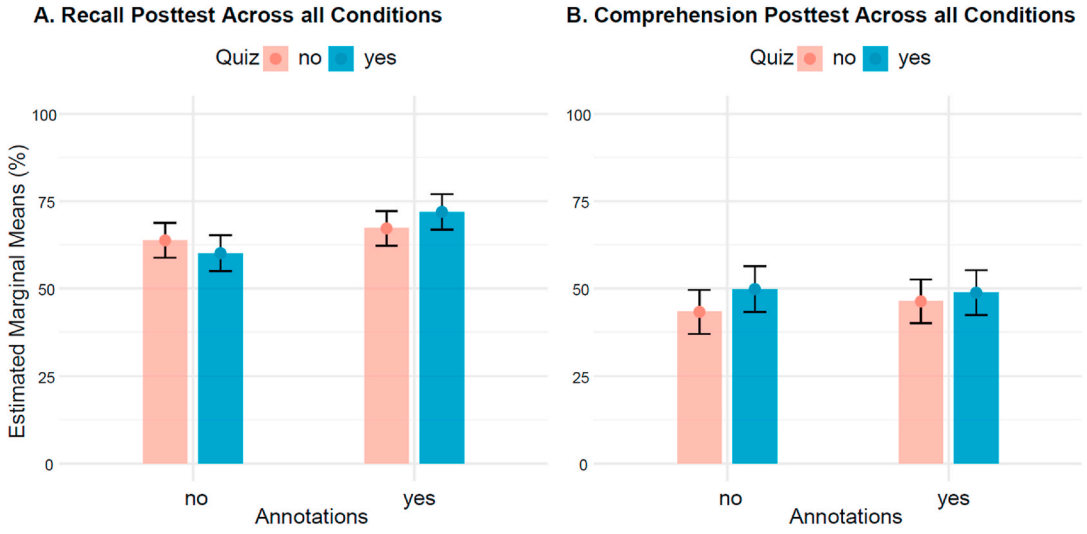

Fig. 3. Estimated Marginal Means from the 2x2-Factorial ANCOVA for the Recall Posttest (Panel A) and the Comprehension Posttest (Panel B).
Note: Post-hoc comparisons and contrasts in the scores between the four groups could be found in the electronic supplement, Tables 1 and 2.

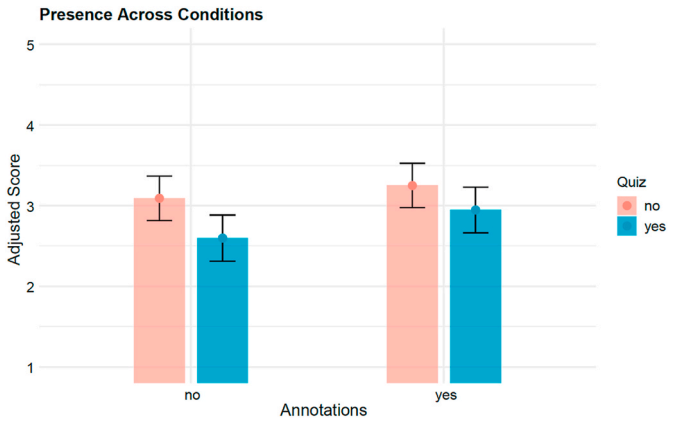

Fig. 4. Adjusted Scores from the 2x2-Factorial ANCOVA for Presence, Displayed Across all Experimental Conditions.
Note: Post-hoc comparisons and contrasts in the scores between the four groups could be found in the electronic supplement, Table 3.

quizzes ($\eta_p^2 = 0.01$, $p = .279$) showed significant effects. Similarly, the interaction between annotations and quizzes ($\eta_p^2 = 0.02$, $p = .127$), comprehension knowledge pretest ($\eta_p^2 = 0.00$, $p = .831$), presence ($\eta_p^2 = 0.00$, $p = .730$) and cyber sickness ($\eta_p^2 = 0.02$, $p = .104$) were not significant in this ANCOVA. The only significant variable was the covariate pretest recall knowledge ($\eta_p^2 = 0.04$, $p = .033$).

Our second ANCOVA model examining extraneous load was significant, $F(7, 116) = 3.90$, $R^2 = 0.18$, $p < .001$. Annotations did not have a significant effect on extraneous load ($\eta_p^2 = 0.00$, $p = .617$), thus, hypothesis H3.1 was not supported. Likewise, quizzes ($\eta_p^2 = 0.00$, $p = .991$) and the interaction between annotations and quizzes ($\eta_p^2 = 0.01$, $p = .326$) were not significant. Also, comprehension knowledge pretest scores ($\eta_p^2 = 0.00$, $p = .494$) didn't show significant effects. In contrast, recall knowledge pretest scores ($\eta_p^2 = 0.05$, $p = .018$), presence ($\eta_p^2 = 0.08$, $p = .002$) and cyber sickness were significant covariates for extraneous load ($\eta_p^2 = 0.06$, $p = .011$).

Our final ANCOVA investigated germane load. The overall model was significant, $F(7, 116) = 2.57$, $R^2 = 0.13$, $p = .017$. While neither annotations ($\eta_p^2 = 0.01$, $p = .301$) nor quizzes ($\eta_p^2 = 0.00$, $p = .731$) or the interaction between annotations and quizzes ($\eta_p^2 = 0.00$, $p = .989$) were significant, we found a significant effect of presence ($\eta_p^2 = 0.08$, $p = .002$). Prior knowledge in recall missed a significance level of $p < .05$ ($\eta_p^2 = 0.03$, $p = .053$) while prior knowledge comprehension ($\eta_p^2 = 0.01$, $p = .288$) and cyber sickness ($\eta_p^2 = 0.01$, $p = .436$) were clearly nonsignificant. Hypothesis H3.2 was therefore not supported by these analyses.

6. Discussion

We enriched an immersive HMD-based VR learning environment by integrating annotations and quizzes. Our objective was to provide participants with instructional support that would promote learning.

6.1. To what extent do annotations and quizzes foster learning? (RQ1)

Annotations increased recall knowledge acquisition but did not affect the acquisition of comprehension knowledge. This indicates that annotations can significantly influence the acquisition of recall knowledge even in a classroom-based IVFT setting. This finding is consistent with IVR studies showing that annotations primarily promote the acquisition of lower levels of Bloom's taxonomy of knowledge (Albus et al., 2021; Huang et al., 2024; Vogt et al., 2021), while they are not effective for fostering higher levels of Bloom's taxonomy (Li et al., 2023). However, the effect of annotations was moderate and smaller than the effect of the covariate cyber sickness. This further underscores how far learning outcomes as well as the effect of interventions in IVFTs are affected by cyber sickness.

The results are quite different for comprehension knowledge, which strongly depended on prior knowledge, showing large effect on the comprehension outcomes. This provides evidence for Matthew effect, originally introduced by Merton (1968) and later applied to education by Stanovich (1986) which describes that individuals with higher prior knowledge can more effectively integrate new knowledge. This Matthew effect primarily affected comprehension knowledge and thereby the higher level of the Bloom et al.'s (1956) taxonomy. Considering that our IVFT was an active learning setting that missed constructive activities according to the ICAP framework of Chi and Wylie (2014), it is reasonable to assume that possible effects of our intervention may have been suppressed by the impact of prior knowledge. Based on these analyses, we would recommend that future IVFT designs should also include constructive activities when targeting comprehensive knowledge outcomes.

Quizzes did not significantly improve the acquisition of recall and comprehension knowledge in our study. However, looking at the descriptive data in Table 1, we see that students in the quiz condition descriptively reached a higher knowledge gain (posttest – pretest) in recall knowledge than students in the control group – and that students with both annotations and quizzes reached the highest posttest score. Research on quizzes in IVR is scarce and although we conducted extensive research, we didn't find a study that implemented quizzes in

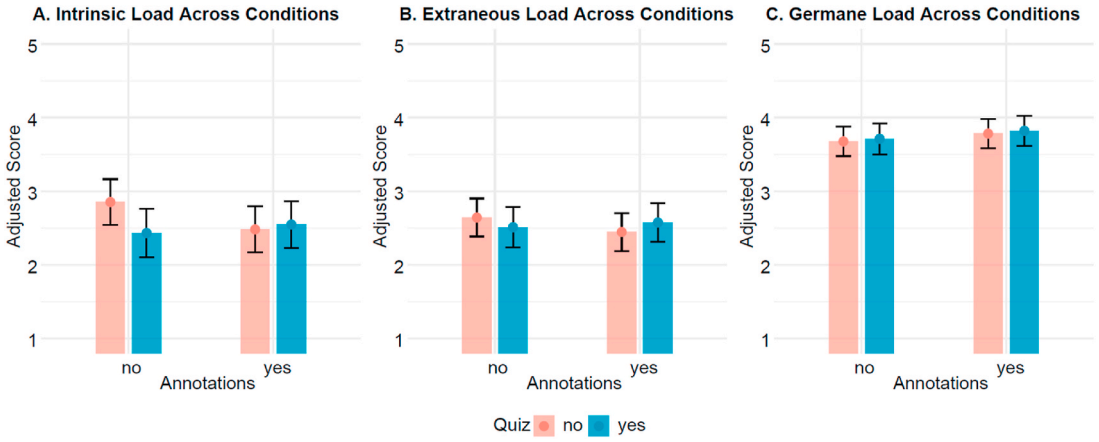

Fig. 5. Adjusted Scores from the 2x2-Factorial ANCOVA for the Intrinsic (Panel A), Extraneous (Panel B), and Germane Load (Panel C) and Displayed Across All Experimental Conditions.
Note: Post-hoc comparisons and contrasts in the scores between the four groups could be found in the electronic supplement, Tables 4–6

IVR presented through an HMD and compared learning outcomes of this group with a control group also using HMDs. Cecotti et al. (2024) applied quizzes in IVR but compared this setting with desktop-based VR; although the difference was not significant, students in the desktop-based VR condition descriptively scored higher in the quiz. Such differences were significant in Rai et al. (2019); however, students in the desktop-based VR received substantially more information than students using HMD.

There is also limited research on the effect of quizzes on learning outcomes in desktop-based VR. For example, Makransky, Mayer et al. (2020) found that learners with quizzes improved their knowledge. However, as they didn't use a control group, one cannot confirm whether learners with quizzes learned better than those without. Their effects are in line with our finding that all learners exposed to quizzes significantly improved in recall and comprehension knowledge. There are more studies that applied quizzes in desktop-based VR but usually these studies compared the desktop-based VR with another setting, e.g. classroom teaching (e.g. Lee & Wong, 2014). The limited available evidence in literature as well as our results raise the question of whether quizzes embedded in IVR offer meaningful additional benefits and to what extent they may stimulate advanced cognitive processes compared to a control group. Here it is important to consider the purpose of the quizzes: One intended function of our quiz was to activate learners during the 30-min IVFT. With 90 % of the quiz questions answered, we can conclude that this activation mechanism worked. After each of the nine stations, that is, approximately every three to 4 min, learners in the quiz condition received a single-choice quiz question. According to ICAP (Chi & Wylie, 2014) learners in the quiz condition found themselves in an active learning setting; however, their activity may not have been substantially higher than that of learners in conditions without quizzes. Relating this back to ICAP, learners in the other conditions may have been active enough, e.g. by progressing from one station to the next, so that quiz-based activation did not provide additional benefits. Such activation effects may emerge more clearly when compared with IVR environments without activation, for instance, when the experience is more film-like or when it provides other kinds of non-content specific interaction.

Another function of quizzes may be that they may provoke cognitive processes that enhance learning outcomes. Quizzes are supposed to promote learning by testing one's understanding and providing feedback (Dunlosky et al., 2013; Fiorella & Mayer, 2016). Feedback about correct answers may increase learners' perceived competence; and indeed, further analyses (Galler et al., 2026), indicate effects of quizzes on learners' competence experience, which are beyond the scope of this paper. Students solved about 75 % of the quiz questions correctly based on the learning materials. It seems reasonable to assume that also students without quizzes would have been able to reach such a high proportion based on the learning materials. The benefit of quizzes may have been that students who answered incorrectly might return to the materials for repetition. This, however, would have required allowing learners with quizzes more time for learning, which would have confounded the comparability of the settings because the quiz group might then have used more learning time. Thus, regarding learning outcomes, quizzes may not have been able to play to their strengths because of the required experimental standardization.

6.2. To what extent do annotations and quizzes influence presence? (RQ2)

Quizzes decreased the feeling of presence, as expected. This finding is consistent with the studies by Seprum et al. (2025) and Ahn and Noh (2025), in which different types of quizzes tended to reduce presence. A reasonable explanation for our results draws on literature suggesting that quizzes reduce presence primarily by interrupting participants (Seprum et al., 2025). This interpretation is also congruent with basic theoretical assumptions about presence. Witmer and Singer (1998) already pointed out that a learner's ability to focus on the VR content can be impaired by external distractions. Similarly, they emphasized that higher interface awareness can also reduce the sense of presence that participants experience (Witmer & Singer, 1998). This point is particularly relevant, as participants in the quiz groups likely spent more time reviewing the quiz questions through the interface, whereas participants in the groups without quizzes were instructed to explore the virtual world.

Considering the benefits of quizzes in less immersive learning settings, this raises the question of how best to integrate quizzes into IVR and to what extend their advantages might may outweigh the associated reduction in presence. Transforming quizzes into game elements and integrating them as game elements into a learning environment (e.g. by an increasing score) may add motivational benefits (Sailer & Homner, 2020), even if they decrease presence to a certain extent. Instructional designers could furthermore mitigate the reduction in presence by embedding quizzes implicitly within virtual environments, rather than presenting them explicitly, for instance, through overlayed survey questions (Frommel et al., 2015). Similarly, instructional designers could reduce the side effects of quizzes by integrating them more seamlessly into the narrative structure of the VR learning environment, or by embedding them in a spoken-response format that utilizes automatic speech recognition and an AI-based response evaluation.

There are indications that annotations may have enhanced the feeling of presence in our study, although this effect was not statistically significant at the $p < .05$ level and was characterized by a small effect size. Such a result aligns with the attentional benefits attributed to the use of signals and cues (Mayer, 2014; Schneider et al., 2018). In relation to Witmer and Singer (1998) and their concept of sensory fidelity, annotations may have contributed to the environmental richness of the learning environment and leaners' active search processes; they may especially have encouraged learners to explore the figures in more detail.

6.3. To what extent do annotations and quizzes affect cognitive load? (RQ3)

In line with the CAMIL model (Makransky & Petersen, 2021), our analysis shows that presence played a stronger role in shaping cognitive load than did instructional support. Presence was a significant covariate for extraneous and germane load, showing medium-sized effects in both ANCOVAs, and also showed significant correlations, whereas annotations and quizzes did not produce measurable effects.

While in traditional e-learning literature, the reduction of cognitive load through annotations is well established (Schneider et al., 2018), our annotations did not reduce extraneous load, which contradicted our expectations. Closer inspection of our ANCOVA results reveals three substantial covariate effects: presence (medium effect), cyber sickness (small to medium effect), and prior knowledge (small to medium effect). It seems reasonable to assume that at least presence and cyber sickness are particularly specific to IVR and therefore influence cognitive load more strongly than instructional characteristics.

In this line, two prior IVR studies found no effect on extraneous load (Albus et al., 2021; Li et al., 2023), whereas one study demonstrated a reduction in extraneous load (Huang et al., 2024). When considering the CAMIL framework again, we must acknowledge that all our learners were already immersed in IVR, working with HMDs in a 360° environment. This means that we did not investigate such dramatic differences in presence, as might be found when comparing desktop-based VR with IVR. The effects that Makransky and Petersen (2021, p. 946) describe in the context of CAMIL and cognitive load are derived from the larger visual field, in which a wider visual field of view can evoke higher levels of extraneous load. In contrast to this argumentation, the visual field does not differ between our four experimental conditions, as all learners used IVR. This context may have two potential implications. As the quizzes reduced presence and annotations slightly increased it, the

effects of both interventions on cognitive load might be mediated by presence, although we didn't find statistical evidence for this. Furthermore, we should reconsider the interaction between presence and extraneous load. If the visual field is same for all learners in IVR, extraneous load may result from cyber sickness and difficulties with orientation in the IVR. Therefore, the path between presence and extraneous load might operate in the opposite direction when all learners are in IVR. Our negative correlation between both variables may indicate such direction although our study was not designed to provide evidence for such an assumption. If we consider the assumptions of CAMIL regarding cognitive load in the context of IVR in more detail, we suggest splitting the concept of extraneous load in IVR. It may be helpful to distinguish between load evoked by the instructional environment, which reflects the traditional concept of extraneous load in paper- and desktop-based learning environments, and load caused by orientation and navigation difficulties in IVR as a second kind. When all learners are in IVR, such navigation and orientations difficulties within the environment may affect presence as well as extraneous load negatively while high presence levels may indicate the absence of such difficulties and are thereby associated with a lower extent of cognitive load.

Annotations did not increase germane load; the descriptively higher means were far away from significance levels ($p = .301$). The literature suggests that annotations can increase germane cognitive load by directing attention to relevant entities and their relationships, or by facilitating the integration and organization of information in short-term memory (Mayer, 2014; Schneider et al., 2018). Descriptively, the means for germane load among learners who received annotations were (slightly) higher than those of learners in other conditions, which may indicate that they processed more information and engaged in more complex tasks. However, not finding the expected effect is consistent with the results of the two available IVR studies, which also found that signals such as annotations did not improve germane load (Albus et al., 2021; Li et al., 2023).

For germane load, presence showed a medium effect and prior knowledge a small, nonsignificant effect. Besides the suppressing effects of the covariates, it is possible that learners were distracted by the seductive details inherent in the IVFT and therefore did not follow the guiding signals with full attention, which were intended to highlight relevant entities and their relationships. Another possibility is that the learning strategies and the described integration and organization processes in short-term memory are more difficult to activate in the IVR environments using HMDs than in multimedia learning without HMDs. Future studies should use eye-tracking methods to identify indicators for the interaction between annotations and germane load in IVR.

6.4. Effects of covariates

Besides the main effects, we found two covariates that had a significant impact on the analyses. The first was cyber sickness. While on a conceptual level the effect of cyber sickness is quite straightforward (learners that feel sick experience cognitive distraction or impairment), it provides significant obstacles for educational practice. Currently, well-documented gender differences exist with respect to cyber sickness, typically to the disadvantage of female learners (Kelly et al., 2023; MacArthur et al., 2021). Applying IVR in classroom learning may therefore impair girls' learning, which becomes even more crucial when we consider fields that already show gender imbalances, such as science classrooms (e.g. Olive et al., 2022). Thus, future research must investigate either new hardware solutions or better learning settings that reduce the phenomenon of cyber sickness, thereby allowing IVR to serve as a more successful tool for learning.

The second covariate was prior knowledge. Prior knowledge has long been recognized as an important factor for learning, as described in the seminal work of John R. Anderson (1981) and Amy Shapiro (2004). Therefore, prior knowledge is assessed in many studies and applied as a covariate in the respective experimental designs to control its effects. There is furthermore the issue of the extent to which prior knowledge moderates the effectiveness of interventions. Han et al. (2023) divided learners by prior knowledge based on a median split. Their students had to manipulate objects in the VR environment and if a learner could not identify the appropriate objects because of insufficient prior knowledge, this learner may have likely experienced difficulties in proceeding with the task appropriately. For learners who were familiar with the environment, the signaling in the Han et al. (2023) study may have (reasonably) been redundant. This is, however, quite different for (the case of) our learning environment, in which the annotations aimed at supporting learners' development of conceptual clarity, while progression through the learning environment was not dependent on learners' ability to identify the objects appropriately or not. This calls for using prior knowledge as covariate rather than as moderating condition. In this line, the effects of prior knowledge in our study are relatively straightforward and indicate that the Matthew effect (Merton, 1968) applies to IVR learning as well.

6.5. Limitations

Our analysis had an effective sample of 126 students, which allowed us to detect medium-sized effects ($f = 0.256$) with a power of 0.80 (calculated using the pwr package; Champely, 2023). Increasing the sample size to detect smaller effects would have been preferable but would have required approximately 791 students, according to the post-hoc power analysis. Such sizes are hardly feasible when working with IVR and most studies in our reference list had sample sizes between 100 and 150 participants, with several reporting fewer than 100. Increasing the sample size might have increased the statistical power of some tests and brought certain effects closer to significance, although this raises the question of whether such small effects would hold practical relevance. The effect of annotations on recall knowledge was medium but provided learners only a small advantage of a few points over the control group. This raises the question of whether it is more beneficial to focus first on interventions that produce larger effects even when based on smaller samples, rather than examining medium and small effects with large samples.

A further limitation of our study is that participants were subject to medium levels of cyber sickness. Although the single-item measure limits diagnostic granularity, single-item or short-form assessments are commonly used in applied IVR studies when time is constrained. At 30 min, our intervention lasted significantly longer than most other IVR interventions (e.g., Albus et al., 2021; Vogt et al., 2021), and it is known that cyber sickness intensifies with increasing duration. Thus, we believe that the long duration may have contributed to these levels of cyber sickness. Longer VR studies, like ours, are needed to assess the effects of VR-based learning in classroom contexts, such as STEM school lessons. In addition, our study showed, in line with theoretical arguments, that female participants suffered more from cyber sickness than male participants (Kelly et al., 2023; MacArthur et al., 2021). This point is interesting and corroborates findings from other studies (Stanney et al., 2020). As a result of these findings, we accounted for cyber sickness in our analysis by including it as a covariate in the ANCOVA models. This approach allowed us to report results that reflect a classroom-based IVR learning scenario of extended duration, in which cyber sickness is naturally more likely to occur.

Another limitation is that we did not use eye-tracking technology in our study. The learning environment was highly interactive, allowing participants to look in all directions, navigate linearly through the stations, as well as interact with the user interface. Consequently, measuring eye movements in such an interactive learning environment would have been a technically demanding and complex task (Ugwitz et al., 2022). Eye-tracking methods would have provided valuable insights into learners' attentional patterns related to annotations and quizzes, yet research on this topic remains scarce within educational IVR studies (Shadiev & Li, 2023). Future studies should utilize eye-tracking

approaches to gain a deeper understanding of learners' attentional processes, particularly when integrating explanatory figures in VR, which, as our study showed, appear to be well-suited for integrated STEM education.

7. Conclusion

We conducted a study with school students learning about a STEM topic in an IVR environment using HMDs. Students participated in an IVFT to a wastewater treatment plant, each using an individual HMD while learning together in the classroom. Experiences from the study implementations showed that this was a feasible setting for school learning and results of the study indicate that all learners received substantial learning gains.

We integrated annotations and quizzes as experimental factors within the IVFT to gain insights into their suitability as instructional support methods in this setting. Annotations fostered the acquisition of recall knowledge, but not comprehension knowledge. This finding emphasizes that annotations can promote lower levels of knowledge in Bloom's taxonomy (Bloom et al., 1956). Quizzes did not significantly promote knowledge acquisition in our study, although they are generally considered to foster learning through feedback and self-testing.

The analyses showed large effects of the framing conditions, especially prior knowledge and cyber sickness, but also presence, which were sometimes clearly greater than the effects of our interventions. This was particularly evident for the cognitive load measures, which were influenced by the framing conditions rather than by the interventions themselves, but also for the learning outcomes. We therefore suggest that these factors should be addressed more intensively in the development and design of IVFTs. While we didn't find interactions between prior knowledge and either intervention, it might be beneficial to target both support methods at students with lower prior knowledge in order to help them catch up with peers who have higher levels of prior knowledge. Such targeted interventions may, however, increase learning time requirements for students with lower prior knowledge. Overall, further research will determine how annotations and quizzes affect real-world systems in their actual operational environments. This may require a conceptual advancement of the implementation of quizzes and annotations so that they their benefits can also unfold in IVFTs.

While the effects of prior knowledge may take place in almost any learning setting, cyber sickness is specific to IVR and IVFTs. In our study, the students spent approximately 35 min in the IVR environment and only a few students (3 % of our sample) withdrew because of cyber sickness. However, female students reported higher levels of cyber sickness than male students. This effect may impair the learning of female students in IVFTs and therefore be a decisive factor in determining whether IVFTs are an appropriate means for classroom learning. However, better hardware, greater user familiarization, and improved instructional design may mitigate or dissolve this effect. Higher resolutions and more accurate adjustment of HMDs to (female) users' interpupillary distance (see Stanney et al., 2020) may be important first steps in this direction. Furthermore, with the increasing availability of HMDs in households, one could expect that individuals will become more familiar with HMD technology and, consequently, experience less cyber sickness. Finally, and this remains the task for us researchers, more research is needed to identify which attributes of IVR may contribute to cyber sickness, such as unanticipated or unfamiliar viewpoints, movements, or abrupt changes in position. To obtain more information about the contribution of such characteristics, eye-tracking (see Shadiev and Li, 2023) and/or electrodermal activity measures may reveal situations that require increased visual search processes and that may relate to stress.

Presence emerged as an ambiguous construct, because it was negatively affected by quizzes, but was an important covariate for the load measures. The CAMIL model (Makransky & Petersen, 2021) hypothesizes positive effects of presence on extraneous load when learning environments differ in their visual field. This wasn't the case in our IVFT as all students worked with HMDs. For such a context, we would hypothesize that learners who are able orient themselves and navigate easily in the IVFT experience higher presence, which, in turn, is associated with lower extraneous load. This hypothesis, however, requires systematic investigation in further studies. For learners completing quizzes, we would consequently hypothesize that increased navigation demands reduced their sense of presence.

CRediT authorship contribution statement

Maximilian C. Fink: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Carina Galler:** Writing – review & editing, Project administration. **Bianca Watzka:** Writing – review & editing, Resources, Project administration, Investigation, Conceptualization. **Bernhard Ertl:** Writing – review & editing, Supervision, Resources, Methodology, Funding acquisition, Formal analysis, Conceptualization.

Statement on open data and ethics

The data can be obtained upon request by contacting the corresponding author.

Ethics declaration

The study involving human participants was reviewed and approved by ethics committee of Universität der Bundeswehr München, project number EK UniBW M 23-52. It was also approved by the school inspection office of Saxony-Anhalt with project number 24-53/23. All procedures were performed in compliance with relevant laws and institutional guidelines, and the appropriate institutional committee(s) have approved them. Participation was voluntary, and all students and their parents provided written informed consent. The privacy rights of human subjects within this study are observed through full anonymization of the participant data and data secure storage of all data associated with this study.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research paper is funded by dtec.bw – Digitalization and Technology Research Center of the Bundeswehr [project RISK.twin]. dtec.bw is funded by the European Union – NextGenerationEU. We acknowledge financial support by Universität der Bundeswehr München. We are grateful to the Kommunalunternehmen Gemeindliche Einrichtungen und Abwasser Holzkirchen (GEA Holzkirchen KU), which made the erKlär-VR learning environment possible. We would also like to thank the Thurau wastewater treatment plant and the Steinhäule wastewater treatment plant association for allowing us to use the explanatory figures in the learning environment. Special thanks to Denis Frischbier, Lukas Hart, Christof Skwara, Maximilian Huisgen, and Leonard Hilbert for their support in developing the learning environment and conducting the lessons.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cexr.2026.100140.

References

Ahn, J., & Noh, G.-Y. (2025). Optimizing the educational potential of virtual reality: The effects of virtual reality and pop quizzes on spatial presence and efficacy perceptions. *International Journal of Human-Computer Interaction, 41*(2), 1107–1118. https://doi.org/10.1080/10447318.2024.2313274

Albus, P., Vogt, A., & Seufert, T. (2021). Signaling in virtual reality influences learning outcome and cognitive load. *Computers & Education, 166*, Article 104154. https://doi.org/10.1016/j.compedu.2021.104154

Alpizar, D., Adesope, O. O., & Wong, R. M. (2020). A meta-analysis of signaling principle in multimedia learning environments. *Educational Technology Research & Development, 68*(5), 2095–2119. https://doi.org/10.1007/s11423-020-09748-7

Andersen, M. S., & Makransky, G. (2020). The validation and further development of a multidimensional cognitive load scale for virtual environments. *Journal of Computer Assisted Learning, 37*(1), 183–196. https://doi.org/10.1111/jcal.12478

Anderson, J. R. (Ed.). (1981). *Cognitive skills and their acquisition*. Psychology Press. https://doi.org/10.4324/9780203728178.

Beege, M., Nebel, S., Schneider, S., & Rey, G. D. (2021). The effect of signaling in dependence on the extraneous cognitive load in learning environments. *Cognitive Processing, 22*(2), 209–225. https://doi.org/10.1007/s10339-020-01002-5

Belland, B. R. (2014). Scaffolding: Definition, current debates, and future directions. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of research on educational communications and technology* (pp. 505–518). Springer. https://doi.org/10.1007/978-1-4614-3185-5_39.

Bloom, B. S., Krathwohl, D. R., & Masia, B. B. (1956). *Taxonomy of educational objectives: The classification of educational goals* (1st ed.). D. McKay.

Cecotti, H., Huisinga, L., & Peláez, L. G. (2024). Fully immersive learning with virtual reality for assessing students in art history. *Virtual Reality, 28*(1), 33. https://doi.org/10.1007/s10055-023-00920-x

Champely, S. (2023). Pwr: Basic functions for power analysis. https://github.com/heliosdrm/pwr.

Chernikova, O., Heitzmann, N., Stadler, M., Holzberger, D., Seidel, T., & Fischer, F. (2020). Simulation-based learning in higher education: A meta-analysis. *Review of Educational Research, 90*(4), 499–541. https://doi.org/10.3102/0034654320933544

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

Coban, M., Bolat, Y. I., & Goksu, I. (2022). The potential of immersive virtual reality to enhance learning: A meta-analysis. *Educational Research Review, 36*, Article 100452. https://doi.org/10.1016/j.edurev.2022.100452

Cummings, J. J., & Bailenson, J. N. (2016). How immersive is enough? A meta-analysis of the effect of immersive technology on user presence. *Media Psychology, 19*(2), 272–309. https://doi.org/10.1080/15213269.2015.1015740

Dalgarno, B., & Lee, M. J. W. (2010). What are the learning affordances of 3-D virtual environments? *British Journal of Educational Technology, 41*(1), 10–32. https://doi.org/10.1111/j.1467-8535.2009.01038.x

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Eisenlauer, V., & Sosa, D. (2022). Pedagogic meaning-making in spherical video-based virtual reality – A case study from the EFL classroom. *Designs for Learning, 14*(1), 129–136. https://doi.org/10.16993/dfl.191

Fink, M. C., Sosa, D., Eisenlauer, V., & Ertl, B. (2023). Authenticity and interest in virtual reality: Findings from an experiment including educational virtual environments created with 3D modeling and photogrammetry. *Frontiers in Education, 8*, Article 969966. https://doi.org/10.3389/feduc.2023.969966

Fiorella, L., & Mayer, R. E. (2016). Eight ways to promote generative learning. *Educational Psychology Review, 28*(4), 717–741. https://doi.org/10.1007/s10648-015-9348-9

Frommel, J., Rogers, K., Brich, J., Besserer, D., Bradatsch, L., Ortinau, I., Schabenberger, R., Riemer, V., Schrader, C., & Weber, M. (2015). Integrated questionnaires: Maintaining presence in game environments for self-reported data acquisition. In *Proceedings of the 2015 annual symposium on computer-human interaction in play* (pp. 359–368). https://doi.org/10.1145/2793107.2793130

Galler, C., Fink, M. C., Watzka, B., & Ertl, B. (2026). *Motivational Effects of Annotations and Quizzes in a VR-Based Field Trip*, 2026 AERA Annual Meeting, Los Angeles, CA.

Han, J., Liu, G., & Zheng, Q. (2023). Prior knowledge as a moderator between signaling and learning performance in immersive virtual reality laboratories. *Frontiers in Psychology, 14*, Article 1118174. https://doi.org/10.3389/fpsyg.2023.1118174

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: Are self-testing and scheduling related to achievement? *Psychonomic Bulletin & Review, 19*(1), 126–134. https://doi.org/10.3758/s13423-011-0181-y

Howard, M. C., & Van Zandt, E. C. (2021). A meta-analysis of the virtual reality problem: Unequal effects of virtual reality sickness across individual differences. *Virtual Reality, 25*(4), 1221–1246. https://doi.org/10.1007/s10055-021-00524-3

Hu, L., Chen, G., Li, P., & Huang, J. (2021). Multimedia effect in problem solving: A meta-analysis. *Educational Psychology Review, 33*(4), 1717–1747. https://doi.org/10.1007/s10648-021-09610-z

Huang, G., Chen, C., Tang, Y., Zhang, H., Liu, R., & Zhou, L. (2024). A study on the effect of different channel cues on learning in immersive 360° videos. *Frontiers in Psychology, 15*, Article 1335022. https://doi.org/10.3389/fpsyg.2024.1335022

Huang, C. L., Luo, Y. F., Yang, S. C., Lu, C. M., & Chen, A.-S. (2020). Influence of students' learning style, sense of presence, and cognitive load on learning outcomes in an immersive virtual reality learning environment. *Journal of Educational Computing Research, 58*(3), 596–615. https://doi.org/10.1177/0735633119867422

Karweit, N., & Slavin, R. E. (1982). Time-on-task: Issues of timing, sampling, and definition. *Journal of Educational Psychology, 74*(6), 844–851. https://doi.org/10.1037/0022-0663.74.6.844

Kassambara, A. (2023). Rstatix: Pipe-friendly framework for basic statistical tests. https://CRAN.R-project.org/package=rstatix.

Kelly, J. W., Gilbert, S. B., Dorneich, M. C., & Costabile, K. A. (2023). Gender differences in cybersickness: Clarifying confusion and identifying paths forward. In *2023 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops* (pp. 283–288). https://doi.org/10.1109/VRW58643.2023.00067

Kennedy, R. S., Lane, N. E., Berbaum, K. S., & Lilienthal, M. G. (1993). Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The International Journal of Aviation Psychology, 3*(3), 203–220. https://doi.org/10.1207/s15327108ijap0303_3

Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and validation of two instruments measuring intrinsic, extraneous, and germane cognitive load. *Frontiers in Psychology, 8*, Article 01997. https://doi.org/10.3389/fpsyg.2017.01997

Klippel, A., Zhao, J., Oprean, D., Wallgrün, J. O., Stubbs, C., La Femina, P., & Jackson, K. L. (2020). The value of being there: Toward a science of immersive virtual field trips. *Virtual Reality, 24*(5), 753–770. https://doi.org/10.1007/s10055-019-00418-5

Kourtesis, P., Linnell, J., Amir, R., Argelaguet, F., & MacPherson, S. E. (2023). Cybersickness in virtual reality questionnaire (CSQ-VR): A validation and comparison against SSQ and VRSQ. *Virtual Worlds, 2*(1), 16–35. https://doi.org/10.3390/virtualworlds2010002

Kourtesis, P., Papadopoulou, A., & Roussos, P. (2024). Cybersickness in virtual reality: The role of individual differences, its effects on cognitive functions and motor skills, and intensity differences during and after immersion. *Virtual Worlds, 3*(1), 62–93. https://doi.org/10.3390/virtualworlds3010004

Krassmann, A. L., Melo, M., Pinto, D., Peixoto, B., Bessa, M., & Bercht, M. (2023). How are the sense of presence and learning outcomes being investigated when using virtual reality? A 24 years systematic literature review. *Interactive Learning Environments, 32*(7), 1–24. https://doi.org/10.1080/10494820.2023.2184388

Lee, E. A.-L., & Wong, K. W. (2014). Learning with desktop virtual reality: Low spatial ability learners are more positively affected. *Computers & Education, 79*, 49–58. https://doi.org/10.1016/j.compedu.2014.07.010

Li, W., Feng, Q., Zhu, X., Yu, Q., & Wang, Q. (2023). Effect of summarizing scaffolding and textual cues on learning performance, mental model, and cognitive load in a virtual reality environment: An experimental study. *Computers & Education, 200*, Article 104793. https://doi.org/10.1016/j.compedu.2023.104793

MacArthur, C., Grinberg, A., Harley, D., & Hancock, M. (2021). You're making me sick: A systematic review of how virtual reality research considers gender & cybersickness. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1–15). https://doi.org/10.1145/3411764.3445701

Makransky, G., & Lilleholt, L. (2018). A structural equation modeling investigation of the emotional value of immersive virtual reality in education. *Educational Technology Research & Development, 66*(5), 1141–1164. https://doi.org/10.1007/s11423-018-9581-2

Makransky, G., Lilleholt, L., & Aaby, A. (2017). Development and validation of the multimodal presence scale for virtual reality environments: A confirmatory factor analysis and item response theory approach. *Computers in Human Behavior, 72*, 276–285. https://doi.org/10.1016/j.chb.2017.02.066

Makransky, G., Mayer, R., Nøremølle, A., Cordoba, A. L., Wandall, J., & Bonde, M. (2020). Investigating the feasibility of using assessment and explanatory feedback in desktop virtual reality simulations. *Educational Technology Research & Development, 68*(1), 293–317. https://doi.org/10.1007/s11423-019-09690-3

Makransky, G., & Petersen, G. B. (2021). The cognitive Affective Model of immersive learning (CAMIL): A theoretical research-based model of learning in immersive virtual reality. *Educational Psychology Review, 33*(3), 937–958. https://doi.org/10.1007/s10648-020-09586-2

Makransky, G., Petersen, G. B., & Klingenberg, S. (2020). Can an immersive virtual reality simulation increase students' interest and career aspirations in science? *British Journal of Educational Technology, 51*(6), 2079–2097. https://doi.org/10.1111/bjet.12954

Makransky, G., Terkildsen, T. S., & Mayer, R. E. (2019). Adding immersive virtual reality to a science lab simulation causes more presence but less learning. *Learning and Instruction, 60*(11), 225–236. https://doi.org/10.1016/j.learninstruc.2017.12.007

Mayer, R. E. (2003). The promise of multimedia learning: Using the same instructional design methods across different media. *Learning and Instruction, 13*(2), 125–139. https://doi.org/10.1016/S0959-4752(02)00016-6

Mayer, R. E. (2014). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 43–71). Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.005.

Merton, R. K. (1968). The matthew effect in science. *Science, 159*(3810), 56–62. https://doi.org/10.1126/science.159.3810.56

Muhanna, M. A. (2015). Virtual reality and the CAVE: Taxonomy, interaction challenges and research directions. *Journal of King Saud University - Computer and Information Sciences, 27*(3), 344–361. https://doi.org/10.1016/j.jksuci.2014.03.023

Olive, K., Tang, X., Loukomies, A., Juuti, K., & Salmela-Aro, K. (2022). Gendered difference in motivational profiles, achievement, and STEM aspiration of elementary school students [Original Research]. *Frontiers in Psychology, 13*. https://doi.org/10.3389/fpsyg.2022.954325, 2022.

Park, J. H. (2020). Correlation between cognitive load, vividness and cyber sickness for 360-degree education video. *International Journal of Advanced Culture Technology, 8*(4), 89–94. https://doi.org/10.17703/IJACT.2020.8.4.89

Parong, J., & Mayer, R. E. (2018). Learning science in immersive virtual reality. *Journal of Educational Psychology, 110*(6), 785–797. Article edu0000241 https://doi.org/10.1037/edu0000241.

Parong, J., & Mayer, R. E. (2021). Learning about history in immersive virtual reality: Does immersion facilitate learning? *Educational Technology Research & Development, 69*(3), 1433–1451. https://doi.org/10.1007/s11423-021-09999-y

Plumed, R., González-Lluch, C., Otey, J. M., & Pérez-Belis, V. (2021). Training engineers in the use of constraints to create quality 2D profiles for 3D models. *Computer-Aided Design and Applications, 18*(3), 612–623. https://doi.org/10.14733/cadaps.2021.612-623

Polcar, J., & Horejsi, P. (2015). Knowledge acquisition and cyber sickness: A comparison of VR devices in virtual tours. *MM Science Journal, 2015*(2), 613–616. https://doi.org/10.17973/MMSJ.2015_06_201516

Poupard, M., Larrue, F., Sauzéon, H., & Tricot, A. (2025). A systematic review of immersive technologies for education: Learning performance, cognitive load and intrinsic motivation. *British Journal of Educational Technology, 56*(1), 5–41. https://doi.org/10.1111/bjet.13503

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Rai, B., Tan, H. S., & Leo, C. H. (2019). Bringing play back into the biology classroom with the use of gamified virtual lab simulations. *Journal of Applied Learning & Teaching, 2*(2). https://doi.org/10.37074/jalt.2019.2.2.7

Reid, D. J. (1990). The role of pictures in learning biology: Part 1, perception and observation. *Journal of Biological Education, 24*(3), 161–172. https://doi.org/10.1080/00219266.1990.9655135

Sailer, M., & Homner, L. (2020). The gamification of learning: A meta-analysis. *Educational Psychology Review, 32*(1), 77–112. https://doi.org/10.1007/s10648-019-09498-w

Schneider, S., Beege, M., Nebel, S., & Rey, G. D. (2018). A meta-analysis of how signaling affects learning with media. *Educational Research Review, 23*, 1–24. https://doi.org/10.1016/j.edurev.2017.11.001

Schrader, C., & Bastiaens, T. J. (2012). The influence of virtual presence: Effects on experienced cognitive load and learning outcomes in educational computer games. *Computers in Human Behavior, 28*(2), 648–658. https://doi.org/10.1016/j.chb.2011.11.011

Schweppe, J., Eitel, A., & Rummer, R. (2015). The multimedia effect and its stability over time. *Learning and Instruction, 38*, 24–33. https://doi.org/10.1016/j.learninstruc.2015.03.001

Sepich, N. C., Jasper, A., Fieffer, S., Gilbert, S. B., Dorneich, M. C., & Kelly, J. W. (2022). The impact of task workload on cybersickness. *Frontiers in Virtual Reality, 3*, Article 943409. https://doi.org/10.3389/frvir.2022.943409

Seprum, P., Chang, S.-C., & Wongwatkit, C. (2025). Exploring the impact of formative and summative assessment approaches in virtual reality emergency response learning. *Interactive Learning Environments*, 1–15. https://doi.org/10.1080/10494820.2025.2503227

Shadiev, R., & Li, D. (2023). A review study on eye-tracking technology usage in immersive virtual reality learning environments. *Computers & Education, 196*, Article 104681. https://doi.org/10.1016/j.compedu.2022.104681

Shapiro, A. M. (2004). How including prior knowledge as a subject variable may change outcomes of learning research. *American Educational Research Journal, 41*(1), 159–189. https://doi.org/10.3102/00028312041001159

Souza, K. A. F. D., & Porto, P. A. (2012). Chemistry and chemical education through text and image: Analysis of twentieth century textbooks used in Brazilian context. *Science & Education, 21*(5), 705–727. https://doi.org/10.1007/s11191-012-9442-z

Stainfield, J., Fisher, P., Ford, B., & Solem, M. (2000). International virtual field trips: A new direction? *Journal of Geography in Higher Education, 24*(2), 255–262. https://doi.org/10.1080/713677387

Stanney, K., Fidopiastis, C., & Foster, L. (2020). Virtual reality is sexist: But it does not have to be. *Frontiers in Robotics and AI, 7*, 4. https://doi.org/10.3389/frobt.2020.00004

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*(4), 360–407. http://www.jstor.org/stable/747612.

Stevens, J. A., & Kincaid, J. P. (2015). The relationship between presence and performance in virtual simulation training. *Open Journal of Modelling and Simulation, 3*(2), 41–48. https://doi.org/10.4236/ojmsi.2015.32005

Sweller, J., van Merrienboer, J. J. G., & Paas, F. G. W. C. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10*(3), 251–296. https://doi.org/10.1023/A:1022193728205

Thibaut, L., Ceuppens, S., De Loof, H., De Meester, J., Goovaerts, L., Struyf, A., Boeve-de Pauw, J., Dehaene, W., Deprez, J., Hellinckx, L., Knipprath, H., Langie, G., Struyven, K., Van De Velde, D., Van Petegem, P., & Depaepe, F. (2018). Integrated STEM education: A systematic review of instructional practices in secondary education. *European Journal of STEM Education, 3*(1). https://doi.org/10.20897/ejsteme/85525

Tsai, F.-H., Tsai, C.-C., & Lin, K.-Y. (2015). The evaluation of different gaming modes and feedback types on game-based formative assessment in an online learning environment. *Computers & Education, 81*, 259–269. https://doi.org/10.1016/j.compedu.2014.10.013

Ugwitz, P., Kvarda, O., Juříková, Z., Šašinka, Č., & Tamm, S. (2022). Eye-tracking in interactive virtual environments: Implementation and evaluation. *Applied Sciences, 12*(3), 1027. https://doi.org/10.3390/app12031027

Venkatakrishnan, R., Venkatakrishnan, R., Anaraky, R. G., Volonte, M., Knijnenburg, B., & Babu, S. V. (2020). A structural equation modeling approach to understand the relationship between control, cybersickness and presence in virtual reality. In *2020 IEEE conference on virtual reality and 3D user interfaces (VR)* (pp. 682–691). https://doi.org/10.1109/VR46266.2020.00091

Vogt, A., Albus, P., & Seufert, T. (2021). Learning in virtual reality: Bridging the motivation gap by adding annotations. *Frontiers in Psychology, 12*, Article 645032. https://doi.org/10.3389/fpsyg.2021.645032

Weech, S., Kenny, S., & Barnett-Cowan, M. (2019). Presence and cybersickness in virtual reality are negatively related: A review. *Frontiers in Psychology, 10*, 158. https://doi.org/10.3389/fpsyg.2019.00158

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), 1686. https://doi.org/10.21105/joss.01686

Witmer, B. G., & Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and Virtual Environments, 7*(3), 225–240.

Woerner, J. J. (1999). Virtual field trips in the Earth science classroom. In P. A. Rubba, J. A. Rye, & P. F. Keig (Eds.), *Proceeding of the annual conference of the association for the education of Teachers in Science* (Eds, pp. 1232–1244).

Xie, Y., Fang, M., & Shauman, K. (2015). STEM education. *Annual Review of Sociology, 41*(1), 331–357. https://doi.org/10.1146/annurev-soc-071312-145659