Fachhochschule Nordwestschweiz
Hochschule für Angewandte Psychologie

Sequencing Immersive Virtual Reality and Physical Laboratory Work for Titration Training: Effects on Learning Outcomes

MASTER-ARBEIT

2025

Autorin / Autor
Chalupny, Urs

Begleitperson
Christ, Oliver

Praxispartner*in
FHNW
Kontaktperson Christ, Oliver

frontiers

Sequencing Immersive Virtual Reality and Physical Laboratory Work for Titration Training: Effects on Learning Outcomes

Urs Chalupny [1,*]

[1] *Angewandte Psychologie, Fachhochschule Nordwestschweiz (FHNW), Olten, Switzerland*

Correspondence*:
Urs Chalupny
urs.chalupny@students.fhnw.ch

ABSTRACT

Higher education institutions preparing students for careers in the chemical industry face mounting challenges in providing quality laboratory training due to increasing student enrollment, limited resources, and safety concerns, prompting exploration of immersive virtual reality as a potential solution for chemistry education. This study investigated the effectiveness of different sequences for integrating virtual reality simulation with traditional laboratory work in teaching titration skills to undergraduate Life Sciences students. Ninety-one students participated in a mandatory course where they completed titration procedures in both virtual reality and physical laboratory environments in counterbalanced order, with knowledge assessed through tests at three timepoints and psychological factors measured using the Cognitive Affective Model of Immersive Learning framework. Results revealed differences in learning outcomes between virtual reality and physical laboratory conditions, nor did the sequence of exposure (virtual reality first versus laboratory first) affect knowledge acquisition. Students demonstrated high baseline knowledge (75% correct) with modest but equivalent gains regardless of learning modality. Path analysis confirmed only one significant relationship from the theoretical model: higher sense of presence in virtual reality was associated with reduced cognitive load. Individual differences among participants accounted for 76% of the variance in learning outcomes, while experimental conditions explained less than 1%. These findings demonstrate that immersive virtual reality can serve as a pedagogically equivalent alternative to traditional laboratory training for procedural chemistry skills, supporting its potential as a scalable complement to physical laboratory experiences when resources are constrained or access is limited.

Keywords: Virtual Reality, higher education, Laboratory, Immersive Learning

1 INTRODUCTION

The development and mastery of practical laboratory skills remain a cornerstone of education in chemistry and related disciplines, crucial for equipping future professionals for the demands of the labor market. A growing emphasis is placed on shifting from traditional, content-focused curricula towards competency-based education (Campbell et al., 2022; European Commission and PwC, 2020). Laboratory environments are pivotal in this shift, fostering not only declarative and procedural knowledge but also transversal skills and higher-order thinking, enabling a holistic educational

approach (Agustian et al., 2022). The recent COVID-19 pandemic underscored the value of hands-on experience, as limited access to physical labs negatively impacted perceived learning outcomes and confidence in practical skills (Finne et al., 2022; Hyde, 2025). However, higher education institutions are facing major challenges due to steadily increasing student numbers. In Switzerland alone, the number of bachelor's students at universities of applied sciences rose by approximately 30% between 2011 and 2021 (Wolter et al., 2023), a trend that is also observed globally (UNESCO, 2024). This development poses significant challenges for institutions aiming to maintain or even expand this form of hands-on training. In the case of laboratory education, it also affects access to lab facilities.

In this situation, providing high-quality learning opportunities in traditional laboratories faces significant hurdles and discussions even among specialists (Bretz, 2019). Restricted lab space, large class sizes (Tee et al., 2018), limited availability of qualified personnel, the complexity of effective pedagogical implementation (Agustian et al., 2022), substantial material costs, and inherent safety concerns (Bennett and O'Neale, 1998) collectively challenge educators. Since the early days of education, when direct experiences essential for learning (Meyer-Drawe, 2003) are not possible, instructional methods have been employed to convey declarative or procedural knowledge or to foster the development of competences (Zierer and Seel, 2012). In this context, virtual simulations have increasingly come into use in recent years. Virtual Reality (VR) represents one specific form along the reality–virtuality continuum (Milgram et al., 1995), in which reality is largely replaced by a simulation. In VR, this replacement is primarily achieved through the use of head-mounted displays (HMD), which block the user's visual access to the physical environment and replace it with a stereoscopic two-dimensional image. In our case, we attempted to create a laboratory simulation with the aim of replicating the physical environment, allowing learners to engage in the intended experiences within a virtual environment. These learning experiences can be accessed independently of the physical location of the educational institution and can be made available asynchronously—essentially anywhere that the necessary technological infrastructure is available. The previously described challenges associated with limited access to laboratory training—along with the core educational intention—could therefore be largely addressed through the use of VR.

Immersive virtual reality (iVR)—digitally created environments offering interactive and "real-feeling" experiences (Bailenson, 2018)—presents a promising technological avenue to address several of these challenges. Specifically, iVR can mitigate ongoing material costs, requires less physical space, allows virtually unlimited repetition of experiments (Dunnagan et al., 2020), and significantly reduces safety risks (Bøg Petersen et al., 2022; Concannon et al., 2019) and can give access to learning experiences otherwise too dangerous or expensive (Tee et al., 2018). Some studies show non-inferor learning outcomes compared to traditional labs. This makes iVR a potentially cost-effective equivalent (Dunnagan et al., 2020). Successful adoption requires careful consideration of virtual environment design, effective instructional pathways, and the development of necessary competencies among both educators and students (Radianti et al., 2020). Even though the design elements required to create effective learning experiences—both in iVR and in physical laboratories—are relatively well researched, the question remains as to how these two environments can be optimally aligned to ensure that learning and the transfer of skills from one setting to the other actually take place, rather than being hindered.

Titration, a fundamental quantitative chemical analysis technique used to determine the concentration of an analyte, serves as a prime example of a core laboratory competency (Sheppard, 2006). Mastering titration demands a blend of skills: students must understand the underlying

chemical principles (declarative knowledge), correctly execute the procedural steps using specific equipment (procedural knowledge), and apply mathematical skills for accurate calculation. This aligns with Johnstone (1993) framework, requiring integration of macroscopic observations, submicroscopic understanding, and representational competence (symbols, equations) and makes the titration an excellent object for learning different skills and comparing their acquisition in different environments. The specific demands differ between environments: physical labs emphasize fine motor skills and safety diligence, while virtual labs require initial proficiency with the iVR interface before engaging with the chemical procedure itself (Johnson-Glenberg et al., 2023).These subtle differences between learning environments are highly relevant when considering the transfer of the competences to be developed. A key question is whether the similarities in the actions performed are sufficient to enable the transfer of procedural titration knowledge across environments. Research on transfer emphasizes underlying principles, identical elements, variability in stimuli, and altered learning conditions as critical factors (Rahman, 2020). In this regard, learning through different environments—but with a shared core task—may prove to be particularly effective.

The potential of iVR in education has been explored for decades, aiming to enhance motivation, engagement, and learning (di Lanzo et al., 2020). However, research reveals a heterogeneous landscape regarding its effectiveness (Won et al., 2023). While iVR often boosts affective outcomes like motivation, engagement, interest, and presence (Liu et al., 2023; Makransky et al., 2021; Parong and Mayer, 2018), its impact on cognitive learning outcomes is ambiguous. Several meta-analyses and systematic reviews suggest that iVR can be particularly advantageous for learning procedural skills compared to less immersive methods like presentations or lectures (Conrad et al., 2024; Hamilton et al., 2021). Studies focusing on procedural tasks often report positive effects (Concannon et al., 2019; Hamilton et al., 2021), and VR training has shown potential as a precursor to real-world task performance, indicating successful transfer in different contexts (Hamilton et al., 2021) and especially in the laboratory (Bøg Petersen et al., 2022) . The fidelity (physical and cognitive similarity to the real task) of the simulation is considered crucial for facilitating this transfer between the simulation and the real-world task (Bøg Petersen et al., 2022; Johnson-Glenberg et al., 2023; Levac et al., 2019). Although a trend appears to be emerging for developping Skills in VR, these findings must be interpreted with caution, as the studies differ considerably in their methodologies, the interventions used, and the operationalization of learning outcomes (Liu et al., 2023; Radianti et al., 2020).

Conversely, when acquiring declarative knowledge, iVR has sometimes led to poorer outcomes compared to traditional methods like slideshows or text (Bøg Petersen et al., 2022; Makransky and Petersen, 2019; Parong and Mayer, 2018). A review by Tusher et al. (2024) stated as well, that the influence of VR on cognitive and affective learning outomes appeared to be “subdued” compared to it’s beneficial effects on motor skills. Numerous empirical investigations (Makransky and Petersen, 2019; Parong and Mayer, 2021a,b) have drawn upon Sweller’s Cognitive Load Theory (Sweller et al., 2011) to explain how environmental factors influence learning outcomes, particularly in the acquisition of declarative knowledge. The richness and interactivity of iVR environments, while potentially engaging, might introduce extraneous cognitive load (processing irrelevant information) or increase intrinsic load (complexity of interaction), overwhelming the learner’s working memory capacity and hindering knowledge construction (Bøg Petersen et al., 2022; Makransky and Petersen, 2019; Parong and Mayer, 2021a). This raises the question of optimal sequencing of the learning experiences; might prior real-world experience reduce the cognitive load in VR or vice-versa? Furthermore, the novelty effect, where unfamiliarity with the technology itself consumes cognitive resources, can negatively impact initial learning sessions (Hamilton et al., 2021; Parong and Mayer, 2021a). Therefore, effective

instructional design within VR, applying principles like signaling (highlighting key information) and spatial/temporal contiguity (presenting related information together) derived from multimedia learning theory (Peeters et al., 2023), is critical but often under-implemented (Liu et al., 2023; Radianti et al., 2020).

To better understand learning mechanisms within immersive environments like iVR, the Cognitive Affective Model of Immersive Learning (CAMIL), developped by Makransky and Petersen (2021), provides a comprehensive framework for understanding the interplay of technological, cognitive, and affective factors that influence learning outcomes. It englobes technological factors, like immersion, that facilitate the feeling of presence in the virtual environment and enable to some degree the agency of the user and the different possibilities to interact and to change and control objects in the simulation. These two concepts, presence and agency, are defined as affordances of iVR and are the factors that influence the affective components of the model. Previous studies found that a high sense of presence was related with higher levels of situational motivation (Makransky and Lilleholt, 2018). Another important cognitive factor is the extraneous cognitive load (CL) depending on the virtual environment and the degree of stimuli and cues that have to be treated by the user when navigating and interacting with the simulation. This CL, as theorized in the CLT, would occupy cognitive resources that would be otherwise free and accessible for the learning tasks. So the design of the virtual environment should be optimized in regards to minimize distractions and maximize cues to engage in the learning task. The CAMIL Model states that a higher motivation and lower CL should have a directed influence on learning (Petersen et al., 2022). The model also states that external factors like low usability scores, measured eventually with the System Usability Scale (Lewis, 2018) or personality-related factors such as technology affinity and behavioral intention toward technology use, as measured by the TAEG questionnaire (Karrer-Gauß et al., 2009) could impair learning with iVR (Makransky and Petersen, 2021).

Given that the CAMIL Model describes external factors as important determinants of the learning process in iVR, this study seeks to elucidate how these influences manifest and interact in educational contexts.These different relationships between technological and intrapersonal factors are crucial when trying to understand the emergence of learning in VR and one aim of this study is to better understand these possible influences. We assume that the acceptance of the use of technology in the specific case could be a predictor of presence in iVR. Links between acceptance and performance when using VR devices were shown (Barrett et al., 2023; Guo et al., 2025) and we hypothzise that the perception of the technology should have a direct influence on the presence in iVR.

H1: Technology affinity (TA) should have a positive relation to percieved presence in iVR. TA is defined as a personality trait that manifests in a positive attitude, enthusiasm, and confidence toward technology (Karrer-Gauß et al., 2009) and can be measured with a short questionaire developped by Karrer-Gauß et al. (2024).

In its original formulation, the CAMIL model posited that presence influences extraneous cognitive load, specifically suggesting that presence would influence extraneous cognitive load stemming from the environment (Makransky and Petersen, 2021). However, this direct relationship has not yet been empirically confirmed and was questioned in the study by Petersen et al. (2022), who found a negative effect of cognitive load on presence rather than the other way around. This indicates that high cognitive load from interaction decreased users' feelings of presence, challenging the presumed direction of influence in that context. It should be noted, however, that their study involved a learning task focused exclusively on declarative knowledge. Furthermore, while not

empirically demonstrated in their specific VR training scenario, it has been theoretically argued and suggested that factors enhancing subjective experience, such as improved fidelity or relevant cues, can contribute to reducing cognitive load (Cooper et al., 2021).Consistent with this, a systematic review of evidence-based design and pedagogical principles in educational VR environments, drawing on frameworks such as multimedia and generative learning, specifically discusses signaling principles, including the use of textual annotations as an adjunct technique, and references studies linking signaling in VR to cognitive load (Oje et al., 2025). Based on these considerations, and acknowledging the need for further empirical evidence on this specific link, we assume that in the context of our study, a skill-learning task supported by explicit visual and textual cues, the original hypothesis that presence reduces cognitive load may still hold. We posit that a strong sense of presence, facilitated by these cues, could minimize the extraneous cognitive load associated with navigating and performing complex procedures in the virtual environment.

H2: Presence (Pre) should have a negative impact on extraneous CL in iVR.

While empirical findings regarding the direct link between presence and intrinsic motivation are not always directly supported (e.g. Petersen et al., 2022) , found physical presence predicted situational interest, which covaried with intrinsic motivation, but not intrinsic motivation directly in their specific model), there is broader support that the use of technologies like VR is associated with a range of positive affective and motivational reactions(Chirico and Gaggioli, 2019). Numerous studies have reported outcomes such as increased motivation(Bohne et al., 2021), self-efficacy(Thisgaard and Makransky, 2017), enjoyment(Cooper et al., 2021), interest (Petersen et al., 2022), and positive affect (Chirico and Gaggioli, 2019) stemming from engagement with VR/AR environments. These positive affective and motivational states are considered important, as theoretical frameworks and empirical evidence suggest they can, in turn, influence learning outcomes.

H3: Presence should have a positive relation to motivation (Mo).

H4: Cognitive Load and motivation should have direct influence on the learning outcome in iVR.

Titration, requiring both procedural execution and conceptual understanding, represents an ideal case study given the mixed findings on iVR effectiveness for different knowledge types. Direct comparisons between VR and real-life training exists (Bøg Petersen et al., 2022; Kaplan et al., 2021), but these studies employed usually a single-session design. Regarding the optimal integration of these modalities within an authentic educational context, the question remains open how to design such learning interventions. Specifically, how does the sequence in which students engage with iVR simulation and traditional laboratory practice impact their learning of a complete procedure like titration? Real laboratory work and learning is crucial (Hyde, 2025), but is there a sequencing or learning condition effect when designing the instruction between the iVR Simulation and the laboratory?

Most studies compare VR against a single alternative or focus on isolated skill components, often outside regular curricula (Hamilton et al., 2021; Radianti et al., 2020). Few have directly compared the sequential application (e.g., VR-first then Lab vs. Lab-first then VR) within an established university course, measuring learning outcomes for the entire procedure. Understanding this sequence effect is crucial for deriving practical guidelines on how best to leverage iVR as a complement to, rather than a replacement for, traditional lab work (Bøg Petersen et al., 2022; Concannon et al., 2019).

This study aims to address this gap by investigating the effectiveness of two different sequential learning approaches for teaching titration to undergraduate Life Sciences students within their mandatory "Basic Laboratory Techniques" module. Specifically, we compare the impact on knowledge acquisition when students first perform the titration procedure in an immersive virtual reality simulation followed by the traditional laboratory (VR-Lab sequence), versus performing the traditional laboratory exercise first followed by the VR simulation (Lab-VR sequence).

Drawing on cognitive load theory and the principle of scaffolding (Oje et al., 2025), we hypothesize that the VR-Lab sequence may offer advantages over the Lab-VR sequence. In the VR-Lab condition, students can first explore the procedure in a low-stakes environment where errors have no material consequences. This initial exposure may serve as cognitive scaffolding, allowing students to focus on understanding the procedural sequence before managing the additional demands of physical manipulation and safety concerns in the real laboratory. Conversely, the Lab-VR sequence may result in the virtual environment being perceived as redundant after real-world experience, potentially reducing engagement and learning gains.

H5: Students in the VR-Lab sequence will demonstrate greater knowledge gains compared to those in the Lab-VR sequence, as the virtual environment provides cognitive scaffolding that prepares students for the complex demands of real laboratory work.

In summary, this study makes both theoretical and practical contributions to understanding VR-based laboratory education. Theoretically, we extend the CAMIL model by testing its applicability to procedural skill learning in chemistry education, specifically examining whether the relationships between technology affinity, presence, cognitive load, and motivation hold in this applied context. Practically, we address a critical gap in instructional design by investigating how the sequencing of VR and physical laboratory experiences affects learning outcomes—a question of immediate relevance to educators facing resource constraints and seeking evidence-based integration strategies for emerging technologies. These two investigations are complementary: understanding the cognitive-affective mechanisms in VR learning (CAMIL) can help explain why certain sequences might be more effective, while the sequencing comparison provides ecological validity for the CAMIL model in authentic educational settings.

2 METHODS

The study was conducted as part of the Bachelor of Science in Life Sciences degree program at the University of Applied Sciences Northwestern Switzerland (FHNW). It was integrated into the foundational module 'Practical Basic Laboratory Techniques' (Praktikum Grundlagen Labortechniken), a mandatory course for all students organized during their first semester.

2.1 Participants

A total of 92 students initially participated in the study. One participant discontinued the VR condition due to experiencing discomfort and was subsequently removed from the sample. The final sample consisted of 91 participants who completed all stages of the study.Of these, 57 participants (62.6%) identified as female, 33 (36.3%) as male, and one (1.1%) as non-binary. The mean age for the entire sample was 21.82 years. There was no significant difference in mean age between the two largest gender groups ($t_{\text{Welch}}(87.48) = -0.17, \quad p = 0.86, \quad n_{\text{obs}} = 90$). Regarding the highest educational qualification attained, 36 participants (39.6%) reported a 'Berufsmatura' (vocational baccalaureate),

44 (48.4%) a 'Fachmatura' or 'gymnasiale Matura' (specialized or academic baccalaureate), 4 (2.2%) a Federal Diploma of Vocational Education and Training ('Eidgenössisches Fähigkeitszeugnis', EFZ), and 2 (2.2%) reported qualifications from higher vocational education. One participant (1.1%) already held a Bachelor's degree. Two participants (2.2%) selected 'Other,' specifying 'Fachhochschulreife' or 'Fachmaturität' (types of university of applied sciences entrance qualifications). Concerning prior experience with VR headsets, 34 participants (37.4%) reported having never used one before. 42 participants (46.2%) indicated having very limited prior experience. 13 participants (14.3%) had used a headset multiple times but indicated needing some support, while 2 (2.2%) described themselves as experienced and largely autonomous users. At the time of the study, 3 participants owned a personal VR headset.

2.2 Materials

The welcome, introduction, and all written tests were administered in a lecture hall on the Muttenz campus, with tables arranged in six rows and eight tables per row, divided by a central aisle. The VR condition was also conducted within this lecture hall. Upon arrival, participants discovered that VR headsets had been strategically placed in pairs on the tables, automatically establishing the two-person teams that would work together for the duration of the experimental session (VR and Lab). Care was taken to maximize the distance between VR stations; consequently, three stations were set up in each half of the room, separated by an unoccupied row of tables. Some research team members were positioned in the front row of the classroom to monitor and support students in the VR condition, observing both the physical classroom environment and the students' virtual activities via computer displays. Other team members were stationed in the chemistry laboratory to instruct and support students in the traditional lab condition (cf. Figure 1 for an overview of the study procedure).

2.2.1 Virtual environment and materials

The virtual environment replicates a real laboratory and includes all necessary tools and objects for performing the titration of an unknown solution. The simulation was developed in Unity 6, supplemented with objects from the Asset Store and some Models were developed in Blender 3.0. The workspace is situated within a fume hood and features a burette, measuring cylinder, pipette, beakers, funnel, various solutions, a magnetic stirrer, and a pipette bulb. Integrated on the right side of the fume hood was a series of tutorial videos explaining and demonstrating the handling of key objects, grasping and holding techniques, and movement within the virtual space. Adjacent to this, a SharePoint interface provided access to an Excel spreadsheet for recording measured and calculated solution volumes during the experiment. The simulation is structured into three distinct sequential phases. The first phase serves as an orientation period ("arrival"), allowing participants to familiarize themselves with the environment, view optional tutorial videos, and interact freely with the various objects. There were no time constraints or restrictions on interaction during this phase. Participants independently decided when to initiate the second phase and commence the guided titration. During the guided titration phase, instructions were displayed solely as text on the wall of the fume hood. Depending on the current instruction, visual cues (such as yellow outlines, arrows, or flashing indicators) highlighted the corresponding objects or relevant controller buttons. Successful completion of an action was acknowledged by an auditory cue, after which the instruction advanced to the next step. Students were thus guided through the entire titration process, which precisely mirrored the procedural protocol of a titration performed in a real laboratory. Throughout this

phase, students were required to transfer three measured burette volumes to the Excel spreadsheet. These entered values were not automatically verified, and no direct feedback on their accuracy was provided within the simulation. Data entry for each volume was confirmed by pressing a designated button, which concluded the specific instruction and displayed the subsequent one. This learning phase concluded with the final completion of the Excel spreadsheet, including the measured volumes and the calculated concentration value. Following the learning phase, a test phase was initiated by pressing another designated virtual button. In this phase, participants had to perform the same sequence of actions again using an unknown solution, transferring the measured values back into the Excel spreadsheet. The experiment concluded with the final transfer of these values. During this phase, no additional guidance or feedback was provided.

2.2.2 Laboratory environment and materials

The titration procedure was also performed by all students in the university's physical laboratory. Analogous to the virtual lab procedure, students received printed instructions that guided them step-by-step through the experiment. All necessary materials and equipment were provided at the laboratory benches. Students worked in pairs and conducted the experiment under the supervision of instructors. The procedure was carried out in accordance with applicable safety regulations and instructor guidelines. Participants were required to record the measured volumes on a protocol sheet and calculate the concentration of the unknown substance. Instructors were available to answer questions and provide assistance as needed.

2.3 Questionaires and scales

All questionaires were created with the Enterprise Feedback Suite from Tivian and made accessible via QR code. Participants completed the tests on their personal smartphones, which allowed for flexible administration. A standardized questionnaire was employed, structured such that specific sections were unlocked depending on the testing timepoint and condition (VR or Laboratory). Data collection was conducted anonymously, and all participants provided informed consent prior to participation. Participants identified themselves using an anonymous code assigned at the beginning of the study, along with their respective group number.

2.3.1 Demographics and prior experience

The first assessment included demographic questions regarding age, gender, and highest educational qualification. Participants' level of experience with VR headsets was assessed using the question: 'How would you currently rate your experience with using VR headsets?' Response options ranged across five levels, from 'I have never worn a VR headset before' to 'I use VR headsets completely autonomously' (the full german questionnaire is provided in the Appendix). Following the demographic and experience questions, the four items of the German short version of the Technology Affinity Scale (Karrer-Gauß et al., 2024) were presented. This was succeeded by the first administration of the knowledge test on titration.

2.3.2 Knowledge Test

The knowledge test was developed by the program instructors and aligned with the expected learning objectives following the completion of the titration procedure. To ensure content validity, the test items were selected by three course instructors, checked for alignment with the module's learning objectives, and reviewed for relevance and consistency with the core concepts of titration.

2.3.2.1 Test Content Categories

The test comprised 13 multiple-choice items with multiple correct selections possible per item, assessing different dimensions of titration competency:

Declarative Knowledge (4 items): Items assessed understanding of fundamental concepts, including:

- Required chemicals for titration (Item 1)
- Essential equipment identification (Item 2)
- Endpoint recognition criteria (Item 6)
- Theoretical understanding of indicator effects (Item 13)

Procedural Knowledge (5 items): Items evaluated knowledge of correct laboratory procedures:

- Correct sequence of titration steps (Item 3)
- Importance of burette rinsing procedures (Item 8)
- Necessity of stirring during titration (Item 9)
- Proper burette filling technique (Item 10)
- Impact of over-titration on results (Item 11)

Problem-Solving and Application (4 items): Items required integration of conceptual understanding with practical application:

- Troubleshooting when burette volume is exhausted (Item 4)
- Solution strategies for concentration issues (Item 5)
- Stoichiometric calculations for concentration determination (Items 7 & 12)

2.3.2.2 Scoring System

Items were scored by awarding one point for each correctly selected option and one point for each correctly unselected distractor, resulting in a maximum score of four points per item. Therefore, a maximum total score of 52 points could be achieved. This scoring method was chosen to reduce guessing probability and provide a more nuanced assessment of student knowledge.

The item order was randomized to minimize potential order effects. The questions had been used previously in the module and were adapted specifically for this experiment. These adaptations were necessary to standardize the answer format across all questions and to facilitate online administration. Specifically, this required all questions to feature four answer options, and a sorting task (original Item 3) was converted to a multiple-choice format due to limitations of the online testing platform. This adapted version was subsequently reviewed, tested, and ultimately approved by the instructors.

2.3.3 Constructs of the CAMIL Model

Scales for Presence (Pre), Motivation (Mo), as well as intrinsic (CL_int) and extrinsic (CL_ext) cognitive load were drawn from the CAMIL model (Makransky and Petersen, 2021). The items for each scale were described by Petersen et al. (2022) and compiled from further primary literature (a complete list of items is provided in appendix 1). The items were literally translated into German and, where necessary, adapted contextually for the current study. All items were rated on a five-point Likert scale ranging from 1 = Strongly disagree to 5 = Strongly agree. The Presence scale consists

of four items, such as 'I was completely captivated by the virtual environement,' and originates from the study by Makransky et al. (2017).

The Motivation scale (Makransky and Petersen, 2019) comprises five items (e.g., 'I enjoy working with the topic of titration'), with the word 'titration' mentioned in each item to establish context. Similarly, the extrinsic Cognitive Load scale (Andersen and Makransky, 2021) consists of two subscales pertaining to the environment (e.g., 'The virtual environment was full of irrelevant content') and control within the virtual space (e.g., 'The interaction technique used in the simulation was difficult to master'). These subscales measure the demand on cognitive resources consistent with Cognitive Load Theory by Sweller et al. (2011). All three scales (Presence, Motivation, Cognitive Load) demonstrate good reliability with Cronbach's Alpha values ranging between .8 and .9, as reported by Petersen et al. (2022).

2.3.4 Technology Affinity

To control for interindividual differences in technology affinity (TA), the short version of the TAEG questionnaire was used (Karrer-Gauß et al., 2024). This scale consists of four items addressing self-assessed domains: competence in using technology (e.g., 'I am knowledgeable about electronic devices'), enthusiasm for new technologies (e.g., 'I get excited when a new electronic device comes onto the market'), and the assessment of positive (e.g., 'Electronic devices make my everyday life easier') or negative consequences (e.g., 'Electronic devices lead to mental impoverishment') associated with using technological applications. Items were rated on a five-point Likert scale. This four items showed a high correlation of .92 in the original study what makes them a suitable control variable for the present study (Karrer-Gauß et al., 2024).

2.3.5 System Usability Scale (SUS)

To assess subjective usability, we utilized the German translation of the System Usability Scale (SUS) (Brooke, 1996) by Gao et al. (2020). This scale comprises 10 items rated on a five-point Likert scale. For the present study, the items were adapted such that the word 'product' was replaced with 'VR environment' in all questions. An example item read: 'I found the various functions in this VR environment were well integrated.'

The final SUS score was calculated according to the standard procedure described by Brooke (1996). For items with odd numbers, 1 was subtracted from the score; for items with even numbers, the score was subtracted from 5. The sum of these adjusted scores for an individual was then multiplied by 2.5 to obtain a final score ranging from 0 to 100. The german version of the scale demonstrated good reliability, with Cronbach's Alpha ranging from $\alpha = .74$ to $\alpha = .88$ (Gao et al., 2020).

2.4 Procedure

The study was conducted on the Muttenz campus over four days between October 9th and 17th, 2024, as part of a mandatory course for the students. Sessions commenced at 7:30 AM, 8:30 AM, and 12:40 PM, with each session lasting approximately 3 hours and 10 minutes on average. During each session, at least two members of the research team were present to administer and support the VR condition, alongside a minimum of four university instructors.

Researchers and instructors welcomed the students in the lecture hall and explained the purpose and procedure of the study. Subsequently, participants signed an informed consent form and received

an anonymized code for completing the questionnaires. The first questionnaire (t_1, see Figure 1) collected demographic data, assessed technology affinity, and included the knowledge test.

After administering the first questionnaire, students formed pairs, which were then divided into two main groups. Pair formation was initiated by the students themselves without direction from the research team, while assignment to the starting condition was determined by the instructors based on participant numbers. One group was guided to the physical laboratory by the instructors, while the other group received an introduction to the VR headsets. This VR introduction involved demonstrating and explaining the software application's user interface, highlighting the tutorial videos available within the virtual lab, and pointing out the function of various buttons. Researchers assisted students with setting up and using the headsets until they successfully reached the virtual lab and grasped the fundamentals of navigation and interaction within the simulation. Subsequently, the pairs conducted the virtual titration experiment independently, following the instructions provided within the application. The progress of each pair was monitored via a separate Windows application, and researchers intervened only if students were unable to proceed.

Concurrently, students assigned to the physical laboratory condition were introduced by the instructors to the equipment and expected procedures. They received a printed version of the experimental instructions and performed the titration independently under instructor supervision. Interventions by the instructors occurred only in response to safety-related issues or when students requested assistance.

Upon completion of their first titration (either VR or physical lab), all students returned to the lecture hall to complete the second questionnaire (t_2). The group that had been in the physical laboratory completed the knowledge test again. The group that had experienced the VR condition completed the knowledge test along with the CAMIL and SUS questionnaires. Following this, there was a short break at the students' disposal. Subsequently, the groups switched conditions. After completing their second assigned condition, students filled out the third and final questionnaire (t_3), concluding the session.

2.5 Data Analysis

The assumptions of the CAMIL model were tested using the Partial Least Squares Structural Equation Modeling (PLS-SEM) approach, as this method, unlike covariance-based SEM models, accommodates smaller sample sizes and does not require normally distributed data (Hair et al., 2019). The analysis and evaluation of the PLS models followed the process described by Hair et al. (2021), and calculations were performed using the SEMinR package (2.3.4, Ray et al., 2024).

The evaluation process for the measurement model involved four steps, beginning with the assessment of indicator reliability, followed by internal consistency reliability, convergent validity, and finally, discriminant validity (Hair et al., 2019, p. 76). The evaluation of the structural model comprised five steps, starting with checking for collinearity, evaluating path significance, assessing the model's explanatory and predictive power, and finally, conducting model comparisons when different structural model specifications were considered.

To assess the main effects of the test condition (VR or Lab), timepoint, and sequence (the order of VR and Lab), a Linear Mixed-Effects Model (LMM) was calculated, specifying participants as a random effect. The LMM was computed using the lme4 package (1.1.36, Bates et al., 2015). The performance package (0.13.0, Lüdecke et al., 2021) was used to check model assumptions and

variance explained, and the effectsize package (1.0.0, Ben-Shachar et al., 2020) was employed to determine the effect sizes of the predictors. The recommendations of Meteyard and Davies (2020) were followed regarding the analysis procedure and the reporting of results.

3 RESULTS

Participants were assigned to the different sessions by the module coordinator; consequently, this allocation was not controlled by the research team. To minimize the potential influence of this assignment process, key group characteristics were compared and tested for significant differences. These characteristics were selected as they were measured prior to the first learning condition and are independent of the experimental manipulation. Table 1 displays descriptive statistics for sample size (n), technology affinity (TAEG), baseline test score, age, gender, and educational level, broken down by test group (session). As homogeneity of variances could not be assumed, a Welch's ANOVA was chosen for these comparisons.

Table 2 summarizes the results of the Welch's ANOVA tests comparing these variables across the test groups; all p-values were greater than .208. Based on the pre-defined alpha level of .05, there were no significant differences between the groups. The effect sizes (ω^2) for baseline knowledge ($\omega^2 = 0.05$) and age ($\omega^2 = 0.06$) were small (Kirk, 1996), while the effect size for technology affinity ($\omega^2 < 0.001$) was negligible. Additionally, Fisher's exact tests revealed no significant association between session group and gender ($p = .262$) or educational level ($p = .083$). Therefore, it can be assumed that the experimental groups did not differ substantially from one another based on these baseline characteristics.

3.1 Knowledge test

Table 3 presents the means and standard deviations for the knowledge test scores across all timepoints and conditions. Overall, at baseline (t_1), students achieved an average score of 38.97 points (SD = 3.77) out of a maximum possible 52 points, corresponding to 74.9%. This suggests a high level of prior knowledge among the participants.

Following the first learning phase (t_2), scores averaged across both conditions increased to 41.64 points (SD = 4.31), representing an improvement of 2.67 points over baseline. After the second learning phase (t_3), the average score reached 42.10 points (SD = 4.09). While this still represents an improvement compared to the baseline test, it reflects only a minimal further increase of 0.46 points from the preceding test score (t_2). Figure 3 visualizes this pattern, plotting the change scores relative to baseline for all participants across all timepoints, differentiated by starting condition (VR or Lab). A clear increase in scores from t_1 to t_2 is evident, which subsequently levels off between t_2 and t_3. Notably, this trend was not uniform; considerable variance in test score changes was observed among participants.

3.1.1 Psychometric Properties of the Knowledge Test

The reliability values for the knowledge test exhibit a specific pattern (cf. Table 4). At the first timepoint t_1 both Cronbach's α and ω are low ($\alpha_{t1} = .485$ and $\omega_{total_{t1}} = .563$). However, they increase to their highest values at the second administration ($\alpha_{t2} = .618$ and $\omega_{total_{t2}} = .672$), before decreasing again at the third administration ($\alpha_{t3} = .579$ and $\omega_{total_{t3}} = .647$), though not below the initial values at t_1 . Simultaneously ω_h increases steadily, reaching its maximum value at the third assessment point ($\omega_{h_{t1}} = .084$, $\omega_{h_{t2}} = .196$ and $\omega_{h_{t3}} = .262$)

The initial increase in reliability likely reflects the consolidation of knowledge following the learning task. The continuous increase in ω_h suggests the development of distinct subdimensional skills or knowledge domains, as this measure, unlike α and ω_{total}, accounts for multidimensionality (McNeish, 2018).

Since the knowledge test likely encompasses multiple facets (e.g., declarative knowledge, mathematical calculations, procedural tasks), the increase in α and ω_{total} – despite the assumption of unidimensionality – provides a useful, albeit cautious, assessment of improved overall reliability.

The subsequent decrease in α und ω_{total} at t_3 could be attributed to factors such as fatigue, decreased motivation, or a ceiling effect. Further investigation is required to determine the statistical significance and practical relevance of these fluctuations. As highlighted by Kelley and Pornprasertmanit (2016) and McNeish (2018) α can vary across repeated measures, particularly if the assumption of tau-equivalence is not met.

3.1.2 Statistical Analysis of Knowledge Test Scores

An overall comparison of mean scores across the three timepoints (t_1, t_2, t_3) showed a significant difference ($F_{\text{Welch}}(2, 179.44) = 17.00$, $p < .001$, $\omega_p^2 = 0.15$, $\text{CI}_{95\%}[0.07, 1.00]$, $n = 273$). Post-hoc pairwise comparisons indicated significant differences between the first timepoint ($M = 38.97$) and both the second timepoint ($M = 41.64$, $p < .001$) and third timepoint ($M = 42.10$, $p < .001$).

When analyzing each experimental condition separately, we found no significant increase in test scores for either the lab condition ($t_{\text{Welch}}(87.8) = -0.06$, $p = 0.95$, $\hat{g}_{\text{Hedges}} = -0.01$, $\text{CI}_{95\%}[-0.42, 0.40]$) or the VR condition ($t_{\text{Welch}}(88.94) = -0.90$, $p = 0.37$, $\hat{g}_{\text{Hedges}} = -0.19$, $\text{CI}_{95\%}[-0.60, 0.22]$). Additionally, a direct comparison between the two conditions at t_2 revealed no statistical difference ($t_{\text{Welch}}(88.92) = 1.01$, $p = 0.32$, $\hat{g}_{\text{Hedges}} = 0.21$, $\text{CI}_{95\%}[-0.20, 0.62]$, $n_{\text{obs}} = 91$). This suggests that the overall effect is only attributed to the first condition, regardless of which condition it was, while the second experimental condition showed no further influence on learning outcomes.

3.2 SUS Score

The overall SUS score for our sample was $M = 71.73 (SD = 13.81, n = 91)$, which indicates that the usability of the lab simulation can be interpreted as good or acceptable (Lewis, 2018). The internal consistency was adequate (Cronbach's $\alpha = .817$), consistent with reliability values reported for the German version of the SUS (Gao et al., 2020).

3.3 Path analysis of the CAMIL Model

The PLS-SEM model (cf. Figure 4) was initially estimated using all available items according to the path-weighting scheme. To determine indicator reliability, the squared factor loadings of all items were compared against the cut-off value of 0.708, although this threshold is not always met in the social sciences, making lower values potentially acceptable (Hair et al., 2021). Following the recommendations of Hair et al. (2022), however, items with squared loadings below 0.4 were excluded from further analysis due to their weak explanatory power for indicator variance. This was the case for the items listed in Table 5, and a subsequent model was estimated excluding these items.

The exclusion of a relatively large number of items (8 out of 17) fundamentally questions the transferability of the original scales to the context of VR-based chemistry education. This challenge cannot be solely explained by negatively worded items or translation issues. Particularly concerning the Motivation scale, where three out of five items had to be removed, the resulting construct likely captures only a limited aspect of the intended motivation. This reduction in construct breadth might partially explain why our structural model exhibited low explanatory power and why only the negative relationship between Presence and Cognitive Load emerged as significant. For future research, this suggests a need to adapt and validate the measurement concepts of the CAMIL model and their german translation, specifically for the application domain of VR learning in science laboratory environments.

The internal consistency reliability and convergent validity values (cf. Table 6) were acceptable for the restricted model. Furthermore, discriminant validity, assessed using the heterotrait-monotrait (HTMT) method, was established, with all values falling below the threshold of 0.95 (Henseler et al., 2015). Therefore, the measurement model of the restricted model was deemed appropriate for subsequent analysis.

To assess the structural model's goodness-of-fit, potential collinearity issues were examined first. Variance Inflation Factor (VIF) values were calculated for the predictors of the endogenous constructs. In our model, this primarily concerns the predictors Cognitive Load (CL) and Motivation (MO) for the Learning Outcome (WIS) construct. The VIF value was 1.019 for both CL and MO, indicating excellent results with no multicollinearity concerns (Becker et al., 2015). Regarding the evaluation of the path coefficients (cf. Table 7), only the negative relationship between Presence (PRE) and Cognitive Load (CL) emerged as statistically significant ($\beta = -0.471, t = -6.491, p < .01, \mathrm{CI}_{95\%}[-0.630, -0.349]$). Table 8 presents the R^2 values for the endogenous constructs, serving as a measure of the model's explanatory power. The results show that all R^2 values fell below the .250 threshold, which is generally considered weak explanatory power. CL had the highest R^2 at .222, while MO ($R^2 = .045$), PRE ($R^2 = .033$), and WIS ($R^2 = .021$) accounted for very little variance, suggesting they did not make a substantial contribution to the explained variance in this sample.

The final two steps for evaluating a PLS-SEM model, as recommended by Hair et al. (2021), were not performed. This decision was based on two reasons: firstly, standard procedures for assessing out-of-sample predictive power (such as PLSpredict) are generally not applicable to models containing higher-order constructs (higher-order composites), CL construct utilized in this study. Secondly, there was insufficient theoretical justification to support the specification of an alternative structural model.

In conclusion, the collected data and the resulting structural model demonstrate limited explanatory power. Of the relationships derived from the CAMIL model, only a significant negative influence of Presence on Cognitive Load was confirmed ($\beta = -0.471, p < .01$). The direction of this relationship aligns with the findings of Petersen et al. (2022), suggesting an interplay between these constructs and supporting the notion that cognitive load can impair performance in immersive VR environments. Notably, this relationship explained approximately 22% of the variance in Cognitive Load ($R^2 = .222$), indicating that the sense of presence might be an important factor in reducing cognitive burden within the iVR environment.

The lack of significant paths from Motivation to Learning Outcome and from Cognitive Load to Learning Outcome contradicts the theoretical propositions of the CAMIL model. This finding might be partially attributable to the aforementioned measurement challenges, but it could also suggest that the learning mechanisms involved in mastering titration within VR differ from those observed for other learning content previously studied in VR contexts. The very low R^2 values for Learning Outcome (.021) and Motivation (.045) suggest that crucial influencing factors, such as prior knowledge, might be absent from the model. Alternatively, the knowledge assessment, with its primary focus on declarative knowledge, may not have adequately captured potential learning gains related to practical laboratory skills.

3.4 Linear Mixed Effects Models

To further investigate the effects of experimental condition, intervention sequence, and participants, a linear mixed-effects model (LMM) was computed. The outcome variable was the change in test score relative to the initial baseline score (t_1). This approach allowed modeling the change directly without including the baseline score as a covariate. The rationale was that the observed changes were relatively small compared to the initial scores, and including the inital test score as a covariate might have masked the potential effects of the other predictors.

Participant was specified as a random effect, estimating a random intercept for each participant. This accounts for the nested data structure (repeated measures within participants) inherent in the study design. Following the recommendation by Barr et al. (2013) to implement a maximal random effects structure, this would ideally involve including random slopes for the effect of Condition (VR vs. Lab) per participant. However, attempting to estimate random slopes for Condition (a factor with only two levels) led to model convergence issues. Therefore, the maximal feasible random effects structure retained only the random intercepts for participants (cf. Eager and Roy, 2017).

To determine the best model fit, a baseline model including only the random intercept for participants was estimated first (cf. Table 9). Fixed effects were then added sequentially. The main predictor of interest, experimental Condition, was added as the first fixed effect (Model 1). Subsequently, Timepoint was added in Model 2, and Sequence (VR-first vs. Lab-first) was added in Model 3. All models met the assumptions of linearity, normality of residuals, and homoscedasticity (Meteyard and Davies, 2020).

Table 10 presents the model comparison results, indicating no statistically significant improvement in model fits including fixed effects compared to the baseline random-intercept-only model.

None of the fixed effects significantly improved model fit or explained additional variance in the outcome variable. It is worth noting, however, that Model 1 (adding Condition) approached statistical significance ($p = 0.06$) and showed a marginal improvement in fit over the baseline model according to the likelihood ratio test ($\chi^2(1) = 3.540$) and AIC comparison ($\Delta AIC = -1.54$), although not BIC ($\Delta BIC = 1.66$).

The lack of significant improvement from adding fixed effects, as determined by the model comparisons, is clearly reflected in the variance explained (R^2) by the final model. The variance accounted for by the random effect of participants was substantial (Conditional $R^2 \approx .761$, indicating ~76% of variance attributed to participant differences), whereas the fixed effects collectively explained very little variance (Marginal $R^2 \approx .008$, or 0.8%). Examining the fixed effect for condition revealed a non-significant trend suggesting potentially lower performance in the VR condition compared to the

Lab condition ($\beta = -0.526$, $SE = 0.289$, 95%-KI $[-1.09, 0.05]$, $t(\text{df}) = -1.817$, $p = .071$, $\eta^2 = 0.036$). However, due to the lack of statistical significance, no firm conclusion can be drawn regarding this difference (see Table 11). Taken together, these results indicate that inter-individual differences in learning performance accounted for the vast majority of the observed variance, while neither the specific intervention condition (VR vs. Lab) nor the sequence of interventions had a statistically detectable impact on the measured learning outcomes.

4 DISCUSSION

The principal goal of this study was to shed some light on the learning efficiency of an iVR training sequence compared to the same activity in a conventional laboratory setting in an actual mandatory first-semester Life-Sciences module. In line with recent research (Johnson-Glenberg et al., 2023; Liu et al., 2023), we found no improvement in the learning outcome between the two conditions, as shown in model fit comparisons when *Condition* (VR vs. Lab) was entered in our mixed-effects analysis, knowledge gains were statistically equivalent across modalities. Simple pair-wise comparisons at the second timepoint corroborated this null effect. Moreover, learning outcomes were unaffected by the order in which students encountered the modalities (VR→Lab vs. Lab→VR), indicating that neither the environment itself nor its sequencing confers an advantage. These results suggest that iVR can serve as a pedagogically equivalent, be potentially more scalable and complementary to traditional laboratories. An insight that should inform future investment decisions in teaching infrastructure and could guide instructional designers when creating new learning activities.

4.1 Implications for Learning in iVR

Our first hypothesis (H1) posited that students' technology affinity would enhance their sense of presence in the simulated laboratory. The path analysis did reveal a positive standardized coefficient, but this effect was not statistically significant. Consequently, within the present sample and measurement precision we cannot claim that technology affinity reliably boosts presence. Because the confidence interval still encompasses small-to-moderate effects, follow-up studies with larger samples or refined instruments remain warranted before ruling out a meaningful relationship.

In line with the CAMIL model and recent meta-analytic evidence (Concannon et al., 2019; Liu et al., 2023; Parong and Mayer, 2018), a stronger sense of presence was associated with lower extraneous cognitive load. Put differently, the more students felt "present" the virtual lab, the fewer cognitive resources were consumed by interface demands or distractions from the simulated environment, leaving greater capacity for task-relevant processing. H2 was therefor confirmed. This finding highlights the importance of designing iVR environments that actively foster presence while minimizing non-essential elements that could compete for learners' cognitive resources. Bøg Petersen et al. (2022) found that learners in VR environments often retain less knowledge compared to more passive learning methods, due to the cognitive load imposed by the simulation. Similarly, Parong and Mayer (2021a) refer to cognitive or affective distraction in such contexts. Another cognitive mechnaism is reported by Oje et al. (2025), who point out, that the use of leading cues and guidance in VR, what we did in our virtual laboratory, can lower the extraneous CL and facilitate the concentration on the learning task. Our result may indicate that the sense of presence could serve as an important predictor of reduced cognitive load—an effect that should, in turn, support improved learning outcomes.

Surprisingly, two further CAMIL paths did not materialize: presence was not a reliable predictor of motivation, H3, and neither cognitive load nor motivation predicted knowledge-test performance (H4). Measurement attenuation likely played a role here; several motivation items fell below reliability thresholds and were removed, narrowing construct coverage and weakening statistical power. While the negative association between motivation and test score can likely be attributed to issues with the measurement instrument, the lack of a relationship between cognitive load and test score is more difficult to explain. As previously discussed, lower cognitive load should generally be associated with better learning outcomes (Parong and Mayer, 2021a; Sweller et al., 2011). However, this connection could not be confirmed in our model. An explanation might be a ceiling effect in the test scores, which could obscure potential effects due to limited variance in the outcome variable. Another consideration is the design of the test instrument itself, which may have focused too heavily on declarative knowledge and insufficiently captured the procedural components of the learning process (Hamilton et al., 2021).

The structural model's fixed effects explained only $\approx 0.8\,\%$ of the variance in test scores (marginal $R^2 \approx .008$), whereas random intercepts for participants accounted for roughly $76\,\%$ (conditional $R^2 \approx .76$). Such dominance of individual differences suggests that prior knowledge or test-taking skills, rather than the experimental manipulations, drove most outcome variability. This interpretation dovetails with the high baseline score (about $75\,\%$ correct) and the very modest absolute gain between the second and third knowledge tests (approximately 0.5 points). Together, the pattern is consistent with ceiling effects, limited test sensitivity to procedural gains, and diminishing returns once the titration protocol has been practiced in either modality.

4.2 Considerations about sequencing and technologies

The fifth hypothesis (H5) examined whether differences exist between the instructional methods employed and their sequencing. Interpretation of this linear mixed-effects model must again consider that the fixed effects — condition, timepoint, and sequence — accounted for only 0.8% of the variance explained by the model. As observed in the path model, the majority of the variance (~76%) was attributable to the random effect of the individual participant. This finding reinforces the conclusion that a substantial proportion of the observed effects can be explained by inherent individual characteristics. Moreover, by employing a second statistical approach, we were able to confirm that there was no significant difference between the two learning activities. The sequence of the activities likewise showed no significant influence, further supporting our assertion that similar learning gains can be achieved when comparable processes or actions are implemented across activities. These findings corroborate previous meta-analyses (Concannon et al., 2019; Liu et al., 2023), which similarly reported no, or only minimal, differences between iVR and other media or instructional methods. The main difference with these studies is that we conducted this study in an actual course for a grounding skill in this curriculum. We can therefor conclude, that in this very realistic setting, the simulation was on the same level than the laboratory learning activity regarding this specific test and his results.

4.3 Limitations

This study has a pilot character, as it represents the first implementation of this learning activity within the module. The scales used to measure motivation, presence, and cognitive load were not validated in advance and only demonstrate satisfactory validity. This likely influenced the statistical results. During the VR condition, some groups received more support and guidance when they

encountered difficulties during the practice phase, which likely introduced unequal treatment between groups. In addition, group assignments were not fully randomized. The influence of collaboration during the simulation phase was neither measured nor explicitly guided. Furthermore, participants in the VR condition often had to wait longer for the laboratory condition participants to return before completing the knowledge test. This delay may have resulted in memory decay effects, as the lab group had the learning content fresher in mind. Conversely, motivational issues may have arisen from the idle waiting time. Finally, the third questionnaire round always took place at the end of the school block, either shortly before lunch or before going home, which—alongside the waiting period for the VR group—may have contributed to fatigue or a tendency to complete this final phase quickly and with reduced attention.

4.4 Practical Implications and future research

The demonstrated equivalence in learning outcomes between the simulation and the real laboratory underscores the potential of immersive technology as a viable means to enrich student education with an additional dimension. However, as Radianti et al. (2020) emphasizes, its implementation must be holistic and grounded in learning theory. Assuming, as Conrad et al. (2024) suggest, that there is no significant difference between various instructional technologies and media, and considering that iVR can be particularly effective for procedural knowledge acquisition (Hamilton et al., 2021), that there is compelling evidence for transfer from VR to real-world workplaces (Bøg Petersen et al., 2022), and that—according to our findings—neither sequence nor condition had a decisive impact, the focus must now shift to the appropriate pedagogical integration of this technology. If VR is to be treated as an equivalent pedagogical tool, curricula must be intentionally designed so that this additional modality contributes meaningfully to the development of targeted competencies, attitudes or skills. Instructional design models such as ADDIE (Branch, 2009)or Constructive Alignment (Biggs and Tang, 2011) may provide valuable frameworks in this regard. These approaches can be combined with evidence-based design and pedagogical principles in VR as reported by Oje et al. (2025), with the aim to create a model for designing learning activities in VR.

Future research should also engage more deeply with the assessment of procedural learning in VR (see also Hamilton et al., 2021) and explore how the resulting data can support personalized and adaptive learning pathways. Which data points could serve as indicators of successful learning? How can gaps in students' knowledge be derived from this data and addressed directly? These questions are likely to become increasingly relevant, particularly as AI models are integrated either in the background to guide and support learning processes or more actively as agents or learning companions within the simulation. It will also be essential to examine how students accept these mechanisms and how they affect cognitive and affective processes.

From an economic perspective, it is important to evaluate whether the acquisition, development, and maintenance of the technological infrastructure required to deliver such simulations offer advantages over traditional real-world learning environments. Equally critical is the initial and ongoing training of educators in the effective use and pedagogical application of these technologies.

4.5 Conclusion

Contemporary educational design increasingly emphasizes competency development and acquisition in the construction of curricula and training programs. Our work aligns with this evolution, and in light of current AI-driven disruptions, the importance of concrete actions in education and

professional development will undoubtedly increase. In connection with this trend, personal, concrete experiential learning opportunities are extremely valuable. However, resources are limited regarding physical spaces, facilities, and personnel when implementing real-world environments for educational purposes.

The application of simulations (iVR especially) is of significant interest in this context. In our specific case, we demonstrated that the implementation of iVR compared to real laboratory experiences in a curricular learning activity and subsequent assessment led to identical outcomes. Based on other studies, we can reasonably assume this finding may be generalizable, suggesting that the application of simulations in educational offerings should undergo more thorough examination, and thoughtful implementation of this technology will certainly become increasingly important in the future.

This approach offers substantial advantages for students, enabling them to engage with learning experiences not only at educational institutions and in real environments but also asynchronously and across various devices. Therefore, the utilization of simulations presents a viable opportunity when seeking to establish educational equity or addressing varying learning rates, offering significant benefits by potentially allowing students to repeat such learning experiences multiple times as needed.

It is worth noting, however, that further research is required to establish a robust and reliable framework for designing these activities in VR (and in near future also in mixed reality settings). The development of virtual learning experiences necessitates evidence-based guidance, and their integration into educational organizations requires a thoughtful approach that involves instructors in the process. Only through such collaborative efforts can we fully realize the potential of immersive technologies in educational contexts.

CONFLICT OF INTEREST

The author declares no conflict of interest. The study was conducted as part of the author's Master thesis in the applied psychology program at the FHNW.

ACKNOWLEDGEMENTS

I would like to thank Dr. Oliver Christ for providing me with the opportunity to participate in this study and to write this thesis. His guidance, feedback, and support throughout the writing process were invaluable. I would also like to thank the teaching team from the Life Sciences Bachelor Course for their collaboration in creating the questionnaire, their assistance in testing it, and their constructive feedback. Their contribution was essential during the conduct of the study, as they adapted and organized their modules to accommodate the iVR condition and integrated it seamlessly into their regular coursework.

Use of Generative AI Technologies: During the writing of this work, various generative AI models were used to assist with translation of German text to English, grammar and syntax correction, brainstorming ideas and arguments, and providing feedback on text sections. No AI models were used to generate original content or write text in the first instance. All writing represents my original intellectual work, with AI tools used solely for correction, translation, and feedback purposes. Portions of the R code used for statistical analysis and results reporting were generated by AI

following my specific instructions. All calculations and reported results are based exclusively on the actual research data processed through R code, not on AI model outputs. The following AI models were utilized: OpenAI GPT-4o, GPT-4.1, GPT-3o; Anthropic Claude 3.5 Sonnet, Claude 3.7 Sonnet, Claude Sonnet 4, Claude Opus 4; Google Gemini 1.5 Flash, Gemini 1.5 Pro, Gemini 2.0 Flash, Gemini 2.0 Pro; Google NotebookLM, GitHub Copilot.

DATA AVAILABILITY

The data that support the fin dings of this study are available from the corresponding author upon reasonable request.

BIBLOGRAPHY

Agustian, H. Y., Finne, L. T., Jørgensen, J. T., Pedersen, M. I., Christiansen, F. V., Gammelgaard, B., et al. (2022). Learning outcomes of university chemistry teaching in laboratories: A systematic review of empirical literature. *Review of Education* 10, e3360, doi: 10.1002/rev3.3360

Andersen, M. S. and Makransky, G. (2021). The validation and further development of a multidimensional cognitive load scale for virtual environments. *Journal of Computer Assisted Learning* 37, 183–196, doi: 10.1111/jcal.12478

Bailenson, J. (2018). *Experience on demand: what virtual reality is, how it works, and what it can do* (New York, NY London: W.W. Norton & Company)

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68, 255–278, doi: 10.1016/j.jml.2012.11.001

Barrett, A., Pack, A., Guo, Y., and Wang, N. (2023). Technology acceptance model and multi-user virtual reality learning environments for chinese language education. *Interactive Learning Environments* 31, 1665–1682, doi: 10.1080/10494820.2020.1855209

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67, 1–48, doi: 10.18637/jss.v067.i01

Becker, J.-M., Ringle, C. M., Sarstedt, M., and Völckner, F. (2015). How collinearity affects mixture regression results. *Marketing Letters* 26, 643–659, doi: 10.1007/s11002-014-9299-9

Ben-Shachar, M. S., Lüdecke, D., and Makowski, D. (2020). effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software* 5, 2815, doi: 10.21105/joss.02815

Bennett, S. W. and O'Neale, K. (1998). Skills development and practical work in chemistry. *University Chemistry Education* 2

Biggs, J. B. and Tang, C. S.-k. (2011). *Teaching for quality learning at university: what the student does*. SRHE and Open University Press Imprint (Maidenhead: McGraw-Hill/Society for Research into Higher Education/Open University Press), 4th edition edn.

Bohne, T., Heine, I., Gurerk, O., Rieger, C., Kemmer, L., and Y. Cao, L. (2021). Perception engineering learning with virtual reality. *IEEE Transactions on Learning Technologies* 14, 500–514, doi: 10.1109/TLT.2021.3107407

Branch, R. M. (2009). *Instructional Design: The ADDIE Approach* (Boston, MA: Springer US), doi: 10.1007/978-0-387-09506-6

Bretz, S. L. (2019). Evidence for the importance of laboratory courses. *Journal of Chemical Education* 96, 193–195, doi: 10.1021/acs.jchemed.8b00874

814 Brooke, J. (1996). Sus: A 'quick and dirty' usability scale. In *Usability Evaluation in Industry*, eds.
815 P. W. Jordan, B. Thomas, I. L. McClelland, and B. Weerdmeester (London: Taylor & Francis).
816 189–194, doi: 10.1201/9781498710411

817 Bøg Petersen, G., Klingenberg, S., and Makransky, G. (2022). Pipetting in virtual reality can predict
818 real-life pipetting performance. *Technology, Mind, and Behavior* 3, doi: 10.1037/tmb0000076

819 Campbell, C. D., Midson, M. O., Bergstrom Mann, P. E., Cahill, S. T., Green, N. J. B., Harris,
820 M. T., et al. (2022). Developing a skills-based practical chemistry programme: an integrated, spiral
821 curriculum approach. *Chemistry Teacher International* 4, 243–257, doi: 10.1515/cti-2022-0003

822 Chirico, A. and Gaggioli, A. (2019). When virtual feels real: Comparing emotional responses and
823 presence in virtual and natural environments. *Cyberpsychology, Behavior, and Social Networking*
824 22, 220–226, doi: 10.1089/cyber.2018.0393

825 Concannon, B. J., Esmail, S., and Roduta Roberts, M. (2019). Head-mounted display virtual
826 reality in post-secondary education and skill training. *Frontiers in Education* 4, 80, doi:
827 10.3389/feduc.2019.00080

828 Conrad, M., Kablitz, D., and Schumann, S. (2024). Learning effectiveness of immersive virtual
829 reality in education and training: A systematic review of findings. *Computers & Education: X*
830 *Reality* 4, 100053, doi: 10.1016/j.cexr.2024.100053

831 Cooper, N., Millela, F., Cant, I., White, M. D., and Meyer, G. (2021). Transfer of training—virtual
832 reality training with augmented multisensory cues improves user experience during training and
833 task performance in the real world. *PLOS ONE* 16, e0248225, doi: 10.1371/journal.pone.0248225

834 di Lanzo, J. A., Valentine, A., Sohel, F., Yapp, A. Y. T., Muparadzi, K. C., and Abdelmalek, M.
835 (2020). A review of the uses of virtual reality in engineering education. *Computer Applications in*
836 *Engineering Education* 28, 748–763, doi: 10.1002/cae.22243

837 Dunnagan, C. L., Dannenberg, D. A., Cuales, M. P., Earnest, A. D., Gurnsey, R. M., and Gallardo-
838 Williams, M. T. (2020). Production and evaluation of a realistic immersive virtual reality organic
839 chemistry laboratory experience: Infrared spectroscopy. *Journal of Chemical Education* 97,
840 258–262, doi: 10.1021/acs.jchemed.9b00705

841 Eager, C. and Roy, J. (2017). Mixed effects models are sometimes terrible, doi:
842 10.48550/arXiv.1701.04858

843 European Commission and PwC (2020). *Skills for industry curriculum guidelines 4.0: future proof*
844 *education and training for manufacturing in Europe : final report.* Tech. rep., Publications Office,
845 LU

846 Finne, L. T., Gammelgaard, B., and Christiansen, F. V. (2022). When the lab work disappears:
847 Students' perception of laboratory teaching for quality learning. *Journal of Chemical Education*
848 99, 1766–1774, doi: 10.1021/acs.jchemed.1c01113

849 Gao, M., Kortum, P., and Oswald, F. L. (2020). Multi-language toolkit for the system
850 usability scale. *International Journal of Human–Computer Interaction* 36, 1883–1901, doi:
851 10.1080/10447318.2020.1801173

852 Guo, H., Ma, F., and Zhou, Z. (2025). Validation of technology acceptance model for virtual reality
853 usage in collaborative learning to enhance learner performance. *Innovations in Education and*
854 *Teaching International* 62, 429–443, doi: 10.1080/14703297.2024.2307994

855 Hair, J. F., Black, W. C., Babin, B. J., and Anderson, R. E. (2019). *Multivariate data analysis*
856 (Andover, Hampshire: Cengage), eighth edition edn.

Hair, J. F., Hult, G. T. M., Ringle, C. M., and Sarstedt, M. (2022). *A primer on partial least squares structural equation modeling (PLS-SEM)* (Thousand Oaks: SAGE Publications, Incorporated), third edition edn.

Hair, J. F., Hult, G. T. M., Ringle, C. M., Sarstedt, M., Danks, N. P., and Ray, S. (2021). *Partial Least Squares Structural Equation Modeling (PLS-SEM) Using R: A Workbook*. Classroom Companion: Business (Cham: Springer International Publishing), doi: 10.1007/978-3-030-80519-7

Hamilton, D., McKechnie, J., Edgerton, E., and Wilson, C. (2021). Immersive virtual reality as a pedagogical tool in education: a systematic literature review of quantitative learning outcomes and experimental design. *Journal of Computers in Education* 8, 1–32, doi: 10.1007/s40692-020-00169-2

Henseler, J., Ringle, C. M., and Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science* 43, 115–135, doi: 10.1007/s11747-014-0403-8

Hyde, J. (2025). A new perspective on chemistry foundation level students laboratory skill development using reciprocal peer-teaching, laboratory simulations, and practical skills portfolio (psp) during covid-19 and post-pandemic in 2024. *Journal of Chemical Education* 102, 984–1003, doi: 10.1021/acs.jchemed.4c01124

Johnson-Glenberg, M. C., Yu, C. S. P., Liu, F., Amador, C., Bao, Y., Yu, S., et al. (2023). Embodied mixed reality with passive haptics in stem education: randomized control study with chemistry titration. *Frontiers in Virtual Reality* 4, 1047833, doi: 10.3389/frvir.2023.1047833

Johnstone, A. H. (1993). The development of chemistry teaching: A changing response to changing demand. *Journal of Chemical Education* 70, 701, doi: 10.1021/ed070p701

Kaplan, A. D., Cruit, J., Endsley, M., Beers, S. M., Sawyer, B. D., and Hancock, P. A. (2021). The effects of virtual reality, augmented reality, and mixed reality as training enhancement methods: A meta-analysis. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 63, 706–726, doi: 10.1177/0018720820904229

Karrer-Gauß, K., Glaser, C., Clemens, C., and Bruder, C. (2009). Technikaffinität erfassen – der fragebogen ta-eg. *Der Mensch im Mittelpunkt technischer Systeme* 8, 196–201

Karrer-Gauß, K., Roesler, E., and Siebert, F. W. (2024). Neuauflage des taeg fragebogens: Technikaffinität valide und multidimensional mit einer kurz- oder langversion erfassen. *Zeitschrift für Arbeitswissenschaft* 78, 387–406, doi: 10.1007/s41449-024-00427-4

Kelley, K. and Pornprasertmanit, S. (2016). Confidence intervals for population reliability coefficients: Evaluation of methods, recommendations, and software for composite measures. *Psychological Methods* 21, 69–92, doi: 10.1037/a0040086

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement* 56, 746–759, doi: 10.1177/0013164496056005002

Levac, D. E., Huber, M. E., and Sternad, D. (2019). Learning and transfer of complex motor skills in virtual reality: a perspective review. *Journal of NeuroEngineering and Rehabilitation* 16, 121, doi: 10.1186/s12984-019-0587-8

Lewis, J. R. (2018). The system usability scale: Past, present, and future. *International Journal of Human–Computer Interaction* 34, 577–590, doi: 10.1080/10447318.2018.1455307

Liu, J. Y. W., Yin, Y.-H., Kor, P. P. K., Cheung, D. S. K., Zhao, I. Y., Wang, S., et al. (2023). The effects of immersive virtual reality applications on enhancing the learning outcomes of undergraduate health care students: Systematic review with meta-synthesis. *Journal of Medical Internet Research* 25, e39989, doi: 10.2196/39989

Lüdecke, D., Ben-Shachar, M., Patil, I., Waggoner, P., and Makowski, D. (2021). performance: An r package for assessment, comparison and testing of statistical models. *Journal of Open Source Software* 6, 3139, doi: 10.21105/joss.03139

Makransky, G., Andreasen, N. K., Baceviciute, S., and Mayer, R. E. (2021). Immersive virtual reality increases liking but not learning with a science simulation and generative learning strategies promote learning in immersive virtual reality. *Journal of Educational Psychology* 113, 719–735, doi: 10.1037/edu0000473

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

Makransky, G., Lilleholt, L., and 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, doi: 10.1016/j.chb.2017.02.066

Makransky, G. and Petersen, G. (2019). Investigating the process of learning with desktop virtual reality: A structural equation modeling approach. *Computers & Education* 134, 15–30, doi: 10.1016/j.compedu.2019.02.002

Makransky, G. and 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, 937–958, doi: 10.1007/s10648-020-09586-2

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods* 23, 412–433, doi: 10.1037/met0000144

Meteyard, L. and Davies, R. A. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language* 112, 104092, doi: 10.1016/j.jml.2020.104092

Meyer-Drawe, K. (2003). Lernen als erfahrung. *Zeitschrift für Erziehungswissenschaft* 6, 505–514, doi: 10.1007/s11618-003-0054-x

Milgram, P., Takemura, H., Utsumi, A., and Kishino, F. (1995). Augmented reality: a class of displays on the reality-virtuality continuum. In *Telemanipulator and Telepresence Technologies*, ed. H. Das (Boston, MA: SPIE), vol. 2351, 282–292, doi: 10.1117/12.197321

Oje, A. V., Hunsu, N. J., and Fiorella, L. (2025). A systematic review of evidence-based design and pedagogical principles in educational virtual reality environments. *Educational Research Review* 47, 100676, doi: 10.1016/j.edurev.2025.100676

Parong, J. and Mayer, R. (2018). Learning science in immersive virtual reality. *Journal of Educational Psychology* doi: 10.1037/EDU0000241

Parong, J. and Mayer, R. E. (2021a). Cognitive and affective processes for learning science in immersive virtual reality. *Journal of Computer Assisted Learning* 37, 226–241, doi: 10.1111/jcal.12482

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

Peeters, H., Habig, S., and Fechner, S. (2023). Does augmented reality help to understand chemical phenomena during hands-on experiments?–implications for cognitive load and learning. *Multimodal Technologies and Interaction* 7, 9, doi: 10.3390/mti7020009

Petersen, G., Petkakis, G., and Makransky, G. (2022). A study of how immersion and interactivity drive vr learning. *Computers & Education* 179, 104429, doi: 10.1016/j.compedu.2021.104429

946 Radianti, J., Majchrzak, T. A., Fromm, J., and Wohlgenannt, I. (2020). A systematic review of
947 immersive virtual reality applications for higher education: Design elements, lessons learned, and
948 research agenda. *Computers & Education* 147, 103778, doi: 10.1016/j.compedu.2019.103778
949 Rahman, A. A. (2020). Tracing the evolution of transfer of training: A review article. *Annals of*
950 *Social Sciences & Management studies* 5, doi: 10.19080/ASM.2020.05.555668
951 Ray, S., Danks, N. P., and Valdez, A. (2024). seminr: Building and estimating structural equation
952 models
953 Sheppard, K. (2006). High school students' understanding of titrations and related acid-base
954 phenomena. *Chem. Educ. Res. Pract.* 7, 32–45, doi: 10.1039/B5RP90014J
955 Sweller, J., Ayres, P., and Kalyuga, S. (2011). *Cognitive Load Theory* (New York, NY: Springer
956 New York), doi: 10.1007/978-1-4419-8126-4
957 Tee, N. Y. K., Gan, H. S., Li, J., Cheong, B. H.-P., Tan, H. Y., Liew, O. W., et al. (2018).
958 Developing and demonstrating an augmented reality colorimetric titration tool. *Journal of*
959 *Chemical Education* 95, 393–399, doi: 10.1021/acs.jchemed.7b00618
960 Thisgaard, M. and Makransky, G. (2017). Virtual learning simulations in high school: Effects on
961 cognitive and non-cognitive outcomes and implications on the development of stem academic and
962 career choice. *Frontiers in Psychology* 8, 805, doi: 10.3389/fpsyg.2017.00805
963 Tusher, H. M., Mallam, S., and Nazir, S. (2024). A systematic review of virtual reality features for
964 skill training. *Technology, Knowledge and Learning* 29, 843–878, doi: 10.1007/s10758-023-09713-2
965 UNESCO (2024). *Higher education: figures at a glance*. Tech. rep., UNESCO
966 Wolter, S. C., Albiez, J., Cattaneo, M. A., Denzler, S., Diem, A., Lüthi, S., et al. (2023).
967 *Bildungsbericht Schweiz 2023* (Aarau: Schweizerische Koordinationsstelle für Bildungsforschung
968 (SKBF))
969 Won, M., Ungu, D. A. K., Matovu, H., Treagust, D. F., Tsai, C.-C., Park, J., et al. (2023).
970 Diverse approaches to learning with immersive virtual reality identified from a systematic review.
971 *Computers & Education* 195, 104701, doi: 10.1016/j.compedu.2022.104701
972 Zierer, K. and Seel, N. M. (2012). General didactics and instructional design: eyes like twins a
973 transatlantic dialogue about similarities and differences, about the past and the future of two
974 sciences of learning and teaching. *SpringerPlus* 1, 15, doi: 10.1186/2193-1801-1-15

FIGURE CAPTIONS

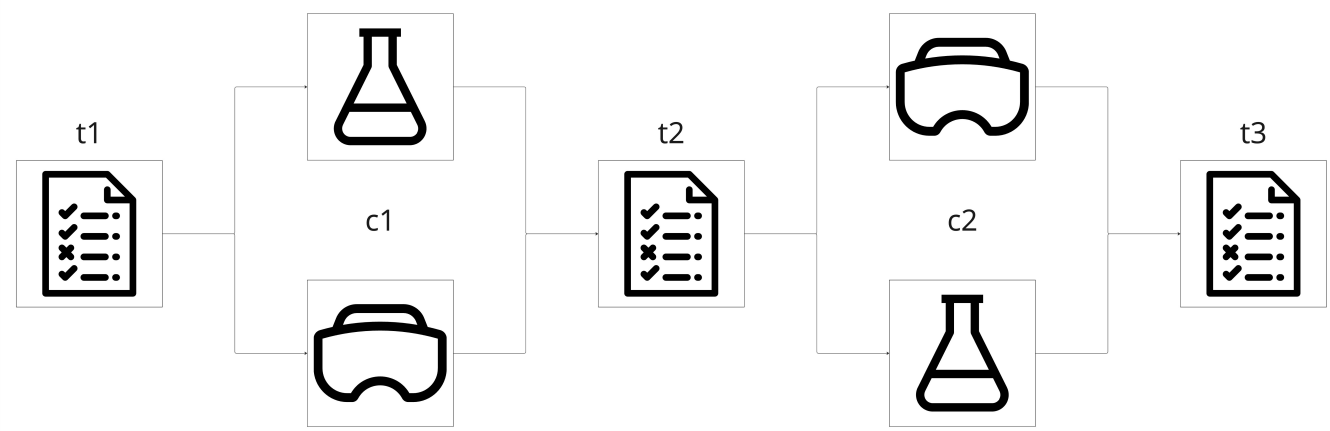
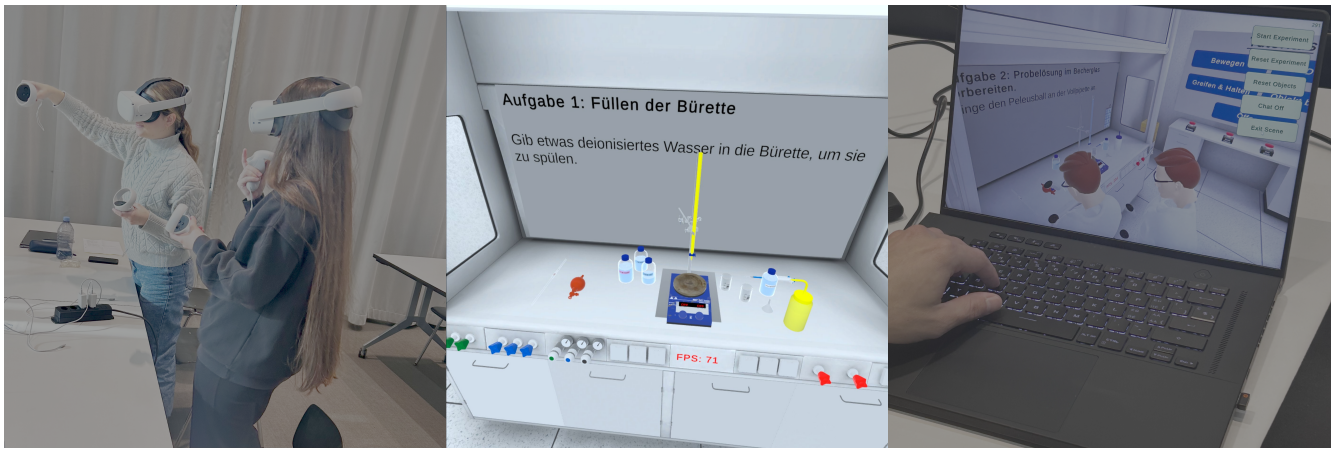


Figure 1. Schematic diagram of the study procedure



(A) (B) (C)

Figure 2. Figures of the VR condition; (A) Students working in iVR, (B) iVR working place, (C) iVR Supervision

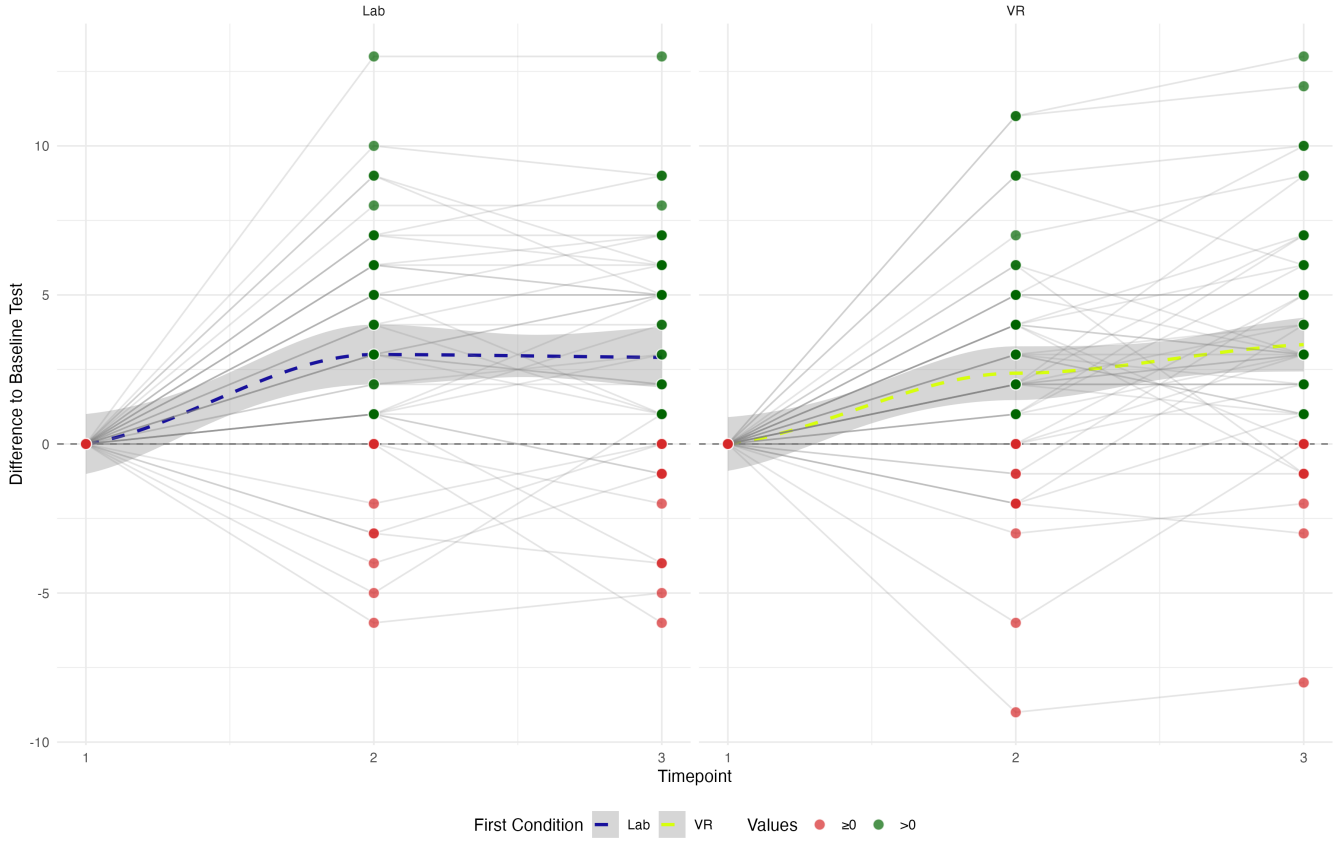


Figure 3. Knowledge test score trends by starting condition and timepoint

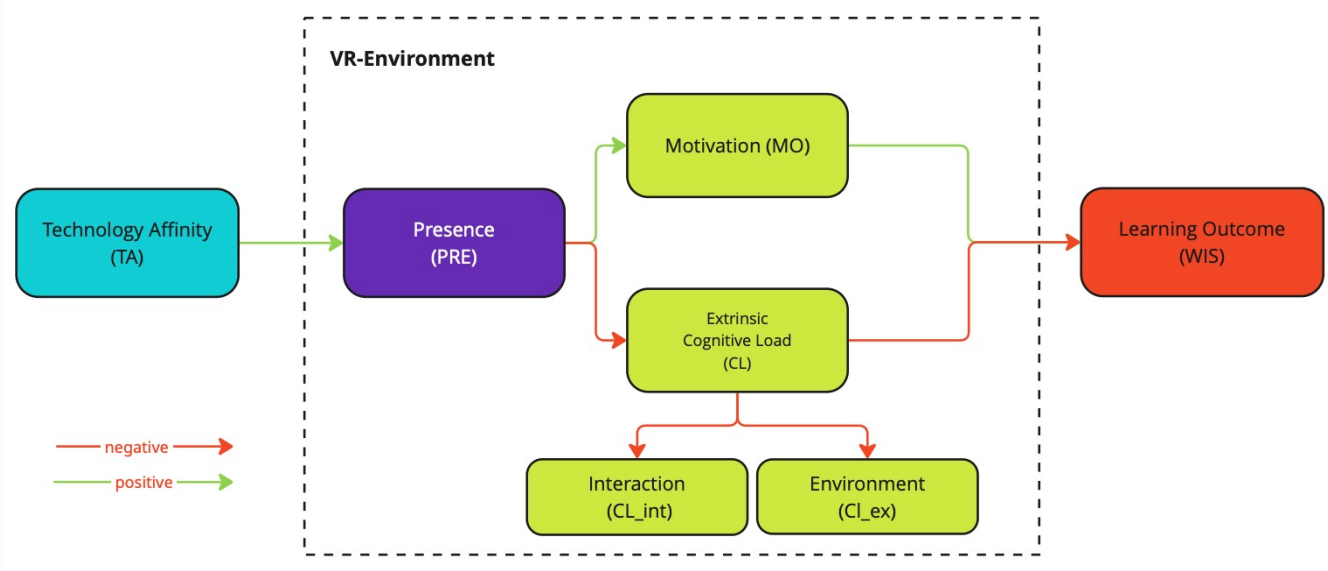


Figure 4. PLS-SEM of the CAMIL model (Makransky and Petersen, 2021)

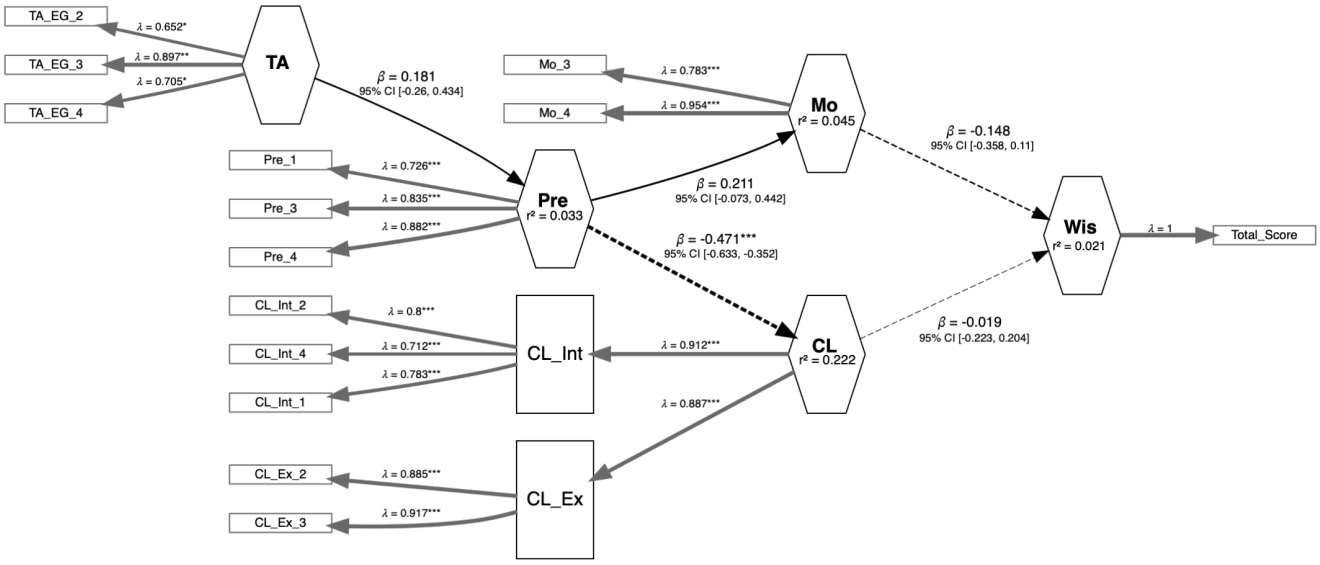


Figure 5. PLS-SEM with path coefficients

TABLES

Table 1 Descriptive statistics of the experimental groups

Session	n	TAEG		Baseline		Age		Sex		
		Mean	Sd	Mean	Sd	Mean	Sd	f	m	d
1	12	3.27	0.76	40.00	4.24	22.17	3.21	9	3	
2	20	3.52	0.44	39.00	3.78	20.80	1.74	11	9	
3	17	3.59	0.63	40.24	2.66	20.88	2.18	11	5	1
4	10	3.45	0.57	37.00	4.00	22.60	2.41	5	5	
5	18	3.18	0.70	38.17	3.65	22.06	3.49	9	9	
6	14	3.55	0.56	38.93	4.25	23.29	6.51	12	2	

Table 2 One-way Welch's ANOVA results comparing session groups

Variable	F	df1	df2	p	ω^2	n
Baseline	1.41	5	35.38	.245	0.05	91
Age	1.52	5	34.53	.208	0.06	91
TA_EG	0.95	5	35.36	.459	0.00	91

Table 3 Results of the knowledge tests

Condition	Timepoint	n	Mean	SD
Baseline	1	91	38.97	3.77
Lab	2	43	42.12	4.12
Lab	3	48	42.17	4.10
VR	2	48	41.21	4.48
VR	3	43	42.02	4.11

Table 4 Reliability values for test items

Scale	Numberof items	CronbachAlpha	OmegaTotal	OmegaHierarchical
TA_EG	4	.650	.713	.427
Mo	5	.822	.860	.643
Pre	4	.700	.765	.353
CL	8	.789	.827	.508
t1	13	.422	.507	.014
t2	13	.632	.678	.052
t3	13	.577	.628	.307

Table 5 Items with factor loadings with values < 0.4

	Squared Loadings
TA_EG_1	0.374
Mo_1	0.336
Mo_2	0.069
Mo_5	0.076
Pre_2	0.364
CL_Ex_1	0.241
CL_Ex_4	0.262
CL_Int_3	0.220

Table 6 Internal consistency reliability and convergent validity metrics

	α	ρ_C	ρ_A	AVE
TA	0.648	0.800	0.790	0.576
Pre	0.758	0.857	0.817	0.667
CL	0.765	0.895	0.773	0.809
Mo	0.719	0.864	1.001	0.762
CL_Ex	0.769	0.896	0.782	0.812
CL_Int	0.653	0.809	0.667	0.587

Table 7 Path Coefficients of the Bootstrapped Structural Model (n=10,000)

	Original Estimate	Bootstrap Mean	Bootstrap SD	T Stat.	p	Confidence Interval 2.5%	Confidence Interval 97.5%
TA → Pre	0.181	0.185	0.180	1.008	.316	-0.255	0.431
Pre → CL	-0.471	-0.497	0.072	-6.538	<.001	-0.630	-0.351
Pre → Mo	0.211	0.221	0.130	1.632	.106	-0.073	0.449
CL → Wis	-0.019	-0.016	0.109	-0.172	.864	-0.222	0.201
Mo → Wis	-0.148	-0.150	0.115	-1.285	.202	-0.354	0.105

Table 8 R^2-Values of the Bootstrapped Structural Model (n=10,000)

	Pre	CL	Mo	Wis
R^2	0.033	0.222	0.045	0.021
adj. R^2	0.022	0.213	0.034	-0.001
TA	0.181			
Pre		-0.471	0.211	
CL				-0.019
Mo				-0.148

Table 9 Model specifications with fit indices

Model	Fixed Effects	AIC	BIC	Log-Likelihood	df
Null	None	947.83	957.44	-470.91	3
Condition	Condition	946.29	959.10	-469.14	4
Timepoint	Timepoint + Condition	946.03	962.05	-468.01	5
Sequence	Timepoint + Condition + Sequence	948.01	967.24	-468.01	6
N total obs = 182; N Subjects = 91					

Table 10 Likelihood Ratio Tests Between Nested Models

Model	Compared to	df	χ^2	p
Condition	Null	1	3.54	.060
Timepoint	Condition	1	2.26	.133
Sequence	Timepoint	1	0.02	.898

Table 11 Comprehensive Overview of Fixed and Random Effects in the Model

	Fixed Effects					
	β	SE	95% CI	t	p	η^2
Intercept	2.135	0.937	0.24 - 3.93	2.279	.024	
Timepoint (1)	0.433	0.289	-0.14 - 1.00	1.496	.136	0.025
Condition (VR)	-0.526	0.289	-1.09 - 0.05	-1.817	.071	0.036
Sequence (VR first)	-0.099	0.781	-1.63 - 1.44	-0.127	.899	0.000

Note: The p-values for the Fixed Effects were calculated using Satterthwaite approximations.
The confidence intervals were calculated using the bootstrap method based on 10000 simulations.
Model equation: Base_Diff_Score ~ Timepoint + Condition + Sequence + (1 | User_Code)