LUNDS UNIVERSITET

DEPARTMENT of PSYCHOLOGY

# *Collaborative Learning in an Immersive Virtual Environment: The Effects of Context and Retrieval Practice*

**Noëlle Bender**

Master's Thesis (30 hp)
Winter 2024

Supervisor: Mikael Johansson

## Acknowledgments

I thank the Virtual Learning Lab from Copenhagen University for supporting this research. Also, I would like to thank my supervisor, Mikael Johansson, for his engagement in this project.

## Abstract

The accessibility of Virtual Reality (VR) enables the investigation of desirable difficulties originating from memory research with increased ecological validity. The two desirable difficulties include contextual variation and retrieval practice. Most studies investigated individual learning in VR and found indirect effects (such as enjoyment and motivation), but knowledge acquisition was not necessarily enhanced (Mayer et al., 2022; Makransky et al., 2019). Therefore, the effects of desirable difficulties on learning in an immersive collaborative setting were investigated.

This study experimentally tested in a 2x2 factorial design (N=159) whether the desirable difficulties of contextual variation and retrieval practice are applicable in a collaborative VR lesson (H1&H2) and if they interact (H3). Students retrieved the encoded information through active vs. passive mapping with a collaboration partner in the same vs. varied context. Hypotheses were tested with a linear mixed model, with the experimental condition as the fixed effect and the dyads as the random effect. Results show a significant main effect of retrieval practice in factual and conceptual learning. Further, results indicate that combining varied contexts and passive retrieval leads to detrimental conceptual learning effects.

Results align with retrieval practice research (Agarwal, 2021) and the cognitive theory of multimedia learning (Mayer, 2014). Further research is necessary to make more reliable inferences about the effect of contextual variation.


*Keywords:* VR, collaborative Mapping, retrieval practice, contextual variation, desirable difficulty, ecological validity

## Introduction

Advancements in technology have made head-mounted displays more affordable, which has increased the accessibility of the Metaverse, a post-reality universe experienced exclusively through Virtual Reality (VR). The Metaverse not only simulates reality as we know it but also enables us to experience situations and environments that are otherwise impossible to reach (Gartner, 2022). Additionally, the Metaverse serves as a platform for socializing, communication, and collaboration. This research focuses on the potential for collaborative learning with this technology. The extension of reality enables a transformation from information-based education to experience-based education (Plechatà et al., 2022). However, learning in VR is different from learning in reality. Therefore, extensive research (Mayer et al., 2022) is needed to support deep and long-term learning. Especially with regard to collaborative learning, most research has focused on individual experiences with a few exceptions (Makransky & Petersson, 2023; Petersson et al., 2023).

To support this transformation to experience-based education and optimize learning in collaborative immersive experiences, I will apply theories of memory research and combine them with contemporary findings of (individual) VR educational psychology. The increased ecological validity (Smith, 2019) allows leveraging the principle of desirable difficulties (Bjork & Bjork, 2019) in collaborative VR. One of the desirable difficulties of interest in this research is contextual variation. The effect of changing contexts on collaborative learning can be examined through controlled and unconstrained access to different environments. The second principle is retrieval practice, potentially resulting in a beneficial testing effect (Butler & Roediger, 2011; Agarwal et al., 2021). Through the active repeated attempt to test learning material (vs. restudy), enhanced learning can be achieved. Specifically in retrieval practice, contemporary findings of VR research (Mayer, 2014) play a crucial role. Both interventions (contextual variation and retrieval practice) have in common the counterintuitive implementation of difficulties during retrieval to enhance learning. Whether these robust findings are also supportive and may interact with each other (Imundo, 2021) in an immersive collaborative setting will be central to this study.

To investigate the impact of context on learning, I will start by discussing the role of episodic and semantic learning for this research; next, context will be defined more specifically as its impact on learning. Further, retrieval practice research and its role in immersive collaborative learning will be discussed. Finally, a remarkable amount of the background is dedicated to creating the simulation according to the findings of learning sciences in VR and multimedia. Afterward, the empirical investigation will be described in detail, resulting in a critical discussion.

**Relevant memory systems: Episodic and Semantic memory**

Retrieving specific past experiences and predicting future experiences is a unique capability of humankind. Episodic memory involves recollecting specific experiences and events (Tulvig, 2002). Episodic memory is a subtype of the declarative memory system and is crucial for knowledge acquisition (Eichbaum, 2000), which will be the focus of this study. The recollection of an event is based on a unique formation of neural code that includes the information of what (the content of the experience itself), where (position of the agent), and when (sequence of the event). This unique ensemble of activities is necessary to avoid interference with memories. When confronted with a specific retrieval cue to the encoding environment, these elements can be associated, integrated, and finally recalled as a whole (Sugar & Moser, 2019). The core of the episodic memory system is represented by the hippocampus, which operates rapidly and where learning occurs due to synaptic changes (Xue, 2022). It is strongly involved specifically in vivid visual representations of scenes (Rolls, 2022), which is why VR technology has become highly relevant in episodic memory research (Smith, 2019).

While episodic memory is characterized by temporal and spatial specificity, the other subtype of declarative memory is called semantic memory (O'Reilly & Norman, 2002) and is concerned with generic and context-free knowledge. It is a network embossed by lifelong learning experience and located in the neocortex. This memory system operates slowly through discovering structures in the ensemble of experiences. Neocortical networks gradually assimilate or accommodate hippocampal memories into structured knowledge. Therefore, the so-called complementary learning systems are somewhat more interconnected (Xue, 2022) than dichotomous (Renoult et al., 2019).

Cognitive flexibility enables deep and abstract learning, which is important for future-focused education. Real-world problems can be solved by employing generic and semantic knowledge and, therefore, be the focus of education (McPhail, 2021).

**Transformation from episodic to semantic memory**

Consolidation is dynamic and generative due to the modification and reconstruction of experience-dependent internal representations. It is episodic if memory depends on the hippocampus and context-specific retrieval cues. Once these cues are not necessary anymore to retrieve memories, they are semantic. Therefore, the transformation is based on an interaction between novel incoming information and existing long-term knowledge (Xue, 2022).

As previously stated, the context in which information is encoded and retrieved impacts the retrieval success and the interaction between the hippocampus and neocortex. Item-related and context-related information is bound together, leading to the consequence of recollection interference in case of a contextual encoding-retrieval overlap. Context drifts can decrease the risk of interference by avoiding contextual overlap (Yonelinas et al., 2019). Caused by a change in context during a learning episode, the post-encoding-related activity differs from the initial encoding activity and adds a new contextual feature to the memory trace. This additional feature then influences the interaction of the hippocampus and neocortex and supports the connectivity of the semantic network (Yonelinas et al., 2019). As more features are added to the trace the more decontextualized the memory (Xue, 2022).

In addition to incorporating new features into the memory trace, the competitive trace theory (Xue, 2022) assumes that competitive traces are built by including more information that is repeatedly retrieved and reconstructed based on the overlaps of the traces. The overlap represents core memory, while the contextual information differs in the competitive traces. New details are either integrated or replaced with every co-activation of a competitive trace. The semantic similarity or overlap enables favorable gist extraction, which leads to the extension of connections in the semantic network (Xue, 2022).

Contextual variation during learning could boost the transformation from episodic to semantic memory; having controlled access to different environments makes VR a promising tool to investigate this further. Building on these findings and theories, I aim to investigate whether VR can increase the chances that an episodic event retrieved in multiple environments becomes more cue-independent.

**The influence of context on learning**

As mentioned earlier, the binding capacity of the hippocampus is triggered by context (Yonelinas et al., 2019). The word context carries a diverse meaning (Roediger et al., 2017); however, it is referred to as the physical environment here. Contrary to the contextual binding or competitive trace theory, which predicts a positive impact of contextual drifts on memory, much literature focuses on a different aspect (Roediger et al.,2017). The context reinstatement literature emphasizes the benefits of contextual overlap between the encoding and retrieval context on memory. The reinstatement process is called the encoding specificity principle (Tulvig & Thompson, 1973). The reinstating aspects of encoding at retrieval serve as a powerful retrieval cue and can enhance recall and recognition (Roediger et al., 2017). This effect was first shown in the impactful Godden and Baddely (1975) experiment, where successful retrieval was determined by the overlap of the distinct (underwater vs. on land) encoding context. Even though this encoding-retrieval paradigm is quite robust, it depends on unique context conditions. Due to the loaded meaning of context, this principle is found across different types of contexts (Roediger, 2017).

The physical context-dependent memory effect contains the encoding-retrieval overlap (e.g., same room with same arrangements); therefore, the physical environment or some features serve as a retrieval cue, which leads to the recollection of the specific event (Smith, 2013; Godden & Baddely, 1975; Smith et al., 1978). No further updates, integration, or differentiation occur. Similar effects were replicated when mentally reinstating the encoding context (Bramão, Karlsson & Johansson, 2017; Smith, 1979).

The background context effect contains quick background variations compared to the slow changeable physical environment (Isadara & Isadara, 2007). For example, Isarida et al. (2014) presented words on different colored background computer screens. A significant context-dependent effect was shown when the background color change was unpredictable for the participant (random) vs. alternating (Isarida et al., 2014).

Another relevant context dependency that does not involve the physical environment is the transfer of appropriate processing (TAP). Suppose the level or depth of encoding matches the mental operations of the test. In that case, higher effects are reported due to the reinstatement of the encoding operation (Roediger, 2017) and the TAP rationale (Agarwal, 2019).

When testing the context-dependency effect, contexts need to be distinct enough; otherwise, the retrieval cue is stronger than the context cue. Memory traces can only be built under the premise that the context has been consciously encoded (Roediger, 2017).

### ***Manipulation of Environment in VR research***

Context manipulation has been the subject of VR experiments. Parker and colleagues found that when the VR content is congruent with the real world, the VR-learned content could be transformed into the real world (Parker et al., 2020).

Recently, Shin et al. (2021) investigated the context-reinstatement effect in VR with conceptually distinct context changes (Mars vs. underwater). The context-dependency effect was replicated in this VR setting. The effect was larger if a conceptual link existed between the to-be-learned content and environment (schema consistency). These findings emphasize the importance of integrating the content into an active schema. This integration into a meaningful context only succeeds if it happens explicitly (i.e., by letting students judge the usefulness of this word in the specific environment/context). Activating only a semantic schema is not enough to reach context-dependent effects; participants must interact with it to show the interaction effects (Shin et al., 2021). Essoe and colleagues (2022) conducted a similar VR study in which participants learned two new languages in either one environment or in two distinct ones. It was found that people in the dual-context condition, who learned each material in its own distinct context, showed reduced interference and better one-week retention, but only if the physical presence was perceived as high (Essoe et al., 2022).

This indicates that context reinstatement can be replicated under specific conditions. The context has different impacts depending on the outcome variable. If the encoding-retrieval context is identical, the content might be recalled, but the transformation from a cue-dependent to an independent semantic memory fails. If it is essential that a person can perform a particular task in the same context, reinstatement is necessary. However, one must use a different approach to reach a decontextualized core semantic memory. This nuanced perspective is embedded in the contextual crutch hypothesis. Smith & Hardy (2016) successfully showed that context enhances learning at different stages of learning. Meaningful context cues might benefit the acquisition of the content, but an overuse during practice maintains context dependency and may lead to forgetting if the contextual crutch is unavailable. The same context condition performs better when testing the same meaningful/supportive context during the test, but a later recall in a different context causes context-dependent forgetting. On the other hand, students who learned in incidental context shifts instead of supportive ones experienced a decreased forgetting rate; however, the acquisition rate was lower (Smith & Hardy, 2016). Forgetting represents a counterintuitive contributor to reaching the core, cue-independent knowledge and will, therefore, be a key to the intervention of the underlying study.

**Forgetting as a contributor to learning**

The "new theory of disuse" by Bjrok and Bjork (1992) assumes that each memory representation consists of two types of strengths: (1) Memory is characterized by storage strength, which describes how interconnected the memory is, (2) Retrieval strength represents the ease of current accessibility of the memory. The higher the level of storage strength, the larger the gain in retrieval strength, which is the result of retrieval practice. This indicates that the higher the current level of storage strength, the smaller the gain in storage. Thus, the loss of retrieval strength, which means forgetting, allows a gain of storage strength to enhance learning. Through desirable difficulties, such as manipulating the context, forgetting can be triggered (reduction of retrieval strength), which can lead to learning in the long run. Contextual variation, one of the unique affordances of VR, allows one to vary the context in which one encodes or retrieves, which is different from where the final recall occurs. Instead of easing the accessibility during retrieval practice (Bjork & Bjork, 2019), which would cause a context effect, forgetting can be seen as a form of neuroplasticity that enables cognitive flexibility (Ryan & Franklin, 2022).

**Retrieval Practice as a consolidation focused strategy**

In addition to contextual variation, the type of retrieval practice before the final recall is the second desirable difficulty that triggers forgetting. Retrieval practice is integral to the consolidation processes of the episodic learning experience. Compared to passive restudying of just learned information, active recall leads to improved long-term memory. The enhanced long-term performance of repeated retrieval practice is called the testing effect. Passive retrieval based on rereading or rewatching the material performs better in the immediate post-test (Roediger & Butler, 2011). Roediger and Karpicke (2006) tested the restudying vs. practicing prose material (free recall) on immediate and delayed memory in a more applied setting. It was found that the restudy groups performed better on the immediate test, but testing groups dramatically outperformed the restudy groups in the later post-test. The authors similarly argue that these results confirm the desirable difficulty prediction (Roediger & Karpicke, 2006).

The retrieval practice activity represents the final retrieval test activity in the classic retrieval-practice paradigm. However, more recent and applied research in a classroom setting describes the practice activity more as an “active attempt” to recall, recognize, and reconstruct memory during initial learning. The process of practicing is at the center of attention rather than the test format (Agarwal, 2019). Similarly, Butler and Roediger (2007) emphasized that production tests produce superior retention compared to recognition tests. Storage strength increases with every effortful practice, but retrieval strength is reduced (forgetting) during that practice. In the post-test, the retrieval and storage strength are increased (Butler, 2010). The bifurcation model can also explain the test-delay interaction. The to-be-learned material is distributed on a memory strength distribution during the study phase, and when restudying, this distribution moves to the right. On the contrary, the retrieval practice bifurcates the distribution into two parts. The successfully retrieved information is strengthened more than the restudied ones. However, the information not recalled is left with the same memory trace strength. This explains why the restudy group only performs better in the immediate test. Immediately after studying, more items are still accessible (see retrieval strength). However, after some delay, this strength will decrease faster than in the tested condition and, therefore, higher performance in the delayed test. In contrast, in the restudy condition, all information is strengthened equally weaker than the practiced ones (Halamish & Bjork, 2011). Following this theory, a lack of feedback during retrieval practice can lower the long-term memory benefits of retrieval practice (Storm et al., 2014). Feedback enhances the testing effect by increasing the mnemonic benefits of testing, allowing the reactivation of the memory trace (Racsmany et al., 2020). The role of feedback and concept learning versus restudy of the concepts has been investigated in a computer-assisted learning situation. The feedback and concept group performed better in the immediate and delayed post-tests (Wiklund-Hörnqvisr et al., 2013). Independently, suppose the retrieval practice result is incorrect. In that case, feedback enhances learning (Roediger & Butler, 2011), especially when it comes to collaborative retrieval practice (Vojdanoska et al., 2010) to avoid collective inhibition. Wissmann and Rawson (2015) showed consistently that collaborative retrieval practice leads to improved results only in the delayed individual post-collaborative test. Nevertheless, the feedback exposure must be conducted adequately to avoid feedback-induced reversal of the testing effect. The feedback should not be too immediate or repetitive and should include a correction instead of just true or false (Racsmany et al., 2020).

In the classic paradigm, testing effects result from multiple choice practice format. However, Agarwal recently tested the different taxonomy types (factual vs. higher-order learning vs. mixed) in combination with retrieval practice. It was shown that higher-order retrieval practice compared to factual learning led to an improved higher-order learning outcome at the final test (Agarwal, 2019). Besides this approach, most studies still fail to test transfer knowledge in the retrieval practice framework. It is described as the “holy grail” of education (Agarwal et al., 2021). The most recent meta-analysis revealed that the most significant effects are in middle school education, STEM-based learning, and between-subject design. In lab settings, more significant results are found in long delays; in a classroom setting, shorter delays between immediate and delayed post-tests were more beneficial (Agarwal et al., 2021). Bjork recently combined both forms of desirable difficulties, retrieval practice, and contextual variation. The contextual variation only benefited the restudy group in the immediate post-test. The participants either restudied or practiced in the same vs. new environment, and the final recall context was new for all conditions 48 hours delayed. Results reveal an Interaction effect between contextual variation and restudy. The contextual variation group benefitted from restudy but not practice. In the same context, the practice group performed better than the restudy group in the final retrieval (Imundo et al., 2021). The retrieval practice group might have shown lower effects due to the missing feedback during practice in this study design.

Achieving active cognitive learning must not necessarily go in hand with high behavioral activity during learning. Specifically, in Multimedia learning, the passive instructional method (reading a passage or watching a presentation) can lead to high cognitive activity as much as active instructional methods (discovering a solution to a problem) do. Passive instruction designed to reduce extraneous processing and manage essential processing can lead to generative processing. A higher behavioral activity during learning is caused by an active instructional design that guides the learner to discovery, so the extraneous load is decreased, which frees up more capacity for essential and generative processing (Mayer, 2009).

***Testing effect in Multimedia learning etc.***

The delay interaction effect of retrieval practice has also been shown in a multimedia learning context. Mayer (2009) investigated transfer knowledge as retrieval practice and demonstrated the transfer-appropriate processing of the testing effect. Adding images to a classic retrieval practice situation increases perceived competence but not necessarily the learning outcome. Images revealing too clearly the answer during practice negatively affected the results. On the other hand, images that partially supported the retrieval were perceived as helpful, especially when they were shown after retrieval as feedback (Van den Broek et al., 2021).

## Collaboration in VR

Besides the effect of the environment and retrieval practice on memory in VR, the interaction with the environment should be considered, especially regarding collaboration in VR. The cognitive and emotional processes during learning are different with this new type of media. The Cognitive Theory of Multimedia Learning (Mayer, 2014) and the Cognitive Affective model of Immersive Learning (CAMIL) have been developed (Makransky & Petersen, 2021). According to cognitive theory of Multimedia Learning by Mayer (2014), a well-structured instructional design can prevent one from being overwhelmed by extraneous processing to provide resources for generative and essential processing. Essential processing describes the cognitive processes necessary for mentally representing the material. Generative processing implies understanding and making sense of the material (Mayer et al., 2018). CAMIL is specialized in IVR (immersive virtual reality) research and assumes that media interacts with the method. For example, the significant interaction between instructional methods and media only led to increased retention, transfer, and self-efficacy in VR compared to desktop VR. According to CAMIL, learning in IVR environments allows unique affordances that interact with instructional methods. The affordances include presence and agency, which influence cognitive-affective factors (Makransky & Petersson, 2021). Theory of Immersive Collaborative Learning (TICOL), the extension of CAMIL, focused on Collaboration in VR and added social presence and body ownership as psychological mediators. These psychological mediators are posited to influence further social interactions and cognitive and socio-emotional quality, which fosters a strong social space and, ultimately, better learning outcomes. In this study, I mainly focus on the learning outcomes, but the psychological mediators play a role in the exploratory analysis. Physical presence represents the psychological experience of virtual physical objects as actual physical objects; social presence the psychological experience of virtual social actors as actual social actors (Lee, 2004); body ownership (the illusion that a virtual body belongs to oneself; Slater et al., 2022) and agency (the sense of being the one who is causing or generating an action; David et al., 2008).

Sweller and colleagues (2018) developed a framework of collaborative learning intending to reduce the extraneous and working memory load through cooperation. According to the mutual cognitive interdependence principle, element interactivity (which is even more complex in a VR environment), and the limited individual working memory capacity extension led to a collective working memory that lowered the extraneous load and improved learning outcomes. Interacting elements can be distributed (distribution advantage), and interindividual communication can form more efficient knowledge structures. This mental advantage of collaborative learning frees up mental capacity and facilitates deeper learning (Janssen, Kirschner & Kirschner, 2022).

Even though collaboration can benefit learning, individual differences in working memory capacity significantly influence spatial abilities in VR (Escamilla et al., 2020). A recent study compared the differences in learning outcomes when adding a generative learning activity to the instructional design in an individual vs. a collaborative VR experience. The results showed that the students in the collaborative condition benefitted more from the generative activity (Peterssen et al., 2023).

Nonetheless, collaborative action is only superior to individual learning under certain conditions. One of them is positive interdependence, which includes the perception that the task can only be solved when collaborating, which results in transactive memory. Students build on each other's comments and insights when engaging in transactive discussions, creating shared knowledge. Only complex tasks can trigger these types of interactions. Further, prior knowledge of the individual learner will influence the collaboration, and therefore, it is essential to provide all members with sufficient knowledge. Supporting self-regulation and collaboration skills can facilitate transactive thoughts and decrease transaction costs (Janssen, Kirschner & Kirschner, 2022). The learner's interaction should be close to the natural face-to-face communication when a collective shared mental model is built initially (Kirschner et al., 2018).

### *Risk factors of collaboration*

Even though learning and retrieving together facilitate extensive benefits, such as collective memory that can result in a supportive, we-mode, collaboration can also have opposite effects. One of these risks, or as Kirschner (2022) would call them, "transaction costs," consists of collective inhibition. It indicates that collective retrieval generally leads to positive results but is lower than the nominal group, which contains the retrieval potential of the individual. This robust phenomenon is explained by the Retrieval strategy disruption hypothesis (RSDH), claiming that individuals organize newly learned information according to their prior experience, schema, and expectations of the retrieval context (Badsen et al., 1997). This subjective strategy is disrupted when collaborating with others. The most recent meta-analysis identified five moderators influencing collective inhibition (Marion & Thorley, 2016), which will be explained in the following.

The lower the group size, the less collective inhibition occurs, and the disruption of others does not decrease productivity. Regarding the study material, organized structures and story-like items allow a clear structurization strategy for all members. For example, highly uncategorized learning tasks such as brainstorming trigger production blocking of the individual and, therefore, is quite sensitive towards collaborative inhibition. Similarly, turn-taking while working on a task disrupts the individual working memory capacity and a standardized time to solve or answer delays generation and production. Further, the more familiar the collaboration partner/s, the lower the risk for collaborative inhibition (Marion & Thorley, 2016).

Regarding the effect of collaboration on individual memory post-test, the meta-analysis shows enhanced results when retrieval practice happens in a team. The rebound effect can explain this post-collaborative advantage, which supports RSDH and predicts a rebound from any potential retrieval strategy disruption experienced during collaboration on later individual remembering (Marion & Thorley, 2016).

### *Strategies to Avoid Collective Inhibition*

To reduce risk factors of collaboration and ensure interdependence between collaborators, scripts that guide the interaction are essential, especially in cognitively demanding practices such as VR learning (Fischer et al., 2013). This section is important for the methods section and the creation of the simulation.

Due to limited internal scripts, high-level interaction cannot be achieved. Learners with constrained prior experience in computer-supported collaborative (CSCL) learning might not interact in ways that benefit the collaboration. The script theory of guidance is based on a dynamic view of memory and a socio-cultural perspective, indicating that discourse activities shape the structure of complex cognitive skills (Fischer et al., 2013). This view assumes external scaffolds can support limited internal scripts to enhance collaborative activity. Both types of scripts consist of a play (knowledge about the main task), scene (single subtask), scriptlets (sequence of the subtasks), and roles (knowledge about the responsibility during collaboration) dynamic components. Perceived situated characteristics and current goals of the internal script influence how people operate during collaboration. Once the internal script does not lead to a successful collaboration, it will be reconfigured with the help of the external script. The better an external script prompts transactive thoughts, the more knowledge acquisition takes place and the best possible learning results by creating situational affordances; a potentially dysfunctional internal script can be inhibited through taking turns, timely restricted problem solving, and a categorized answering behavior (Fischer et al., 2013).

Research on the efficiency of scripts shows that providing external scaffolding versus none leads to higher-quality collaboration (Ratkowitsch et al., 2020). Nevertheless, the risk of over-scripting needs to be considered. If there are too many repeated restrictions during collaboration by the script, the application of self-regulated internal scripts is inhibited, lowering the chances of knowledge acquisition. To avoid this consequence, some prompts should fade out. Vogel and colleagues (2022) investigated in a recent study adaptable versus non-adaptable scripts in CSCL learning on socio-discursive skills and domain knowledge as well as the role of self-regulation skills during the collaboration. Adaptable scripts showed small benefits compared to non-adaptable ones on direct learning outcomes. Learners with high levels of self-regulation skills benefited mainly from adaptive scripts. Therefore, self-regulation skills should be supported (Vogel et al., 2022). Wang and colleagues (2016) found that adaptable scripts enhance self-regulation skills during learning. Adaptable scripts enhance the transfer performance of the learner. Further, there was no difference between static or adaptable scripting in its effect on perceived autonomy and social relatedness. However, adaptable scripts showed higher ratings in perceived competence (Ratkowitsch et al., 2020). A meta-analysis shows that CSCL scripts positively impact domain knowledge and strongly affect collaboration skills. Interestingly, scripts that prompt transactive activities between learners through content-specific scaffolds on the scene level by choosing learning activities such as worked examples or concept maps increase domain knowledge (Vogel et al., 2021). Scripts that prompt dynamic feedback are helpful for domain knowledge and collaboration skills. It prompts students to engage and enhances critical thinking, reasoning skills, and conceptual learning (Hmelo-Silver & Heiswan Jeong, 2022).

**A powerful interplay of consolidation-focused and construction-focused strategies**

Research has shown that implementing instructional design principles that aim to reduce the individuals' cognitive load, which is made possible using VR, allows learners to engage more in generative processing (Mayer et al., 2022). The Generative Learning Theory assumes that selecting, organizing, and integrating learning content deeper leads to more elaborative learning. The learner uses incoming information and transfers it into usable information to further transform that knowledge (Parong & Mayer, 2018). Recall the testing effect, which facilitates knowledge consolidation – in the past, these two paradigms opposed each other. Roelle and colleagues (2022) promote combining these two different but also complementary types of learning, providing retrieval practice but with generative activities. They acknowledge that these are fundamentally different as retrieval practice leads to lower forgetting rates than generative learning, while generative learning leads to deeper understanding. Nonetheless, deeper comprehension can also reduce forgetting rates and increase the depth of understanding, and consolidation not only leads to lower forgetting and increases the depth of understanding. By engaging learners in constructing and consolidating mental representations through meaningful retrieval practice tasks that require explanations, problem-solving, or transfer, retrieval practice's direct and indirect effects are more challenging to reach. Direct effects indicate that at least 75% must be retrieved during practice to avoid bifurcation. This is much harder to reach with tasks that require deeper comprehension as learners are more prone to errors, which reduces factual knowledge gain. Therefore, including measures that improve correct retrieval through instructional support is crucial. Indirect effects include the practice's effect on metacognition, motivation, or activation of prior knowledge and that learners get supported to monitor their state of knowledge. In a generative retrieval practice, this monitoring becomes more complex and demanding, which could lead to an overload. By providing precise feedback monitoring, indirect benefits could be reached. For example, drawing is often used in STEM domains to enhance understanding. Providing learners with instructor-generated drawings after creating their own showed promising feedback benefits to update their knowledge construction (Fiorella & Zhang, 2018).

Most studies that combined retrieval practice and generative learning strategies focused on finding the most beneficial sequence of the strategies and inferred underlying cognitive processes (Waldeyer et al., 2020; Ortega-Tedula et al., 2019). Some recommend construction-before-consolidation, but more recently, also the other way around (Roelle et al., 2022). This approach is comparable to Bloom's Taxonomy of Learning (Krathwohl, 2002), which suggests that factual (lower-order learning) needs to be ensured before applying the knowledge. On the other hand, the Transfer-appropriate processing approach emphasizes the encoding-retrieval activity match, explicitly explaining the long-term testing effects. When comparing the effects of factual retrieval vs. higher-order or mixed order and restudy, it was shown that when engaging in higher-order quizzes and delayed tests, higher-order tests were performed by the matching practice group. These results emphasize the importance of the desirable difficulty framework and the transfer-appropriate processing. Furthermore, when retrieval-based concept mapping and paragraph writing were compared to the passive restudy condition, both practice activities showed significantly enhanced delayed post-test results (Blunt & Karpicke, 2014). Following these indications, the consolidation-focused method of retrieval practice will be combined with a generative learning strategy established in multimedia learning.

## Collaborative Concept Mapping

The generative learning strategy of choice in this research is represented by Collaborative Concept Mapping, which can be seen as the entailment of the cognitive theories of generative learning (Adesope, 2021). When creating concept maps, information is integrated into a combined verbal and visuospatial format. According to meta-analyses (Schroeder et al., 2017), better learning results follow if people jointly translate and integrate information. Spatial proximity or direction is used to convey semantic similarity or relationship. While vertical dimensions indicate generality, horizontal ones indicate relations and details (Adesope, 2021). Nodes represent the concepts, and the labeled lines indicate the relationship between nodes (Farrokhina et al., 2019). Images can also replace concepts. Concept mapping prompts students to identify and select content based on the learning goal (Concept selection stage). Information is understood through building connections to prior knowledge (Gist extraction stage). They organize knowledge in specificity and hierarchy structures and relate concepts via directional arrows or labels (Adesope, 2021). Schroeder and colleagues (2018) found overall beneficial effects on individual learning ($g = .58$). Most concept mapping studies have been conducted in STEM-based education in secondary and university education. Adesope (2022) re-analyzed the results and differentiated collaborative concept mapping effects. The findings showed a large *(g=1.20)* overweighted mean on retention and transfer in collaborative concept mapping. Testing collaborative mapping in connection to the Signalling, Interleaving, and Spacing principles was also recommended.

To reach the best possible learning results, providing examples of concept maps before starting the procedure is suggested (Blunt & Karpicke, 2014; Farrokhina, 2019). Usually, the turn-taking takes place by allowing switching “ownership” of building parts of the map. Through the knowledge co-construction with peers in CSCCM, externalization of own mental models, elicitation and accepting partners' reactions as well as integration of partners' recommendations into the team's map, conflict-oriented consensus can be elicited, which represents divergence thinking of the group (Farrokhina et al., 2019).

Specific risk factors of CSCCM consist of ill-structured maps that allow too many possible solutions. Therefore, a moderately structured concept that decreases cognitive load is recommended (Hattami, Farrokhina & Hassanzadeh, 2016; Chin, 2003). Not all studies confirm the positive effect of collaborative mapping. It was reported that students rarely reached the explanatory level (Van Boxtel et al., 2002), and on-task discussions were mainly devoted to process-oriented themes (Chiu, 2003). The risk factors emphasize the importance of scripts described earlier.

## The current study

This study investigates whether the difficulties of contextual variation and retrieval practice via Mapping are transformable to VR and lead to enhanced learning. A factorial design with the factor context (same vs. varied) and type of retrieval (active/Mapping vs. passive) was implemented in a VR lesson. The post-tests took place in VR immediately after the lesson and one week later in the classroom. With the intervention of contextual variation (change of physical environment) during learning, the study will investigate whether the decontextualization of episodic memory is reachable with VR technology. Based on the previous literature review, the context effect could go either direction.

H1: The VR context in which the retrieval practice occurs affects learning performance when controlling for subjective estimates of prior knowledge on the topic.

H1a: Based on the context reinstatement findings, the same context groups are predicted to perform better in the immediate post-test in VR than the varied context groups, controlling for subjective estimates of prior knowledge on the topic.
H1b: Based on the contextual binding, competitive trace theory, and beneficial contextual variation (desirable difficulty), it is predicted that the contextual variation groups perform better at the immediate post-test in VR than the same context groups, controlling for subjective estimates of prior knowledge on the topic.
The ecological validity of the Testing effect via Mapping in VR will be examined. Contrary to Imundo et al. (2021), feedback will be included, which is why a non-directional hypothesis is expected.
H2: Based on the literature on the testing effect and collaborative Concept Mapping, it is expected that students in the active retrieval condition will perform better on the immediate post-test in VR than the passive groups, controlling for the estimate of subjective prior knowledge on the topic.
Based on the findings of Imundo et al. (2021) and Bjork & Bork (2019)., the interaction of both desirable difficulties is scrutinized.
H3: There is an interaction between context (same vs. variation) and retrieval type (active vs. passive) on the immediate post-test in VR, controlling prior knowledge on the topic.
Due to consolidation processes, delayed effects are also subject to this research to study whether there is an interaction with time on the main factor and their potential interaction.
H4: The contextual variation (vs. same context) positively interacts with time in its effect on the delayed post-test when controlling for the subjective estimate of prior knowledge on the topic.
H5: The active retrieval (vs. passive) positively interacts with time in its effect on the delayed post-test when controlling and subjective estimate of prior knowledge on the topic.
H6: Context (same vs. variation) and type of retrieval (active vs. passive) interact with change over time when controlling for subjective estimate of prior knowledge on the topic.

The impact of the psychological mediator's agency, physical presence, social presence, and body ownership on the learning outcomes will be investigated as an exploratory analysis. According to the TICOL framework (Makransky & Peterssen, 2023) and findings from Essoe et al., (2022), these constructs should positively affect learning.

# Method

## Participants

159 Danish high school students in the capital region of Denmark fully finished the experiment (*Mdn* = 16). 89 identified as male, 62 as female, four as non-binary, and another four preferred not to answer. Consistent with the fact that Danes are ranked among the top on the EF English Proficiency Index (EF Education First, 2022), 89 of the participants rated their ability to comprehend English as very high; 55 rated it as somewhat high, 33 as average, five as somewhat low, and four selected the "very low" option. Generally, students were not very experienced with VR (*M*=2.67, *SD*= 1.41). 18 students had never experienced VR before, 78 between 1-3 times, 36 between 4-10 times, eight students 11-20 times, eight between 21-50 times, five students between 51 and 100, and only six students said they used it more than 100 times. Students scored very low on a self-reported science interest, experience, and ability measure (SIEA), with a median of 1 and a mean of 1.40 (SD=1.49); the reachable SIEA scores were between 0 and 10. The median of subjective prior knowledge of the human body was three, corresponding to "average."

A priori power analysis was conducted for a multiple regression model (initially with two moderators) in R with the pwrss package to estimate the required sample size. Even though this study was based on paradigms and findings from retrieval practice and collaborative concept mapping, which have shown medium to large effect sizes (Agarwal et al., 2021; Fiorella & Mayer, 2016), a lower expected effect size was chosen due to the novel design (VR research and combination with effect of context). Therefore, the $r^2$ was set to a weak effect ($r^2$ = .05), power of .80, and alpha level of $p$ = .05. This gives us a minimum sample size of 187. The final sample size was not reached due to technical issues with a few headsets that failed to download the data, students not answering the post-test in VR, and students dropping out after the pre-test.

### *Ethics and transparency*

The study design, hypotheses, sampling plan, and analysis plan were preregistered(https://osf.io/345rm). The Ethics Committee of the Psychology Department of Copenhagen University has approved the study.

## Materials

### *Instructional Materials -VR lesson*

Depending on the condition, participants were given instructional material consisting of different parts of a VR lesson. The base version of the educational VR experience contains "The Body VR: Journey Inside a Cell," in which the learner travels with a first-person perspective through the bloodstream and discovers the different organelles and how they work together. The Body VR video (The Body VR, 2016) was used as it has previously been used as the main learning material in studies investigating the impact of generative activity on learning and memory (Parong & Mayer, 2018; Meyer et al., 2019).

The original version was shortened in the beginning (explanation of different types of blood cells) as well as the last part (Virus attack) due to practical reasons of total time in VR and potentially extraneous cognitive overload; however, mainly due to the generative learning strategy of Concept Mapping, which requires concepts being relatable to one another. The two mentioned parts cut out were too isolated in a concept map, which led to a lack of information integration. The video's final version was 9 minutes 16 seconds with 668 spoken words. The video was in English and equipped with English subtitles to support the comprehension of the biological explanations.

Before the video started, students were given a choice of ten avatars with different appearances and gender. They were instructed to choose an avatar representing them in the virtual world. Participants saw themselves in the mirror, and by using their virtual hands, they could switch their appearance and choose their avatars. The avatars were sourced from the MQuest unity avatar library (Gonzalez-Franco et al., 2020). After the avatar was picked, students saw each other for the first time in a neutral environment, where they were given further instructions on the procedure of the virtual lesson.

The adapted video started in the bloodstream, with the first investigation consisting of the cell membrane supplied with receptor proteins to experience how crucial elements enter the cell or get refused access. Further, students experienced the structure of cytoplasm and skeleton, the creation of ATP, and the release of ADP. Next, the VR experience guided the learners toward the nucleus. Here, students learned and experienced how DNA and RNA are related and their impact on protein synthesis. This crucial protein creation is explained in the last part of the video. It showed the learner how the ribosomes, which are located around the

Rough Endoplasmic Reticulum (RER), receive instructions from RNA about protein creation and their path of transportation. Students saw each other constantly due to the potential positive effect of joint attention on learning (Kim & Mundy, 2012).

After the video, the experimental condition was exposed to a tutorial (see appendix A) in which learners were guided through the different hand movements and buttons on their controllers necessary to interact with objects in VR. Each person was allowed to try the different functionalities to feel capable of the following learning activities.

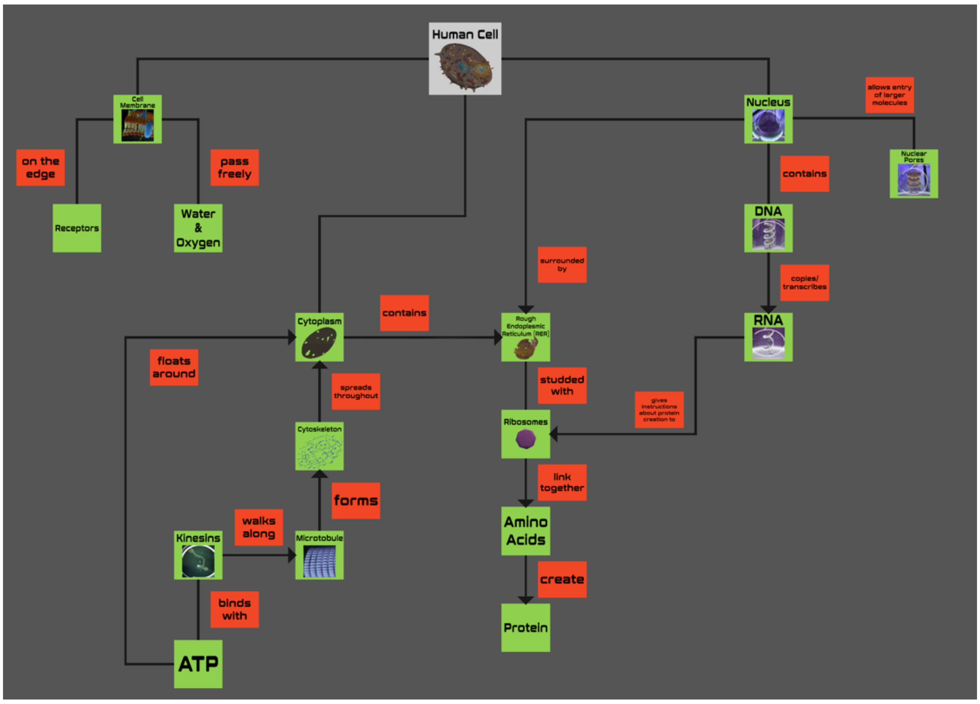
***Retrieval practice in VR lesson***

The generative learning method of collaborative concept mapping followed either the tutorial or right after the video (depending on the condition). The students were exposed to a moderately structured concept map that required 19 slots to be filled by the participants (in the active condition). In total, the map consisted of 32 components and links; figure 1 shows the concept map and the script information that was always visible to the learners.

To generate beneficial transactive discussions, an adaptable external script was implemented. Students were instructed to take turns selecting and moving the concepts into the determined slots. The laser attached to the virtual hand indicated which player's turn it was. As described in the theoretical background, turn-taking is not only a typical strategy in collaborative mapping but also triggers transactive discussions that are also supported by the emphasis on encouraging participants to elaborate and consult each other during the learning activity (Janssen et al., 2022). Furthermore, situational affordances to inhibit dysfunctional internal scripts (Fischer et al., 2013) were achieved by including a 5-minute time restriction per map. In total, students were exposed to three maps. In the experimental conditions, the maps were differently structured in terms of which components were filled out by default to ensure that every component had been retrieved by an “active attempt” to avoid bifurcation. Students were instructed to give each other a high-five with their virtual hands to receive dynamic feedback. The feedback included once on each map, which signaled misplaced components, and participants were given one chance to correct themselves. The fully corrected map proceeded if this was not corrected, ensuring equal retrieval practice across conditions. After, the verbal instructions faded out, while the non-adaptable ones were constantly visible in the script.

**Figure 1**

*Concept Map on a neutral background*




*Note.* The screenshot shows the completed map with all the components students had to actively retrieve (active conditions) or restudy (passive conditions).

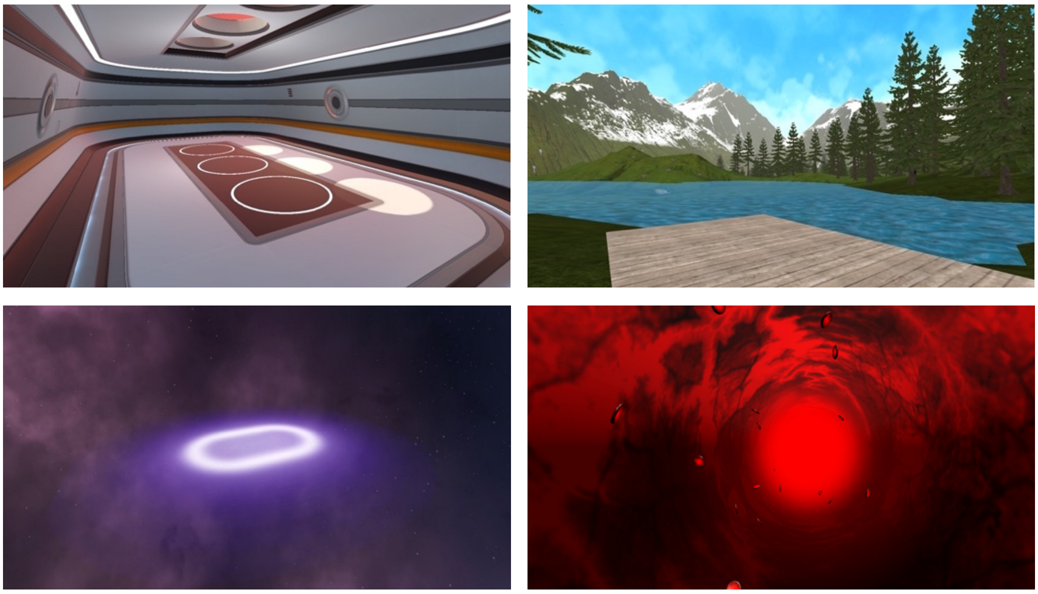
The concept maps presented to the control groups were the same corrected maps shown to the experimental condition. The differences between the control and experimental groups were that participants in the control did not interact with the concept map required, and no actions were taken. To proceed in the control condition, students were also instructed to give each other a high-five (see figure 4).

### *Contextual Variation in VR Lesson*

During retrieval, students in the contextual variation condition were transported to three different environments in randomized order (see Figure 2). The same context groups retrieved or restudied the map in the bloodstream.

**Figure 2**

*Contexts of retrieval*



*Note.* Screenshots show the three different environments that participants encountered in the variation groups. The image on the lower right shows the retrieval environment of the same context groups.

## Measures

### *Pre-questionnaire*

A pre-test that consisted of 16 items covering demographics (age and gender), VR experience (Never; 1-3; 4-10; 11-20; 21-50; 51-100; More than 100), and English proficiency (on a 5-point-Likert scale from very high until very low) was given to participants to control for potential confounders and other variables that could influence the results. Interest in experience with science-related topics and classes was assessed with ten dichotomous (Yes, No) statements, for example: “I earn mostly A´s (12) and B´s (10) in my science classes in high school” or “I would like to have a career in a science-related field. By combining all 10 item answers, a science interest in and experience with science-related topics score was computed. Cronbach´s Alpha of SIEA scale had a Cronbach’s Alpha of .61. Finally, the perceived knowledge about the human body was assessed on a five-point Likert scale from very low to very high (“Please rate your knowledge of the human body”). These questions were oriented on a similar study design by (Parong & Mayer, 2018).

***Immediate Post-test in VR***

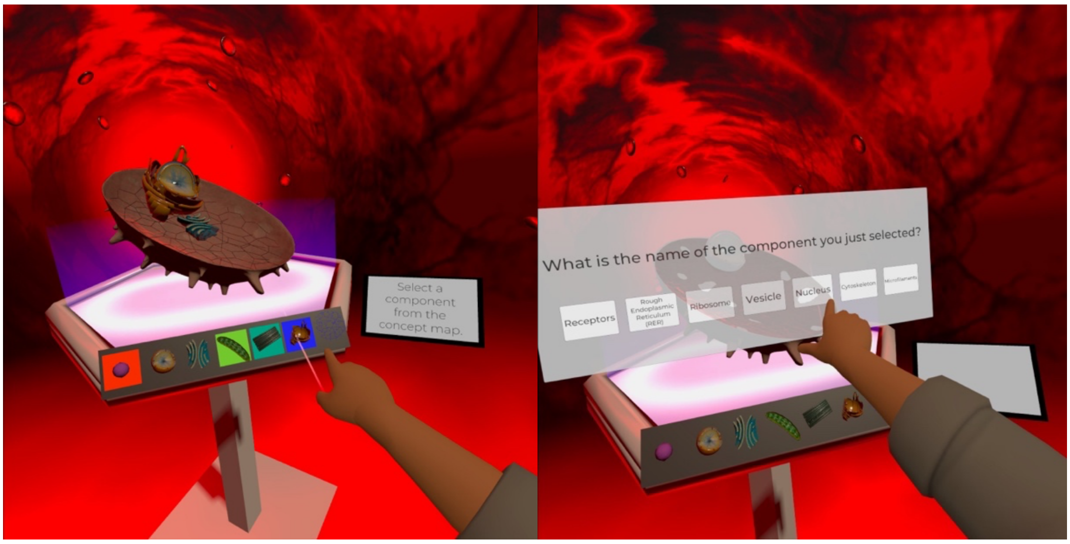
After the main part of the study, students completed a post-test in the bloodstream VR environment. The immediate VR test assessed factual, conceptual, and transfer knowledge. First, ten factual multiple-choice questions, with four potential correct and just one correct answer, were asked (e.g., “On what structure does Kinesin walk along?”; “Which of the pictures shows the Nuclear Pore?” or “What is RER studded with?”). All the items with screenshots can be found in the appendix B and C. The questions were strictly focused on the concept maps they were exposed to but not necessarily the video. The Cronbach´s Alpha scale indicated weak reliability with .28.

For the conceptual question, the process-related connection between cytoplasm and ATP was asked (“How is ADP released?  Choose the elements and links that you believe are necessary and bring them in the right order”). Participants were offered nine components that had to be placed in the correct sequence, consisting of seven slots. The sequence's start (Cytoplasm) and endpoint (ATP) were given to scaffold the task. The components were movable in the same manner as during the collaborative concept mapping. Cronbach´s Alpha on this scale was .59.

For the transfer task, students found themselves in front of a workstation. Right underneath the workstation, potential components belonged either to the animal cell or not (see Figure 3). Students were asked to select components of the human cell and name them once they selected them. When students hovered over the miniature visualization of the cell component, they could observe on the workstation the 3D visualization of that component related to other selected components. Once selected, students had to name each component (in a multiple-choice format). Out of seven components, only four were relevant for this task. This task can be considered as near transfer (Barnett & Ceci, 2002).  The task was still focused on the topic of cell components of animal cells, but they have never been presented with 3D components as they have only experienced it “inside” the body in VR. The 3D cell components were imported from the Unity Asset Store. Cronbach´s Alpha was not computed, as the sequence was not tracked, so the way items were selected was not comparable across participants.

**Figure 3**

*Near Transfer task in VR post-test*




*Note*. Screenshots of post-test in VR. (Left) transfer selecting, (right) transfer naming.

***Post-questionnaire – outside of VR***

Right after the VR experience, the post-test focused on evaluating the VR experience and the psychological mediators relevant to individual and collaborative VR experiences. To measure the individual physical presence, the scale from Makransky et al. (2017) was used with a 5-point-Likert scale from completely agree to completely disagree (e.g., “I had a sense of “being there” in the virtual environment.”) The quality of the scale had a Cronbach’s Alpha of .79. Further, three items for the agency were borrowed to assess how much control students felt they had in terms of their actions in VR (Polito et al., 2013). The same answer format was used as physical presence (e.g., “During the lesson, I felt that I did not cause my experiences and actions”). Here Cronbach´s Alpha was .49. For body ownership, the subscale from Peck & Gonzalez-Franco (2021) embodiment questionnaire was used to measure how much the students perceived the avatars body as their own. Here the recommended 7-point Likert scale from strongly disagree to strongly agree was used (e.g., “At some point, it felt as if my real body was starting to take on the posture or shape of the virtual body”). Cronbach´s Alpha indicated a quality of .70.

The Social Presence Scale (Makransky, Lilleholt & Aaby, 2017) measured th unique psychological phenomenon of perceiving other social persons as being physically real in a virtual environment. The scale included the sense of co-existence attribute, the participant's perception of the avatar representations' credibility (Human realness), the effect on the social interaction between the artificial humans, and the unawareness of the social mediation of the interaction. This scale was tested in a VR learning simulation and indicated a high Cronbach's alpha value of .90 (Makransky, Lilleholt & Aaby, 2017). The social presence scale in this investigation led to a Cronbach's Alpha of .67.

Further, the cognitive load was measured with the extraneous cognitive load (4 items) in an environment specifically for immersive learning experiences (Andersen & Makransky, 2021), with a 5-point Likert scale from completely agree to completely disagree (e.g., “The elements in the virtual environment made the learning very unclear.”). Cronbach´s Alpha of .74 was computed.

Motion sickness was measured using the Virtual Reality Sickness Questionnaire (VRSQ) created by Kim et al. 2018, which rated VR users’ experience of nine symptoms on a 4-point Likert scale (e.g., “General discomfort” or “Blurred vision”). It consisted of two subscales: oculomotor and disorientation. The disorientation subscale had a Cronbach’s Alpha of .79, and the oculomotor subscale .68.

Further, to account for collective inhibition, the perceived closeness to the collaboration partner was measured on a 7-point-Likert scale from very close to not close at all (“How close are you with your team partner you collaborated with in VR in the real world?”).

***Delayed Post-test – outside of VR***

The structure of the one-week delayed post-test was the same as the post-test in VR; however, it was conducted using the online tool formr. Ten different factual questions that covered the topics of the VR test equally were asked (e.g., “Which protein can walk the microtubule?” or “Which image shows the Nuclear Pore?”). For the conceptual question, another process-related sequence had to be matched with the same number of components (7) and extra components (2), which made a total of 9 movable components (“How are Nucleus and Ribosomes involved in protein synthesis? Choose the components and connections you believe are relevant and create a sequence starting with the Nucleus. Do so by choosing the correct elements for each number.”). For the transfer task, the participants had to name four components of a visualization of a bacteria cell, which were not discussed in the VR lesson. Out of seven components, only four were relevant.

***Working memory***

The individual working memory capacity (WMC) was planned to be collected at the follow-up as it represents a covariate. The WMC was measured by the forward digit span task, implemented in the post-test questionnaire. The participants had to recall a sequence of digits that increased subsequently. The task and WMC scores were assessed by the mean span score based on Woods et al. (2011).

***Apparatus***

The VR experience was conducted with Oculus Quest 1 and 2 and Oculus Pro VR headsets. Participants held an appertaining controller in each hand. The audio material was delivered via Sony on-ear headphones. Participants responded to the questionnaires given outside of VR using smartphones or computers.

***Context***

An Agreement with the Danish High School Niels Brock enabled us to recruit full classes of the 3rd year high-school students. Each class was visited one week later in their classroom to remind students to fill out the post-test. The students were not paid, nor did they receive incentives of some kind. However, agreements were made regarding follow-up visits to present results so the teachers could incorporate experimental research into their teaching.

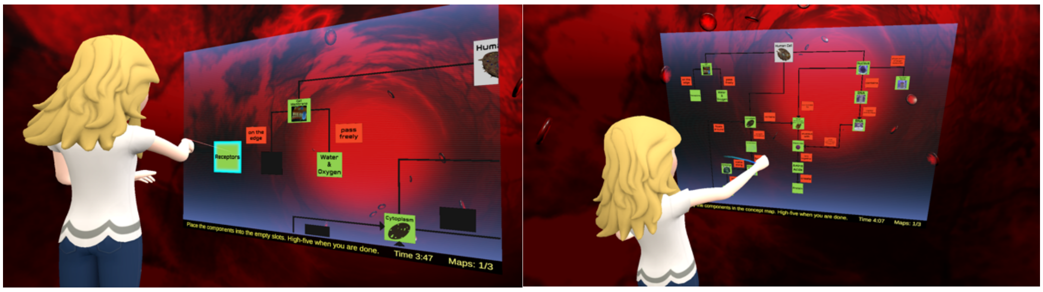
**Design & Procedure**

A $2 \times 2$ factorial design with two experimental manipulations resulted in 4 different VR lessons. The manipulation consisted of contextual variation vs. same context and active collaborative mapping vs. passive collaborative mapping. The contextual variation consisted of a physical environmental change in VR. Compared to the control groupswho stayed in the bloodstream for the entire time of collaborative retrieval practice (active and passive), students in the experimental condition experienced the retrieval practice (each map in one environment) in three different environments. The second manipulation consisted of the type of retrieval. Students were either tasked with the generative learning method of collaborative concept mapping or passively presented with the same concept map they were tasked to study. The different types of Maps between the conditions can be found below (see figure 4).

**Figure 4**

*Active vs. passive condition in varied contexts*

*Active vs. passive condition in same context*



*Note.* The upper row shows the (left) active retrieval and passive (right) in the contextual variation condition, exemplary in the sci-fi environment. The lower row shows active (left) and passive (right) retrieval in the same context condition.

After the students arrived at campus, they were instructed to the full procedure. After signing the participant information and the consent form, they filled out the pre-test questionnaire and were randomly assigned to pairs and conditions. Each pair was then taken to a separate room by one study runner. The headsets were adjusted to each person individually, and they were given a quick tutorial on holding the controllers and using their virtual hands. The study runners informed students before they started the simulation.

## Statistical Analysis

All analyses were conducted with R version 4.2.1. P values lower than .05 were considered significant. Participants with missing values in the VR were excluded from the dataset. However, the missing values in other questionnaires were kept depending on the analysis, omitted, or included in the analysis. All hypotheses were tested with a linear mixed model, with fixed effects representing the manipulation of the experiment (type of retrieval and context). The clustering variable and random effect consisted of the dyads, the pair ID that indicates the collaboration teams. To report an overall test of significance of the fixed effects, an ANOVA type 3 (Satterthwaite's degrees of freedom method) was computed based on the mixed model; results can be found in Appendix E. For the computation of the linear mixed model, the lme4 package was used (Bates et al., 2015). The *R2* for the regression models is used as an effect size measure. To examine the proportion of variation in the learning outcomes accounted for by the dyads (random effect), the intraclass coefficient (ICC) is reported (Schnaubert, 2018). The fixed effects consist of categorical predictors. Therefore, dummy coding in the mixed model allowed a pairwise comparison of the individual group contributions to the outcome variable as a post hoc measure. As reported in the methods sections, Cronbach's alpha is very low, and further correlations confirmed that the knowledge scales (factual, conceptual, and transfer) are only weakly correlated (Appendix G). Therefore, all models were built with each learning scale as a separate outcome variable. As described in the hypotheses and as preregistered, the prior knowledge and individual working memory capacity were supposed to be included as moderators. The latter was not collected due to technical issues, and prior knowledge did not differ significantly between groups. Hence, it was not included in the model. The exploratory analysis showed no significant impacts of the psychological mediators on the learning outcomes. The results can be found in Appendix H.

## Results

### Differences between groups

As a preliminary analysis, differences in characteristics between groups were investigated. The final sample consisted of 159 participants who went through VR. Out of the 216 pretest questionnaires, only data of 168 participants in VR were collected from headsets. Seven participants were eliminated due to a mean of 4 in either of the cybersickness subscales. Duplicates and invalid IDs were also excluded from the analysis. Fisher's exact tests indicated that the group did not differ significantly concerning gender ($p$ = .309) and VR experience ($p$ = .197).

According to the Kruskal-Wallis test, there were also no significant differences found in prior knowledge $\chi^2(2) = 3.58$, $p$ = .309; English comprehension experience $\chi^2(2) = 6.63$, $p$ = .084; SIEA $\chi^2(2) = 0.91$, p = .824 and age $\chi^2(2) = 0.36$, p = .305 (*M*= 16,4, *SD*= 0.6). Hence, it can be concluded that there are no significant differences between the groups in basic characteristics.

Regarding the experience itself, the Kruskal-Wallis test was also conducted here because, like the demographics analysis, the condition of normality distribution for an

ANOVA was violated. No significant differences were found in the motion sickness subscale misorientation $\chi^2(3) = 7.66$, $p = .053$, while the oculomotor subscale was significant $\chi^2(3) = 7.82$, $p = .049$. However, Dunn test with Bonferroni correction was computed, resulting in non-significant between varied & active (*Mdn*=2.25, *SD* = 0.629) and varied & passive groups (*Mdn* = 2.00, SD = 0.54) as well as same & active (*Mdn* = 2.00, SD=0.62) and varied & passive. Regarding cognitive load $\chi^2(3) = 2.17$, $p = .537$, and the familiarity of the collaborator, the Kruskal-Wallis test indicated no significant differences between groups $\chi^2(3) = 1.65$, $p = .648$.

**Effects of Manipulation during retrieval practice on Factual Learning in VR post-test**

In the first model, the total score of factual learning represented the outcome variable, and the dyads the grouping variable. The null model of factual learning indicated an ICC of .14, the proportion of variation in learning outcomes accounted for by dyads (Kenny et al., 2006). 13.8% of the variance in factual learning can be explained by active retrieval. The explorative analysis showed no violations of the normality of residuals and the normality of random effects. However, some deviations within clusters can be reported, but according to Knief & Forstmeier (2021), mixed models are robust when it comes to deviations of normality. All other assumptions about Linearity and Homoscedacity were met.

**Table 1**

*Means and Standard Deviations of learning outcomes across conditions.*

| | | | *M (SD)* | |
|---|---|---|---|---|
| Condition | *N* | *Factual* | *Conceptual* | Transfer |
| Active & varied | 47 | 4.00 (1.82) | 1.85 (1.60) | 4.87 (1.19) |
| Varied & passive | 42 | 3.23 (1.51) | 0.83 (1.66) | 5.21 (1.14) |
| Same & active | 36 | 4.08 (1.86) | 1.61 (1.74) | 5.08 (0.97) |
| Same & passive | 34 | 3.65 (1.59) | 1.59 (1.39) | 4.65 (1.39) |

*Note.* The table describes the sample distribution in conditions and descriptives of outcome variables separately.

The model's fixed effects indicate a lack of support for H1 in factual outcomes. On the contrary, results show a positive significant main effect of active retrieval practice on factual learning; find the regression results in Table 2. Post hoc analysis supports this main effect. When comparing all the conditions to the varied passive condition, the varied and active condition (*b* = 0.79, 95% CI [0.05; 1.54], *t*(86.94) = 2.06, *p* = .042), and same and active (*b*= 0.82, 95% CI [0.02;1.62], *t*(96.63)=1.99*, p*=.005, performed significantly better in factual learning.

This allows a rejection of the null hypothesis of H2. As shown in Table 2, there are no significant interactions between the type of retrieval and context; therefore, H3 cannot be supported. The mixed model built can account for only 16.9% of the total variability in factual learning.

**Effects of Manipulations on conceptual learning in VR post-test**

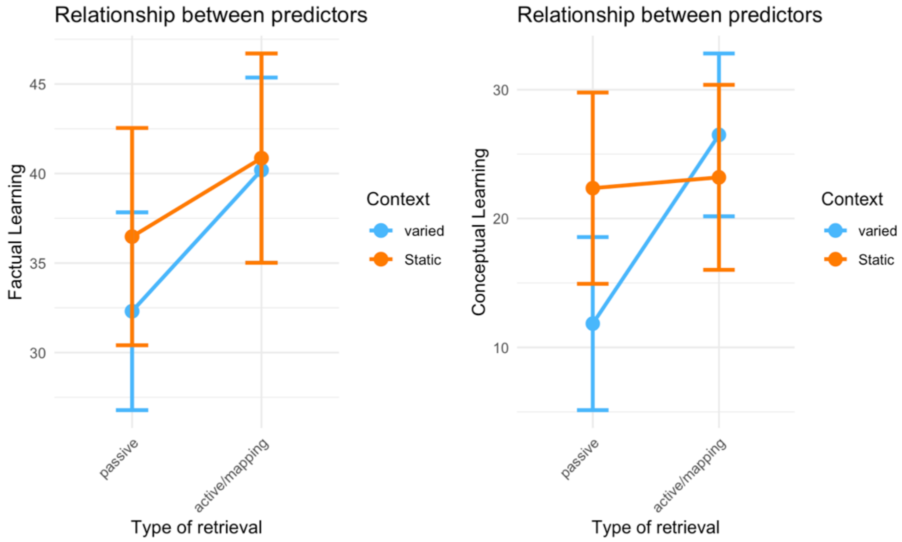
The null model of conceptual learning revealed that the random intercept accounted for 10.2% of the variability in conceptual learning (ICC=.10). The model diagnostics indicated that all assumptions were met except the within-cluster normal distribution.

The output of the mixed model regression did not strictly indicate a significant interaction (*p*= *.515).* However, when also considering the means of the groups (see Table 1) and figure 5, which illustrates the relationship between variables, I will examine the between groups differences. The contextual variation seems to influence the active vs. the passive retrieval.

Only when combined with active retrieval in VR does this manipulation positively affect conceptual learning (Table 2). When combined with passive retrieval, conceptual learning decreases. On the other hand, the static component does not seem to be as sensitive towards the type of retrieval. Post-hoc analysis confirms that varied and active (*b* = 1.03, 95% CI [0.39; 1.65], *t* (83.32) = 3.14, *p* = .001), same and active (*b* = 0.79, 95% CI [0.11; 1.47], *t* (91.69) = 2.28, p = .002) and same passive groups (*b* = 0.74, 95% CI [0.05; 1.42], *t* (88.24) = 2.08 *p* = .004) predicts conceptual learning significantly better. The main effect of active retrieval leads to rejecting the null hypothesis of H2 regarding conceptual knowledge. H3 cannot fully be supported, but there is some indication that contextual variation decreases learning in passive retrieval. Including the ICC, the model accounted for 13.2% of the variance in conceptual scores. This *R2* indicates a rather weak relationship between predictor and outcome.

**Figure 5**

*Effects of conditions on factual and conceptual learning*




*Note.* The figure shows the relationship between predictors on factual learning (left) and conceptual learning (right) in percentage, accounting for the random effect of dyads.

**Table 2**

*Results of linear mixed models*

| Outcome | Predictor | *B* | *SE* | *t* | *p* | CI |
|---|---|---|---|---|---|---|
| Factual | (Intercept) | 3.23 | 0.28 | 11.55 | <.001 | [2.68, 3.77] |
| | Active retrieval | 0.79 | 0.38 | 2.06 | .043 * | [0.04, 1.53] |
| | Static Context | 0.42 | 0.41 | 1.00 | .317 | [-0,38, 1.22] |
| | Retrieval* Context | -0.35 | 0.57 | -0.61 | .541 | [-1.46, 0.76] |
| Conceptual | (Intercept) | 0.83 | 0.24 | 3.49 | .001 | [0.37, 1.29] |
| | Active retrieval | 1.02 | 0.33 | 4.14 | .002* | [0.39, 1.66] |
| | Static Context | 0.74 | 0.35 | 2.08 | .041* | [0.05, 1.43] |
| | Retrieval * Context | -0.97 | 0.49 | -1.97 | .052 | [-1.92, -0.01] |
| Transfer | (Intercept) | 5.19 | 0.19 | 26.94 | <.001 | [4.81, 5.57] |
| | Active retrieval | -0.32 | 0.26 | -1.2 | .235 | [-0.83, 0.2] |
| | Static Context | -0.55 | 0.29 | -1.94 | .055 | [-1.11, -0.01] |
| | Retrieval * Context | 0.76 | 0.4 | 1.91 | .058 | [-0.01, 1.53] |

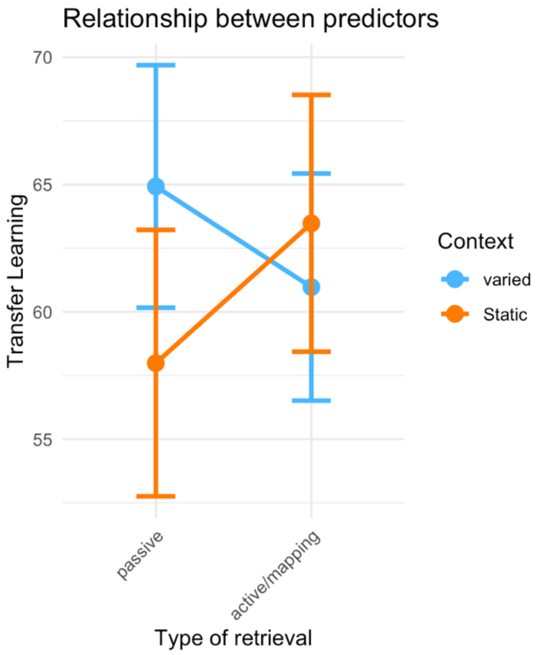
*Note.* Results of the linear mixed model of (H1-H3) with random effect of dyads. Analysis was run on the separated knowledge scales. *N=159.*

**Effect of Manipulation on Transfer Learning in VR post-test**

In transfer learning, the dyads accounted for 14.8% of the variability (*ICC*= .15). The mixed model results did not support H1 and H2, even though the static environment was close to being significant. Similar to the results in conceptual learning, there is a close to significant interaction between active retrieval and context (see figure 6). Post hoc analyses were conducted to explore the nature of the potential interaction. No clear statistical indication of significant differences was found between groups. Therefore, none of the hypotheses regarding main and interaction about transfer learning can be supported.

**Figure 6**

*Effects of conditions on transfer learning*




*Note.* Figure shows the relationship between predictors on transfer learning in percentage accounting for the random effect of dyads.

**Hypotheses 4-6 delayed post-test in classroom**

After investigating the quality of the data of the delayed post-test, I decided to exclude the results of this questionnaire due to the following data exploration. 72 of the registered 119 participants of the delayed post-test did not spend more than 3 minutes on the questionnaire, indicating a maximum of seven seconds per question. Further, the means indicate a floor effect (of the total sample), especially on conceptual (*M*= 0.99, *SD*= 1.00) and transfer learning (*M*= 0.72, *SD*= 0.77). Even though encouraging results regarding immediate retrieval practice effects were found, a clear testing effect in VR is not reportable as the data on long-term memory are missing.

# Discussion

## Empirical Contribution

This present study aimed to investigate whether the robust findings, such as the Testing effect, are replicable in an applied educational VR lesson. The retrieval practice literature findings were combined with approaches of the cognitive theory of multimedia learning to build on recent findings in VR research. Combining a consolidation-focused strategy like retrieval practice with a generative learning method of collaborative mapping compared to the passive control groups led to the general finding of active retrieval practice via mapping being more beneficial than the passive groups, specifically in factual and conceptual learning. The second aim of this research consisted of implementing the potential positive effect of contextual variation on learning with the affordances of VR by experimentally manipulating the context changes during retrieval practice compared to a static environment. Due to the novelty of VR, the hypotheses regarding the consequences of context on learning were nondirectional, and no clear direction was found. However, there is a remarkable negative effect of passive retrieval and contextual variation to build on in the future. To the best of my knowledge, this is the first time these theories were tested in a collaborative, immersive setting. However, the effects are very small and should not be overgeneralized and instead results should serve as a starting point to investigate further to optimize learning with new technology.

## Theoretical contributions

The positive main effects of active retrieval support the desirable difficulty framework by Bjork and Bjork (2019) in an immersive environment. The active attempt to retrieve during practice could have increased storage strength by reconstructing the initially retrieved information. Compared to the classical testing effect, the immediate feedback on the effortful reconstruction could have yielded a beneficial result in the immediate post-test. By providing immediate feedback to the participants, bifurcation was avoided, which could be why the restudy condition did not perform better in the immediate post-test. This finding aligns with Wiklund-Hörnqvisr (2013), who provided feedback after the concept mapping activity. However, this study did not fulfil the expected strong beneficial effects on individual learning as Adeopse (2018) and Schroeder and colleagues (2017) reported. Possibly due to too high transactive costs in the highly interactive VR environment, most collaboration could have been process-oriented rather than organizing and integrating information into a shared mental model (Janssen et al., 2022). Even though the feedback was provided constructively, the mistakes were highlighted, and students had the opportunity to correct their mistakes; the feedback exposure was quite repetitive. After each map, after one round of correction, students were given the corrected map that always looked identical. Potentially, this could have supported the feedback-induced reversal of the testing effect in the active conditions (Racsmany et al., 2020). This aligns with Vogel's (2022) negative prediction of a decrease in self-regulated internal scripts and, hence, a decrease in knowledge acquisition when using non-adaptable scripts. Even though parts of the script were designed to be adaptable, the way feedback was provided was rather non-adaptable.

The second aim of this study was represented by investigating the impact of contexts (static vs. varied) on learning during retrieval. Contrary to the main effect of retrieval practice, no clear main effect in either direction indicating a context dependency or benefits of variation was found.

The beneficial main effect of collaborative active retrieval extends results from Petersen and colleagues (2023), in which students in the collaborative generative group outperformed the individual and non-generative method groups. It contributes to that finding as this underlying study ran all the conditions in collaboration, allowing to conclude the beneficial effect of the generative method compared to the non-generative in a collaborative VR environment.

However, post-hoc analysis revealed that contextual variation combined with passive retrieval practice leads to significantly lower conceptual learning. Further, the contextual drifts were only higher in groups where participants retrieved actively in the conceptual learning outcome. In the passive retrieval groups, the variation in context had negative consequences on learning. This finding contrasts Imundo´s (2021) results, which showed an interaction between variation and restudy in the immediate post-test. One reason for these opposing results could be the feedback that was also provided to the active retrieval group in this experiment; the restudy groups did not have that advantage as in the Imundo (2021) study. In Imundo´s study, the environmental changes might not have been as distinct as in this study, which might have allowed more contextual cues to be linked to the memory trace. Students, especially in the active group, might have interacted more with the environment as the environmental change was less predictive. Actively building relations between concepts while confronted with different contextual cues might be supportive to bind item-related and context-related information together. The passive groups might have been less likely to bind information and context information. Most importantly, the environment distracted them from the to-be-learned items. This is in line with the prediction from VR research that learners engage more in extraneous processing and less in generative processing in case of a lack of instructional methods implemented in the VR lesson (Mayer, 2022).  Parong & Mayer (2018) used the very same VR simulation and compared it to a self-directed slideshow. Due to too much extraneous processing, the slideshow group scored higher on learning outcomes. The extraneous processing resulting from the distracting environment changes could explain the varied passive conditions performing so weak. However, it is not observable in the results (no significant differences in cognitive load between groups) but could still be an explanation, as the cognitive load was only measured with four items and was based on self-reports. In the second part of Mayer's study, the generative method of summarizing was added to the design, leading to increased learning. This finding also aligns with this study; collaborative mapping as a generative learning strategy helped students in this highly interactive environment, while passive groups suffered from distraction. As Mayer, Makransy, and Parong (2022) conclude, immersion and the resulting presence is a promise of VR research, but the Pitfall of this richness of irrelevant information can go in line with distraction and decrease resources left for generative processing. In the present study, this pitfall occurred in passive and varied conditions. Bjork and Bjork (2019) would describe this pitfall as an "undesirable difficulty," which is the case if the learners are not equipped to respond successfully to the difficulty of contextual variation. The perceptual richness is even more drastic with the presence of another avatar. The passive conditions did not engage in high behavioral activity, what Mayer would call unprincipled presentations, and therefore, no deeper cognitive processing followed. Further, the type of presentation might not match the mental model of how the participants would have organized the knowledge. Similar to the retrieval strategy disruption hypothesis (Marion & Thoreley, 2016).

Koh, Lee, and Kim (2018) compared the generative teaching and retrieval practice strategy to passive controls and found that both active retrieval practice and teaching outperformed their controls, concluding that the effortful attempt to retrieve information benefits learning. Besides the fact that in this present experiment, the interventions took place in VR, the generative strategy implemented by Mapping was combined with retrieval practice and not separated in conditions. It supports the suggestion of Roelle et al. (2022) to bring together retrieval practice and generative processing to optimize learning, and there is some indication that this is even more relevant in highly distracting environments such as VR. The teaching and retrieval practice study (Koh et al., 2018) and this work both indicated that both processes are somewhat similar in their influence on learning. Still, both can´t make inferences about the underlying mechanisms that could explain the results.

While contextual variation depended on the type of retrieval, the static context manipulation was not affected as strongly by the type of retrieval. There are almost no differences in whether their active or passive retrieval practices occurred. However, this pattern is only observable in conceptual learning. The encoding retrieval match according to the transfer-appropriate processing (Agarwal, 2019) potentially helped reinstate the context. The conceptual post-test was the most similar to the retrieval activity (in both active and passive).

**Practical Contributions**

Two primary practical contributions emerge from the present investigation for practitioners. First, when implementing a collaborative VR lesson, minimizing any distractions is even more important than in an individual learning experience. Increased presence (Makransky et al., 2019b; Makransky et al., 2021; Parong & Mayer, 2018) and, in this case, social presence does not necessarily lead to more learning. In a highly interactive environment, as in this study, it is even more important to implement instructional design principles and not assume that complex content and a collaborative condition enhance learning outcomes. On the contrary, specific interventions that trigger the collaborator's interdependence (Janssen et al., 2022), as was done in the active conditions here, help exploit immersive media's affordances and avoid the pitfalls.

Secondly, combining consolidation-focused learning with construction-focused methods is attractive for practitioners. Both types of knowledge are relevant and specifically implementable in immersive learning. Further, it is worth considering encoding-retrieval activity matches to reach the best possible learning results. Factual learning and transfer, for example, did not benefit as much from the generative learning strategy in this research.

**Study limitations**

Testing theories and findings that originate from laboratory studies in a field experiment naturally come with limitations. The main limitations consist of the very limited generalizability of the results due to data collection and analysis restrictions.

First, the low internal consistency indicates a low instrument quality that limits the interpretability of the results. The items were too broad in the constructs they were measuring. Separating the learning scales did not lead to better reliability. Especially the near-transfer task measures do not allow to make inferences outside of VR. The near transfer task specifically included two types of retrieval strategies: selecting and naming in a nonsequential order, making inferences unreliable. Moreover, the total scores were extremely low, indicating a mismatch between the sample and the degree of difficulty of the post-test. The extremely low SIEA score aligns with that explanation. There was no incentive for students to engage with the simulation and retrieval. For students to engage in generative processing, the motivation to learn is necessary, and if the effort to learn is not graspable to the students, cognitive processing won´t occur (Fiorella & Mayer, 2021). The material participants interacted with was irrelevant to their education, which could account for the small effects.

This research allowed to test in a more ecologically valid setting, especially the contextual variation consequences. However, this and the testing effect rely on more repetitions and trials. For example, in Isadara's study, the session lasted 15 minutes, and the background changes occurred between 3 and 6 seconds each (Isadara & Isadara, 2007). The contextual variation with retrieval practice by Imundo and colleagues (2021) took place on several days. The initial study phase consisted of 36 words presented for 5 seconds, each word seven times per session, and there were three sessions in total. A recent review found that students should not spend more than 50 minutes in VR due to motion sickness symptoms (Pellas et al., 2021). This simulation lasted between 35 and 45 min in this study. Including the post-test in VR led to only 15 minutes for retrieval practice; therefore, the number of context changes was limited to only three. It is, therefore, questionable whether these effects are even capturable with the current opportunities of VR technology. More repetitions and environmental changes are necessary to make more reliable conclusions of these effects.

Even though Smith describes VR as an ecologically valid promising tool for episodic memory research, he also warns that, for example, physical presence and motion sickness can lead to deteriorating memory effects (Smith, 2019). Including Physical Presence and Social Presence as moderators to the mixed model did not significantly impact the outcome, but there is still the possibility that increased physical presence across all conditions (which is the case in this study) due to high immersion led to negative effects on memory as interventions that contribute to the immersion of VR also increase cognitive resources spend on additional task-irrelevant information. The episodic memory retrieval is core to this present study; therefore, attention is probably divided by the immersion of the simulation itself but also the collaboration partner. Ataseven (2023) and colleagues describe in their review how divided attention impairs episodic memory retrieval in all the six essential stages of memory retrieval. Considering this high risk of distraction and the limited time in VR, it is questionable whether such a design is applicable.

Another limitation of this study is the amount of data lost during the data collection. Initially, the individual working memory capacity should have been measured to account for intra-individual differences. Investigating the ICC per group, adding this moderator would have been valuable, as the ICC was quite low when considering that they retrieved together, inferring that intraindividual differences could have been stronger than the collaboration activity itself. Further, due to technical issues, lots of data was lost on the headsets, causing a loss of Power. The missing values in dyad IDs were included in the analysis to mitigate these power losses, causing a decrease in the explainability through the random effect.

Finally, the unreliable, delayed post-test led to a loss of explainability of the effects of the two types of desirable difficulties. The delayed post-test was conducted in a new environment and would have allowed to make insightful conclusions about retrieval if no retrieval cue had been present. The delayed effects would have been interesting, especially for the interaction effects, as some reverse effects have been reported (Imuno et al, 2021).

**Future Research**

Even though the underlying results suffer from limited generalizability, the learnings and small effects give reason to investigate the effect of contextual variation and retrieval practice in VR. Instead of combining them in one collaborative experimental design, research would benefit from isolating each. Moreover, contextual variation should be tested individually due to the already distracting environmental changes, and more environment switches should be implemented. The environment cues could be embedded more into the retrieval activity to increase the likelihood that context-related information is bound to content-related information. For example, in the forest environment, leaves and sticks could have been integrated into the map. The environment could have been more meaningful to the learners to build more powerful retrieval cues (Essoe et al., 2022)

The main effect of active retrieval builds a solid foundation to investigate the combination of retrieval practice and generative learning in VR in more detail. Firstly, separating retrieval practice and Mapping in one experiment like Koh et al. (2018) in a VR setting would give a more in-depth understanding of the mechanisms of both strategies. Secondly, combining retrieval practice and generative learning in a sequential order instead of intertwined, as is the case here.

The RSDH hypothesis was essential for creating the simulation. However, a measure that assesses collaborative inhibition between groups could have given more insights into the mechanisms behind the collaboration quality.

Further, a similar design but a within design or the implementation of individual working memory capacity would allow controlling for intraindividual differences of immersive learning. Another controlling measure consists of the motivational component during learning; this would permit the exclusion of participants who are less likely to show the general learning effects of such sensitive interventions. Controlling for the headset type is recommended when using different types of headsets that go in line with different degrees of visual fidelity.

Generally, when testing such implicit effects as contextual variation, a sample that pays attention and is motivated to participate in a study is necessary. Only if the material is relevant to the students and matches the participants' cognitive capabilities transactive interactions occur. Lastly, investigating the delayed memory effects of the Testing effect would give valuable insights into the long-term effects on memory by VR interventions.

## Conclusion

In conclusion, this was the first attempt to test the desirable difficulties of retrieval practice and contextual variation on learning in a collaborative immersive environment. Consistent with the findings of the testing effects outside of VR, the results indicate a positive effect of active retrieval on factual and conceptual learning. Inconsistent with the limited research on contextual variation and retrieval practice, results show a negative effect of combining passive retrieval with varied contexts in conceptual learning. When learning in an immersive collaborative environment, avoiding distractions and providing structured scripts is essential to avoid undesirable difficulties. However, these small effects should be further investigated to test and scrutinize theories like context reinstatement, Competitive Trace Theory, and the testing effect in a more ecologically valid setting. When designed adequately, VR can be useful for future education and episodic memory research. Future research should focus on testing these effects separately, and when investigating the effect of context on learning, the context should be more meaningfully embedded into the environment. The current study should be considered an optimizable starting point that integrates applied and fundamental research with the overreaching goal to enhance learning.

**References**

Adesope, O., Nesbit, J. C. & Sundararajan, N. (2021). The mapping principle in multimedia learning. In *Cambridge University Press eBooks* (S. 351–359). https://doi.org/10.1017/9781108894333.037

Adesope, O. O., Nesbit, J. C., & Sundararajan, N. (2022). The mapping principle in multimedia learning. In R. E. Mayer & L. Fiorella (Eds.), *The Cambridge handbook of multimedia learning* (3rd ed., pp. 351–359). Cambridge University Press.

Agarwal, P. (2019). Retrieval Practice & Bloom's Taxonomy: Do students need fact knowledge before higher order learning? *Journal of Educational Psychology*, *111*(2), 189–209. https://doi.org/10.1037/edu0000282

Agarwal, P., Nunes, L. & Blunt, J. R. (2021). Retrieval practice consistently benefits student learning: A Systematic review of applied research in schools and classrooms. *Educational Psychology Review*, *33*(4), 1409–1453. https://doi.org/10.1007/s10648-021-09595-9

Andersen, M. S., & Makransky, G. (2021). The validation and further development of a multidimensional cognitive load scale for virtual environments. *Journal of ComputerAssisted Learning, 37*(1), 183–196. https://doi.org/10.1111/jcal.12478

Ataseven, N., Ünver, N. & Günseli, E. (2023). How does divided attention hinder different stages of episodic memory retrieval? *Current Research in Behavioral Sciences*, *5*, 100139. https://doi.org/10.1016/j.crbeha.2023.100139

Barnett, S. M. & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological Bulletin*, *128*(4), 612–637. https://doi.org/10.1037/0033-2909.128.4.612

Basden, B. H., Basden, D. R., Bryner, S., & Thomas, R. L. (1997). A comparison of group and individual remembering: Does collaboration disrupt retrieval strategies? *Journal of Experimental Psychology: Learn- ing, Memory, and Cognition, 23,* 1176–1189. http://dx.doi.org/10.1037/ 0278-7393.23.5.1176 Bates D, Mächler M, Bolker B, Walker S (2015). "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software*, **67**(1), 1–48. doi:10.18637/jss.v067.i01.

Bjork, R. A., & Bjork, E. L. (2019). Forgetting as the friend of learning: Implications for teaching and self-regulated learning. *Advances in Physiology Education*, *43*(2), 164–167.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to* *cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Hillsdale, NJ: Erlbaum.

Blunt, J. R. & Karpicke, J. D. (2014). Learning with retrieval-based concept mapping. *Journal of Educational Psychology*, *106*(3), 849–858. https://doi.org/10.1037/a0035934

Bramão, I., Karlsson, A. & Johansson, M. (2017). Mental reinstatement of encoding context improves episodic remembering. *Cortex*, *94*, 15–26. https://doi.org/10.1016/j.cortex.2017.06.007

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *36*(5), 1118–1133. https://doi.org/10.1037/a0019902

Butler, A. C. & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, *19*(4–5), 514–527. https://doi.org/10.1080/09541440701326097

Chiu, C. (2004). Evaluating system-based strategies for managing conflict in collaborative concept mapping. *Journal of Computer Assisted Learning*, *20*(2), 124–132. https://doi.org/10.1111/j.1365-2729.2004.00072.x

David, N., Newen, A., & Vogeley, K. (2008). The “sense of agency” and its underlying cognitive and neural mechanisms. *Consciousness and Cognition, 17*(2), 523–534. https://doi.org/10.1016/j.concog.2008.03.004

Escamilla, J. C., Castro, J. J. F., Baliyan, S., Ortells-Pareja, J. J., Rodríguez, J. J. O. & Cimadevilla, J. M. (2020). Allocentric spatial memory performance in a virtual Reality-Based task is conditioned by visuospatial working memory capacity. *Brain Sciences*, *10*(8), 552. https://doi.org/10.3390/brainsci10080552

EF Education First. (2022). *The world’s largest ranking of countries and regions by English skills: 2022 edition*. https://www.ef.com/wwen/epi/

Eichenbaum, H. (2000). A cortical–hippocampal system for declarative memory. *Nature Reviews Neuroscience*, *1*(1), 41–50. https://doi.org/10.1038/35036213

Essoe, J., Reggente, N., Ohno, A., Baek, Y., Dell’Italia, J., & Rissman, J. (2022). Enhancing learning and retention with distinctive virtual reality environments and mental context reinstatement. *npj Science of Learning*, *7*(1), 1-14

Farrokhnia, M., Pijeira-Díaz, H. J., Noroozi, O. & Hatami, J. (2019). Computer-supported Collaborative concept mapping: The effects of different instructional designs on conceptual understanding and knowledge co-construction. *Computers & Education*, *142*, 103640. https://doi.org/10.1016/j.compedu.2019.103640

Fiorella, L. & Mayer, R. E. (2015). Eight ways to Promote Generative Learning. *Educational Psychology Review*, *28*(4), 717–741. https://doi.org/10.1007/s10648-015-9348-9

Fiorella, L. & Mayer, R. E. (2021). Principles based on social cues in multimedia learning. In *Cambridge University Press eBooks* (S. 277–285). https://doi.org/10.1017/9781108894333.029

Fiorella, L. & Zhang, Q. (2018). Drawing boundary conditions for learning by drawing. *Educational Psychology Review*, *30*(3), 1115–1137. https://doi.org/10.1007/s10648-018-9444-8

Fischer, F., Kollar, I., Stegmann, K. & Wecker, C. (2013). Toward a script Theory of Guidance in Computer-Supported Collaborative Learning. *Educational Psychologist*, *48*(1), 56–66. https://doi.org/10.1080/00461520.2012.748005

Gartner, Inc. (2022, February 7). *Gartner predicts 25% of people will spend at least one hour per day in the metaverse by 2026*. Retrieved January 13, 2023, from https://www.gartner.com/en/newsroom/press-releases/2022-02-07-gartner-predicts-25-percent-of-people-will-spend-at-least-one-hour-per-day-in-the-metaverse-by-2026

Godden, D. & Baddeley, A. (1975). CONTEXT-DEPENDENT MEMORY IN TWO NATURAL ENVIRONMENTS: ON LAND AND UNDERWATER. *British Journal of Psychology*, *66*(3), 325–331. https://doi.org/10.1111/j.2044-8295.1975.tb01468.x

Gonzalez-Franco, M., Ofek, E., Pan, Y., Antley, A., Steed, A., Spanlang, B., Maselli, A., Banakou, D., Pelechano, N., Orts-Escolano, S., Orvalho, V., Trutoiu, L., Wojcik, M., Sanchez-Vives, M. V., Bailenson, J., Slater, M., & Lanier, J. (2020). The rocketbox library and the utility of freely available rigged avatars. *Frontiers in Virtual Reality*, *1*. https://doi.org/10.3389/frvir.2020.561558

Halamish, V. & Bjork, R. A. (2011). When does testing enhance retention? a distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *37*(4), 801–812. https://doi.org/10.1037/a0023219

Hatami, J., Farrokhnia, M. & Hassanzadeh, M. (2016). Select-and-Fill-In concept maps as an evaluation tool in science classrooms. In *Communications in computer and information science* (S. 169–180). https://doi.org/10.1007/978-3-319-45501-3_13

Hmelo-Silver, C.E., & Jeong, H. (2022). Synergies Among the Pillars. In Zawacki-Richter, O., & Jung, I (Eds.), *Handbook of Open, Distance and Digital Education* (pp. 1-16). Springer, Singapore.

Imundo, M. N., Pan, S. C., Bjork, E. L. & Bjork, R. A. (2021). Where and how to learn: The interactive benefits of contextual variation, restudying, and retrieval practice for learning. *Quarterly Journal of Experimental Psychology*, *74*(3), 413–424. https://doi.org/10.1177/1747021820968483

Isarida, T., Isarida, T.K., 2014. Environmental Context-Dependent Memory. Advances in Experimental Psychology Research. NOVA Science Publishers, New York, pp. 115–151.

Isarida, T. & Isarida, T. (2007). Environmental context effects of background color in free recall. *Memory & Cognition*, *35*(7), 1620–1629. https://doi.org/10.3758/bf03193496

Janssen, J., Kirschner, F., & Kirschner, P. A. (2022). The collaboration principle in multimedia learning. In R. E. Mayer & L. Fiorella (Eds.), *The Cambridge handbook of multimedia learning* (3rd ed., pp. 304–312). Cambridge University Press.

Johnson, C. & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology*, *101*(3), 621–629. https://doi.org/10.1037/a0015183

Krathwohl, D. R. (2002). A revision of Bloom's Taxonomy: An Overview. *Theory IntoPractice*, *41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). *Dyadic data analysis*. The Guilford Press.

Kim, H. K., Park, J., Choi, Y., & Choe, M. (2018). Virtual reality sickness questionnaire (VRSQ): Motion sickness measurement index in a virtual reality environment. *Applied Ergonomics*, *69*, 66–73. https://doi.org/10.1016/j.apergo.2017.12.016

Kim, K. & Mundy, P. (2012). Joint attention, Social-Cognition, and recognition memory in adults. *Frontiers in Human Neuroscience*, *6*. https://doi.org/10.3389/fnhum.2012.00172

Kirschner, P. A., Sweller, J., Kirschner, F. & Zambrano, R. J. (2018). From cognitive load theory to collaborative cognitive load theory. *International Journal of Computer-Supported Collaborative Learning*, *13*(2), 213–233. https://doi.org/10.1007/s11412-018-9277-y

Knief, U., & Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01587-5

Koh, A. W. L., Lee, S. C. & Lim, S. W. H. (2018). The Learning Benefits of Teaching: A retrieval Practice hypothesis. *Applied Cognitive Psychology*, *32*(3), 401–410. https://doi.org/10.1002/acp.3410

Lee, K. M. (2004). Presence, explicated. *Communication theory, 14*(1), 27–50.

Makransky, G., Lilleholt, L. & Aaby, A. (2017). Development and Validation of the Multimodal Presence Scale for Virtual Reality Environments: A Confirmatory Factor Analysis and Item Response Theory approach. *Computers in Human Behavior*, *72*, 276–285. https://doi.org/10.1016/j.chb.2017.02.066

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

Makransky, G. & Petersen, G. B. (2023). The Theory of Immersive Collaborative Learning (TICOL). *Educational Psychology Review*, *35*(4). https://doi.org/10.1007/s10648-023-09822-5

Marion, S. B. & Thorley, C. (2016). A meta-analytic review of collaborative inhibition and postcollaborative memory: testing the predictions of the retrieval strategy disruption hypothesis. *Psychological Bulletin*, *142*(11), 1141–1164. https://doi.org/10.1037/bul0000071

Mayer, R. E. (2009). Constructivism as a theory of learning versus constructivism as a prescription for instruction. In *Routledge eBooks* (S. 196–212). https://doi.org/10.4324/9780203878842-18

Mayer, R. E. (2014). *The cambridge handbook of multimedia learning*. Cambridge University Press.

Mayer, R. E. (2018). Thirty years of research on online learning. *Applied Cognitive Psychology*, *33*(2), 152–159. https://doi.org/10.1002/acp.3482

Mayer, R. E., Makransky, G. & Parong, J. (2022). The promise and pitfalls of learning in immersive virtual reality. *International Journal of Human-Computer Interaction*, *39*(11), 2229–2238. https://doi.org/10.1080/10447318.2022.2108563

McPhail, G. (2020). The Search for Deep Learning: A curriculum coherence model. *Journal of Curriculum Studies*, *53*(4), 420–434. https://doi.org/10.1080/00220272.2020.1748231

Meyer, O. A., Omdahl, M. K. & Makransky, G. (2019). Investigating the effect of pre-training when learning through immersive virtual reality and video: a media and methods experiment. *Computers & Education*, *140*, 103603. https://doi.org/10.1016/j.compedu.2019.103603

O'Reilly, R. C. & Norman, K. A. (2002). Hippocampal and Neocortical Contributions to Memory: Advances in the Complementary Learning Systems Framework. *Trends in Cognitive Sciences*, *6*(12), 505–510. https://doi.org/10.1016/s1364-6613(02)02005-3

Peck, T. C. & González-Franco, M. (2021b). Avatar Embodiment. a standardized questionnaire. *Frontiers in virtual reality*, *1*. https://doi.org/10.3389/frvir.2020.575943

Parker, J., Kaplan, A. D., Volante, W. G., Abich, J. & Sims, V. K. (2020). Exploring the encoding specificity principle and Context-Dependent recognition in virtual reality. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *64*(1), 1481–1485. https://doi.org/10.1177/1071181320641353

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

Peck, T. C. & González-Franco, M. (2021). Avatar Embodiment. a standardized questionnaire. *Frontiers in virtual reality*, *1*. https://doi.org/10.3389/frvir.2020.575943

Petersen, G. B., Stenberdt, V. A., Mayer, R. E. & Makransky, G. (2023). Collaborative generative learning activities in immersive virtual reality increase learning. *Computers & Education*, *207*, 104931. https://doi.org/10.1016/j.compedu.2023.104931

Plechatá, A., Makransky, G. & Böhm, R. Can extended reality in the metaverse revolutionise health communication?. *npj Digit. Med.* **5**, 132 (2022). https://doi.org/10.1038/s41746-022-00682-x

Polito, V., Barnier, A. J. & Woody, E. Z. (2013). Developing the Sense of Agency Rating Scale (SOARS): an empirical measure of agency disruption in hypnosis. *Consciousness and Cognition*, *22*(3), 684–696. https://doi.org/10.1016/j.concog.2013.04.003

Racsmány, M., Szöllősi, Á. & Marián, M. (2020). Reversing the testing effect by feedback is a matter of performance criterion at practice. *Memory & Cognition*, *48*(7), 1161–1170. https://doi.org/10.3758/s13421-020-01041-5

Radkowitsch, A., Vogel, F. & Fischer, F. (2020). Good for learning, bad for motivation? A meta-analysis on the effects of computer-supported collaboration scripts. *International Journal of Computer-Supported Collaborative Learning*, *15*(1), 5–47. https://doi.org/10.1007/s11412-020-09316-4

Renoult, L., Irish, M., Moscovitch, M. & Rugg, M. D. (2019). From Knowing to Remembering: The Semantic–Episodic Distinction. *Trends in Cognitive Sciences*, *23*(12), 1041–1057. https://doi.org/10.1016/j.tics.2019.09.008

Roediger, H. L. & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., Tekin, E. & Üner, Ö. (2017). Encoding–Retrieval interactions. In *Elsevier eBooks* (S. 5–26). https://doi.org/10.1016/b978-0-12-809324-5.21036-5

Roediger, H. L. & Karpicke, J. D. (2006). Test-Enhanced learning. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Roelle, J., Schweppe, J., Endres, T., Lachner, A., Von Aufschnaiter, C., Renkl, A., Eitel, A., Leutner, D., Rummer, R., Scheiter, K. & Vorholzer, A. (2022). Combining retrieval practice and generative learning in educational contexts. *Zeitschrift Fur Entwicklungspsychologie Und Padagogische Psychologie*, *54*(4), 142–150. https://doi.org/10.1026/0049-8637/a000261

Rolls, E. T. (2022). The hippocampus, ventromedial prefrontal cortex, and episodic and semantic memory. *Progress in Neurobiology*, *217*, 102334. https://doi.org/10.1016/j.pneurobio.2022.102334

Ryan, T. J. & Frankland, P. W. (2022). Forgetting as a form of adaptive engram cell plasticity. *Nature Reviews Neuroscience*, *23*(3), 173–186. https://doi.org/10.1038/s41583-021-00548-3

Schroeder, N. L., Nesbit, J. C., Anguiano, C. & Adesope, O. (2017). Studying and Constructing Concept Maps: A Meta-Analysis. *Educational Psychology Review*, *30*(2), 431–455. https://doi.org/10.1007/s10648-017-9403-9

Shin, Y. S., Masís-Obando, R., Keshavarzian, N., Davé, R. & Norman, K. A. (2020). Context-dependent memory effects in two immersive virtual reality environments: on Mars and underwater. *Psychonomic Bulletin & Review*, *28*(2), 574–582. https://doi.org/10.3758/s13423-020-01835-3

Slater, M., Banakou, D., Beacco, A., Gallego, J., Macia-Varela, F., & Oliva, R. (2022). A separate reality: an update on place illusion and plausibility in virtual reality. *Frontiers in Virtual Reality*, *3*, 914392.

Smith, S.M., 1979. Remembering in and out of context. J. Exp. Psychol. Hum. Learn. Mem. 5 (5), 460.

Smith, S. M. (2014). Effects of environmental context on human memory. In *SAGE Publications Ltd eBooks* (S. 162–182). https://doi.org/10.4135/9781446294703.n10

Smith, S.M., Glenberg, A., Bjork, R.A., 1978. Environmental context and human memory. Mem. Cogn. 6 (4), 342–353.

Smith, S. M. & Handy, J. D. (2016). The Crutch of context-dependency: Effects of contextual support and constancy on acquisition and retention. *Memory*, *24*(8), 1134–1141. https://doi.org/10.1080/09658211.2015.1071852

Smith, S. A. (2019). Virtual reality in episodic memory research: A Review. *Psychonomic Bulletin & Review*, *26*(4), 1213–1237. https://doi.org/10.3758/s13423-019-01605-w

Schnaubert, L., & Bodemer, D. (2018). What interdependence can tell us about collaborative learning: A statistical and psychological perspective. *Research and Practice in Technology Enhanced Learning*, *13*. https://doi.org/10.1186/s41039-018-0084-x

Storm, B. C., Friedman, M. P., Murayama, K. & Bjork, R. A. (2014). On the transfer of prior tests or study events to subsequent study. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *40*(1), 115–124. https://doi.org/10.1037/a0034252

Sugar, J. & Moser, M. (2019). Episodic memory: neuronal codes for what, where, and when. *Hippocampus*, *29*(12), 1190–1205. https://doi.org/10.1002/hipo.23132

Tudela, J. M. O., Lechuga, M. T. & Gómez-Ariza, C. J. (2018). A specific benefit of retrieval-based concept mapping to enhance learning from texts. *Instructional Science*, *47*(2), 239–255. https://doi.org/10.1007/s11251-018-9476-y

The Body VR. (2016). *The Body VR: Journey inside a cell - 360 gameplay*. https://www.youtube.com/watch?v=9zTsDXMyBEY

Tulving, E. (2002). Episodic memory: from mind to brain. *Annual Review of Psychology*, *53*(1), 1–25. https://doi.org/10.1146/annurev.psych.53.100901.135114

Tulving, E. & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*(5), 352–373. https://doi.org/10.1037/h0020071

Van Boxtel, C., Van Der Linden, J., Lambin, P. & Erkens, G. (2002). Collaborative concept mapping: provoking and supporting meaningful discourse. *Theory Into Practice*, *41*(1), 40–46. https://doi.org/10.1207/s15430421tip4101_7

Van Den Broek, G. S. E., Van Gog, T., Jansen, E., Pleijsant, M. & Kester, L. (2021). Multimedia effects during retrieval practice: images that reveal the answer reduce vocabulary learning. *Journal of Educational Psychology*, *113*(8), 1587–1608. https://doi.org/10.1037/edu0000499

Vogel, F., Kollar, I., Fischer, F., Reiss, K. & Ufer, S. (2022). Adaptable scaffolding of mathematical argumentation skills: the role of self-regulation when scaffolded with CSCL scripts and heuristic worked examples. *International Journal of Computer-Supported Collaborative Learning*, *17*(1), 39–64. https://doi.org/10.1007/s11412-022-09363-z

Vogel, F., Weinberger, A. & Fischer, F. (2021). Collaboration scripts: guiding, internalizing, and adapting. In *Springer eBooks*(S. 335–352). https://doi.org/10.1007/978-3-030-65291-3_18

Vojdanoska, M., Cranney, J. & Newell, B. R. (2009). The testing effect: The role of feedback and collaboration in a tertiary classroom setting. *Applied Cognitive Psychology*, *24*(8), 1183–1195. https://doi.org/10.1002/acp.1630

Waldeyer, J., Heitmann, S., Moning, J. & Roelle, J. (2020). Can generative learning tasks be optimized by incorporation of retrieval practice? *Journal of applied research in memory and cognition*, *9*(3), 355–369. https://doi.org/10.1016/j.jarmac.2020.05.001

Wang, X., Kollar, I. & Stegmann, K. Adaptable scripting to foster regulation processes and skills in computer-supported collaborative learning. *Intern. J. Comput.-Support. Collab. Learn* **12**, 153–172 (2017). https://doi.org/10.1007/s11412-017-9254-x

Wiklund-Hörnqvist, C., Jonsson, B. & Nyberg, L. (2013). Strengthening concept learning by repeated testing. *Scandinavian Journal of Psychology*, *55*(1), 10–16. https://doi.org/10.1111/sjop.12093

Wissman, K. T. & Rawson, K. A. (2015). Why does collaborative retrieval improve Memory? Enhanced relational and item-specific processing. *Journal of Memory and Language*, *84*, 75–87. https://doi.org/10.1016/j.jml.2015.05.003

Woods, D. L., Kishiyama, M. M., Yund, E. W., Herron, T. J., Edwards, B., Poliva, O., Hink, R. F. & Reed, B. R. (2010). Improving digit span assessment of short-term verbal memory. *Journal of Clinical and Experimental Neuropsychology*, *33*(1), 101–111. https://doi.org/10.1080/13803395.2010.493149

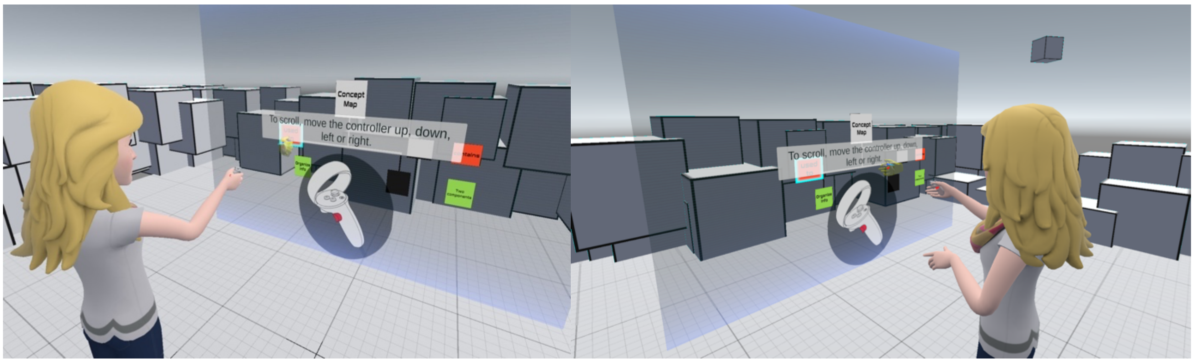
Xue, G. (2022). From Remembering to Reconstruction: the transformative neural representation of episodic memory. *Progress in Neurobiology*, *219*, 102351. https://doi.org/10.1016/j.pneurobio.2022.102351

Yonelinas, A. P., Ranganath, C., Ekstrom, A. D. & Wiltgen, B. J. (2019). A Contextual Binding Theory of Episodic Memory: Systems Consolidation Reconsidered. Nature Reviews Neuroscience, 20(6), 364–375. https://doi.org/10.1038/s41583-019-0150-4

Πέλλας, N., Mystakidis, S. & Kazanidis, I. (2021). Immersive virtual reality in K-12 and Higher Education: A Systematic review of the last decade Scientific literature. *Virtual Reality*, *25*(3), 835–861. https://doi.org/10.1007/s10055-020-00489-9

## Appendix A

*VR Tutorial*




Appendix A. Shows screenshot of tutorial for participants in active conditions how to interact with objects in VR with their controllers.

# Appendix B

*VR Transfer task*


Appendix B. Screenshots of post-test in VR. The left shows a screenshot of an example of a factual multiple-choice question; right shows a conceptual question.

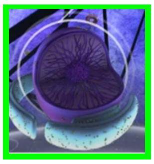
# Appendix C

*Factual questions in VR*

| Item | Question | Correct response |
|---|---|---|
| Q1 | Which objects freely pass the membrane? | Oxygen |
| Q2 | What element is located on the edge of the cell? | Protein Receptors |
| Q3 | What is the name of the energy source that floats around the cytoplasm? | ATP |
| Q4 | On what structure does Kinesin walk along? | Microtubule |
| Q5 | What is shown in this picture?  | Kinesien |
| Q6 | Which molecule is involved in the process of transcription and contains a single recipe for protein creation? | RNA |
| Q7 | Which of the pictures shows the Nuclear Pore? |  |
| Q8 | What is RER studded with? | Ribosomes |
| Q9 | What is RNA? | The copy of DNA |
| Q10 | Which structure contains the RER (Rough Endoplasmic Reticulum)? | Cytoplasm |

# Appendix D

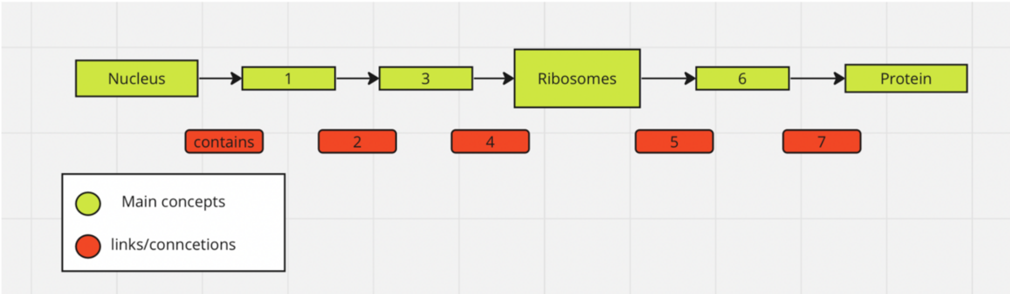
*Factual questions delayed post-test*

| Item | Question | Correct response |
|---|---|---|
| Q1 | What is shown in this picture?  | Receptors |
| Q2 | How does Water enter the cell? | Freely pass the membrane |
| Q3 | Which protein can walk along the microtubule? | Kinesien |
| Q4 | What does Kinesin bind with? | ATP |
| Q5 | What creates partly the structure of the cytoskeleton? | Microtubule |
| Q6 | Which of the images shows the Nucleus? |  |
| Q7 | What does the process of transcription involve? | Copy of DNA |
| Q8 | What element allows the entry of larger molecules into the nucleus? | Nuclear Pores |
| Q9 | What does the Nucleus surround? | RER (Rough Endoplasmic Reticulum) |
| Q10 | Which element do the ribosomes link together to create the protein? | Amino Acid |

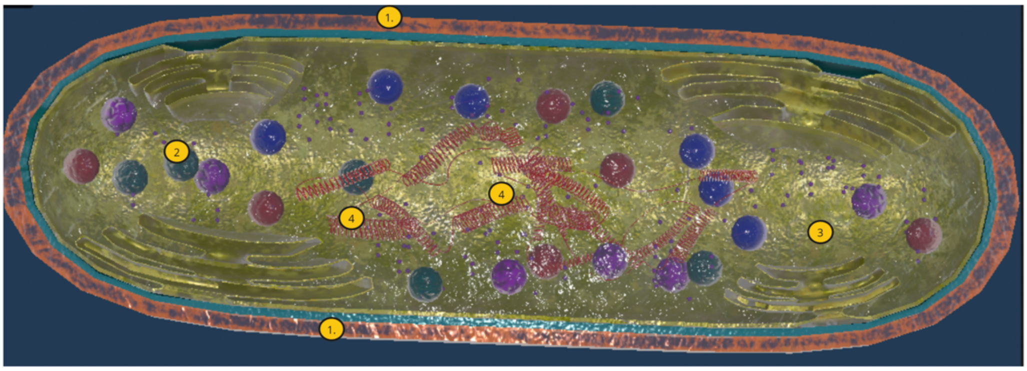
# Appendix E

*Conceptual and Transfer questions delayed post-test*

How are Nucleus and Ribosomes involved in protein synthesis? Choose the components and connections you believe are relevant and create a sequence starting with the Nucleus. Select the correct name for each number by clicking on the drop-down menu of each component/connection.




*Below you can see a bacteria cell. Apply your knowledge from last week about the human cell. Name the 4 components highlighted.*




Appendix E. Upper row shows the conceptual question of the delayed post-test. The lower row shows the transfer question of the delayed post-test.

## Appendix F

*ANOVA table factual learnig*

| Measures | ANOVA | | |
|---|---|---|---|
| | $F(1, 95)$ | $p$ | $\eta^2$ |
| Retrieval Practice | 4.57 | .003* | .05 |
| Context | 0.71 | .400 | .00 |
| Retrieval Practice* Context | 0.37 | .54 | .00 |

Appendix F. Shows ANOVA table with Satterthwaite's degrees of freedom method for factual learning outcome. $N = 159$

*ANOVA table conceptual learning*

| Measures | ANOVA | | |
|---|---|---|---|
| | $F(1, 93)$ | $p$ | $\eta^2$ |
| Retrieval Practice | 4.87 | .029* | .05 |
| Context | 1.06 | .305 | .01 |
| Retrieval Practice* Context | 3.88 | .051 | .04 |

Appendix G. Shows ANOVA table with Satterthwaite's degrees of freedom method for conceptual learning outcome. $N$=159

## Appendix G

*Correlation of learning outcomes*

| Variable | 1 | 2 | 3 |
|---|---|---|---|
| 1. F_total | 1.00 | | |
| 2.C_total | .10 | 1.00 | |
| 3. T_total | .14 | .05 | 1.00 |

Appendix H. The table shows the correlation matrix between the separate total scores of factual, conceptual, and transfer outcomes. *N*=159

## Appendix H

| Outcome | Predictor | *B* | *SE* | *t* | *p* | CI |
|---|---|---|---|---|---|---|
| Factual | Agency | -0.14 | 0.14 | -0.99 | .32 | [-0.41, 0.15] |
| | BO | -0.02 | 0.03 | -0.58 | .78 | [-0.07, 0.04] |
| | SP | 0.05 | 0.21 | 0.26 | .77 | [-0.07, 0.04] |
| | PP | 0.01 | 0.19 | 0.03 | .98 | [-0.07, 0.04] |
| Conceptual | Agency | -0.12 | 0.12 | -1.03 | .30 | [-0.17, 0.22] |
| | BO | 0.01 | 0.02 | -0.01 | .93 | [-0.05, 0.04] |
| | SP | 0.08 | 0.18 | 0.43 | .66 | [-0.28, 0.43] |
| | PP | -0.03 | 0.17 | -0.19 | .84 | [-0.37, 0.29] |
| Transfer | Agency | 0.02 | 0.09 | 0.21 | .83 | [-0.17, 0.22] |
| | BO | -0.01 | 0.02 | .77 | .44 | [-0.39, 0.18] |
| | SP | -0.15 | 0.15 | -0. 71 | .45 | [-0.39, 0.18] |
| | PP | -0.02 | 0.14 | -0.17 | .87 | [-0.29, 0.24] |

Appendix G. Shows the results of the exploratory analysis. Fixed effect of psychological mediators on learning outcomes. BO = Body ownership, SP = Social Presence, PP = Physical Presence. ICC see results from main analysis (H1-H3). *N* = 159.