Computers & Education 166 (2021) 104154

Contents lists available at ScienceDirect

Computers & Education

journal homepage: http://www.elsevier.com/locate/compedu

Signaling in virtual reality influences learning outcome and cognitive load

Patrick Albus *, Andrea Vogt, Tina Seufert

Ulm University, Institute of Psychology and Education, Department Learning and Instruction, Albert-Einstein-Allee 47, 89081, Ulm, Germany

A R T I C L E I N F O

Keywords:
Virtual reality
Media in education
Learning outcome
Cognitive load
Signaling

A B S T R A C T

Virtual reality (VR) learning environments are highly visual and need instructional aid to help the learner. Presenting too much visual information can impair selection and organization processes by overloading the learner's cognitive capacity. Signaling could support these processes by highlighting relevant information. In this study we used signals in form of textual annotations. Research on multimedia design principles provides evidence that such signals can foster learning, but we address the question whether the signaling principle is also applicable in a virtual reality learning environment. Particularly, we investigate the effects on learning by measuring different learning outcomes as well as different types of cognitive load, i.e., extraneous, and germane load in a between-subjects design. Participants ($n = 107$) were randomly assigned to a group with or without annotations in a VR learning environment. We assumed that learning material with integrated annotations lead to higher learning outcome, lower extraneous cognitive load (ECL), and higher germane cognitive load (GCL) than the control group. Results show that annotations improved learners' recall performance and GCL compared to a control group but had no effects on deeper levels of processing and on ECL. This indicates that annotations in VR can help learners to process the information on a recall level, but not due to a relief in ECL. When VR learning environments should support learners in recall and at the same time foster GCL, then adding annotations can be an appropriate approach.

1. Introduction and theoretical background

Virtual Reality (VR) is a promising learning tool that allows learners to immerse themselves in three-dimensional environments. It has the capability to enable interactive learning experiences since it can actively involve the learner in the learning process by reacting dynamically to the learner's movements and behavior (Chen, 2016; Christou, 2010). Due to the constant development of technical equipment, digital media such as VR are increasingly being used in education in addition to conventional learning methods such as textbooks (Freina & Ott, 2015; Radianti, Majchrzak, Fromm, & Wohlgenannt, 2020).

VR has different technical and pedagogical requirements than classic multimedia learning material (Fowler, 2015). While current design guidelines in VR could also be transferred to VR learning environments (Bowman, Kruijff, LaViola Jr, & Poupyrev, 2004; Jerald, 2015; Sherman & Craig, 2018), these guidelines are often driven by technology or by design and do not give indications on how these guidelines may affect cognitive learning processes (Radianti et al., 2020; Sutcliffe et al., 2019). Although there are many stimuli in VR, for cognitive information processing in VR the relevant focus is on text and image, in varying sensory modality. Therefore, the

* Corresponding author.
E-mail addresses: patrick.albus@uni-ulm.de (P. Albus), andrea.vogt@uni-ulm.de (A. Vogt), tina.seufert@uni-ulm.de (T. Seufert).

https://doi.org/10.1016/j.compedu.2021.104154
Received 29 April 2020; Received in revised form 27 January 2021; Accepted 30 January 2021
Available online 5 February 2021

Cognitive Theory of Multimedia Learning (CTML; Mayer, 2005) might provide the theoretical foundation to help understand the underlying cognitive processes when using VR learning environments (Parong & Mayer, 2020). They recommend how to design multimedia learning material based on principles of human cognitive processing. These principles have been shown to have a meaningful effect on learning in real non-virtual learning environments. Few studies on the use of multimedia design principles in VR have already been implemented (Baceviciute, Mottelson, Terkildsen, & Makransky, 2020; Makransky, Andreasen, Baceviciute, & Mayer, 2020; Parong & Mayer, 2018) as will be discussed below in greater detail. Overall, they show that although the different design variants are also meaningful in VR, the effects are nevertheless less straightforward as in classical multimedia design settings. In order to better understand the psychological processes behind this, a more differentiated view of the learning process would be necessary. Even more important, studies on how to design VR should consider the specific requirements of VR learning environments. One particular challenge for learning in VR, is the potential overload of visual input that would generate unnecessary cognitive load and may hinder learning (Makransky et al., 2019; Meyer, Omdahl, & Makransky, 2019). This presentation of learning material, which is often new to learners can have more degrees of freedom, allowing more visual stimuli to be presented than on a stationary screen. In order to support the learner to find out what information is currently relevant we can provide attentional guidance. A promising design principle for this is the signaling principle. Signaling could help the learner to focus on the important information in highly visual 3D worlds that can offer many distractions. Large effect sizes for signaling on learning outcome can be found in classical multimedia experiments (Mayer, 2005). However, it is unclear whether these effects can also be found in VR learning environments.

The present paper investigates whether the signaling principle in a VR learning environment, by adding textual annotations, has an influence on learning outcome and cognitive load. The signaling principle states that, deeper understanding processes in multimedia learning can occur when signals direct the learner's attention to relevant information or emphasize the organizational structure of the core content (Mayer, 2005). In interactive games mostly visual highlights are used as signals to indicate that an interaction with an object is possible, or to direct attention. In VR, such signals could also provide necessary attentional guidance and therefore relieve cognitive resources, which otherwise would be needed for unnecessary visual search (Mautone & Mayer, 2001; Ozcelik, Arslan-Ari, & Cagiltay, 2010). To encourage the learner to engage in deeper learning processes, we have chosen signals in form of textual annotations. As these signals are also visual and pop-up at the exact moment when learners should pay attention to this visual entity, they also function as a guidance of attention. However, annotations also contain semantic information. Thus, they should foster learning processes in working memory, as they are described in theories on learning with multimedia such as the Cognitive Theory of Multimedia Learning (Mayer, 2005) and the Cognitive Load Theory (Sweller, 2005). Based on these theories, it is expected that positive effects on learning outcome and cognitive load will be achieved through the use of signals. For learning outcomes as well as for cognitive load we differentiate between different levels or types in order to specify the effects of annotations. Moreover, with these differentiations we strive for a better understanding of the cognitive processes while learning in VR.

1.1. *Virtual reality and learning*

VR enables the learner to explore and manipulate computer-generated multimedia learning environments in real time. VR is capable of providing immersive learning environments in which learners can experience otherwise dangerous situations in real life that present no risk to themselves or others. It also enables users to experience situations to which there is limited or no access in the real world (Freina & Ott, 2015). One prominent feature of VR head-mounted-displays is their capability to make learning environments highly immersive. Radianti et al. (2020) define immersion as "involvement of a user in a virtual environment during which his or her awareness of time and the real world often becomes disconnected, thus providing a sense of being in the task environment instead" (p. 2). The immersive experience depends highly on technical features of the VR device (Radianti et al., 2020; Selzer, Gazcon, & Larrea, 2019). Generally, VR-Head mounted displays (VR-HMDs) can be categorized as low-end or high-end systems based on their hardware specifications. The low-end devices can be smartphone-based (e.g., Google Cardboard), or self-sufficient (e.g., Oculus Go). The inbuilt accelerometers, gyroscopes, and magnetometers process head movements in real time allowing for three degrees of freedom (3DOF; roll, pitch, yaw). They are usually suited for 360° videos, since less computing power is needed here. The high-end VR-HMDs often come with motion tracking. This allows for six degrees of freedom (6DOF). These devices can additionally track the translational motion of the movement, processing if the device has moved forward, backward, sideways, or vertically. This requires additional sensors and more processing power, and in many cases they have to be connected to an external computer (e.g., Valve Index). Although these systems differ in their level of immersion, it has been shown that low-end VR HMDs can compare to high-end VR systems with regard to retention (Juan, García-García, Mollá, & López, 2018; Slater & Sanchez-Vives, 2016). Considering the substantial price differences between low-end and high-end VR systems, low-end systems seem to be a suitable way to provide virtual learning environments to a larger group of people, for example classrooms (Olmos, Cavalcanti, Soler, Contero, & Alcañiz, 2018).

In general, one can ask how virtual learning environments can become a promising method of learning in education (Maas & Hughes, 2020; Merchant, Goetz, Cifuentes, Keeney-Kennicutt, & Davis, 2014; Radianti et al., 2020). The novelty of VR technology could bring excitement and fun to learning environments. On the one hand this excitement could be attributed to the technological novelty of VR and thus may distract from the learning itself. On the other hand, it could raise motivation of the learners (Makransky, Terkildsen, & Mayer, 2019; Packer, 2006). Previous studies have found that using new technology for learning can foster learners' motivation in comparison to conventional learning materials (Akçayır, Akçayır, Pektaş, & Ocak, 2016). In another study of Parong and Mayer (2018), participants were more motivated and interested in learning material when learning in a virtual learning environment when compared to other learning methods, and it was seen that their commitment was therefore increased. Participants who learned with only a desktop presentation showed less motivation and interest.

Besides these positive effects on motivation, can VR also enhance learning outcomes? When learning with a VR-HMD, the learning material is presented directly in front of the eyes using a display. This way, the learner can interact strongly with the material and experience immersive sensory experiences (Kozhevnikov, Gurlitt, & Kozhevnikov, 2013; Parong & Mayer, 2018; Wu, Yu, & Gu, 2020). Jensen and Konradsen (2018) reviewed the use of immersive VR technologies particularly for skill acquisition. They focused on immersion and presence and found that VR is useful for training in cognitive skills that are related to spatial and visual knowledge, visual scanning, observational skills, psychomotor skills that involve head-movement, and affective control of emotional response in stressful or difficult situations (Jensen & Konradsen, 2018). They also found that using an immersive HMD can lead to greater engagement and more time spent on the learning task. A more recent review from Radianti et al. (2020) found that VR can be an appropriate tool for enhancing procedural, practical, and declarative knowledge in higher education (Radianti et al., 2020). Another study of Webster (2016), who compared students who learned with either VR-HMDs or desktop presentations showed that students with VR-HMDs achieved higher scores on a post-test for learning performance. These findings are in line with other research that also found that learning in VR can improve learning outcomes (Alhalabi, 2016; Meyer et al., 2019; Passig, Tzuriel, & Eshel-Kedmi, 2016).

However, these studies did not further differentiate the levels of processing on which these positive effects occur, i.e., whether VR affected recall, comprehension, or transfer performance (Bloom, 1956). That such a differentiation nevertheless is important is also underlined by a recent study, which highlights the importance of using different assessment methods to investigate learning in immersive VR (Makransky, Borre-Gude, & Mayer, 2019a). Thus, our goal is to differentiate the levels of processing according to Bloom (1956). The first level of processing represents knowledge. Information can be recalled without having a deeper understanding of it. The next level is comprehension, where learners are able to recognize connections and are able to transfer them to analogous situations. When learners can transfer the acquired knowledge to new, i.e., previously unknown situations, they have reached the application level (Bloom, 1956). Up to now, there are no studies which have differentiated all these levels of processing in a single study. In the study conducted by Allcoat and von Mühlenen (2018) at least one level is assessed and specified. They found positive effects on recall when learning with VR glasses as compared to a desktop presentation or a textbook. Makransky, Terkildsen, and Mayer (2019b) analyzed two levels of processing in their study, namely recall and transfer for learning with either a VR or a desktop system. They found a negative effect for VR on the recall level and no effect on the transfer level. Baceviciute et al. (2020) also found positive effects on recall but not on retention when learning with different modalities in VR. Another study from Petersen, Klingenberg, Mayer, and Makransky (2020) showed that pre-training before a VR learning experience can increase declarative knowledge and transfer. Makransky, Terkildsen, and Mayer (2019) found no difference between groups in retention but favored VR on transfer when compared to a text group. Makransky et al. (2020) found that enactment in VR can lead to higher procedural knowledge and transfer. These heterogeneous findings emphasize that it is in fact worthwhile differentiating these levels of learning outcome.

Besides all the studies that have reported positive effects, there are nevertheless some studies that revealed no effect or even negative effects of learning in VR. Richards and Taylor (2015) compared a classical teacher-based instruction with either a 2D presentation or a 3D presentation via VR-HMDs. Learners with the 2D presentation outperformed those with the 3D version, which they suggest is explained by the cognitive overload owing to the visually demanding 3D animation in VR. Further, as mentioned above, the study of Makransky et al. (2019b) revealed negative effects of VR compared to desktop presentation on recall, but not for transfer.

Recent studies on the effects of cognitive load in VR show how important it is to distinguish between different types of cognitive load (Andersen & Makransky, 2020; Baceviciute, Mottelson, Terkildsen, & Makransky, 2020; Parong & Mayer, 2020). Especially when conducting instructional design studies, a differentiation of cognitive load is important, because it can help to understand the underlying mechanisms of multimedia design principles (Klepsch & Seufert, 2020). A differentiated approach in VR can help to understand whether there are hindering effects in terms of extraneous cognitive load or fostering effects in terms of increasing germane processes (Ayres, 2020; Huang, Luo, Yang, Lu, & Chen, 2020). These possible effect patterns will be discussed below in further detail. But before asking which cognitive load effects the processing of information of VR can have, we would like to discuss these processes in greater detail.

1.2. Cognitive processes while learning in VR

The learning content in VR is often presented to the learners aurally via spoken texts and visually via animations (Parong & Mayer, 2018). Based on the CTML (Mayer, 2005) such a multimedia design of learning material can lead to meaningful learning as it addresses two different sensory channels as well as two different code systems. By referring to Paivio's dual coding theory (Paivio, 1991), the CTML describes that people have two separate information-processing systems for verbal and pictorial information. In each of these systems a separate mental model for either text or picture is formed, which eventually has to be integrated into a combined mental model. Based on Baddeley's working memory model (1992) the CTML also assumes different processing channels for either visual or auditory information. As cognitive capacity is limited, Mayer argues that the design of multimedia should reflect these structural characteristics of working memory. For example, when presenting highly visual and spatial information as would be the case in VR learning environments the cognitive resources can best be used when the accompanying text is presented aurally. Thus, the visual information will be processed in the visuo-spatial sketchpad and the auditory text information can be processed in the independent and separate phonological loop (Mayer, Heiser, & Lonn, 2001). This is also known as the modality effect (Mayer, 2005).

Mayer (2005) not only describes the structural characteristics of working memory but also the processes which are necessary for meaningful learning. This occurs when learners actively engage in cognitive processes during learning, so they can generate coherent mental representations of the information, which involves the processes of selecting, organizing, and integrating the relevant information. *Selection* is achieved through an attention process. The necessity to select only a part of the presented learning content arises due to capacity limitations in each channel of the cognitive system. In VR, the selection processes might require even more capacity because there are more sensory stimuli overall. Due to high degrees of immersion and 360° all-round visibility, learners have to filter more to select the relevant elements. As a result, the search and orientation processes may also become more complex. The selected information will then be *organized* to create a coherent structure in form of a verbal model or a pictorial model. In order to organize the selected content, the elements have to be linked to each other within the working memory. The learner does not proceed randomly but tries to recognize the cause-effect chain while trying to make sense of the elements to create these simple structures. The more complex selecting process could also affect the organization process in VR. This could mean that much more selected information has to be discarded in order to link only the relevant with each other. The *integration* process is about combining the verbal and the pictorial model to a coherent model while mapping the respective elements and relations. This is a challenging process and it is cognitively demanding (Mayer, 2005). While integrating, the learner has to focus on the underlying structures of visual and verbal representations. In addition, prior knowledge from long-term memory can be transferred into working memory to help to coordinate the integration process while connecting the new structures. In VR, integration could be increasingly challenging, since the difficult selection and organization processes might lead to incomplete or faulty verbal or pictorial models, which makes it even more difficult to map them onto each other.

1.3. Cognitive load while learning in VR

As stated above the CTML assumes a limited working memory capacity (Mayer, 2005). The Cognitive Load Theory (CLT; Sweller, van Merrienboer, & Paas, 1998); specifies more precisely the demands on working memory by differentiating between three types of load.

The *intrinsic cognitive load (ICL)* characterizes the type of load caused by the difficulty and complexity of the learning material. The complexity arises from the number of elements and their interrelations that have to be processed in working memory at the same time (Sweller et al., 1998). Using interactive virtual elements in VR could lead to high element interactivity and thus to an increased ICL (Frederiksen et al., 2020).

The *extraneous cognitive load (ECL)* arises from unnecessary processes that do not contribute directly to learning, often caused by distractions or inadequate instructional design. When, for example, learners have to split their attention between visually presented text and visual pictures where they have to look back and forth, this would result in search and orientation processes which are extraneous to the learning goal itself (Ayres & Sweller, 2005). In visually rich VR learning environments learners definitely have to split their attention between different visual entities. Through its nature of being immersive, a VR learning environment could also offer irrelevant but interesting visual information, which then would lead to extraneous distraction (Frederiksen et al., 2020; Parong & Mayer, 2018). While the intrinsic load of an environment is given, based on the issue of element-interactivity in learning, the ECL can be reduced by an optimized design. Due to the limited capacity of working memory, this would be necessary to free up resources for germane processes (Sweller et al., 1998). Therefore, the main implication for designing multimedia learning materials is that the learning material should be designed in such a way that ECL is minimized while simultaneously giving maximum potential for germane cognitive processing (Brünken, Steinbacher, Plass, & Leutner, 2002; Chen, Grierson, & Norman, 2015).

The *germane cognitive load (GCL)* arises from the effort, which is germane for learning, i.e., for schema construction (Sweller et al., 1998). This type of load is positively related to learning outcomes (Chandler & Sweller, 1991). In learning environments such germane processes could be elicited by prompting deeper processing or generative tasks (Seufert, 2018). The use of VR could therefore be supportive, since the increased immersion and interactivity might stimulate such an active engagement. Prior research actually showed that VR can have an influence on GCL (Frederiksen et al., 2020).

1.4. Multimedia design principles

In light of these complex processes of selecting, organizing, and integrating information in VR that a learner has to execute within a limited working memory, one must ask how VR learning environments should best be designed to support learners. Mayer (2014) presents multimedia principles that have robust effects in a real learning environment in order to optimally utilize working memory so that learning may be successful, and learners are not cognitively overloaded. These principles offer possible starting points for successful learning in VR by providing sufficient capacity for processes of understanding (Frederiksen et al., 2020; Makransky, Terkildsen, & Mayer, 2019). The most relevant for the present study is the signaling principle. The signaling principle refers to the finding that people learn better when cues are added to guide attention to the relevant elements in the learning material (Mayer, 2005).

But why did we choose this? Owing to the more difficult selection and organization processes that can result from the visually demanding VR learning environment, the signaling of relevant parts of the learning material could support the learner in these processes. If the learner receives signals in these visually demanding environments as to what information is now relevant and where he needs to look more closely, this reduces unnecessary search processes and draws attention to relevant aspects. Reducing these unnecessary search processes also helps to reduce the extraneous cognitive load. This can also facilitate the organization process and leave greater capacity of the working memory for learning processes. Signals support the process of selection because they highlight relevant elements to help extract the essential information. This is particularly important as the complexity of the visual content increases in VR. Moreover, emphasizing the important elements can reduce ECL (Van Gog, 2014) as it reduces unnecessary search processes and draws attention to relevant aspects (Kalyuga, Chandler, & Sweller, 1999; Lorch, 1989). With a decreased ECL the freed-up capacity could be used to actively invest in GCL, e.g., to organize and integrate the relevant information in VR (Chandler & Sweller, 1991). As the selection process is the crucial first step that needs to be supported, signals can be seen as an effective means to foster the overall learning process. This is also supported by a perspective of human-computer interaction, in which the selection of objects is considered an essential task in VR (Bacim, Kopper, & Bowman, 2013). Argelaguet and Andujar (2013) argue that the selection of a virtual object in a VR environment is significantly different from that in the real world. In VR the eyes try to focus on a distant object that is actually shown on a display immediately in front of the eyes. These vergence-accommodation conflicts, the stereoscopic vision, and the limited field of vision in VR differ from real world selection processes (Lubos, Bruder, & Steinicke, 2014). In these visually suboptimal conditions, Teather and Stuerzlinger (2014) demonstrated that by signaling target stimuli the selection time was extended, but the accuracy improved and the error rates were significantly reduced compared to target stimuli without signaling.

Regarding the question of how to signal, we assume that textual annotations could be an effective type of signal in VR, based on research on text-picture-comprehension and the above mentioned CTML. In VR the most salient information is the visually presented environment. The auditory narration nevertheless is an anchor for learner's attention as it is transient, and it guides where learners look in the visual display. When there are important aspects in the narration that the learner should particularly pay attention to, it can be helpful to insert textual annotations that are temporarily aligned with the narration and spatially aligned with the visual entity they refer to (temporal and spatial contiguity principle; Mayer, 2005). Thus, the annotations first help the selection as also demonstrated by Wallen, Plass, and Brünken (2005). They further provide support to organize and integrate the information, as the annotations make the relations between the individual elements more salient (De Koning, Tabbers, Rikers, & Paas, 2009).

Concerning the different levels of processing, one might assume that the annotations would be particularly helpful for recall as they support double encoding with the auditory narration and the added visual keywords in textual form. The temporal and spatial arrangement nevertheless also can be seen as a hint to the relation of verbal and pictorial entities, and thus should also be beneficial for deeper learning processes like comprehension and transfer. Based on such a differentiation of processing levels we can learn about the specific effects of annotations and also about the learning processes in VR itself. This could show on which process level the learners have difficulties to process the VR learning material. There is an entire body of research indicating that using signaling is helpful to improve learning (Lai, Tsai, & Yu, 2011; Martin & Betrus, 2019; Wallen et al., 2005; Yeh & Lo, 2009). The mentioned studies have focused mainly on established learning tools, like PowerPoint or web-based environments, and to the best of our knowledge, there is no research on the signaling principle via annotations in a VR learning environment.

With regard to cognitive load, the spatially close placement of corresponding text and visual elements can reduce visual search and thus reduce ECL. Making the relations between the individual elements more salient might also function as a prompt for organizing and integrating them (De Koning et al., 2009), thus increasing learning related GCL (Mayer & Fiorella, 2014). This interplay between different types of cognitive load nevertheless needs to be investigated in one study, which has not been done so far.

But is the transfer of multimedia design principles from classical multimedia learning environments to VR learning environments a promising approach? Prior research investigating multimedia design principles in VR has shown heterogeneous findings. Makransky, Terkildsen, and Mayer (2019) investigated the *redundancy principle* in VR. Subjects learned via VR-HMDs or desktop presentations while receiving learning simulations with either on-screen text or on-screen text with narration. They found no effect for the redundancy principle, so both groups learned equally well when the material was presented redundantly compared to on-screen text only. In another study by Parong and Mayer (2018), the authors examined the *segmenting principle*. They compared a self-paced slideshow to a continuous VR animation. The results show that learners who received the slideshow scored higher on the factual questions then learners in the VR group but not on conceptual questions. Bacevuciute and colleagues (2020) explored the *modality principle* in VR by comparing audiovisual and visual-only presentations. The results show an inverse modality effect, where reading leads to higher learning outcomes than listening. The *pre-training principle* in VR was investigated by Petersen et al. (2020). In a virtual tour through Greenland about the consequences of climate change, the first group received a narrated pre-training followed by the VR exploration tour, while the second group received the same narrated training material integrated within the VR exploration tour. While both groups showed increased declarative knowledge in a pre to posttest assessment, there was a significant increase in transfer scores ($d = .46$) for the pre-training group. Makransky, Wismer, and Mayer (2019) investigated the *embodiment principle* and found that a gender-specific design in VR learning environment can have an impact on performance, retention, and transfer. Female students (between the age of 13 and 16) learning about lab safety procedures in VR, learned better from a female on-screen avatar, while males performed better when they received the information from a superhero avatar. There are other studies that have investigated multimedia design principles in VR such as the generative learning strategy of enacting (Makransky et al., 2020) where they found that enactment can lead to significantly better procedural knowledge and transfer when using VR compared to video.

These first studies on multimedia design principles in VR give primary indications of the transferability and effectiveness to VR learning environments, but currently do not allow any clear conclusions. Although the emphasis on information in VR is commonly found in applications in VR environments, there is no empirical research on their effects on learning outcomes and cognitive load as yet. It is important to investigate whether visual signals overstrain the learner in the already visual VR environment or whether they provide the necessary orientation to foster learning. Hence, the present paper investigates the signaling principle in form of textual annotations in a VR learning environment by differentiating effects on learning performance and cognitive load.

2. Hypotheses

The present study examines the effects of annotations in virtual learning environments in one group compared to a control group without annotations. We postulate that learning material that applies the signaling principle could lead to higher learning outcome. Thus, we hypothesize that learners who receive annotations reach higher recall (H1), comprehension (H2), and transfer (H3) scores than learners who did not receive annotations.

The signaling principle should also affect cognitive load. According to the cognitive load theory the ICL should not differ since the element-interactivity of the learning material does not change (H4). On the basis of the empirical evidence presented, it is expected that learners with annotations integrated into the learning material experience lower extraneous cognitive load (H5) and higher germane load (H6) than subjects in the control group. We tested our hypotheses in a virtual learning environment developed for headsets with 3DOF.

3. Method

3.1. A priori power analysis

To estimate the necessary sample size, we performed an a priori power analysis. A recent review on studies investigating signaling on learning outcomes reports a median effect size of $d = 0.46$ (Mayer, 2017). For this effect size, with an alpha = .05 and power = .95, the projected total sample size needed is approximately $N = 64$ (G*Power 3.1; Faul, Erdfelder, Lang, & Buchner, 2009).

3.2. Participants and experimental design

We collected data from 107 students from a German secondary school between the 8th and 10th grades. We excluded four students due to technical errors during the experiment. Participants' age ranged from 13 to 17 years with a mean of $M = 15.3$ ($SD = 0.95$), and 57 percent of them were males. In our one-factorial between-subject-design with two groups, the experimental group received visual annotations in form of text in their VR learning environment while the control group did not receive any annotations or additional help. To ensure proper randomization we implemented random permuted blocks of size two and four to ensure similar sample sizes across conditions. The program used for randomization was www.sealedenvelope.com with an allocation ratio of 1:1. As dependent variables we measured learning performance of the participants, differentiated in measures of recall, comprehension, and transfer, as well as learner's perceived cognitive load, differentiated in measures for ICL, ECL, and GCL. As potential covariates, we considered motivation, technology affinity, working memory capacity, spatial visualization, spatial abilities, and prior knowledge.

3.3. Materials and measures

The learning material consisted of a 3D animation on the subject of seawater desalination. The topics revolved around the desalination plant, distillers, evaporation, ion exchange, reverse osmosis, and electrodialysis. The animation was presented with a low-end VR-HMD setup on an Honor 8X smartphone, which was integrated into the Zeiss ONE Plus VR HMD. This phone has a 2.2 GHz Octa-core HiSilicon Kirin 710 CPU, ARM Mali-G51 MP4 1000 MHz GPU, and 4 GB Ram. The 6.50-inch display has a resolution of 2340 × 1080 and a refresh rate of 60 Hz. The HMDs binocular field of view is approximately 100° and has an adjustable interpupillary distance

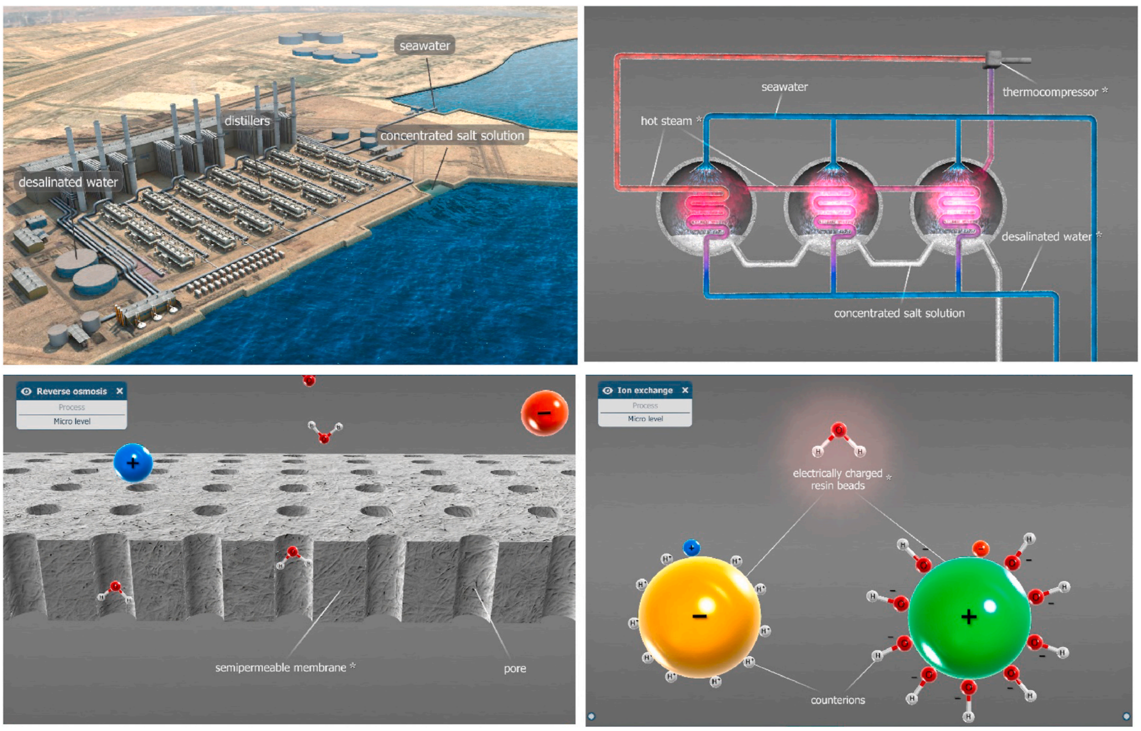

Fig. 1. Visual annotations shown as white text with lines.

knob (53–77 mm). The procedure of seawater desalination, the chemical processes as well as the advantages and disadvantages were explained visually and auditorily. All auditory information was provided by a narrator while the visual input consisted of 3D animations of the current topic. This learning unit can be accessed via the app mozaWeb3D (2018) and had a total duration of about 7 min. During the animation, learners could rotate the 3D models to a certain degree by moving their head around. The learning material could not leave the learner's field of vision but resonated with the field of vision at the end of the possible rotation. Both experimental groups saw the same animation and heard the same narration. Additional textual annotations (as seen in Fig. 1), were visually presented to the subjects of the experimental group. These annotations consisted of short words written directly on the corresponding picture elements to facilitate the orientation towards and the integration of the information.

All other tests were presented in an online questionnaire using laptops provided to the participants. Prior knowledge was measured with six open-ended questions, and one multiple choice question on domain specific knowledge, developed by a domain expert to ensure content validity. More specifically, this test contained seven knowledge questions (e.g., "What is electrodialysis?") with a total score of 17, and three questions to assess comprehension (e.g., "How is osmosis relevant?") with a maximum score of three. They could achieve between 1 and 5 points for each answer, depending on the question's difficulty. Partial-correct answers were scored with partial-points. Participants could reach a maximum of 20 points. The questions for prior knowledge and the post-test reflect different concepts and processes and not one unique psychological construct. It would therefore not be appropriate to report the internal consistency of these tests. Instead we used the Pearson product moment correlation coefficient, to calculate an inter-rater reliability between two independent ratings ($r = 0.92$, $p < .001$, $CI = 0.86$ - 0.94).

For learning outcomes, a post-test with ten open-ended and three multiple choice questions about the process of sea water desalination was developed by a domain expert. In order to better understand cognitive processes, we differentiated learning outcomes according to Bloom's taxonomy (Bloom, 1956). There are six hierarchical levels in the taxonomy, each requiring a higher level of abstraction. These levels represent reasoning skills required in educational settings and we focused on the first three levels. The test contained three recall questions in which a total score of seven could be obtained (e.g., "What do you call the process in which the sea water flows under high pressure through layers with electric current?"). In these questions, learners had to name or recognize short facts from the learning material. Comprehension was measured with four questions with a total score of five (e.g., "Explain why the process of desalination by ion exchange pollutes the environment"). In these questions, learners had to explain facts and their relationships in their own words. Transfer was measured with three questions in which learners could obtain a total score of three (e.g., "Which everyday object works similar to reverse osmosis?"). In these questions, learners had to draw inferences from the learning content provided and apply them in a new domain or field of application. Answers were compared with predefined solutions and participants could score partial points. Participants could reach a maximum of 15 points and results were reported as percentages. As the test measured diverse aspects of the overall content instead of a unique construct, once again we did not calculate an internal consistency measure, but calculated the inter-rater reliability, which revealed a very high consistency between the two raters ($r = 0.95$, $p < .001$, $CI = 0.93$–0.97).

We measured cognitive load with the differentiated cognitive load questionnaire (Klepsch, Schmitz, & Seufert, 2017) with 2 items for ICL, 3 items for ECL, and 3 items for GCL. They are subjectively rated on a 7-point Likert scale from 1 (not at all) to 7 (completely). The results are shown as percentages. Reliability was $\alpha = 0.74$ ($CI = 0.62$ - 0.82) for ICL, $\alpha = 0.73$ ($CI = 0.62$ - 0.81) for ECL, and $\alpha = 0.65$ ($CI = 0.52$ - 0.75) for GCL.

Demographical data about their age, gender, class level, and their previous contact with VR, rated on a five-point Likert Scale from 1 (never) to 5 (more than 6 times), were assessed as well. Furthermore, qualitative questions on their well-being while wearing the HMD, defective vision (glasses), and their general opinion on pros and cons of learning with HMDs were assessed by self-developed online questions during the posttest. As a manipulation check, we showed them two pictures (one with annotations and one without annotations) and asked them, which version of the learning material they received.

To measure *motivation*, two subscales of the questionnaire on current motivation (Rheinberg, Vollmeyer, & Burns, 2001) were used, and their means were aggregated. The subscale "current interest" was measured with two items (e.g., "I don't need a reward for tasks like this because I enjoy it anyway"), and four items measured the subscale "current challenge" (e.g., "This VR learning lesson is a real challenge for me"). The construct was measured with a 7-point Likert scale that ranged from 1 (totally disagree) to 7 (totally agree). Reliability for the subscale "current interest" was $\alpha = 0.98$ ($CI = 0.98$ - 0.99) and for "current challenge" $\alpha = 0.99$ ($CI = 0.98$ - 0.99).

To measure *working memory capacity* we developed an online version of the Numerical Memory Updating Test, using the same algorithms as in the original version (Oberauer, Süß, Schulze, Wilhelm, & Wittmann, 2000). Participants were first shown a 1×3 grid. Going cell by cell, they saw a number ranging from 1 to 9 which they had to remember. Then, participants had to perform simple mathematical operations with arrows appearing in each cell either pointing up (adding 1) or down (subtracting 1). In the end, they were required to report the result in each cell. When they completed the current grid with no errors, an additional cell was added with the grid expanding to a maximum of 3×3 (9 cells). When they made a mistake, the level repeated itself a maximum of two times until the participant made no more errors or the test was over.

To measure *spatial abilities*, the Paper Folding Test Vz-2, and the Card Rotation Test S-1 (Rev.) (Ekstrom, French, Harman, & Dermen, 1976) in paper-pencil format were used. For the paper folding test subjects were presented pictures of a paper that was folded several times and then perforated. They were required to then decide which of the five presented options was represented by the unfolded sheet of paper. The total score for the test was determined by adding the number of correct answers. In the Card Rotation Test (S-1 Rev.) from Ekstrom et al. (1976), participants were presented with a two-dimensional geometric figure in each task. For eight similar images (rotated and/or mirrored), they were required to decide whether or not the images resembled the presented figure. We calculated an overall score for spatial abilities by calculating the mean of the percentages of correct answers of both tests. The time limit for both tests according to Ekstrom et al. (1976) was 3 min each.

To measure *technology affinity*, the TA-EG was used (Karrer et al., 2009). We used two items of the subscale enthusiasm (e.g., "I love owning new electronic devices"), and one item for negative attitude (e.g., "Electronic devices help to access information"). Items were rated on a 5-point-scale ranging from 1 = "not at all true" to 5 = "very true". Cronbach's alpha showed a reliability of $\alpha = 0.79$ (CI = 0.69 - 0.86) for enthusiasm.

3.4. Procedure

As data collection took place in a school, all participants and their parents received an information letter about the study in advance. This letter included a description of the involved tasks, the anonymous use of all collected data, and the assurance that the experiment had no consequences on their school grades. They were also informed that they could quit the experiment at any time. Permission was sought and granted in writing from the participants parents since they were all under the age of 18. During the 90 min data collection, participants first completed the prior knowledge test and the demographical questions. After this, participants had to fill in the questionnaire on current motivation and technology affinity. Participants were then randomly allocated to either the experimental or control group. Following this, participants put on their headphones and placed the HMD on their head. After adjusting the straps for a proper fit, the VR learning phase began. Following the learning phase, participants were asked to answer the questions to measure learning outcomes. At the end, they completed the Cognitive Load Questionnaire, spatial ability tests, the qualitative questions as stated above, and the working memory test. All data analyses were performed with the Statistical Package for the Social Sciences Version 26 with an α-error set to $\alpha = 0.05$ for all calculations. The given interpretations of the effect size eta squared refer to Cohen (1988). According to Cohen (1988), the limits for the effect size are 0.01 (small effect), 0.06 (medium effect) and 0.14 (large effect).

4. Results

4.1. Descriptive data

The analysis of descriptive statistics of demographic data showed that subjects of the experimental group and the control group did not differ. We conduced t-tests, with regard to age ($p = .053$), class level ($p = .680$), prior knowledge ($p = .407$), working memory capacity ($p = .928$) and frequency of contact with VR ($p = .349$) but no significant differences were found. A χ^2-test revealed no differences between groups regarding gender (X^2 (1, N = 107) = 1.31, $p = .253$).

Descriptive data for all control variables per condition are listed in Table 1.

We analyzed if the potential control variables (prior knowledge, technology affinity, motivation, working memory capacity, spatial ability) had an influence on learning outcomes or cognitive load. Prior knowledge correlated significantly transfer ($r = 0.24$, $p = .006$). We have therefore included them as covariates in the further calculations concerning transfer. Motivation ($r = 0.45$, $p < .001$) and technology affinity ($r = 0.25$, $p = .009$) correlated significantly with GCL. These are included as a covariate for analysis regarding GCL. Further correlations can be taken from the correlation matrix in Table 2.

To test our hypotheses, we performed an ANCOVA since the main objective of the analysis is to examine a group effect and additionally control for prior knowledge. Since prior knowledge is not part of the hypotheses but has an influence on the dependent variable, this influence must be controlled. Residuals of critical variables were normally distributed by assessing Shapiro-Wilk tests ($p > .16$). Variances were homogenous based on Levene's Test which showed that equal variances could be assumed ($p > .45$). The descriptive statistics and ANOVA results for learning outcomes and cognitive load are shown in Table 3.

4.2. Learning outcomes

An ANCOVA showed a statistically significant difference for recall (H1: $F(1, 105) = 2.79$, $p = .048$, $\eta^2 = 0.026$). The effect size for H1 ($\eta^2 = 0.026$) can be interpreted as a small effect (Cohen, 1988). Participants, who had annotations present, received higher scores. For comprehension, we found no significant differences between the groups (H2: $F(1, 105) = 0.41$, $p = .524$, $\eta^2 = 0.004$). No difference for transfer was found between the groups after controlling for prior knowledge (H3: $F(1, 104) = 1.96$, $p = .08$, $\eta^2 = 0.019$) (see Fig. 2). The covariate, prior knowledge, was significantly related to transfer ($F(1, 104) = 6.94$, $p = .005$, $\eta^2 = 0.063$).

Table 1
Descriptive data for all variables in the conditions with or without annotations.

Baseline Characteristics	Experimental Group ($n = 51$)		Control Group ($n = 56$)		Full Sample ($N = 107$)	
Gender (female): *N* (%)	19	(37)	27	(48)	46	(43)
Age (years): *M* (*SD*)	15.22	(1.05)	14.86	(.82)	15.03	(.95)
Prior-knowledge (max = 20)	3.77	(1.57)	3.99	(1.24)	3.88	(1.40)
Contact to VR: *N* (%)	33	(64)	35	(62)	68	(63)
Visual aids (e.g. glasses): *N* (%)	19	(37)	16	(29)	35	(33)
Motivation: *M* (*SD*)	4.69	(.91)	4.63	(.74)	4.66	(.82)
Working memory capacity: *M* (*SD*)	3.75	(1.12)	3.73	(1.07)	3.74	(1.09)
Spatial ability: *M* (*SD*)	6.72	(2.08)	6.92	(2.00)	6.86	(2.03)
Technology affinity: *M* (*SD*)	4.15	(.85)	4.04	(.67)	4.09	(.77)

Table 2
Pearson correlation matrix (n = 107).

	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Age	1.00												
2. Gender	.09	1.00											
3. Recall	.02	-.09	1.00										
4. Comprehension	.04	.22*	.25*	1.00									
5. Transfer	.02	.12	.19*	.22*	1.00								
6. ICL	-.15	-.22*	-.21*	-.26**	-.02	1.00							
7. ECL	.01	-.17	-.20*	-.38**	-.35**	.23*	1.00						
8. GCL	.15	-.06	.34**	.18	.30**	.14	-,32**	1.00					
9. PK	-.08	-.07	.17	.15	.24*	-.19	-.17	-.07	1.00				
10. Paper folding	-.08	-.18	.18	-.07	.13	-.16	.01	.13	.10	1.00			
11. Figure rotation	-.14	-.14	.00	.09	.10	-.15	-.14	-.09	.11	.25*	1.00		
12. Motivation	.07	-.06	.16	.13	.07	.15	-.11	.45**	-.08	.01	-.14	1.00	
13. TA	.18	.22*	-.04	.06	.12	.12	-.06	.25**	-.06	.03	-.14	.41**	1.00

ICL = intrinsic cognitive load; ECL = extraneous cognitive load; GCL = germane cognitive load; PK = Prior knowledge; TA = Technology affinity; *p > .05; **p > .01.

Table 3
Means, standard deviations and ANOVA results in the different experimental conditions.

	Condition					
	Without Annotations (n = 51)		With Annotations (n = 56)		p	η^2
Outcome	*M*	*SD*	*M*	*SD*		
Post-test (%)						
Recall	52.29	24.29	60.29	24.86	.048	.026
Comprehension	39	21	42	28	.524	.004
Transfer	48.33	32.34	55.67	33	.08	.019
Cognitive load (%)						
ICL	60.71	19.57	59	18.71	.637	.002
ECL	60.14	14.6	58.1	21	.542	.004
GCL	64.3	14.7	69.9	16.6	.028	.035

ICL = intrinsic cognitive load; ECL = extraneous cognitive load; GCL = germane cognitive load.

4.3. Cognitive load

As expected, ICL did not differ between the groups (H4: $F(1, 105) = 0.22$, $p = .637$, $\eta^2 = 0.002$). Against our expectations, we found no differences in ECL between the two groups (H5: $F(1, 105) = 0.38$, $p = .542$, $\eta^2 = 0.004$). However, GCL was higher in the group with annotations compared to the control group (H6: $F(1, 102) = 3.74$, $p = .028$, $\eta^2 = 0.035$) (see Fig. 3). The covariate, motivation ($F(1, 102) = 19.57$, $p < .001$) was significantly related to GCL, while technological affinity was not ($F(1, 102) = 0.48$, $p = .49$). The effect size for H6 ($\eta^2 = 0.035$) can be classified as a small effect (Cohen, 1988).

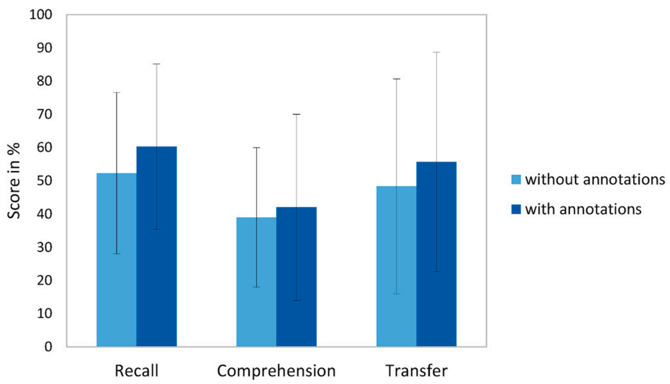

Fig. 2. Mean scores in the post-test.

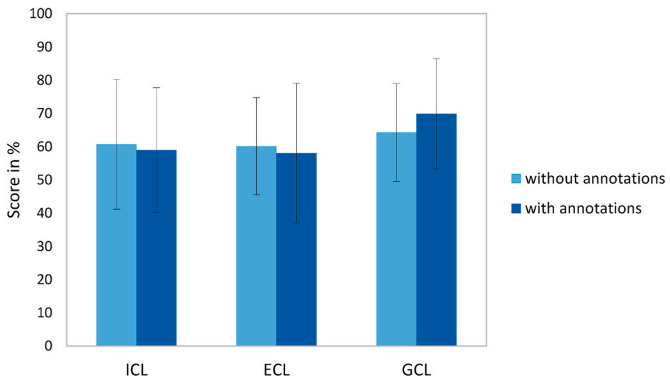

Fig. 3. Mean cognitive load. ICL = intrinsic cognitive load; ECL = extraneous cognitive load; GCL = germane cognitive load.

5. Discussion

The primary purpose of this study was to examine the effects of signaling in a virtual learning environment. We assumed that we could apply instructional design research with classical multimedia settings and their underlying cognitive models for learning in VR environments (Mayer, 2005). Therefore, we analyzed whether signals in form of textual annotations in a VR learning environment could increase learning outcomes, reduce ECL, and foster GCL.

5.1. *Effects of annotations on learning outcome*

For our first hypotheses we analyzed whether learning outcomes, differentiated into the levels of recall (H1), comprehension (H2), and transfer (H3), are affected by annotations. In line with our first hypothesis (H1), we found that subjects who received the VR learning material with annotations reached higher *recall* scores compared to the control group that learned without annotations present. This might be explained by three different reasons: First, relevant information of the auditive narration was repeated by the annotations. All participants heard verbal instructions. However, the experimental group additionally read the keywords and main statements, which were visually presented by the annotations. Thus, the information was repeated and could thus be encoded dually. Second and most obvious, the annotations helped the learner to guide their attention to relevant parts of the learning material (Clark, Nguyen, & Sweller, 2011). This reduces unnecessary visual search. Third, as the labels were attached spatially near the respective graphical entity it could additionally foster the integration of text and video, at least on a surface level. Although these explanatory approaches can also be found in classical learning environments, it is especially the attention-direction and spatial proximity that could be important in VR. While the dual encoding of content takes place as a cognitive process independent of the learning medium, the spatial component in particular must be taken into account in design decisions in VR. To give signals an attention-guiding function, they should not be obscured by other objects in the VR environment. At the same time, the placement of signals in virtual space also becomes more complex, since three-dimensional orientations of the signals to the learner must be taken into account.

However, both *comprehension and transfer* outcomes were not raised through the implementation of annotations (H2 and H3). We expected increased learning on all levels of learning outcomes, as support of the integration of text and video should benefit the entire learning process. We may have supported integration, but as just mentioned, this could have been merely on a surface level, without an understanding of the deeper relationship between text and video. Thus, it seems that the main supportive mechanism was the repetition of labels and main statements that supported the surface level processing of isolated elements. If the textual annotations had consisted more of relations, they might have been more helpful for comprehension and transfer. The mere fact that the annotations hinted towards a connection of textual labels and pictorial elements seemed to be insufficient to actually foster text-picture-integration which would have been the prerequisite for comprehension. Additional explanations of how the elements could be connected in terms of a deep-level mapping help would have been necessary (Brünken & Seufert, 2006). Such a phenomenon is also known for cognitive strategies that aim at a simple repetition (Annis & Annis, 1987; Brünken & Seufert, 2006; Dinsmore & Alexander, 2012). According to Bloom (1956) the different levels are hierarchical. Therefore, it is not surprising that if we have not found differences between the groups on comprehension, that the transfer performance also could not be enhanced by textual annotations.

Another reason for the missing effects on comprehension and transfer might also be the difficulty of the learning material. Although care was taken in planning the study to ensure that the learning content matched the classroom levels and that basic knowledge of the subjects had possibly already been acquired, participants showed difficulties in comprehension questions, even when they were supported by annotations. Overall, the previous knowledge of learners examined in this study was low. Therefore, it is also possible that participants were unable to use the annotations effectively for these deeper levels of processing owing to insufficient previous knowledge (Seufert, 2003).

An additional reason for the missing effects on deeper levels of processing might be, that the annotations in the learning material are kept very short and simple. They aim at guiding attention and promoting semantic processing. However, particularly this deeper processing was not triggered. In order to actually encourage learners to process the content more deeply, signals should rather initiate generative activities, like prompts with explicit tasks (Makransky et al., 2020). Practically, these signals could be integrated dynamically into the VR learning environment, e.g., by fading in and out adaptively to the position of the learner. Whether the annotations actually had an attention guidance effect, additional information would be needed. Tracking eye movements during learning might be a suitable approach. Consequently, future studies might want to explore the use of signals combined with eye-tracking measures.

It would also be interesting to analyze the learners feeling of being present in the environment, which could positively support deep learning. The annotations used in this study are simple texts floating in virtual space. This representation may have negatively influenced the learners experience of presence, as the annotations make learners aware that they are not physically present in the learning environment. One consideration could be to integrate the texts in a way that they resemble how texts occur in the real world. Although Baceviciute et al. (2020) found no difference in knowledge acquisition when presenting texts in book form compared to text overlay, they found a lower cognitive load and higher self-efficacy among learners.

Another reason why the effects on the deeper levels of processing do not occur may be due to the sensitivity of the measurements. In order to determine the different learning levels, we have based our tests on learning tests used in classic multimedia settings. Depending on the content of the learning environment, which was rather knowledge-based, these tests seem to be a suitable method. However, at the deeper levels of processing, such as comprehension or transfer, there might be a need for more realistic tests with more diverse and sophisticated items to address the different levels and aspects of processing.

Perhaps the device itself led students to a more play-like attitude, thus focusing on surface features rather than on deep processing. Whether this surface orientation owing to distraction or play-like attitudes is an issue of learning in VR in general needs to be addressed in further studies that differentiate the levels of processing as we proposed in our study.

5.2. Effects of annotations on cognitive load

In line with our hypotheses (H4) we found no differences between the groups for ICL. The perceived difficulty of the learning material was in the medium range and no differences were seen between the two groups. That is because element interactivity remains the same in both groups, only one element is repeated in the annotations group.

Against our expectation (H5) we found no differences between the two groups for ECL. This result is inconsistent with the assumptions of the CTML and CLT, which state that signaling the relevant parts of the learning material should reduce ECL (Mayer & Fiorella, 2014). However, such a relieving effect for the visual search would only pay off when the visual requirements are considerable. In line with this argument, annotations can be more effective when the visual display of the learning material is complex rather than simple (Jeung, Chandler, & Sweller, 1997). Even though our learning material was presented in VR, the visual display of the learning material may not have been visually overwhelming for the learners as could be inferred from the rather low scores for ECL and ICL of about 60%. Therefore, no cognitive capacity needed to be freed up by adding annotations. The annotations were not necessary to reduce the effort of visually scanning the learning material, and thus failed to reduce extraneous processing (De Koning et al., 2009; Mayer & Fiorella, 2014).

A between-subject design was chosen for the study. In a within-subject design, the learners might have noticed the difference between the conditions and therefore the annotations might have had the desired impact on ECL. However, this would be challenging to realize, as comparable learning environments would be necessary.

Findings concerning GCL are in line with our hypotheses (H6). We found higher GCL when learners received annotations compared to the control group. It can be assumed that processing and understanding processes can be supported by adding annotations (Kalyuga, 2009). Based on the theoretical considerations of the additive composition of the types of cognitive load (Sweller et al., 1998), we assumed this is due to the reduction of ECL. Since this was not the case in this study, we have to discuss a different explanation. We have referred to the original cognitive load theory which consists of three different load types. There are also other perspectives on cognitive load that understand GCL more as a resource to deal with ICL (Kalyuga, 2011). This means that if the complexity of the learning material is sufficient to challenge the learner and the learner wants to process the material more deeply, then more germane resources must be made available than originally necessary. These additional resources can then be used to foster learning, in our case on the surface level. Kalyuga (2011) also states that motivation plays a role in how much the learner is willing to invest GCL. Following on from this, another potential alternative explanation is the motivation of learners.

We found that GCL and motivation correlate substantially ($r = 0.45$, $p < .001$). Previous studies on motivation found that motivation can activate better search strategies (Baranes, Oudeyer, & Gottlieb, 2014). Thus, learners might be encouraged by annotations to activate more learning processes leading to more GCL. These active processes could also be activated since learners already have developed better learning strategies for the familiar presentation form (see ECL). The correlation between GCL and technology affinity ($r = 0.25$, $p = .009$) suggests that this new technology may be particularly exciting for children with a high affinity for technology. Since technological affinity and motivation correlate ($r = 0.41$, $p < .001$), the enthusiasm for this technology could lead to children being more motivated to invest greater load. However, further research is needed to draw conclusions. Why the increased germane processes nevertheless failed to affect learning process on a deeper level would need further investigation. Mediation analyses with motivational factors as mediator for example, could shed more light on these relations between cognitive and motivational or even affective processes.

Concerning our cognitive load measure in general, we conducted the cognitive load questionnaire after the post-test which could have biased the results. A study by Schmeck, Opfermann, Van Gog, Paas, and Leutner (2015) compared immediate and delayed ratings of cognitive load and found higher reported ratings for difficulty and effort in the delayed condition than the average score collected with multiple measures during the learning session. Perhaps the difficulty of the post-test tasks affected the load ratings and thus we have no direct measure of the load caused by the learning material itself. The delayed measurement could therefore have had an influence on the results obtained. Based on the results of Schmeck et al. (2015) we would assume that the scores for the load ratings would be higher than directly after learning. However, since we measured the data at the same time in both groups, the bias in the data should be in the same direction and should not alter the differences we found. Although we used a differentiated questionnaire, it has been shown that using a one-time measurement of cognitive load using a single item at the end of a series of tasks can lead to higher reported scores than multiple repeated measurements (Van Gog, Kirschner, Kester, & Paas, 2012). Learning materials are also not always equally difficult over time and are subject to fluctuations, which can also affect a retrospective measurement of cognitive load (Leppink, Paas, Van der Vleuten, Van Gog, & Van Merriënboer, 2013). For future studies it could be illustrative to investigate how cognitive load changes over time in VR learning environments, and which measurement instruments are particularly suitable for measuring cognitive load in virtual environments.

5.3. Practical implications

When educational institutions or companies decide to develop learning material in VR, they currently have little empirical findings or guidelines that could help them design effective learning environments. Especially in VR, learners can have a rich learning experience through a multimedia and interactive learning environment, where they can immerse themselves in the learning material. Since many companies and educational institutions are already providing VR learning experiences, we need further empirical research, both in terms of cognitive processes and how to design appropriate learning material, so that learning in VR can be successful.

The major findings of the present study suggest that using annotations in a VR learning environment can not only improve learning outcome, at least on a surface level, but also increase GCL. When the task is to recall information, annotations can help learners. Other studies suggest that VR as a new popular technology can be more motivating to students than conventional media and can spark interest with appropriate design of the learning material (Parong & Mayer, 2018).

The annotations in this study were placed close to the image. In previous studies this spatial contiguity principle was investigated in a real, non-virtual learning environment and a robust effect was found (Schroeder & Cenkci, 2018). In most of these studies the experiment was designed by comparing visually close annotations to visually distant annotations. In this experiment we did not manipulate the distance of the annotations, but it might be an interesting research question for future studies.

The features of VR learning environment include three-dimensional and stereoscopic presentation of the learning material. This can make the learning material appear visually more complex, as it can make spatial orientation more difficult for the learner. In terms of the field of view, there is more visual information in VR due to the change of perspective, even if such changes have been very limited in our case. Signals in VR could reduce these difficulties by serving as orientation aids. If learners need a three-dimensional understanding of the learning content, then the creation of a mental model can be facilitated if the learning content is presented directly in 3D, because the material does not have to be mentally translated from 2D to 3D. However, whether signals can assist in these processes requires further investigation, as we did not find any effects of signals in VR for deeper processing of learning content in this study.

This study investigated VR in a realistic school context. While the technical progress with new VR systems is making rapid advances (Radianti et al., 2020), this might not impact schools or similar educational settings for the time being. The practicability is limited by the cost of high-end VR HMDs and high-performance computers required per student, as high-end HMDs were not originally intended for educational classroom use (Jensen & Konradsen, 2018). Furthermore, the training of teachers, technical problem solving, and suitable environments (thinking of the motion sensors with multiple HMDs in the same room) are challenging. However, low-end VR can also be used to create immersive learning environments that are more likely to find practical applications. Therefore, we wanted to investigate a low-end VR setting that could be realistically applied, both in terms of feasibility and acquisition cost.

5.4. Limitation and recommendations for future studies

In this study we used an experimental design to investigate if annotations can improve learning outcome in a VR learning environment.

Our study was conducted in an authentic school setting. The participants in the study were pupils of different grades in a regular Bavarian grammar school. Compared to a laboratory setting, the experiment took place in a familiar environment that students associate with learning. As a result, we can generalize the effects found in a realistic school context. However, it should be noted that even though we collected data in the field, it was a guided setting. To generalize the effects found to a setting with more degrees of freedom by the learner (e.g., learning from home) it must be further researched. Although we have selected a number of classes in which the subject is part of the curriculum, our sample is limited to a specific age group. Whether similar effects are observed in students or adults is an open question.

Although the power-analysis based on former findings about the signaling effect revealed an a priori sample size of 64 participants, there were two reasons why we collected more. First, the effect size for the calculation was taken from a laboratory study under controlled conditions with adults. Since we collected data in grades 8 to 10 in a realistic school setting, we had to expect dropouts caused by technical problems or by the absence of children. Since parents had to sign the consent form, concerns could also have led to higher dropout rates. These dropouts did not occur in the practical realization of the study. The second reason is an ethical one. We have chosen a learning topic in close consultation with the school, so it fits into the Bavarian curriculum. In order not to withhold the VR learning unit from certain students in a common class, we always included all eligible children in a class.

Our study design was also conducted in a very controlled environment. Results of a pilot study showed that initial contact with VR HMD can be challenging for learners. In order to overcome these technical challenges, the researchers pre-configured the HMD's to ensure a seamless experience for the learners. Thus, the results of the study contribute empirical findings on how to design VR learning material.

A few limitations need to be taken into consideration. First, the textual annotations were merely word repetitions of the auditory narration. The text and picture integration can be of different depth and elaboration. These annotations have only managed to establish an element-to-element mapping basis and therefore the image integration remains on surface level processing.

The words in the app were predetermined and could not be replaced by more comprehensive and relation-based explanations that might have been more beneficial to learning. Although the words were subject to these restrictions, it is even more striking that we found a significant learning effect for the annotations. The effect sizes found ($\eta^2 = 0.026$ - 0.035) support the assumption that annotations can be an important design choice to support learners' recall performance. Although the informational content of the texts was limited, the learners were able to increase their knowledge by an average of 8%.

Secondly, the presented learning material was presented via a low-end VR HMD with 3DOF and stereoscopic vision. In a 360° 3D environment it could happen that the learners cannot follow the learning material because it is currently not visible in their field of view. To make sure that the learners have perceived the learning material, it could not leave their field of view. This allowed the students to look around while the 3D models can rotate to a certain degree, but it was always ensured that they had the learning material in their field of vision. In future studies it would be preferable to remove this fixation to see if the effects are still replicable. We assume that in an open-world learning setting, learners might benefit even more from annotations because in this setting annotations could bring out the advantage of drawing attention. At the same time, however, one should consider how to ensure that the learners' focus of attention is on the learning material. To ensure to what extent the annotations guide attention and how long the annotations are actually being viewed, future studies could collect process data using eye tracking. This could provide further insights into the visual search patterns and may suggest how the annotations could be improved. This should also be considered if the results should be applied to learning environments used with 6DOF HMDs. The focus of this study was on cognitive processes. These cognitive processes should not differ with more degrees of freedom, because they take place in working memory and the differences of DOF would rather be noticeable in perceptual or motivational processes in advance. Nevertheless, differences could occur in the workload of the working memory. While we already found a high ECL in 3DOF, there might be additional affordances in 6DOF that could further increase the ECL. However, it would be interesting to see if the found effects can be replicated with 6DOF HMDs.

Another limitation could have been that the texts were difficult to read through VR HMDs. Due to an improper fit of the glasses or visual impairments, some learners might have found it difficult to read the annotations. With technological advance, manufacturers offer increasingly higher resolutions in the HMDs which also result in better readability of texts. In future research it would be interesting to investigate how much text is appropriate for the learner at any given time. Therefore, it could also be interesting to look at the modality effect in VR.

In the future, learning environments in VR will become more complex with further development of technology and at the same time appeal to a wider audience. The learning possibilities go beyond the classical learning materials. Dangerous situations can be simulated realistically to learn safety rules without danger. Companies can improve the learning process of new employees by rehearsing processes in advance in 3D. In this way, learners can experience the tasks of an industry in VR, such as milling, sawing or welding. Doctors can learn techniques for difficult operations with the help of medical VR operation robots, which simultaneously adapt to the skills of the learner. Although all these examples are different in their application, it can be useful for all of them to present information with the help of textual annotations. Therefore, it is important to investigate the underlying psychological processes of learning with VR, so that learning environments create optimal preconditions for different learners. It would be important to investigate other forms of signals in VR as well. In a meta-analysis on the signaling-principle it is suggested that dynamic signals have advantages over static signals (Alpizar, Adesope, & Wong, 2020). Future research should therefore go beyond textual annotations and focus on how different signals can influence different learning processes like allocation of attention or effort, surface or deep learning and also motivational processes.

6. Conclusion

The aim of this study was to investigate if the signaling principle can help learners in a VR learning environment. In comparison to previous research, we have focused on cognitive processes in a differentiated approach (Maas & Hughes, 2020; Makransky, Terkildsen, & Mayer, 2019; Parong & Mayer, 2018). The differentiated analysis was important because we have found that, as in classical multimedia settings, the signaling principle in the form of textual annotations can help learners to improve their learning outcome in recall, but not in comprehension or transfer questions. This might be because annotations help selecting the relevant elements and in integrating text and images but only on the surface level of mapping elements. With the help of the annotations, the learners also invested more GCL than learning without annotations. Since this effect is not due to a relief in ECL, the question arises as to which effects contribute to an increase of GCL. Thus, our study might be one first initiative towards a refined multimedia learning theory with regard to the different levels of processing while learning in VR. In further studies the motivation of the learners as possible mediators will therefore be analyzed to complement the rather cognitive perspective of this study. Overall, when designing VR learning environments and wanting to increase both GCL and the learning outcome on a recall level, we recommend supporting the learners with annotations.

Ethics statement

This study was exempt from an ethic committee approval due to the recommendations of the German Research Association: All participants were in no risk of physical or emotional pressure. We fully informed all participants about the goals and process of this study and none of the participants were patients or persons with disabilities. Minors were required to bring a written consent of their parents to participate in this study. In all studies participation was voluntary and all participants signed a written informed consent and were aware that they had the chance to withdraw their data at any point of the study.

Credit author statement

Albus Patrick: Conceptualization, Methodology, Investigation, Formal analysis, Writing – original draft, Writing – review & editing; Andrea Vogt: Conceptualization, Methodology, Investigation; Tina Seufert: Conceptualization, Supervision, Funding acquisition, Writing – review & editing.

References

mozaWeb3D. (2018). *Seawater desalination [video Game]*. Szeged, HUN: Mozaik Education Ltd.

Akçayır, M., Akçayır, G., Pektaş, H. M., & Ocak, M. A. (2016). Augmented reality in science laboratories: The effects of augmented reality on university students' laboratory skills and attitudes toward science laboratories. *Computers in Human Behavior, 57*, 334–342. https://doi.org/10.1016/j.chb.2015.12.054

Alhalabi, W. (2016). Virtual reality systems enhance students' achievements in engineering education. *Behaviour & Information Technology, 35*(11), 919–925. https://doi.org/10.1080/0144929X.2016.1212931

Allcoat, D., & von Mühlenen, A. (2018). Learning in virtual reality: Effects on performance, emotion and engagement. *Research in Learning Technology, 26*. https://doi.org/10.25304/rlt.v26.2140

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

Annis, L. F., & Annis, D. B. (1987). *Does practice make perfect? The Effects of Repetition on student learning (ED281861)*. ERIC. https://eric.ed.gov/?id=ED281861.

Argelaguet, F., & Andujar, C. (2013). A survey of 3D object selection techniques for virtual environments. *Computers & Graphics, 37*(3), 121–136. https://doi.org/10.1016/j.cag.2012.12.003

Ayres, P., & Sweller, J. (2005). The split-attention principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 135–146). Cambridge University Press.

Baceviciute, S., Mottelson, A., Terkildsen, T., & Makransky, G. (2020). Investigating representation of text and audio in educational VR using learning outcomes and EEG. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–13). https://doi.org/10.1145/3313831.3376872

Bacim, F., Kopper, R., & Bowman, D. A. (2013). Design and evaluation of 3D selection techniques based on progressive refinement. *International Journal of Human-Computer Studies, 71*(7–8), 785–802. https://doi.org/10.1016/j.ijhcs.2013.03.003

Baddeley, A. (1992). Working memory. *Science, 255*(5044), 556–559. https://doi.org/10.1126/science.1736359

Baranes, A. F., Oudeyer, P.-Y., & Gottlieb, J. (2014). The effects of task difficulty, novelty and the size of the search space on intrinsically motivated exploration. *Frontiers in Neuroscience, 8*. https://doi.org/10.3389/fnins.2014.00317

Taxonomy of educational objectives. In Bloom, B. S. (Ed.), *Handbook 1: Cognitive domain*, (1956). Longman.

Brünken, R., Steinbacher, S., Plass, J. L., & Leutner, D. (2002). Assessment of cognitive load in multimedia learning using dual-task methodology. *Experimental Psychology, 49*(2), 109–119. https://doi.org/10.1027//1618-3169.49.2.109

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction, 8*(4), 293–332. https://doi.org/10.1207/s1532690xci0804_2

Chen, Y.-L. (2016). The effects of virtual reality learning environment on student cognitive and linguistic development. *The Asia-Pacific Education Researcher, 25*(4), 637–646. https://doi.org/10.1007/s40299-016-0293-2

Chen, R., Grierson, L., & Norman, G. (2015). Manipulation of cognitive load variables and impact on auscultation test performance. *Advances in Health Sciences Education, 20*(4), 935–952. https://doi.org/10.1007/s10459-014-9573-x

Christou, C. (2010). Virtual reality in education. In A. Tzanavari, & N. Tsapatsoulis (Eds.), *Affective, interactive and cognitive methods for e-learning design: Creating an optimal education experience* (pp. 228–243). IGI Global. https://doi.org/10.4018/978-1-60566-940-3.

Clark, R. C., Nguyen, F., & Sweller, J. (2011). *Efficiency in learning: Evidence-based guidelines to manage cognitive load*. John Wiley & Sons.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale: L. Erlbaum Associates.

De Koning, B. B., Tabbers, H. K., Rikers, R. M., & Paas, F. (2009). Towards a framework for attention cueing in instructional animations: Guidelines for research and design. *Educational Psychology Review, 21*(2), 113–140. https://doi.org/10.1007/s10648-009-9098-7

Dinsmore, D. L., & Alexander, P. A. (2012). A critical discussion of deep and surface processing: What it means, how it is measured, the role of context, and model specification. *Educational Psychology Review, 24*(4), 499–567. https://doi.org/10.1007/s10648-012-9198-7

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Manual for kit of factor-referenced cognitive tests*. Educational testing service.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160. https://doi.org/10.3758/BF03193146

Fowler, C. (2015). Virtual reality and learning: Where is the pedagogy? *British Journal of Educational Technology, 46*(2), 412–422. https://doi.org/10.1111/bjet.12135

Frederiksen, J. G., Sørensen, S. M., Konge, L., Svendsen, M. B., Nobel-Jørgensen, M., Bjerrum, F., et al. (2020). Cognitive load and performance in immersive virtual reality versus conventional virtual reality simulation training of laparoscopic surgery: A randomized trial. *Surgical Endoscopy, 34*(3), 1244–1252. https://doi.org/10.1007/s00464-019-06887-8

Freina, L., & Ott, M. (2015). A literature review on immersive virtual reality in education: State of the art and perspectives. *The International Scientific Conference eLearning and Software for Education, 1*, 133.

Jensen, L., & Konradsen, F. (2018). A review of the use of virtual reality head-mounted displays in education and training. *Education and Information Technologies, 23*(4), 1515–1529. https://doi.org/10.1007/s10639-017-9676-0

Jeung, H.-J., Chandler, P., & Sweller, J. (1997). The role of visual indicators in dual sensory mode instruction. *Educational Psychology, 17*(3), 329–345. https://doi.org/10.1080/0144341970170307

Juan, M., García-García, I., Mollá, R., & López, R. (2018). Users' perceptions using low-end and high-end mobile-rendered HMDs: A comparative study. *Computers, 7*(1), 15. https://doi.org/10.3390/computers7010015

Kalyuga, S. (2009). *Managing cognitive load in adaptive multimedia learning. Information Science Reference*. https://doi.org/10.4018/978-1-60566-048-6

Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology, 13*(4), 351–371. https://doi.org/10.1002/(SICI)1099-0720(199908)13:4<351::AID-ACP589>3.0.CO;2-6

Karrer, K., et al. (2009). Technikaffinität erfassen - der Fragebogen TA-EG [Measure affinity to technology – the questionnaire TA-EG]. In A. Lichtenstein, C. Stößel, & C. Clemens (Eds.), *Der Mensch im Mittelpunkt technischer Systeme 8. Berliner Werkstatt Mensch-Maschine-Systeme 7. bis 9. Oktober 2009* (Vol. 29, pp. 196–201). ZMMS Spektrum Band 22 (Düsseldorf: VDI).

Kozhevnikov, M., Gurlitt, J., & Kozhevnikov, M. (2013). Learning relative motion concepts in immersive and non-immersive virtual environments. *Journal of Science Education and Technology, 22*(6), 952–962. https://doi.org/10.1007/s10956-013-9441-0

Lai, Y.-S., Tsai, H.-H., & Yu, P.-T. (2011). Integrating annotations into a dual-slide PowerPoint presentation for classroom learning. *Journal of Educational Technology & Society, 14*(2), 43–57.

Leppink, J., Paas, F., Van der Vleuten, C. P., Van Gog, T., & Van Merriënboer, J. J. (2013). Development of an instrument for measuring different types of cognitive load. *Behavior Research Methods, 45*(4), 1058–1072.

Lorch, R. F. (1989). Text-signaling devices and their effects on reading and memory processes. *Educational Psychology Review, 1*(3), 209–234. https://doi.org/10.1007/BF01320135

Lubos, P., Bruder, G., & Steinicke, F. (2014). March). Analysis of direct selection in head-mounted display environments. In *2014 IEEE symposium on 3D user interfaces (3DUI* (pp. 11–18). IEEE. https://doi.org/10.1109/3DUI.2014.6798834.

Maas, M. J., & Hughes, J. M. (2020). Virtual, augmented and mixed reality in K–12 education: A review of the literature. *Technology, Pedagogy and Education, 29*(2), 231–249. https://doi.org/10.1080/1475939X.2020.1737210

Makransky, G., Andreasen, N. K., Baceviciute, S., & Mayer, R. E. (2020). Immersive virtual reality increases liking but not learning with a science simulation and generative learning strategies promote learning in immersive virtual reality. *Journal of Educational Psychology*. https://doi.org/10.1037/edu0000473

Makransky, G., Borre-Gude, S., & Mayer, R. E. (2019a). Motivational and cognitive benefits of training in immersive virtual reality based on multiple assessments. *Journal of Computer Assisted Learning, 35*(6), 691–707. https://doi.org/10.1111/jcal.12375

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

Makransky, G., Wismer, P., & Mayer, R. E. (2019c). A gender matching effect in learning with pedagogical agents in an immersive virtual reality science simulation. *Journal of Computer Assisted Learning, 35*(3), 349–358. https://doi.org/10.1111/jcal.12335

Martin, F., & Betrus, A. K. (2019). *Digital media for learning: Theories, processes, and solutions*. Springer International Publishing. https://doi.org/10.1007/978-3-030-33120-7

Mautone, P. D., & Mayer, R. E. (2001). Signaling as a cognitive guide in multimedia learning. *Journal of Educational Psychology, 93*(2), 377. https://doi.org/10.1037/0022-0663.93.2.377

Mayer, R. E. (2005). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (1st ed., pp. 31–48). Cambridge University Press.

Mayer, R. E. (Ed.). (2014). *The Cambridge handbook of multimedia learning* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.

Mayer, R. E. (2017). Using multimedia for e-learning. *Journal of Computer Assisted Learning, 33*(5), 403–423. https://doi.org/10.1111/jcal.12197

Mayer, R. E., & Fiorella, L. (2014). Principles for reducing extraneous processing in multimedia learning: Coherence, signaling, redundancy, spatial contiguity, and temporal contiguity principles. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 279–315). Cambridge University Press.

Mayer, R. E., Heiser, J., & Lonn, S. (2001). Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of Educational Psychology, 93*(1), 187–198. https://doi.org/10.1037//0022-0663.93.1.187

Merchant, Z., Goetz, E. T., Cifuentes, L., Keeney-Kennicutt, W., & Davis, T. J. (2014). Effectiveness of virtual reality-based instruction on students' learning outcomes in K-12 and higher education: A meta-analysis. *Computers & Education, 70*, 29–40. https://doi.org/10.1016/j.compedu.2013.07.033

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

Oberauer, K., Süß, H.-M., Schulze, R., Wilhelm, O., & Wittmann, W. W. (2000). Working memory capacity—facets of a cognitive ability construct. *Personality and Individual Differences, 29*(6), 1017–1045. https://doi.org/10.1016/S0191-8869(99)00251-2

Olmos, E., Cavalcanti, J. F., Soler, J. L., Contero, M., & Alcañiz, M. (2018). Mobile virtual reality: A promising technology to change the way we learn and teach. In *Mobile and ubiquitous learning* (pp. 95–106). Singapore: Springer.

Packer, J. (2006). Learning for fun: The unique contribution of educational leisure experiences. *Curator: The Museum Journal, 49*(3), 329–344. https://doi.org/10.1111/j.2151-6952.2006.tb00227.x

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/Revue canadienne de psychologie, 45*(3), 255–287. https://doi.org/10.1037/h0084295

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

Parong, J., & Mayer, R. E. (2020). Cognitive and affective processes for learning science in immersive virtual reality. *Journal of Computer Assisted Learning*. https://doi.org/10.1111/jcal.12482

Passig, D., Tzuriel, D., & Eshel-Kedmi, G. (2016). Improving children's cognitive modifiability by dynamic assessment in 3D Immersive Virtual Reality environments. *Computers & Education, 95*, 296–308. https://doi.org/10.1016/j.compedu.2016.01.009

Petersen, G. B., Klingenberg, S., Mayer, R. E., & Makransky, G. (2020). The virtual field trip: Investigating how to optimize immersive virtual learning in climate change education. *British Journal of Educational Technology*. https://doi.org/10.1111/bjet.12991

Radianti, J., Majchrzak, T. A., Fromm, J., & Wohlgenannt, I. (2020). A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Computers & Education, 147*, 103778. https://doi.org/10.1016/j.compedu.2019.103778

Rheinberg, F., Vollmeyer, R., & Burns, B. D. (2001). FAM: Ein fragebogen zur erfassung aktueller motivation in lern- und leistungssituationen [fam: A questionnaire to assess current motivation in learning and performance situations]. *Diagnostica, 47*(2), 57–66. https://doi.org/10.1026//0012-1924.47.2.57

Richards, D., & Taylor, M. (2015). A Comparison of learning gains when using a 2D simulation tool versus a 3D virtual world: An experiment to find the right representation involving the Marginal Value Theorem. *Computers & Education, 86*, 157–171. https://doi.org/10.1016/j.compedu.2015.03.009

Schmeck, A., Opfermann, M., Van Gog, T., Paas, F., & Leutner, D. (2015). Measuring cognitive load with subjective rating scales during problem solving: Differences between immediate and delayed ratings. *Instructional Science, 43*(1), 93–114.

Schroeder, N. L., & Cenkci, A. T. (2018). Spatial contiguity and spatial split-attention effects in multimedia learning environments: A meta-analysis. *Educational Psychology Review, 30*(3), 679–701. https://doi.org/10.1007/s10648-018-9435-9

Selzer, M. N., Gazcon, N. F., & Larrea, M. L. (2019). Effects of virtual presence and learning outcome using low-end virtual reality systems. *Displays, 59*, 9–15. https://doi.org/10.1016/j.displa.2019.04.002

Slater, M., & Sanchez-Vives, M. V. (2016). Enhancing our lives with immersive virtual reality. *Frontiers in Robotics and AI, 3*, 74. https://doi.org/10.3389/frobt.2016.00074

Sutcliffe, A. G., Poullis, C., Gregoriades, A., Katsouri, I., Tzanavari, A., & Herakleous, K. (2019). Reflecting on the design process for virtual reality applications. *International Journal of Human-Computer Interaction, 35*(2), 168–179. https://doi.org/10.1080/10447318.2018.1443898

Sweller, J. (2005). Implications of cognitive load theory for multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (1st ed., pp. 19–30). Cambridge University Press.

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

Teather, R. J., & Stuerzlinger, W. (2014). October). Visual aids in 3D point selection experiments. In *Proceedings of the 2nd ACM symposium on Spatial user interaction* (pp. 127–136). https://doi.org/10.1145/2659766.2659770

Van Gog, T. (2014). The signaling (or cueing) principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 263–278). Cambridge University Press.

Van Gog, T., Kirschner, F., Kester, L., & Paas, F. (2012). Timing and frequency of mental effort measurement: Evidence in favour of repeated measures. *Applied cognitive psychology, 26*(6), 833–839. https://doi.org/10.1002/acp.2883

Wallen, E., Plass, J. L., & Brünken, R. (2005). The function of annotations in the comprehension of scientific texts: Cognitive load effects and the impact of verbal ability. *Educational Technology Research & Development, 53*(3), 59–71. https://doi.org/10.1007/BF02504798

Webster, R. (2016). Declarative knowledge acquisition in immersive virtual learning environments. *Interactive Learning Environments, 24*(6), 1319–1333. https://doi.org/10.1080/10494820.2014.994533

Wu, B., Yu, X., & Gu, X. (2020). Effectiveness of immersive virtual reality using head-mounted displays on learning performance: A meta-analysis. *British Journal of Educational Technology*. https://doi.org/10.1111/bjet.13023

Yeh, S.-W., & Lo, J.-J. (2009). Using online annotations to support error correction and corrective feedback. *Computers & Education, 52*(4), 882–892. https://doi.org/10.1016/j.compedu.2008.12.014

Brünken, R., & Seufert, T. (2006). Aufmerksamkeit, Lernen, Lernstrategien [Attention, learning, learning strategies]. *Handbuch Lernstrategien* (pp. 27–37). Hogrefe.

Seufert, T. (2003). Supporting coherence formation in learning from multiple representations. *Learning and instruction, 13*(2), 227–237. https://doi.org/10.1016/S0959-4752(02)00022-1

Seufert, T. (2018). The interplay between self-regulation in learning and cognitive load. *Educational Research Review, 24*, 116–129. https://doi.org/10.1016/j.edurev.2018.03.004

Klepsch, M., Schmitz, F., & Seufert, T. (2017). Development and Validation of Two Instruments Measuring Intrinsic, Extraneous, and Germane Cognitive Load. *Frontiers in Psychology, 8*, 1997. https://doi.org/10.3389/fpsyg.2017.01997

Klepsch, M., & Seufert, T. (2020). Understanding instructional design effects by differentiated measurement of intrinsic, extraneous, and germane cognitive load. *Instr Sci, 48*, 45–77. https://doi.org/10.1007/s11251-020-09502-9

Ayres, P. (2020). Something old, something new from cognitive load theory. *Computers in Human Behavior, 113*, 106503. https://doi.org/10.1016/j.chb.2020.106503

Huang, C. L., Luo, Y. F., Yang, S. C., Lu, C. M., & Chen, A.-S. (2020). Influence of Students' Learning Style, Sense of Presence, and Cognitive Load on Learning Outcomes in an Immersive Virtual Reality Learning Environment. *Journal of Educational Computing Research, 58*(3), 596–615. https://doi.org/10.1177/0735633119867422

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

Ozcelik, E., Arslan-Ari, I., & Cagiltay, K. (2010). Why does signaling enhance multimedia learning? Evidence from eye movements. *Computers in Human Behavior, 26*(1), 110–117. https://doi.org/10.1016/j.chb.2009.09.001

Bowman, D., Kruijff, E., LaViola Jr, J. J., & Poupyrev, I. P. (2004). *3D User interfaces: theory and practice. CourseSmart eTextbook*. Addison-Wesley.

Jerald, J. (2015). *The VR book: Human-centered design for virtual reality*. Morgan & Claypool.

Sherman, W. R., & Craig, A. B. (2018). *Understanding virtual reality: Interface, application, and design.*. Morgan Kaufmann.

Kalyuga, S. (2011). Cognitive load theory: How many types of load does it really need? *Educational Psychology Review, 23*(1), 1–19. https://doi.org/10.1007/s10648-010-9150-7