Learning and Instruction 60 (2019) 225–236

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc

Adding immersive virtual reality to a science lab simulation causes more presence but less learning

Guido Makransky[a,*], Thomas S. Terkildsen[a], Richard E. Mayer[b]

[a] Department of Psychology, University of Copenhagen, Copenhagen, Denmark
[b] Psychological and Brain Sciences, University of California Santa Barbara, CA, USA

ARTICLE INFO

Keywords:
Virtual reality
EEG
Cognitive load
Simulation
Presence
Redundancy principle

ABSTRACT

Virtual reality (VR) is predicted to create a paradigm shift in education and training, but there is little empirical evidence of its educational value. The main objectives of this study were to determine the consequences of adding immersive VR to virtual learning simulations, and to investigate whether the principles of multimedia learning generalize to immersive VR. Furthermore, electroencephalogram (EEG) was used to obtain a direct measure of cognitive processing during learning. A sample of 52 university students participated in a 2×2 experimental cross-panel design wherein students learned from a science simulation via a desktop display (PC) or a head-mounted display (VR); and the simulations contained on-screen text or on-screen text with narration. Across both text versions, students reported being more present in the VR condition ($d = 1.30$); but they learned less ($d = 0.80$), and had significantly higher cognitive load based on the EEG measure ($d = 0.59$). In spite of its motivating properties (as reflected in presence ratings), learning science in VR may overload and distract the learner (as reflected in EEG measures of cognitive load), resulting in less opportunity to build learning outcomes (as reflected in poorer learning outcome test performance).

1. Introduction

1.1. Objective and rationale

Recently, there has been a surge in attention and hype surrounding immersive Virtual Reality (VR), and how it is predicted to create a paradigm shift in several fields including entertainment, gaming, and education (e.g., Belini et al., 2016; Blascovich & Bailenson, 2011; Greenlight & Roadtovr, 2016). This excitement is partly driven by high-volume business analyses, popular reports, and heavy investment by some of the biggest technology companies like Google, Apple, Facebook, Microsoft, and Samsung. As a consequence, many companies and educational institutions are investing significant resources in adapting standard educational tools that have traditionally been used on a desktop computer to immersive VR involving head-mounted displays, with the expectation that a higher level of immersion will increase student motivation and learning (Bodekaer, 2016). With little existing research evidence available to either support or contradict this assumption, instructional design decisions are often made based on practical or economic considerations rather than evidence-based arguments because there is limited research available in this rapidly developing field.

The main objective of this study is to assess the influence of the role of immersive technologies on learning outcomes (i.e. how media influences learning). In other words we explore how porting a learning simulation designed for a low-immersive environment to a highly-immersive environment influences subjective and objective learning outcomes. A secondary objective is to investigate whether the principles of multimedia learning (Mayer, 2009) generalize to immersive VR. These research questions are highly relevant because most large scale VR learning implementations are currently taking a technology-rather than a learner-centered approach which has historically lead to limited impact of technology in educational practice (Cuban, 1986). A final objective is to use cognitive neuroscience methodologies to obtain a direct measure of cognitive processing during learning. This is in line with a report by the National Research Council that highlights “the need to examine the mediating processes within the individual that influence science learning with simulations and games with the aim to illuminate what happens within the individual—both emotionally and cognitively—that leads to learning and what design features appear to activate these responses” (NRC, 2011 p. 122). Many instructional design studies investigate posttest results, or indirectly assess the cognitive processing during learning through self-report measures. In line with recent research that has used cognitive neuroscience to measure overload (e.g.,

* Corresponding author. University of Copenhagen, Oester Farimagsgade 2A 1353 Copenhagen K, Denmark.
E-mail address: gm@psy.ku.dk (G. Makransky).

https://doi.org/10.1016/j.learninstruc.2017.12.007
Received 23 May 2017; Received in revised form 17 November 2017; Accepted 21 December 2017
Available online 26 December 2017
0959-4752/ © 2017 Elsevier Ltd. All rights reserved.

Antonenko, Paas, Grabner, & Van Gog, 2010; Gerjets, Walter, Rosenstiel, Bogdan, & Zander, 2014; Mills et al., 2017) we investigate cognitive processing during learning directly with electro-encephalogram (EEG) to get a better understanding of how immersion affects the learning process in this study.

A distinction between low immersion (also referred to as desktop VR) and high immersion VR (generally involving a head-mounted display) is typically made in the literature (Lee & Wong, 2014; Limniou, Roberts, & Papadopoulos, 2007). In desktop VR, the virtual reality environment (VRE) is displayed on a conventional PC monitor with sound coming through speakers and the interaction is controlled through a regular computer mouse. This is the type of VR that is generally referenced in literature reviews on VR, and is regarded as a low-immersion medium (e.g., Merchant, Goetz, Cifuentes, Kenney-Kennicutt, & Davis, 2014; Moreno & Mayer, 2002; NRC, 2011). The second type of VR is often refered to as high-immersion VR, and is characterised by using a head-mounted-display in which a high graphical fidelity screen is mounted in front of one's eyes with separate lenses for each eye and with sound delivered through earphones. The interaction in this type of VR is controlled through head-motion tracking in conjunction with a computer system, so when users move their heads to look around they correspondingly move their field of view inside of the virtual 360-degree environment (Moreno & Mayer, 2002). The present study examines the effects of moving a science simulation from learning in a low-immersion VR (also referred to as PC condition in this study) environment to a high-immersion VR environment (also referred to as VR condition in this study).

1.2. Virtual learning simulations

The use of science labs has a long history in science education dating back decades, so it is reasonable that the advances in computer-based learning would include the development of computer-based simulations of science labs and learning experiences (Honey & Hilton, 2011; Klopfer, 2008; Slotta & Linn, 2009). Computer-based simulations for science learning can be used to promote procedural knowledge for carrying out lab procedures as well as conceptual knowledge for understanding and explaining the demonstration, but research on the instructional effectiveness of simulated science environments is needed (Honey & Hilton, 2011).

An important issue for the broader field of learning and instruction concerns whether the motivational benefits of simulated labs can be enhanced with virtual reality in a way that promotes learning. In particular, a field where the value of immersive VR may be specifically relevant is in designing virtual learning simulations. Virtual learning simulations are designed to replace or amplify real-world learning environments by allowing users to manipulate objects and parameters in a virtual environment. This has the advantage of allowing students to observe otherwise unobservable phenomena, reduce the time demand of experiments, and provide adaptive guidance in a virtual world that provides a high sense of physical, environmental, and social presence (De Jong, Linn, & Zacharia, 2013; De Jong, 2017; Makransky, Lilleholt, & Aaby, 2017). Some empirical studies and meta-analyses have shown that low-immersion simulations result in better cognitive outcomes and attitudes toward learning than more traditional teaching methods (e.g., Bayraktar, 2001; Bonde et al., 2014; Clark, Tanner-Smith, & Killingsworth, 2016; Merchant et al., 2014; Rutten, van Joolingen & van der Veen, 2012; Sitzmann, 2011; Vogel et al., 2006). There is also research supporting the motivational value of low-immersion VR simulations (e.g., Makransky, Thisgaard, & Gadegaard, 2016; Makransky, Bonde, et al., 2016; Thisgaard & Makransky, 2017).

There is less research investigating whether high-immersion VR technology increases cognitive and motivational outcomes as compared to low-immersion VR. One study by Moreno and Mayer (2002) investigated the role of method and media by introducing multimedia learning material based on different learning principles from the cognitive theory of multimedia learning (CTML; Mayer, 2014) in desktop VR, immersive VR while sitting, and immersive VR while walking. They found a method effect based on the redundancy principle of multimedia learning (Mayer, 2009), but the media did not affect performance on measures of retention, transfer, or program ratings. Similarly, Richards and Taylor (2015) compared the knowledge of students after a traditional classroom lecture about a biological principle known as Marginal Value Theorem with their knowledge after they were exposed to simulations of two- and three-dimensional models. They found that the two-dimensional model worked better than the three-dimensional model, presumably due to additional cognitive load imposed by the three-dimensional model. In contrast, other studies have found positive results favoring high-immersion VRE's (e.g., Alhalabi, 2016; Passig, Tzuriel, & Eshel-Kedmi, 2016; Webster, 2016). Therefore, there is limited and inconclusive research investigating whether the added immersion offered by high-immersion VREs leads to higher levels of presence, and ultimately better learning and transfer outcomes; and little is known about how different levels of immersion affect cognitive load and ultimately learning and transfer outcomes. This type of research is specifically relevant for highly realistic educational material such as virtual learning simulations.

1.3. Theoretical background

What is the theoretical basis for predicting that more highly immersive VREs would lead to better or worse learning outcomes? Similar to cognitive load theory (CLT, Sweller, Ayres, & Kalyuga, 2011), the CTML (Mayer, 2009) suggests that there are three types of cognitive processing that can occur **during** multimedia instruction: *extraneous processing*–cognitive processing that does not support the instructional goal, caused by poor instructional design or distractions during learning; *essential processing*–cognitive processing required to mentally represent the essential material, caused by the complexity of the material for the learner; and *generative processing*–cognitive processing aimed at making sense of the material, caused by the learner's motivation to exert effort. Given that processing capacity is limited, if a learner engages is excessive amounts of extraneous processing, there will not be sufficient capacity available for essential and generative processing (which cause meaningful learning outcomes). Thus one goal of instructional design is to reduce extraneous processing, because to the extent that the perceptual realism of high immersion causes extraneous processing, such environments will diminish learning. On the other hand another goal of instructional design is to foster generative processing, because to the extent that highly immersive environments motivate learners to process the material more deeply, they will increase learning.

From one perspective the theories suggest that immersive VREs could foster generative processing by providing a more realistic experience which would result in a higher sense of presence (Slater & Wilbur, 1997). This would cause the learner to put in more effort and to actively engage in cognitive processing in order to construct a coherent mental representation of the material and the experience, which would lead to learning outcomes that are better able to support problem-solving transfer. This expectation is consistent with interest theories of learning such as initially offered by Dewey (1913), who believed that students learn through practical experience in ecological situations and tasks by actively interacting with the environment. The expectation the increased immersion can lead to learning may be specifically relevant for VR because the sense of presence experienced by the user can have a very powerful emotional impact (Milk, 2015). Models by Salzman, Dede, Loftin, and Chen (1999) and Lee, Wong, and Fung (2010) also suggest that immersive environments create a stronger sense of presence, which leads to higher engagement and motivation and a deeper cognitive processing of educational material. Therefore, based on these motivational arguments, it would be expected that immersive VR would provide a higher level of presence and generative cognitive processing which should lead to higher levels of learning and transfer.

An alternative line of reasoning which is also based on the CTML and CLT suggests that any stimuli not absolutely necessary to understanding what needs to be learned is redundant and may decrease learning. These theories suggest that any material that is not related to the instructional goal should be eliminated in order to eliminate extraneous processing (Moreno & Mayer, 2002). Therefore, immersive VR could simply be triggering situational interest through the process of taking a boring topic and spicing it up in an attempt to make it interesting. This is just the first step in promoting academic achievement, and by itself may not foster deep learning. Situational interest can but does not always develop into later phases involving individual interest development which have been found to promote positive long term educational outcomes (Renninger & Hidi, 2016). Alternatively, added immersion could be categorized as a seductive detail (i.e., interesting but irrelevant material) which could distract students by priming the wrong schema (e.g., Harp & Mayer, 1997). Immersive environments that offer a high level of presence can interfere with reflection during learning (Norman, 1993), because these seductive details create extraneous processing that can distract the learner's process of building a cause-and-effect schema based on the material.

Van Der Heijden (2004) provides a complementary perspective on why highly immersive environments might not result in higher learning and transfer outcomes. This theory proposes that information systems can be perceived as either hedonic or utilitarian (Van Der Heijden, 2004). While utilitarian systems provide instrumental value (e.g., productivity and increased task or learning performance), hedonic systems provide self-fulfilling value (e.g., fun or pleasurable experiences; Van Der Heijden, 2004). The distinction between utilitarian and hedonic systems is not always clear (Van Der Heijden, 2004), which can lead students who use immersive VREs to treat them as hedonic systems. This could lead them to disregard the instrumental value and concentrate on the entertainment value of the system, resulting in them focusing their cognitive effort on irelevant material that is not part of the instructional goal of the lesson. Therefore, these theoretical perspectives suggest that the increased immersion in VREs would lead to higher levels of extaneous cognitive load and lower learning and transfer. They also suggest that directly and simply porting a simulation designed for desktop environments to a VRE could in itself either hinder or facilitate learning.

1.4. Does the level of immersion impact the redundancy principle?

A secondary issue related to the level of immersion in VREs is whether multimedia design principles apply in low and high immersion VR environments. That is, does media affect method, or do the learning principles that were originally developed for less immersive multimedia environments generalize to highly immerse interactive environments like VR? Investigating whether media affects method is important because there is limited research examining learning principles within simulations and VR, and thus few evidence based guidelines for deloping learning content in highly immersive environments. The redundancy principle has previously been investigated in VR contexts (Moreno & Mayer, 2007). The redundancy principle is that people learn better from graphics or illustrations and narration than from graphics, narration, and redundant on-screen text (Mayer, 2009). Adding a redundant form of the verbal material can create extraneous overload, which has a negative effect on learning. This effect occurs when identical information is presented to learners in two or more different forms or media simultaneously; or when redundant material in general is presented and selected for processing (Kalyuga & Sweller, 2014). For instance, having identical and concurrent written and spoken text (through narration) was demonstrated to be redundant, and was shown to interfere with learning (Kalyuga, Chandler, & Sweller, 1999). Even if the identical information is presented concurrently across both modalities, it will still cause a redundancy effect because it then requires unnecessary co-referencing between the two channels; and if it's a question of identical on-screen text and narrated text, then they are both processed in the phonological channel even though they are presented across different sensory modalities.

The redundancy principle is consistent with the interference theory, which dates back to Dewey's warning (1913) against regarding extra embellishments that can be added to an otherwise boring lesson to try to motivate the students as increasing their interest level, since this extra material will need to be processed, which will in turn interfere with essential processing. This is in contrast to general arousal theory which advocates adding entertaining additions to make learning more interesting and enjoyable, resulting in higher levels of attention (Mayer, Heiser, & Lonn, 2001; Moreno & Mayer, 2002). Moreno and Mayer (2000) found that adding extraneous music or sounds to a desktop VR system hurts students' understanding of a multimedia explanation. They argue that this shows that adding what they refer to as “bells and whistles” can hurt the sense-making process in the same way as redundant on-screen text can (Moreno & Mayer, 2000). Based on this, it is possible that the redundancy principle from the CTML applies differently across different media, such that the redundancy principle might apply differently with immersive vs. low-immersive VR.

1.5. Main research questions and predictions

In the current study we investigate two main research questions. The first investigates whether a higher level of immersion in the VR learning simulations leads to higher levels of student learning, self-report ratings, and brain-based measures of overload. If increased immersion serves to increase extraneous processing, we predict that it will lead to less learning as measured by tests of learning outcome and more overload as measured by EEG. If increased immersion serves to foster generative processing, we predict that it will lead to more learning as measured by tests of learning outcome and appropriate levels of brain activity as measured by EEG.

The second research question is to determine if the redundancy principle of multimedia learning is present in low and high-immersion VREs. Here we investigate the consequences of adding narration to a science lab simulation that presents words as printed text, particularly on the same outcomes of self-report ratings, student learning, and brain-based measures of overload. If the redundancy principle applies, we predict that students will learn better when the simulation includes text only rather than text and concurrent narration in the PC and VR conditions.

2. Method

2.1. Participants and design

The participants were 52 (22 males and 30 females) students from a large European university with ages ranging from 19 to 42 (M = 23.8 years, SD = 4.5). The experiment employed a 2 × 2 mixed design, in which participants learned from two simulation lessons. The first factor was a between subjects factor, in which 28 participants were randomly assigned to receive two versions of a simulation lesson that had on-screen text (T condition); and 24 participants were randomly assigned to receive two versions of a simulation that had both text and corresponding narration (T + N condition). The second factor was a within subjects factor wherein the students were administered the head-mounted display VR version of the simulation (immersive VR condition) followed by the desktop VR version of the simulation (PC condition), or vice versa. The order of the two versions of the simulation was counterbalanced, with half the participants in each group receiving the immersive VR condition first, and half receiving the PC condition first.

2.2. Materials

The materials used in the study included four different versions of a virtual laboratory simulation, participant questionnaire, knowledge test, transfer test, and self-report survey designed to measure presence, learning beliefs, and satisfaction. All versions of the simulation were in English, but the surveys and posttests were in Danish.

2.2.1. Virtual lab simulation

The virtual simulation used in this experiment was on the topic of mammalian transient protein expression and was developed by the simulation development company, Labster. It was designed to facilitate learning within the field of biology at a university level by allowing the user to virtually work through the procedures in a lab by using and interacting with the relevant lab equipment and by teaching the essential content through an inquiry-based learning approach (Bonde et al., 2014; Makransky, Bonde et al., 2016). The main learning goal for the simulation is to develop an understanding of mammalian transient protein expression. In the simulation the student experiences using techniques such as cell culturing, cell transfection, and protein expression.

Labster supplied four versions of this simulation with identical instructional design and method for each version: PC with text, PC with text and narration, immersive VR with text, and immersive VR with text and narration. The PC versions were displayed on a desktop computer screen as shown in the top of Fig. 1, whereas the immersive VR versions were displayed using a head-mounted display that allow the students to move their heads and see around the virtual laboratory environment as shown in the bottom of Fig. 1. The text versions presented words as onscreen printed text, whereas the text and narration versions presented words as onscreen printed text and simultaneous narration using a voice to read the text aloud.

In every version of the simulation the virtual lesson starts off with the learner being presented with a brief introduction to their primary in-game tool "The lab pad". The lab pad is a tablet that is used to provide written information and illustrations and is also the display medium for the multiple-choice questions that the learner is required to answer correctly in order to progress (see panels A in Fig. 2). After this brief tutorial, the learner is introduced to the virtual agent Marie. Marie serves as an AI instructor, who guides the learner through the essential material, such as lab procedures and lab equipment. She also functions as the source of both the verbal narration and the on-screen text, depending on which version of the simulation is being presented (see panel A in Fig. 3). Generally, the simulation consists of four different kinds of tasks: (1) receiving information (see panel A in Fig. 3 for an example), (2) answering multiple-choice questions (see panels A in Fig. 2 for an example), (3) getting feedback, and (4) doing interactive lab procedures such as mixing specific compounds with a serological pipette and discarding the used pipette tip after use (see panels B in Figs. 2 and 3).

2.2.2. Tests

Two multiple-choice tests were developed for evaluating the participants' learning outcomes – a knowledge test and a transfer test. A

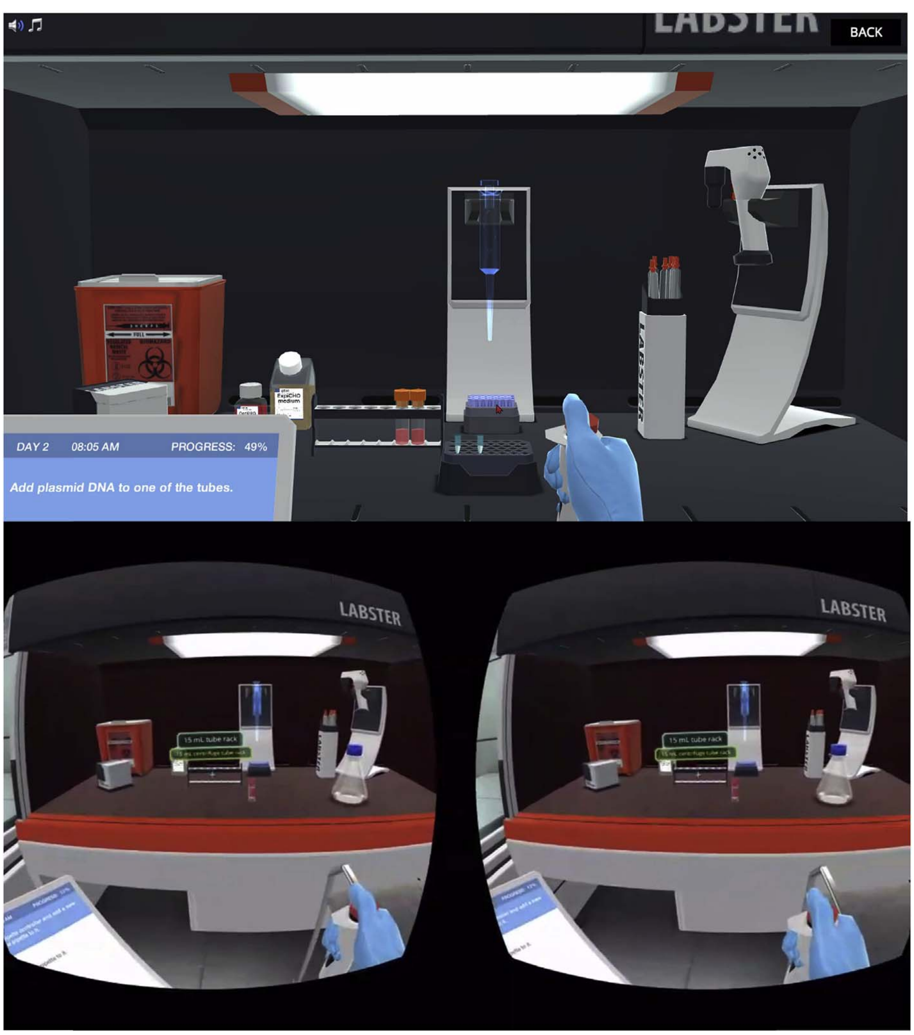

Fig. 1. Screenshot of the simulation used in this study: The top picture is a screenshot from the PC version and the one below it is a screenshot from the VR, which shows the stereoscopic display technology used in this version. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

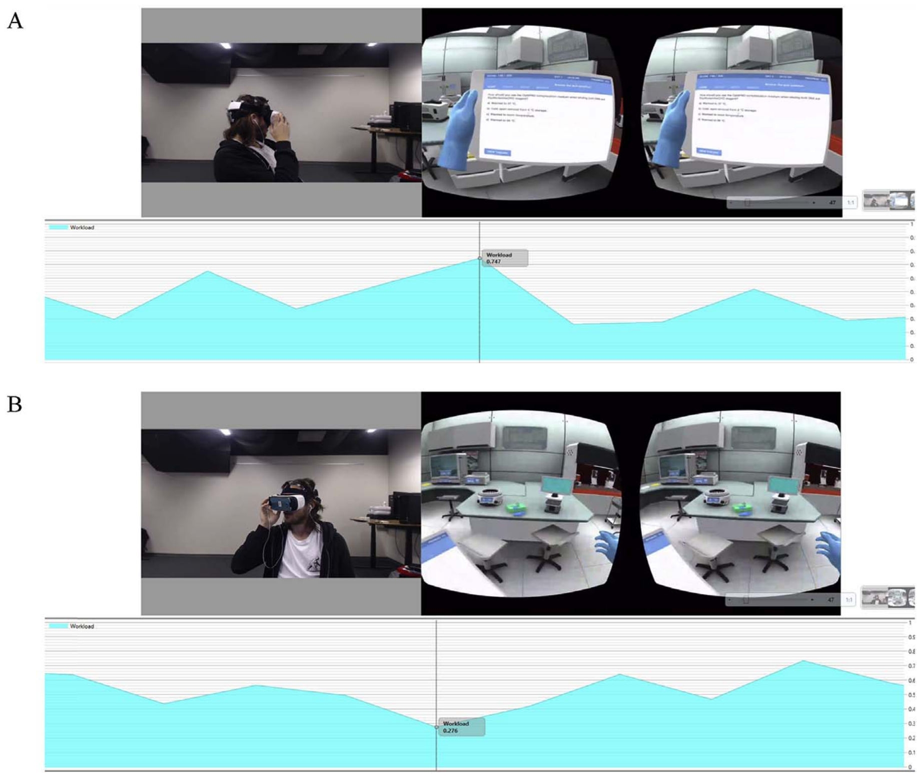

Fig. 2. Screen shots of the VR condition where the iMotions system simultaneously shows the student working through the simulation on the top left panel; the stimulus they are experiencing in the top right panel; and the continuous EEG workload measure in the bottom panel of each screen shot. Screen shot A shows a student answering a multiple-choice question; screen shot B shows a student doing interactive lab procedures in the VR condition. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 3. Screen shots of the PC condition where the iMotions system simultaneously shows the student working through the simulation on the top left panel; the stimulus they are experiencing in the top right panel; and the continuous EEG workload measure in the bottom panel of each screen shot. Screen shot A shows a student getting information from the virtual agent Marie; and screen shot B shows a student doing interactive lab procedures in the PC condition. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

group of subject matter experts, including two scientists who had developed the virtual simulation from Labster, two psychologists, and a psychometrician, developed these questions. The knowledge test consisted of 10 multiple-choice questions designed to assess conceptual and procedural knowledge of essential material presented in the simulation (e.g., How should you use your OptiPro medium for complex formation, when both DNA and ExpifacterminCHO reagent is diluted? A) Heated to room temperature; B) Heated to 56 °C; C) Heated to 37 °C; D) Cold, taken from storage at 4 °C). The transfer test consisted of 10 multiple-choice questions designed to assess the participants' ability to apply what they had learned to new situations (e.g., A delivery company is delivering frozen cells to you, but you have a meeting with your boss at the time of delivery. What is your best chance to ensure the cell's survival? A) Ask your boss to wait 20 min. Thaw the cells and put them in liquid nitrogen; B) Ask the delivery company to leave the cells at room temperature. This is the best temperature for thawing frozen cells, and they can be stored later; C) Ask the delivery company to put them in a water bath at 37 °C that you've prepared. The cells can survive until you are back; D) Ask the delivery company to put them in a water bath at 56 °C that you've prepared. This is the optimal temperature for thawing frozen cells). The questions required that students had a deep knowledge of the content and that they could apply that knowledge to a realistic context. Students received one point for each correct answer and 0 points for selecting an incorrect answer. The posttests were delivered on a computer.

2.2.3. Participant questionnaire

Information on participants' age, gender, and major were collected through iMotions software along with the other measures used in the study (iMotions, 2016).

2.2.4. Survey

The self-report survey asked participants to rate their level of presence, learning beliefs, and satisfaction. These constructs have previously been used as dependent variables in VR research (e.g., Moreno & Mayer, 2007). Presence was measured with 10 items adapted from Schubert, Friedmann, and Regenbrecht (2001; e.g., "The virtual world seemed real to me"). **L**earning beliefs was measured with eight items adapted from Lee et al. (2010; e.g., "I gained a good understanding of the basic concepts of the materials"). Satisfaction was measured with seven items adapted from Lee et al. (2010; e.g., "I was satisfied with this type of virtual reality/computer-based learning experience"). All of these used a five-point Likert scale ranging from (1) strongly disagree to (5) strongly agree.

2.2.5. Apparatus

The PC condition version of the simulation was administered on a high-end laptop computer and presented to the participants on an external 23-inch computer monitor. A standard wireless mouse was used by the participants to control input in the PC condition. The participants used this mouse to both navigate from the different static points of view and to select answers to multiple-choice questions. In general, the mouse functioned as a way to select which object the participant wanted to interact with through cursor movement and left-clicks.

In the immersive VR condition the simulation was administered using a Samsung Galaxy S6 phone, and stereoscopically displayed through a Samsung GearVR head-mounted display (HMD). This condition requires the participants to use the touch pad on the right side of the HMD to emulate the left-click function of a wireless mouse in order to select which objects to interact with. In this condition, however, head movement is used to move the participant's field of view and the centered dot-cursor around the dynamic 360-degree VRE. All versions included a visible pedagogical agent, named Marie, who did not speak in the T version and who narrated the text in the T + N version.

2.2.6. Measurement of cognitive load with EEG

An electroencephalogram (EEG) was chosen to assess students' workload brain activity while using the different versions of the simulation. There is some evidence from previous studies to suggest that EEG has potential as a valid and objective measure of mental workload (e.g., Gerjets et al., 2014; Sterman & Mann, 1995). In the present study, EEG data was collected using an Advanced Brain Monitoring (ABM) X-10, wireless 9-channel EEG system running at 256hz. The X-10 records data in real-time from nine sensors that are positioned in accordance with the International 10–20 system (as shown in Fig. 4), along with two reference signal sensors that are attached to the mastoid bone behind each ear (ABM, 2016).

The 256 EEG signals per second were processed and decontaminated for excessive muscular activity, fast and slow eye blinks, and excursions due to movement artifacts by ABM's proprietary software in order to produce classifications of cognitive load in epochs of one second (Berka et al., 2004). The workload classifier was developed by Berka et al. (2007) using a linear DFA with two classes, low and high mental workload. Absolute and relative power spectra variables were derived using stepwise regression from channels C3-C4, Cz-PO, F3-Cz, Fz-C3, and Fz-PO. The workload metric computation is based on 30

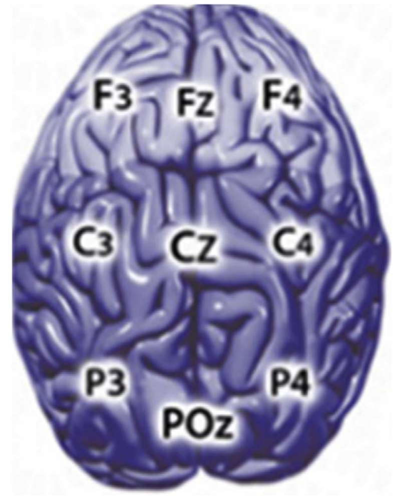

Fig. 4. EEG sensor locations (ABM, 2016). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

distinct variables across all frequency bands within 1–40 Hz (an overview of the variables used for the calculation of workload can be found in Berka et al., 2007, Table 1, p.235). The classifier was evaluated and trained based on data obtained from testing different combinations of low and high difficulty levels of mental arithmetic, grid location, trail making, and digit-span tasks (forward and backwards; Berka et al., 2007). These tasks are often used in standardized batteries for neuropsychological assessment of working memory (such as the Working Memory Index portion of WISC-IV, which includes forward and backwards digit-span, trail making and mental arithmetic; Colliflower, 2013) and as such, the workload metric is developed specifically to be sensitive to executive processes involving working memory. As a result, the metric value increases when working memory load and task demands increases, and decreases when resource demand lessens (Berka et al., 2004). In other words, the workload metric is a continuous measure of resource allocation and cognitive activity in response to task demands.

The workload metric ranges numerically from 0 to 1, with larger values representing increased workload; and it is divided into three different range classifications: boredom (up to 0.4), optimal workload (0.4–0.7) and stress and information overload (above 0.7; iMotions, 2016). This metric and the methods behind it have been validated by several empirical studies across various fields (military, industrial and educational research; Stevens et al., 2011). It has been shown to significantly correlate with both subjective self-reports of cognitive load and objective performance on tasks with varying levels of difficulty and cognitive demand such as the ones mentioned above (Berka et al., 2007; Galán & Beal, 2012; Sciarini, Grubb, & Fatolitis, 2014). The definition of ABM's workload metric is consistent with how cognitive load is described in the cognitive theory of multimedia learning (Mayer, 2009) and cognitive load theory (Sweller et al., 2011). By having identical learning material (i.e., identical demands for essential processing) in both versions of the simulation, the cognitive load metric is intended to examine the difference in extraneous and generative cognitive processing during learning between the two platforms.

A requirement for this mental state metric to be valid and accurate across different participants is to run an impedance test to ensure that the recordings are within the recommended impedance tolerances, and to provide a 9-min individualized baseline benchmark profile for each participant based on three distinct cognitive assessment tasks: (1) 3-choice Vigilance Task, (2) Visual Psychomotor Vigilance Task, and (3) Auditory Psychomotor Vigilance Task (for further documentation see BIOPAC, 2016). For the EEG measures, average workload was calculated for each respondent within each media condition by taking the

Table 1
Chi-squared fit statistics to the Rasch model and Cronbach's Alpha Reliability Coefficients for the Scales in the Study.

Scale	Number of items	Rasch chi-squared fit statistic		Cronbach's alpha	
		After 1st intervention	After 2nd intervention	After 1st intervention	After 2nd intervention
Learning beliefs	8	.81	.85	.84	.87
Satisfaction	7	.07	.39	.77	.91
Presence	10	.27	.34	.72	.85
Knowledge	8	.28	.13	.68	.68
Transfer	9	.52	.35	.36	.55

average level of workload while using the simulation, and overload was the percentage of time the respondent was over the threshold value of 0.7 on the workload measure.

Data from the surveys and the EEG data were collected using the iMotions research software platform, which permits synchronization of the brain-based EEG measures and allows for accessible data analysis of these measures (see iMotions.com for further information regarding the platform). The data was exported to IBM SPSS version 23.0 for statistical analyses.

2.3. Procedure

Participants were tested individually in a VR learning lab at a European university. The lab is sound-proofed and the lighting is stable and controlled since there are no windows. The participants were randomly assigned to either the T or the T + N simulation condition. Additionally, participants were randomly assigned to receive the VR version first followed by the PC version or vice versa. The first step in the study design was preparation, which is shown in Fig. 5. Participants were fitted with the EEG sensors and subsequently data quality tests were run, such as the EEG impedance test, to ensure that the equipment was functioning properly. Then the experimenter gave oral instructions on how to complete the following EEG benchmark. The experimenter left the room each time after instructions were provided, so the participant was alone in the room when the experimental tasks were performed. The next step was to complete the participant questionnaire which included the knowledge and transfer tests. These served as pretests to determine whether participants knew any of the answers before being exposed to the simulations. This information was subsequently used as a covariate in the analyses. Next, the participant received the first simulation (based on the randomly assigned condition) for 15 min, and then retook the knowledge and transfer tests, and completed the self-report survey. Next, the participant received the second simulation (based on the randomly assigned condition) for 15 min, and then retook the knowledge test, transfer tests, and self-report survey. Instructions for each component of the experiment were given when relevant in order not to overload the participant with extraneous information. In order to ensure equal time on task, participants had 15 min with each of the two versions of the virtual lab simulation, and they dynamically interacted with the simulations at their own pace. There was no time limit for the pre- and post-questionnaires and learning outcome tests. The average run time for each participant was about an hour and a half. Each participant was compensated for their time with a gift card valued at 100 Danish crowns (about 13 Euros) upon completion. We followed standards for ethical treatment of human subjects and obtained IRB approval for the study.

A cross-panel study design was selected because a preliminary pilot study showed that students were very enthusiastic about the use of all versions of the virtual lab simulation, and that it was not until the students had used both versions (PC and VR) that they could accurately compare their experiences. The cross-panel design provided a true experiment after the first intervention, but extra information about the comparison across media after using both the PC and VR version of the experiment.

3. Results

3.1. Are the instruments valid and reliable?

The first analyses evaluated the validity of the outcome variables used in the study by testing the fit of the data to the Rasch model (Rasch, 1960). Results indicated that two items in the knowledge test, one item in the transfer test, and three items in the presence scale had positive fit residuals above the critical value of ± 2.5 which is an indication that the items do not measure the intended construct appropriately (Pallant & Tennant, 2007; Makransky et al., 2017). Therefore, these items were eliminated from the total and gain scores reported and analyzed in this paper. The chi-squared fit statistics reported in Table 1 indicate that the remaining scales fit the Rasch model (values over 0.05 indicate acceptable fit; Pallant & Tennant, 2007). Table 1 also reports the reliability of the scales used in the study based on Cronbach's alpha. The reliability coefficients for the self-report scales were acceptable with values of 0.72 and 0.85 for presence; 0.84 and 0.87 for learning beliefs; 0.77 and 0.91 for satisfaction (see top of Table 1). The reliability coefficients were 0.68 and 0.68 for the knowledge test; and 0.32 and 0.55 for the transfer test following the first and second interventions respectively (see bottom of Table 1). Although the transfer test had low internal consistency reliability, this could be expected because the

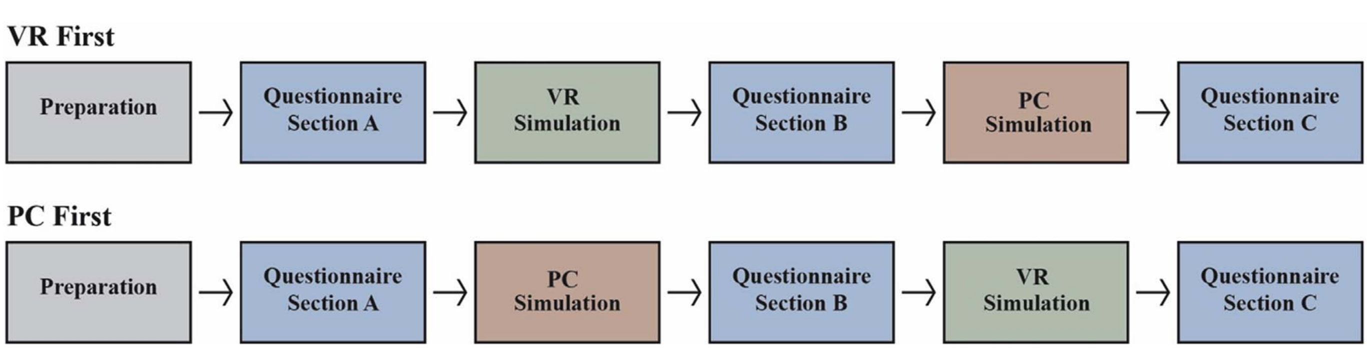

Fig. 5. An overview of the overall counterbalanced design. Half of the participants used the simulation with text and narration (redundancy condition), and the other half used the simulation with screen text alone. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Means and standard deviations for the VR and PC conditions on eight measures.

	Outcome	1st intervention		2nd intervention	
		VR M (SD)	PC M (SD)	VR M (SD)	PC M (SD)
Test	Knowledge gain	**1.81 (2.12)**	**2.92 (2.53)**	**1.54 (1.39)**	**2.69 (1.49)**
	Transfer gain	0.96 (1.18)	1.46 (1.70)	0.38 (1.17)	0.58 (1.06)
Survey	Presence	**3.50 (0.46)**	**2.77 (0.50**	**3.72 (0.49)**	**2.52 (0.61)**
	Learning beliefs	3.68 (0.66)	3.59 (0.70)	3.96 (0.70)	3.82 (0.64)
	Satisfaction	3.96 (0.44)	3.74 (0.63)	3.95 (0.75)	4.11 (0.61)
EEG	Work load	0.63 (0.12)	0.63 (0.11)	**0.67 (0.10)**	**0.60 (0.13)**
	Overload time	48.75 (21.36)	48.80 (19.81)	**55.21 (20.53)**	**41.44 (24.13)**

Note. Bold font indicates significant differences at $p < .05$.

items were designed to measure a very broad domain with different content, namely assessing whether students were able to apply their knowledge to novel and different problems. The average score on the knowledge pre-test was 2.15 out of 8 (SD = 1.35), and the transfer test was 3.88 out of 9 (SD = 1.64) across the groups, indicating that the students did not have a high level of prior-knowledge of the material before using the simulations.

3.2. Media effects

The main objective of this study is to determine the consequences of adding immersive virtual reality to a science lab simulation, particularly on student learning, self-report ratings, and brain-based measures of overload.

3.2.1. Do students learn better with immersive VR or conventional media?

The primary issue addressed in this paper concerns whether students learn better with immersive VR (VR group) or with conventional media (PC group). The top two lines of Table 2 show the mean gain score (and standard deviation) on the knowledge test and transfer test for the VR group and the PC group. ANCOVAs were conducted with the pre-test score as a covariate, media (VR vs. PC) and method (text versus text + narration) as independent variables, and gain scores on knowledge and transfer (i.e., difference between pre-test score and post-test) as the dependent variables for the first and second intervention, respectively. The PC group gained significantly more knowledge than the VR group, both for the first intervention, $F(1, 47) = 4.45$, $p = .040$, $d = 0.48$, and the second intervention, $F(1, 47) = 8.45$, $p = .006$, $d = 0.80$. The advantage of the PC group over the VR group on the transfer test gain did not reach statistical significance for the first intervention, $F(1, 47) = 0.89$, $p = .350$, or the second intervention, $F(1, 47) = 0.43$, $p = .513$. There were no significant interactions with method for any of the ANCOVAs. We conclude that students learned more when the material was presented via a PC than via immersive VR. This is a major empirical contribution of this study.

3.2.2. Do students give more positive self-report ratings to immersive VR or conventional media?

Another important issue addressed in this study concerns whether students produce more positive self-report ratings when they learn with immersive VR (VR group) or with conventional media (PC group). The next four lines of Table 2 show the mean and standard deviation on the ratings of presence, learning beliefs, and satisfaction for the VR group and the PC group. ANOVAs were conducted with media (VR vs. PC) and method (text vs. text + narration) as independent variables, and each of the three rating scales as the dependent variables for the first and second intervention, respectively. The VR group produced significantly higher ratings of presence than the PC group, both for the first intervention, $F(1, 48) = 28.67$, $p < .001$, $d = 1.30$, and the second intervention, $F(1, 48) = 59.37$, $p < .001$, $d = 2.20$, indicating that the immersive VR medium was highly successful in creating a sense of presence for learners. The advantage of the VR group over the PC group failed to reach statistical significance on the rating of learning beliefs for the first intervention, $F(1, 48) = 0.24$, $p = .618$, and the second intervention, $F(1, 48) = 0.54$, $p = .467$; or on the rating of satisfaction for the first intervention, $F(1, 48) = 1.94$, $p = .170$, and the second intervention, $F(1, 48) = 0.60$, $p = .443$. There were no significant interactions with method for any of the ANOVAs. We conclude that students reported greater sense of presence when the material was presented via immersive VR than via a desktop computer, thus validating the power of immersive VR to create a sense of presence in learners. This is another major empirical contribution of this study.

3.2.3. Do students show greater workload brain activity with immersive VR or conventional media?

In addition to behavioral measures of learning outcome and self-ratings, we included an EEG-based measure of workload in order to determine whether the VR environment created greater workload than the PC environment. The second-to-last line of Table 2 shows the mean workload score and standard deviation (with higher scores showing higher workload) for the VR and PC groups based on EEG data recorded during learning. ANOVAs were conducted with media (VR vs. PC) and method (text vs. text + narration) as independent variables, and mean workload as the dependent variable, for the first and second intervention, respectively. There was no significant difference between the groups on the first intervention, $F(1, 48) = 0.001$, $p = .978$, but the VR group scored higher on average workload than the PC group on the second intervention, $F(1, 48) = 5.0$, $p = .030$, $d = 0.59$.

The final line of Table 2 shows the average proportion of time the participants in each group scored above the overload level of 0.7 (which indicates cognitive overload) for the first and second interventions, respectively. Students were overloaded an average of 48.78% of the time, indicating that the science lab simulation was a difficult learning task for most students. There was no significant difference between the groups on the first intervention, $F(1, 48) = 0.007$, $p = .933$, but the VR group was overloaded significantly more than the PC group on the second intervention, $F(1, 48) = 5.51$, $p = .028$, $d = 0.62$. There were no significant interactions with method for any of the ANOVAs. We conclude that students were more overloaded during learning later in the session when they were learning in immersive VR than when they were learning with a desktop computer. This is a preliminary piece of brain-based evidence suggesting that VR environments may be overstimulating.

3.3. Method effects

A secondary objective of this study is to determine the consequences of adding narration to both media versions of the science lab simulation that presents words as printed text, particularly on student learning, self-report ratings, and brain-based measures of cognitive overload.

3.3.1. Do students learn better when words are presented as text and narration or as text alone?

The top two lines of Table 3 show the mean gain scores (and standard deviations) on the knowledge test and transfer test for the T and the T + N groups. ANCOVAs were conducted with the pre-test score as a covariate, media (VR vs. PC) and method (text vs. text + narration) as independent variables, and gains on knowledge and gains on transfer (difference between pretest score and posttest score) as the dependent variables for the first and second interventions, respectively. There was no significant difference between the text and the text + narration groups on the amount of knowledge gained (as measured by the knowledge test) for the first intervention, $F(1, 47) = 0.14$, $p = .706$, and the second intervention, $F(1, 47) = 3.70$, $p = .060$, $d = 0.51$. There

Table 3
Means and standard deviations for text and Text + Narration conditions on eight measures.

Source	Outcome	1st intervention		2nd intervention	
		T M (SD)	T + N M (SD)	T M (SD)	T + N M (SD)
Test	Knowledge gain	2.29 (2.79)	2.46 (1.84)	2.46 (1.71)	1.71 (1.23)
	Transfer gain	1.25 (1.51)	1.17 (1.46)	0.57 (1.26)	0.38 (0.92)
Survey	Presence	3.15 (0.70)	3.11 (0.52)	3.08 (0.79)	3.18 (0.86)
	Learning beliefs	3.67 (0.64)	3.59 (0.72)	3.88 (0.67)	3.89 (0.67)
	Satisfaction	3.82 (0.60)	3.88 (0.49)	4.02 (0.85)	4.03 (0.44)
EEG	Work load	**0.66 (0.10)**	**0.59 (0.13)**	0.66 (0.10)	0.60 (0.13)
	Overload time	53.29 (20.0)	43.51 (20.0)	53.71 (22.6)	42.04 (22.8)

Note. Bold font indicates significant differences at $p < .05$.

was also no significant difference between the text and text + narration groups for the gain on the transfer test for the first intervention, $F(1, 47) = 0.00$, $p = 1.00$, or the second intervention, $F(1, 47) = 0.34$, $p = .562$. There was no significant interaction with media for any of the ANCOVAs. We conclude that there was no redundancy effect and that students learned equally well when the material was presented with text as when it was presented with text and concurrent narration.

3.3.2. Do students give more positive self-report ratings when words are presented as text and narration or as text alone?

Another important issue addressed in this study concerns whether students produce more positive self-report ratings when the material is presented as text and narration or as text alone. The next four lines of Table 3 show the mean and standard deviation on the ratings of presence, learning beliefs, and satisfaction for the text and the text + narration groups. ANOVAs were conducted with media (VR vs. PC) and method (text vs. text + narration) as independent variables, and each of the three rating scales as the dependent variables for the first and second interventions, respectively. There were no significant differences between the two groups for any of the self-report measures used in the study. That is, the two groups did not differ significantly on their ratings of presence for the first intervention, $F(1, 48) = 0.09$, $p = .772$, or the second intervention, $F(1, 48) = 0.41$, $p = .524$; on the ratings of learning beliefs for the first intervention, $F(1, 48) = 0.17$, $p = .678$, and the second intervention, $F(1, 48) = 0.001$, $p = .972$; or on the rating of satisfaction for the first intervention, $F(1, 48) = 0.18$, $p = .671$, and the second intervention, $F(1, 48) = 0.00$, $p = .982$. There was no significant interaction with media for any of the ANOVAs. We conclude that there is no evidence of a redundancy effect involving students' self-report ratings on any of the scales in the study.

3.3.3. Do students show greater workload brain activity when words are presented as text and narration or as text alone?

The second to last line of Table 3 shows the mean and the standard deviation of the EEG-based measure of workload (with higher scores showing higher workload) for the T and T + N groups. ANOVAs were conducted with media (VR vs. PC) and method (text only vs. text + narration) as independent variables, and average workload as the dependent variable for the first and second intervention, respectively. The T group scored significantly higher than the T + N group on the first intervention, $F(1, 48) = 4.99$, $p = .030$, $d = 0.61$, but the difference did not reach statistical significance on the second intervention, $F(1, 48) = 3.27$, $p = .077$. The final line shows the average proportion of time the participants in each group scored above the overload level of 0.7 for the first and second interventions, respectively. There was no significant difference between the groups on the first intervention, $F(1, 48) = 3.03$, $p = .088$, or the second intervention, $F(1, 48) = 3.63$, $p = .063$. There was no significant interaction with media for any of the ANOVAs. We conclude that students were more overloaded during learning in the first intervention when they were in the text condition compared to the text and narration condition, and the difference was not significant in the second intervention.

Overall, across all the dependent measures there is not strong and consistent evidence that the T and T + N groups differed.

4. Discussion

4.1. Empirical contributions

Many companies and public institutions are deciding to adapt educational and training material to immersive VR even though there is a lack of theoretical or scientific evidence to guide this decision and adaption. The major empirical contribution of this study is the finding that students felt a greater sense of presence when they used the high-immersion VR science lab simulation involving a head-mounted display, but they actually learned less as compared to the low-immersion version of the simulation on a desktop computer. This finding is consistent with previous studies by Moreno and Mayer (2002) and Richards and Taylor (2015) who also found lower levels of learning with more immersive technology. However, the results differ from newer research that has found that high-immersion VRE's lead to more learning (e.g., Alhalabi, 2016; Passig et al., 2016; Webster, 2016).

A second empirical finding in this paper was that the addition of narration to the simulation that presents words as printed text did not significantly affect student learning or self-report ratings. There was a significant difference in cognitive load after the first intervention which showed that the text-only group was more overloaded than the group that had text and narration. This result is contradictory to the redundancy principle which states that people learn more deeply from graphics and narration than from graphics, narration, and on-screen text (Mayer, 2009). The explanation for the redundancy principle is that the added text competes for visual processing capacity with the graphics and the learner wastes precious processing capacity trying to reconcile the two verbal streams of information. It should be noted that the comparison between text versus text and narration used in this study is not the way that the redundancy effect has been tested in most previous research, which used a comparison between narration versus narration and text (Mayer & Fiorella, 2014).

Observations of the students in this study showed that rather than reading and listening to the same text, some students (specifically in the immersive-VR condition) simply listened to the narration without reading the text, while others did both. Listening to text rather than reading it is classified as the modality principle, which has been found to increase learning and transfer by decreasing cognitive load (Moreno & Mayer, 1999). Therefore, the lack of significant results related to the method effect of the redundancy principle, and the unexpected result related to cognitive load in this study, could be the consequence of a combination of the redundancy principle and modality principle which concurrently occurred based on the student's specific behavior.

4.2. Theoretical implications

Our predictions based on CLT and CTML were that a more immersive VR environment could increase learning by increasing generative processing because students are more present in this environment; but that it could also limit learning due to added extraneous load to the extent that added perceptual realism is distracting and not relevant to the instructional objective. The results of the study could be an indication that the effect of added immersion in the VRE was stronger in terms of increasing extraneous load, and that the added immersion acted as a kind of seductive detail, or what is referred to as "bells and whistles" by Moreno and Mayer (2002). This finding supports previous research and theory which proposes that added immersion can interfere with reflection as the entertainment value of the environment does not give the learner ample time to cognitively assimilate new information to existing schemas.

Similarly, from Van der Heijden's (2004) perspective the results could suggest that students viewed the high-immersion VR simulation as hedonic, which could cause them to focus on enjoying the environment rather than focusing on learning the material. It is possible that some students were overwhelmed by the excitement and fun of being in immersive VR for the very first time, as the technology used in the study is very new. The novelty of the VR technology and its control scheme and interface could have impeded the participants' learning processes through an overall increase in extraneous workload as they would lack the familiarity and the automaticity that comes with practice and experience in comparison to the more commonly used desktop environment.

An overarching perspective that combines both the affective and cognitive aspects of multimedia learning is needed in order to obtain a better understanding of how to build instructional matarial for immersive VR, which uses a seeks to use a high level of presence to increase learning. Consistent with advances in motivational theory (Renninger & Hidi, 2016; Wentzel & Miele, 2016) the present study examines the role of affect in science learning by building on the cognitive affective model of learning with media (Moreno & Mayer, 2007) and the model of emotional design in game-like environments (Plass & Kaplan, 2015). Understanding how to harness the affective appeal of virtual environments is a fundamental issue for learning and instruction because research shows that initial situational interest can be a first step in promoting learning (Renninger & Hidi, 2016) and the learner's emotional reaction to instruction can have a substantial impact on academic achievement (Pekrun, 2016).

When asked about her experience after the experiment one student said: “The first simulation on the computer was boring, but then when I was in the lab it was fun.” The reaction is an example of how realistic immersive VR can feel, inasmuch as she had experienced the immersive VRE as real in comparison with the PC version. The sense of presence that immersive VR provides can be powerful if the physical and psychological fidelity of the experience can be channeled into proper cognitive processing to promote learning. In short, current cognitive theories of learning need to be expanded to include the role of affective and motivational factors, including a better understanding of the link between affective factors (include a feeling of presence) and appropriate cognitive processing during learning. This work has implications for the broad field of learning and instruction because it helps expand cognitive theories of learning and instruction to make them more applicable to highly immersive environments.

4.3. Practical implications

The results of this study and others in this developing field suggest that it is not appropriate to take a technology-centered approach and expect that the adaptation of learning material to immersive VR will automatically lead to better learning outcomes. If the goal is to promote learning (rather than simply to promote a sense of presence), it appears that science lab simulations need not be converted from a desktop-computer medium to an immersive VR medium. Just because an exciting, cutting-edge technology is available does not necessarily mean it should be used in all education and training situations without taking into consideration and utilizing the unique affordances that comes with this new technology. Conversely, it is too early to write off immersive VR as it still has the potential to be a viable educational platform if instructional designers take a learner-centered approach which focuses on how the technology fosters knowledge acquisition (Mayer, 2009; Moreno & Mayer, 2002) in an attempt to find the boundary conditions under which added presence is imperative to learning and transfer.

4.4. Methodological implications

A methodological contribution of this paper was the use of EEG to obtain a direct measure of cognitive processing during learning, and thus extend the domain of the emerging field of educational neuroscience (Mayer, 2017). The brain-based measure of workload showed that students were more overloaded during learning later in the session when using the immersive VR simulation as compared to the PC version of the simulation. This is preliminary brain-based evidence suggesting that the reason for a lower level of learning with immersive VR is that these environments may be overstimulating. The use of EEG to measure cognitive load is promising because it could provide learning scientists with the potential of examining the mediating processes within the individual that influence science learning. The EEG results also showed that students were overloaded an average of 48.78% of the time during learning which suggests that all versions of the science lab simulation were too challenging for the sample in the study. This is a good example of the value of objective cognitive measures because they can give information about the process by which learning takes place, and can provide specific data about the particular points within a multimedia lesson that are overloading students (see Figs. 2 and 3). This work encourages the idea that brain-based measures can ultimately be used to help design multimedia educational materials optimally.

4.5. Limitations and future directions

One of the research questions in this study was to investigate if the CTML also applied to immersive VR. The findings did not suggest that there were any differences between the low- and high-immersion VREs regarding the redundancy principle. However, more research is needed which compares a narration only condition to a condition with text and narration. Future research should also investigate if other principles from CLT and CTML generalize to immersive VR environments. In particular, it would be interesting to investigate the consequences of the modality principle because reading text can be more cognitively demanding in immersive VR, whereas spoken words might not cause extra cognitive load.

In this study we used an experimental design to investigate the differences between the low- and high-immersion VR simulations. However, this controlled environment might not be the best way to assess the potential value and impact of immersive VR for education and training. If immersive VR can engage students more deeply in the content of a science lab, it is possible that students would use this technology more and thus learn more. The ultimate idea of using immersive VR simulations in education could be giving the students a head-mounted display at the beginning of a term which they can use at home at their discretion with their smartphones. Therefore, given that enough high quality educational material is available, a fairer way to assess the value of immersive VR could be to have a longitudinal study which follows students across a longer period of time. Future research should investigate whether students in real educational environments would use immersive VR technology more and if this added use leads to more learning. More field research is also needed to understand how immersive VR might actually be implemented in different educational settings. In addition, in future work, instead of measuring engagement by self-report, it would be useful to use online behavioral measures such as number of mouse clicks.

One limitation of this study was that the technology used was the Samsung Gear VR, which required the participants to use a touch pad on the right side of the HMD to emulate the left-click function of a wireless mouse in order to select which objects to interact with in the lab. On the other hand, the control panel in the PC version was a mouse (with which the students already had a lot of experience), so the control panel in the immersive VR condition was new and not very intuitive. The simulation in this study was designed to create a setting wherein students could perform an experiment where they had to manipulate different items in a lab using two hands which are guided by the touch pad. Therefore, they were given a situation in which they were supposed to be active; but they were not given the tools to do so (rather than being able to manipulate the environment with their hands it was necessary to use a control panel that was not very intuitive). Therefore, future research should investigate the value of immersive VR with more advanced technology that affords a more natural control system. The sample size was also relatively small in this study because it is so time consuming to conduct this type of research. Future studies should use larger and different samples and different VR content to investigate the generalizability of the results.

The use of EEG to measure cognitive load is quite novel in educational settings. A simple EEG set-up was used in this study as this type of measure could easily be used by instructional designers who do not have expertise in cognitive neuroscience to measure cognitive load continuously and use this information to design learning material optimally. Furthermore, an ultimate instructional goal would be a moment-to-moment assessment of cognitive load leading to an immediate online adaptation of instructional material when learners are overwhelmed by the difficulty; or bored because the material is too easy compared to their working memory capacity (Gerjets et al., 2014). However, more research is needed that investigates different combinations of raw EEG data. Specifically, studies have shown that a drop in alpha and increased theta waves is associated with cognitive load (Gevins et al., 1998; Sauseng et al., 2005; Antonenko et al., 2010), but more research is needed to identify optimal combinations in order to provide a robust measure of cognitive load that is valid across learning settings. More research is also needed that combines EEG with other process measures in real time, such as eye tracking and pupil dilation in order to assess the validity of EEG measures of cognitive load (Mills et al., 2017).

There are several elements within this simulation that could potentially be improved in an attempt to make the immersive VR platform more successful. One is that the content in the simulation was difficult (as shown by the previously mentioned overload average) and might have imposed a heavy intrinsic load on the participants as the sample in this study was made up of novices. Finally, the immersive VR simulation was adapted from the PC version, so the specific advantages of immersive VR were not optimized. There are likely settings where the added presence that VR affords increases learning and transfer. The National Research Council report (2011) suggests that more evidence is needed about the value of simulations for developing science process skills, understanding of the nature of science, scientific discourse and argumentation, and identification with science and science learning. Immersive VR might be more suited for these advanced science learning goals, particularly when realistic visualizations of scientific material are important for gaining a deeper understanding of the subject matter. Higher immersion is also likely to make a difference in settings where the learning goal is to teach specific performance skills in realistic settings to an experienced group of students or practitioners. Furthermore, it seems essential that the design of VR educational content be developed from the start with the understanding of how this platform can support the given learning objectives. Therefore, the results in this study suggest that rather than porting educational content to VR, it is necessary to develop content specifically for VR, with an understanding of the unique advantages of the technology and how it will impact the learner.

5. Conclusion

Overall, the present study offers a step in assessing the educational value of low-cost immersive VR for improving student learning. In line with calls for rigorous experiments on learning with science simulations (NRC, 2011), the present study provides evidence for the idea that "liking is not learning"–that is, instructional media that increase the fun of a simulation–such as the sense of presence–do not necessarily increase student learning. To the contrary, cutting-edge high-immersion VR can create an increase in processing demands on working memory and a decrease in knowledge acquisition, as compared to conventional media. Therefore, considerations of the specific affordance of immersive VR for learning should be considered in designing learning content for this media.

References

iMotions A/S (2016). *EEG pocket guide.* Retrieved from https://imotions.com/guides/on November 26th, 2016.

ABM (2016). *B-alert X10 EEG headset system.* Retrieved on the 3rd of December, 2016 from http://www.advancedbrainmonitoring.com/xseries/x10/.

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

Antonenko, P., Paas, F., Grabner, R., & Van Gog, T. (2010). Using electroencephalography to measure cognitive load. *Educational Psychology Review, 22*(4), 425–438.

Bayraktar, S. (2001). A Meta-analysis of the effectiveness of computer-assisted instruction in science education. *Journal of Research on Technology in Education, 34*(2), 173–188. https://doi.org/10.1080/15391523.2001.10782344.

Belini, H., Chen, W., Sugiyama, M., Shin, M., Alam, S., & Takayama, D. (2016). *Virtual & Augmented Reality: Understanding the race for the next computing platform. Goldman Sachs report*. Retrieved on the 1st of March, 2017 from http://www.goldmansachs.com/our-thinking/pages/technology-driving-innovation-folder/virtual-and-augmented-reality/report.pdf.

Berka, C., Levendowski, D. J., Lumicao, M. N., Yau, A., Davis, G., Zivkovic, V. T., et al. (2007). EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation Space & Environmental Medicine, 78*(5), 231–244.

Berka, C., Levendowski, D. J., Petrovic, M. M., Davis, G., Lumicao, M. N., Zivkovic, et al. (2004). Real-time analysis of EEG indexes of alertness, cognition, and memory acquired with a wirelessEEG headset. *International Journal of Human-Computer Interaction, 17*(2), 151–170.

BIOPAC (2016). *B-ALERT with AcqKnowledge quick guide. Benchmark acquisition and cognitive states analysis.* Retrieved on 10th November, 2016 https://www.biopac.com/wp-content/uploads/b-alert-with-acqknowledge-quick-guide.pdf.

Blascovich, J., & Bailenson, J. (2011). *Infinite reality.* New York: HarperCollins.

Bodekaer, M. (2016). *Michael Bodekaer: The virtual lab will revolutionize science class.* Retrieved from https://www.ted.com/talks/michael_bodekaer_this_virtual_lab_will_revolutionize_science_class.

Bonde, M. T., Makransky, G., Wandall, J., Larsen, M. V., Morsing, M., Jarmer, H., et al. (2014). Improving biotech education through gamified laboratory simulations. *Nature Biotechnology, 32*(7), 694–697. http://doi.org/10.1038/nbt.2955.

Clark, B. D., Tanner-Smith, E. E., & Killingsworth, S. S. (2016). Digital games, design, and learning: A systematic review and meta-analysis. *Review of Educational Research, 86*(1), 79–122. https://doi.org/10.3102/0034654315582065.

Colliflower, Talya J. (2013). *Interpretation of the WISC-IV working memory Index as a measure of attention* Theses, Dissertations and Capstones. Paper 699.

Cuban, L. (1986). *Teachers and machines: The classroom use of technology since 1920.* Teachers College Press.

De Jong, T. (2017). Instruction based on computer simulations and virtual laboratories. In R. E. Mayer, & P. A. Alexander (Eds.). *Handbook of research on learning and instruction* (pp. 502–521). (2nd ed.). New York: Routledge.

De Jong, T., Linn, M. C., & Zacharia, Z. C. (2013). Physical and virtual laboratories in science and engineering education. *Science, 80*(340), 305–308. http://dx.doi.org/10.1126/science.1230579.

Dewey, J. (1913). *Interest and effort in education.* Cambridge, MA: Houghton Mifflin.

Galán, F. C., & Beal, C. R. (2012). EEG estimates of engagement and cognitive workload predict math problem solving outcomes. In J. Masthoff, (Ed.). *UMAP 2012, LNCS 7379* (pp. 51–62). Berlin: Springer-Verlag.

Gerjets, P., Walter, W., Rosenstiel, W., Bogdan, M., & Zander, T. O. (2014). Cognitive state monitoring and the design of adaptive instruction in digital environments: Lessons learned from cognitive workload assessment using a passive brain-computer interface approach. *Frontiers in Neuroscience, 8*, 386. http://dx.doi.org/10.3389/fnins.2014.00385.

Gevins, A., Smith, M. E., Leong, H., McEvoy, L., Whitfield, S., Du, R., et al. (1998). Monitoring working memory load during computer-based tasks with EEG pattern recognition methods. *Human Factors, 40*(1), 79–91.

Greenlight, V. R., & Roadtovr (2016). *2016 virtual reality industry report*. Retrieved from http://www.greenlightinsights.com/reports/2016-industry-report on March 3rd, 2017.

Harp, S. F., & Mayer, R. E. (1997). The role of interest in learning from scientific text and illustrations: On the distinction between emotional interest and cognitive interest. *Journal of Educational Psychology, 89*(1), 92–102. https://doi.org/10.1037/0022-0663.89.1.92.

Honey, M. A., & Hilton, M. L. (2011). *Learning science through computer games and simulations.* Washington, DC: National Academies Press.

Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology, 13*, 351–371.

Kalyuga, S., & Sweller, J. (2014). The redundancy principle in multimedia learning. In R. E. Mayer (Ed.). *The Cambridge handbook of multimedia learning* (pp. 247–262). (2nd ed.). New York: Cambridge University Press.

Klopfer, E. (2008). *Augmented learning: Research and design of mobile educational games.* Cambridge, MA: MIT Press.
Lee, E. A., & Wong, K. W. (2014). Learning with desktop virtual reality: Low spatial ability learners are more positively affected. *Computers & Education, 79*, 49–58.
Lee, E. A.-L., Wong, K. W., & Fung, C. C. (2010). How does desktop virtual reality enhance learning outcomes? A structural equation modeling approach. *Computers & Education, 55*(4), 1424–1442. http://doi.org/10.1016/j.compedu.2010.06.006.
Limniou, M., Roberts, D., & Papadopoulos, N. (2007). Full immersive virtual environment CAVEtm in chemistry education. *Computers & Education, 51*(2), 584–593. http://dx.doi.org/10.1016/j.compedu.2007.06.014.
Makransky, G., Bonde, M. T., Wulff, J. S. G., Wandall, J., Hood, M., Creed, P. A., et al. (2016). Simulation based virtual learning environment in medical genetics counseling: An example of bridging the gap between theory and practice in medical education. *BMC Medical Education, 16*(1) 98. 6.
Makransky, G., Lilleholt, L., & Aaby, A. (2017). Development and validation of the multimodal presence scale for virtual reality environments: A confirmatory factor analysis and item response theory approach. *Computers in Human Behavior, 72*, 276–285. http://doi.org/10.1016/j.chb.2017.02.066.
Makransky, G., Thisgaard, M. W., & Gadegaard, H. (2016). Virtual simulations as preparation for lab exercises: Assessing learning of key laboratory skills in microbiology and improvement of essential non-cognitive skills. *PLoS One, 11*(6), e0155895. http://dx.doi.org/10.1371/journal.pone.0155895.
Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). New York: Cambridge University Press.
Mayer, R. E. (2014). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.). *The Cambridge handbook of multimedia learning* (pp. 43–71). (2nd ed.). New York: Cambridge University Press.
Mayer, R. E. (2017). How can brain research inform academic learning and instruction? *Educational Psychology Review, 29*, 835–846.
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* (pp. 279–315). (2nd ed.). New York: Cambridge University Press.
Mayer, R. E., Heiser, J., & Lonn, S. (2001). Cognitive constraints on multimedia learning: When presenting more material results. *Journal of Educational Psychology, 93*(1), 187–198.
Merchant, Z., Goetz, E. T., Cifuentes, L., Kenney-Kennicutt, W., & Davis, T. J. (2014). Effectiveness of virtual reality-based instruction on students' learning outcomes in K-12 and higher education. *Computers & Education, 70*, 29–40.
Milk, C. (2015). *Chris Milk: How virtual reality can create the ultimate empathy machine.* Retrieved from https://www.ted.com/talks/chris_milk_how_virtual_reality_can_create_the_ultimate_empathy_machine.
Mills, C., Fridman, I., Soussou, W., Waghray, D., Olney, A. M., & D'Mello, S. K. (2017, March). Put your thinking cap on: Detecting cognitive load using EEG during learning. *Proceedings of the seventh international learning analytics & knowledge conference* (pp. 80–89). ACM.
Moreno, R., & Mayer, R. E. (1999). Cognitive principles of multimedia learning: The role of modality and contiguity. *Journal of Educational Psychology, 91*, 358–368.
Moreno, R., & Mayer, R. E. (2000). A coherence effect in multimedia learning: The case for minimizing irrelevant sounds in the design of multimedia messages. *Journal of Educational Psychology, 92*, 117–125.
Moreno, R., & Mayer, R. E. (2002). Learning science in virtual reality multimedia environments: Role of methods and media. *Journal of Educational Psychology, 94*(3), 598–610.
Moreno, R., & Mayer, R. E. (2007). Interactive multimodal learning environments. *Educational Psychology Review, 19*, 309–326.
National Research Council (2011). *Learning science through computer games and simulations. committee on science learning: Computer games, simulations, and education.* Washington, DC: National Academies Press.
Norman, D. A. (1993). *Things that make us smart: Defending human attributes in the age of the machine.* Reading, MA: Addison-Wesley.
Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the hospital anxiety and depression scale (HADS). *British Journal of Clinical Psychology, 46*, 1–18. https://doi.org/10.1348/014466506X96931.
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.
Pekrun, R. (2016). Emotions at school. In K. R. Wentzel, & D. B. Miele (Eds.). *Handbook of motivation at school* (pp. 120–144). (2nd ed.). New York: Routledge.
Plass, J. L., & Kaplan, U. (2015). Emotional design in digital media for learning. In S. Y. Tettegah, & M. Gartmeier (Eds.). *Emotions, technology, design, and learning* (pp. 131–161). San Diego: Academic Press.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.
Renninger, K. A., & Hidi, S. E. (2016). *The power of interest for motivation and engagement.* New York: Routledge.
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.
Rutten, N., van Joolingen, W. R., & van der Veen, J. T. (2012). The learning effects of computer simulations in science education. *Compuers & Education, 58*(1), 136–153. https://doi.org/10.1016/j.compedu.2011.07.017.
Salzman, M. C., Dede, C., Loftin, R. B., & Chen, J. (1999). A model for understanding how virtual reality aids complex conceptual learning. *Presence: Teleoperators and Virtual Environments, 8*(3), 293–316. http://doi.org/10.1162/105474699566242.
Sauseng, P., Klimesch, W., Doppelmayr, M., Pecherstorfer, T., Freunberger, R., & Hanslmayr, S. (2005). EEG alpha synchronization and functional coupling during top-down processing in a working memory task. *Human Brain Mapping, 26*(2), 148–155.
Schubert, T., Friedmann, F., & Regenbrecht, H. (2001). The experience of presence: Factor analytic insights. *Presence: Teleoperators and Virtual Environments, 10*(3), 266–281. June 2001 https://doi.org/10.1162/105474601300343603.
Sciarini, L. W., Grubb, J. D., & Fatolitis, P. G. (2014). Cognitive state assessment: Examination of EEG-based measures on a stroop task. *Proceedings of the human factors and ergonomics society 58th annual meeting. Vol. 58. Proceedings of the human factors and ergonomics society 58th annual meeting* (pp. 215–219). Sage CA: Los Angeles, CA: SAGE Publications No. 1.
Sitzmann, T. (2011). A meta-analytic examination of the instructional effectiveness of computer-based simulation games. *Personnel Psychology, 64*(2), 489–528. http://dx.doi.org/10.1111/j.1744-6570.2011.01190.x.
Slater, M., & Wilbur, S. (1997). A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments. *Presence: Teleoperators and Virtual Environments, 6*, 603–616.
Slotta, J. D., & Linn, M. C. (2009). *WISE Science: Web-based inquiry in the classroom.* New York: Teachers College Press.
Sterman, M. B., & Mann, C. A. (1995). Concepts and applications of EEG analysis in aviation performance evaluation. *Biological Psychology, 40*, 115–130.
Stevens, R., Galloway, T., Berka, C., Behneman, A., Wohlgemuth, T., Lamb, J., et al. (2011). Linking models of team neurophysiologic synchronies for engagement and workload with measures of team communication. In: *Proc. 20th conf. Behavioral representation in modeling and simulations* (pp. 122–129). .
Sweller, J., Ayres, P. L., & Kalyuga, S. (2011). *Cognitive load theory.* New York: Springer.
Thisgaard, M., & Makransky, G. (2017). Virtual learning simulations in high School: Effects on cognitive and non-cognitive outcomes and implications on the development of STEM academic and career choice. *Frontiers in Psychology, 8*.
Van Der Heijden, H. (2004). User acceptance of hedonic information systems. *MIS Quarterly, 28*(4), 695–704.
Vogel, J. J., Vogel, D. S., Cannon-Bowers, J., Bowers, C. A., Muse, K., & Wright, M. (2006). Computer gaming and interactive simulations for learning: A meta-analysis. *Journal of Educational Computing Research, 34*(3), 229–243.
Webster, R. (2016). Declarative knowledge acquisition in immersive virtual learning environments. *Interactive Learning Environments, 24*(6), 1319–1333.
Wentzel, K. R., & Miele, D. B. (2016). *Handbook of motivation at school* (2nd ed.). New York: Routledge.