Journal of Science Education and Technology (2021) 30:862–876
https://doi.org/10.1007/s10956-021-09925-0

Exploring Differences in Student Learning and Behavior Between Real-life and Virtual Reality Chemistry Laboratories

Elliot Hu-Au[1] · **Sandra Okita**[1]

Accepted: 4 July 2021 / Published online: 21 July 2021

Abstract
Recent global events and educational trends have led schools to heavily rely on digital media to educate their students. Science classes, in particular, stand to lose substantial learning opportunities without the ability to provide physical laboratory experiences. Virtual reality (VR) technology has the potential to resolve this issue, but little is known if VR environments can produce similar results to real-life (RL) science learning environments. This 2×1, between-subjects study compares students' learning results and safety behaviors in VR and RL chemistry laboratories. The study attempts to identify differences in learning experience (i.e., general chemistry content, experiment comprehension, laboratory safety knowledge) and laboratory safety behavior. Results indicate learning general content knowledge, laboratory skills, and procedure-related safety behaviors were comparable between RL and VR conditions, but clean-up behaviors were less frequent in VR. Also, the exploratory, risk-free nature of VR environments may have allowed the learners to elaborate and reflect more on general chemistry content and laboratory safety knowledge than in the RL environment.

Keywords Chemistry education · Immersive learning environments · Science education · Virtual reality

Introduction

Laboratory exercises have been an integral part of US chemistry and physical science classes for over a century (Hofstein, 2004). Research has shown that active learning methods, such as chemistry laboratory exercises, are significantly more effective at increasing student performances in science, technology, engineering, and math courses (STEM) than traditional lecture-style teaching methods (Baragona, 2009; Freeman et al., 2014; Morell, 1994). They are also connected with generating higher levels of student interest in the subject matter (Halim et al., 2018; Jones & Stapleton, 2017) and introduce students to the methods and procedures that experts use (Tsaparlis, 2009). Through hands-on activities like laboratory exercises, students gain valuable skills and knowledge that can support them through future careers in STEM.

A recent trend in education, exacerbated by the COVID-19 pandemic, has borne witness to many schools transitioning hands-on laboratory exercises into online or digital experiences (Crippen et al., 2013). The National Research Council (2006) identified reasons for this transition and the decline of physical, hands-on experiences as the lack of financial resources, preparation of teachers, and difficulties in maintaining a safe and well-stocked science laboratory. Additionally, time pressures due to the high-stakes testing climate have led to minimal opportunities for effective, but time-consuming, hands-on exercises. Thus, many chemistry classes revert into purely lecture-based experiences (Anderman et al., 2012; Tatli & Ayas, 2013). With no end in sight, there is a need to explore alternatives that can be efficient and continue to provide students with the benefits of traditional science laboratory experiences. The paper will first review the relationship between learning and virtual reality (VR) technology, draw attention to VR's attributes that support science laboratory experiences, and finally explain the goals of our research study.

✉ Elliot Hu-Au
elliot.hu-au@tc.columbia.edu

Sandra Okita
okita@tc.columbia.edu

[1] Communications, Media, and Learning Technologies Design, Teachers College, Columbia University, NY, New York, USA

Virtual Reality Technology and Learning

Virtual reality technology represents a potential innovation that could provide rigorous content and immersive experiences to support hands-on science learning (Castelvecchi, 2016; Chao et al., 2016). Like traditional hands-on learning, the learning mechanisms available to VR technology often depend "upon the kinds of experience that come from having a body with various sensorimotor capacities" (Varela et al., 1991). VR's head-, position-, and hand-tracking allow the user to use natural bodily movements to explore and increase perception in digital environments. Thus, the sensorimotor capabilities of the user and contextual information are utilized to create knowledge. In this respect, the foundations of learning experiences in VR can be readily drawn from the embodiment of virtual avatars and the affordances of VR learning environments (Shin, 2017). The following are some of the immersive features of VR technology that support these learning mechanisms and have similar characteristics to hands-on learning in real-life.

First, VR enables the user to control their visual senses in a similar way to their natural function. High-accuracy head-tracking allows the user to move and rotate their head, changing the position of their perceptual view. Research has observed how enhanced visualizations using VR technology provide a range of perspectives that help students' cognition and knowledge development of complex information (Bailey et al., 2016; Salzman et al., 1999). Bailey et al. (2016) posit that it is possible that the level of immersion in VR grounds the user's experiences in their virtual body, giving them access to sensorimotor capacities for cognition. The user can mimic how they would physically move their head and bodies to gain a better perspective on an object. VR taps into that same natural action to enable the user to gain more knowledge about a desired object or environment. The ability to support what Gibson (1966) calls "active" knowledge-seeking (p. 5), gives VR users the capability to use their bodies as perceptual systems, proactively gathering information in the digital world. An example of this is when a user is confronted with a three-dimensional (3D) model within VR. Often, they will swivel their head around, move their body around to a new position, and physically move their head closer or away from the object to get more detail. VR technology brings a person's natural information-gathering movements into the digital realm (i.e., head and body movement) to enable the user to perceive different aspects of the object and visually receive more information than what is originally available.

Secondly, hand-tracking controllers in today's high-end VR systems enable the user's hands to be tracked and seen as they move within the digital world. Hands can be used for exploration and knowledge-creation gestures as they would in real life. The use of natural movements in their learning environment is integral to forming strong knowledge concepts and memories (Brown et al., 1989). For example, in a physics education study where students used text, video, or gestures with an *Xbox Kinect*, researchers discovered that gestures involving greater levels of embodiment led to increased amounts of learning (Johnson-Glenberg & Megowan-Romanowicz, 2017). They observed that "if the learner is induced to manipulate the content on screen and control the content with representational gestures that are congruent to what is being learned… they may learn the content faster or in a deeper manner" (p. 17). Similarly, Goldin-Meadow (2011) reports that children who watched a teacher gesture during a lesson and responded with their own gestures had greater learning than children who did not gesture in return. The children's gesturing lessened the cognitive load during the learning experience and enabled them to more readily process new information. In each of these studies, participants' ability to move and see their own hand gestures strengthened their learning. VR's capability to provide similar access to the hand and gesture modality is a promising step toward knowledge creation in digital realms (Weisberg & Newcombe, 2017).

Lastly, the highly immersive nature of VR is also exceptionally well-suited to situate the learner in context and environment. As mentioned before, the immersion a user feels in VR can ground them in the virtual body they are embodying (Bailey et al., 2016). This causes them to create a mental model of their body based on the new affordances of the virtual body, physically and socially. Identifying with this new body can often change the perspective of the user (Loke, 2015; Oviatt, 2013). Early STEM-focused examples of this type of environment are virtual worlds like *Quest Atlantis* and *River City* (Barab et al., 2005; Dede, 2009). Within these contexts, users take on the identity of a scientist, collecting information, sharing with other users, and conducting exploratory missions. Dialogue and interactions reinforce the storyline of the user as a scientist. These immersive worlds encourage identification as the character one portrays (e.g., a scientist), enabling the shedding of negative self-references and encouraging empathy with multiple perspectives (Oviatt, 2013). In situated learning-type activities, players can learn experientially (Lave & Wenger, 1990) through joint quests and social interactions with others who are not physically near, providing a unique learning experience that is a step closer to immersion than virtual worlds.

Virtual Reality's Potential Benefits Over Real-Life Laboratory Environments

For science laboratory work, a few affordances of VR provide advantages over what can be done in real life. Some features can augment the immediate experience of a student, giving information or guidance in timely fashion or a relevant position. Other features of VR can enable easier implementation of experiences that maximize certain learning strategies (Tatli & Ayas, 2013). Features, such as the ability to reify abstract concepts, enabling faster and more frequent repeatability of experiences, and a more forgiving environment for mistakes, present a learning experience that is ideal for learners (Dalgarno & Lee, 2010; Lau & Lee, 2015).

Particularly valuable for chemistry and other sciences, VR can provide clear visual and spatial representations of concepts that are typically difficult to visualize (Crosier et al., 2000; Ferrell et al., 2019). The 3D capability, interactivity, and flexibility of how things can be displayed give VR a unique capacity for presenting topics in novel ways. For instance, manipulating molecular structures and pulling methane molecules through carbon nanotubes in virtual environments can improve students' motivation, understanding of critical chemistry concepts, and develop greater spatial awareness (Merchant et al., 2013; Ferrell et al., 2019). The ability to easily make abstract and intangible concepts visual and concrete is a significant asset for science experiences in VR.

The ease of implementation and repeated practice afforded by VR has led to benefits for many STEM-related subjects, such as military, aeronautical, architectural, and medical fields (Lau & Lee, 2015; Psotka, 1995). Seymour et al. (2002) demonstrated that VR training significantly improved operating room performance of residents doing laparoscopic surgery. Butt et al. (2018) also observed similar results in their study of student nurses practicing urinal catheterization in VR. Both studies determined that the ability to train as frequently as desired and without time restrictions brought about the increased performance. This was due to the relative ease in turning on a VR program as opposed to using the traditional box trainer or mentor-trainee model during surgical procedures (Seymour et al., 2002). When compared with the amount of effort needed to implement real hands-on learning experiences, VR provides a quicker and easier alternative, one that is rapidly proving to be just as effective as well.

The freedom to learn through the virtual experience without the fear of injury or major consequence is also an important advantage to VR learning environments (Standen & Brown, 2006). This experience is similar to observations of students in virtual and game-like environments where exploratory or risky behavior is often rewarded, even when resulting in failure (Vogel et al., 2006; Dickey, 2005). Research in many virtual training programs supports this notion of enhanced learning due to an environment free of fear of consequence. Medical schools have been using virtual training for decades, allowing students to practice on virtual patients before moving to real subjects (Rubio-Tamayo et al., 2017; Seymour et al., 2002). Construction safety training programs have also demonstrated increased learning when participants can make errors virtually and learn from them (de-Juan-Ripoll et al., 2018). Thus, the negligible consequences in VR learning environments act as an effective tool for knowledge training and skill gain.

VR environments also reveal that levels of anxiety in students are far less compared to the amount they may experience in a real-life setting (Lindner et al., 2017). VR learning environments can be designed without many of the extraneous stimuli that normal learning environments may contain. Without these distractions, students can focus on the task and concepts at hand. For example, a review on VR treatments of public speaking anxieties reports that most studies show positive effects from their participants (Hinojo-Lucena et al., 2020). A primary reason for these benefits is that VR provides an environment where participants can have control, develop without fear, and feel safe. Nursing and medical school research has also found that VR training mitigates the anxiety many students feel in clinical settings (Jenson & Forsyth, 2012; Butt et al., 2018). VR scenarios provide students with anxiety-relieving benefits like immediate feedback and unlimited practice in risk-free environments.

Study Hypotheses

Our research question is exploratory: what are the differences in learning results when conducting a chemistry laboratory experience in virtual reality or real life?

We have two hypotheses. The first hypothesis is that there will be no significant difference in learning content knowledge between VR and RL chemistry laboratory environments. The VR lab simulation is designed to be as accurate as possible to real-life laboratory environments. As both conditions rely on similar hands-on learning experiences, similar learning outcomes are expected. In a study by Winkelmann et al. (2014) findings showed that students produced similar quality lab reports for virtual experiments done in Second Life versus students that engaged in hands-on experiments in a RL laboratory. Thus, we expect learning results to not vary significantly between VR and RL.

The second hypothesis is that there will be an observable difference in demonstrated lab behaviors. Participants in the VR condition may not exhibit the same level of care and attention to their actions and behaviors due to the lack of physical contact with laboratory materials (Lau & Lee,

2015). In addition, the game-like nature of VR may prime participants to see the VR setting as a casual learning environment. This could result in participants becoming more relaxed, a mentality Itō (2010) describes as one where "game outcomes do not transfer to the real-life economies of academic achievement and playing the role of the good student" (p. 201). Therefore, participants in the VR chemistry laboratory may exhibit atypical lab behaviors due to the less-consequential, game-like environment of VR.

Methodology

Participants and Context

The study consisted of 40 graduate students ranging from 20 to 42 years old from a private university in New York City. To avoid participants majoring in chemistry, the participants were recruited from graduate programs in education and the arts. Notably, there were significantly more female participants than male participants (37 female and 3 male).

Study Design and Procedure

This study was a 2 × 1, between-subjects design that compared the learning experience in a real-life Chemistry Lab (control) vs. a Virtual Reality Chemistry Lab (experimental). Participants were randomly assigned to one of two conditions (see Fig. 1). Participants in the RL condition conducted the laboratory exercise in a traditional science lab classroom, where the experimenter remained out-of-sight but present for safety reasons. Participants in the VR condition were physically in an office space but conducted the laboratory exercise in VR, again, the experimenter being present in the room throughout the study for safety purposes. All participants first watched, in real life, a 9-min video on lab safety information and general chemistry content usually covered in class lectures prior to a laboratory exercise (e.g., what is a chemical vs. physical change). This was followed by a pretest conducted on a laptop computer.

Next, all participants were given a 5-min training session to familiarize themselves with their particular lab space, safety equipment, tools, and materials. The RL and VR chemistry lab environments had a similar equipment set up. Those in the VR condition wore a VR headset, operated hand controllers, and entered a custom-designed VR chemistry lab space. For the next 10–25 min, participants in both conditions engaged in the first exercise, a single-mixture chemistry experiment, then completed the first Exercise Comprehension Assessment (ECA) using a laptop (10–15 min). Immediately following the ECA, they returned to the lab setup and conducted the second exercise, a multiple-mixture chemistry experiment. This normally took between 10 and 25 min. Finally, on the same laptop as before, the participant completed the second ECA along with the posttest. The total time between pretest and posttest averaged between 45 and 60 min. Participants were video recorded for analysis of lab behaviors and verbal comments.

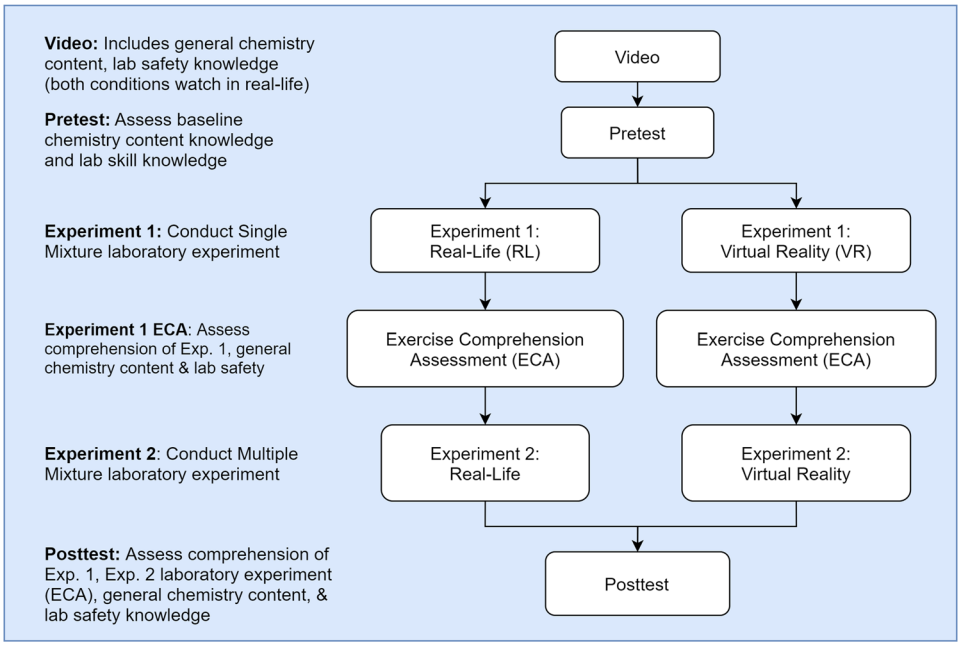

Fig. 1 Study procedure

Materials

Study Materials

The chemistry content was based on the NGSS (2013) Middle School Physical Science Standard MS-PS1-2 that focuses on chemical or physical reactions (NGSS, 2013, p. 42). The first experiment (*Single Mixture Exercise)* involved mixing anhydrous copper (II) chloride with water, then introducing a piece of aluminum metal to the aqueous solution. This lab exercise is recommended for middle to high school (ages 11–18 years) chemistry students by the American Association of Chemistry Teachers (AACT, 2020). The second experiment (*Multiple Mixture Exercise*) involved observing chemical reactions from a series of mixtures using hydrochloric acid, lead nitrate, copper sulfate and mossy zinc (PSI Chemistry, 2018). Both experiments are of the "cookbook" style of laboratory design where procedural steps are detailed and meant to be followed closely (National Research Council, 2006). This is unlike inquiry-based labs where results are more open-ended and students have more chances for exploration and interpretation. While research has shown that inquiry-based labs may promote a more active and meaningful learning process (Zacharia et al., 2015), the "cookbook" method was chosen for two reasons: it is still a frequently used classroom pedagogy (Akuma & Callaghan, 2019; Keiner & Graulich, 2021), and it is currently easier to develop a digital learning environment with a more rigid procedure.

Experimental Environment

The real-life chemistry lab space was a university classroom outfitted like a typical high school science lab (see Fig. 2). The VR chemistry lab was built in *Unity3D* and replicated the lab classroom environment (i.e., No extra-immersive or extraordinary features were added) (see Fig. 3). The *HTC Vive* VR headset digitally tracked participants as they walked around and engaged in the experiments using the hand controllers. Interactions between objects and substances were designed to reflect real-world experiences and based on real-life trials of the experiments.

Fig. 2 The real-life chemistry lab environment

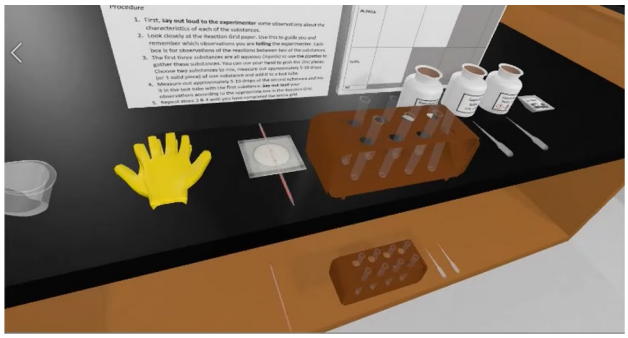

Fig. 3 The VR chemistry lab environment

Measures

This study includes three measures that differ in focus (a) general chemistry content (e.g., *Define a chemical change*), (b) laboratory safety knowledge (e.g., *Can you return unused copper sulfate into the original container?*), and (c) experiment exercise comprehension assessment (e.g., *Is heat generated when mixing lead nitrate & copper sulfate?*). Within these measures, there were different types of questions: basic, inference, or application questions. The categorization of the questions derives from a revision of Bloom's Taxonomy (Krathwohl, 2002) where different levels of understanding a concept were delineated based on action words instead of Bloom's original nouns. Basic questions recall facts (e.g., *define physical change*). Inference questions require interpretation drawing from multiple forms of evidence or information (e.g., *Does creating cement involve physical or chemical changes?*). Application questions require knowledge to be applied to real-world situations (e.g., *Boiling an egg. Is this a chemical change?*).

The pretest and posttest contained 40 questions each: 23 questions cover general chemistry content, 17 focused on laboratory safety knowledge. A subject-matter expert was consulted on the construction and breadth of the questions. All questions on the posttest were identical to the questions on the pretest. The questions were a mix of multiple-choice and some free-response questions. All multiple-choice questions had a value of one point, and each free-response question, based on its content, had one to three points possible. For the free-response questions, five cases were chosen at random, and two raters independently scored them.

Following is a breakdown of the question types for each assessment. The pre-posttest contained two sections, general content with eight basic, six inference, and nine application questions and lab safety knowledge with five basic, 12 inference, and no application questions. There were two ECAs given, one completed immediately after each experiment. The first assessment was on the *Single Mixture Exercise* consisting of 10 questions (7 basic, 3 inference) and was administered twice (i.e., once after the exercise was completed and again after the second exercise was completed) for recall after some delay. The second assessment consisted of 19 questions (4 basic, 15 inference) and was administered once after the *Multiple Mixture* experiment.

Video was used to record participants' laboratory safety behaviors during the two chemistry experiments. Evaluated laboratory safety behaviors were chosen from a commonly used American Chemical Society training video (Wickstrom, 1991). To provide a common foundation for all participants, parts of this safety video were shown during the training video at the beginning of the study. During the study, participants were observed to see whether they followed the demonstrated safety procedures, skills (i.e., wearing lab gloves, eye goggles, correctly using scale, and handling thermometers), and laboratory clean-up behaviors (i.e., dispose solids in trash, liquids in the sink). Scoring safety behaviors and skills from video observations required a specific action to be completed, which could then be marked as "observed." For instance, the behavior *Wore safety eye goggles* would be confirmed if the participant placed the eye goggles on their face before handling chemicals. Some behaviors, such as thermometer safety, required a participant to demonstrate safe thermometer placement (i.e., on a non-smooth surface away from the edge of the table) at least once during the exercise. Even if the participant reverted to an unsafe thermometer placement at a different time, they were characterized as demonstrating the safety behavior because they did it at least once.

Video data was also used to analyze the prevalence of comments that belied anxious feelings for participants during the exercises. A similar rubric for determining anxious comments was used as Bourne (2015) where "negative self-talk" was the primary consideration (p. 187). To quantify the comment score, the occurence of any anxious comment was counted. The strength or content of the comment was not judged, other than it being an example of negative self-talk.

Two independent raters reviewed the video data for the existence of safety behaviors, lab cleanup behaviors, and anxiety comments. An interrater reliability analysis using Cohen's κ statistic was performed to determine consistency among the raters.

Results

Due to reporting multiple contrasts using the same data, Bonferroni corrections were used to account for the chance of accidentally finding a false positive (Maxwell et al., 2017). For the contrasts of general chemistry content, we tested for pre-posttest differences between total scores, basic, inference, and application questions. The four comparisons resulted in a correction of the typical significance level of $\alpha = .05$ being reduced to $\alpha = .0125$. The same method was used for the Bonferroni correction values for the three contrasts tested in the Single Mixture ECA (see Table 2), Multiple Mixture ECA (see Table 3), and Lab Safety Knowledge (see Table 4). The Bonferroni correction reduced the standard for significance for these measures to $\alpha = .0167$.

Interrater reliability measures were employed for the free-response questions as described in the "Measures" section above. The process was completed twice, with the second set of five cases being rated with a Cohen's $\kappa = 0.985$ ($p < 0.001$), 95% CI (.958, 1.0).

General Chemistry Content

The pretest showed no difference between the two conditions prior to the study. An independent sample t-test was conducted that compared the difference in mean score from pretest to posttest between the two conditions. There was no significant difference between the mean differences from pretest to posttest at $t(38) = 0.518$, $p = 0.607$. A paired-sample t-test on scores from pretest to posttest was performed. The VR condition showed significant increase overall from pretest ($M = 70.2\%$, $SD = 0.133$) to posttest ($M = 76.5\%$, $SD = 0.14$) at $t(19) = 3.031$, $p = .007$ (see Table 1). No significant increase for VR's basic questions at $t(19) = 0.00$, $p = 1.00$ or inference questions at $t(19) = 2.459$, $p = .024$ was observed. The application questions at $t(19) = 3.107$, $p = .006$ showed a significant increase and drove the overall increase in general chemistry content learning for the VR condition. For the RL condition, the increase from pretest ($M = 72.4\%$, $SD = .131$) to posttest ($M = 76.5\%$, $SD = .119$) was not significant at $t(19) = 1.582$, $p = .130$. Looking closely by question type, there was no significant increase for RL on basic questions at $t(19) = -1.798$, $p = .088$, inference questions at $t(19) = 2.604$, $p = .017$, or application questions at $t(19) = 2.027$, $p = .057$.

A post hoc power analysis was computed on the VR condition's pretest–posttest general content results, with an $N = 20$, mean difference = 6.3, standard deviation = 9.3, and Bonferroni-corrected significance level of $\alpha = .0125$.

Table 1 Comparison of assessment percent scores of VR ($n=20$) and RL ($n=20$) conditions

Variable	Pretest		Posttest		*Diff*	*t*	*p*	*Effect size (Cohen's d)*
	M (SD)	*SE*	*M (SD)*	*SE*				
General content—Total								
Real life	72.4 (13.0)	2.9	76.5 (11.9)	2.7	4.1	1.582	.130	.354
Virtual reality	70.2 (13.3)	3.0	76.5 (13.6)	3.0	6.3	3.031	.007*	.678
General content—Basic								
Real life	87.5 (16.2)	3.6	82.5 (16.9)	3.8	−5.0	−1.798	.088	.402
Virtual reality	78.1 (23.6)	5.3	78.1 (24.3)	5.4	0.0	.000	1.000	.000
General content—Inference								
Real life	80.0 (19.2)	4.3	89.2 (16.5)	3.7	9.2	2.604	.017	.582
Virtual reality	75.0 (23.9)	5.3	85.8 (17.3)	3.9	10.8	2.459	.024	.550
General content—Application								
Real life	53.9 (14.1)	3.2	62.8 (17.8)	4.0	8.9	2.027	.057	.453
Virtual reality	60.0 (13.2)	3.0	68.9 (17.1)	3.8	8.9	3.107	.006*	.695

*This value is significant at the Bonferroni-adjusted value of α = .0125

A medium power of 0.612 was achieved with effect size, d = .678. This demonstrated that the VR lab intervention provided a learning increase with a medium-large effect size.

Experiment 1 Exercise Comprehension Assessment (Single Mixture)

For the *Single Mixture* assessment following the first experiment exercise, no significant differences were found between the mean differences of RL (M = 76.5%, SD = .123) and VR (M = 72%, SD = .120) at t(38) = 1.175, $p = .247$. When the same assessment was administered a second time following the second experiment (to check for retention), there was no significant difference between RL (M = 59%, SD = .168) and VR (M = 57.5%, SD = .148) at t(38) = 0.299, $p = .766$. Interestingly, both conditions significantly decreased in score from the first to the second assessment. A paired-sample t-test revealed a significant decrease was seen from the first assessment (M = 76.5%, SD = .123) to the second assessment (M = 59%, SD = .168), at t(19) = 4.937, $p < .0001$ for the RL condition (see Table 2). The significant decrease was seen in both the basic questions at t(19) = 4.721, $p < .0001$, and inference questions at t(19) = 2.698, $p = .0143$. Similarly, the VR condition also showed a significant decrease from first assessment (M = 72%, SD = .120) to the second assessment (M = 57.5%, SD = .148), at t(19) = 4.313, $p = .0003$. The significant decrease was seen only in basic questions at t(19) = 4.172, $p = .0005$.

The VR condition did not show the same significant decrease in scores for inference questions at t(19) = .295, $p = .772$. It seemed the VR condition was more effective at retaining inference-related content over time. To test this possible conclusion, an independent-samples t-test was then performed between conditions' scores on the second

Table 2 Comparison of Exercise Comprehension Assessment percent scores of VR ($n=20$) and real-life (RL) ($n=20$) conditions

Variable	After Single-Mixture experiment		After Multiple-Mixture experiment		*Diff*	*t*	*p*	*Effect size (Cohen's d)*
	M (SD)	*SE*	*M (SD)*	*SE*				
Exercise Comprehension Assessment (Single Mixture)—Total								
Real life	76.5 (12.3)	2.7	59.0 (16.8)	3.8	−17.5	−4.937	.000*	1.0
Virtual reality	72.0 (12.0)	2.7	57.5 (14.8)	3.3	−14.5	−4.313	.000*	.964
Exercise Comprehension Assessment (Single Mixture)—Basic								
Real life	74.3 (15.1)	3.4	57.9 (17.6)	3.9	−16.4	−4.721	.000*	1.0
Virtual reality	70.0 (17.9)	4.0	50.0 (18.2)	4.1	−20	−4.172	.001*	.933
Exercise Comprehension Assessment (Single Mixture)—Inference								
Real life	81.7 (17.0)	3.8	61.7 (34.7)	7.8	−20	−2.698	.014*	.603
Virtual reality	76.7 (19.0)	4.3	75.0 (32.2)	7.2	−1.7	−.295	.772	.066

*This value is significant at the Bonferroni-adjusted value of α = .0167

Table 3 Comparison of Exercise Comprehension Assessment (Multiple Mixture) percent scores of VR ($n=20$) and real-life (RL) ($n=20$) conditions

Variable	Real-Life		Virtual Reality		*Diff*	*t*	*p*	*Effect Size (Cohen's d)*
	M (SD)	*SE*	*M (SD)*	*SE*				
Total	64.5 (10.2)	2.3	58.4 (14.6)	3.3	6.1	1.520	.137	.481
Basic	96.3 (9.2)	2.0	90.0 (20.5)	4.6	6.3	1.244	.221	.393
Inference	56.0 (12.5)	2.8	50.0 (16.1)	3.6	6.0	1.316	.196	.416

instance of this assessment. The difference between the mean scores of each condition was found not significant at $t(38)=-1.260$, $p=.215$, and the effect size was medium, Cohen's $d=0.398$. In essence, there is a lack of evidence for the VR intervention to create a significant difference on inferential knowledge when compared to the RL condition.

Experiment 2 Exercise Comprehension Assessment (Multiple Mixture)

No significant difference between learning performance was observed between RL (M = 64.5%, SD = .102) and VR (M = 58.4%, SD = 0.146) for the *Multiple Mixture* assessment immediately following the second exercise at $t(38)=1.52$, $p=0.137$ (see Table 3). There was no difference between basic questions at $t(38)=1.244$, $p=.221$ and inference questions at $t(38)=1.316$, $p=.196$. There were no application questions.

Laboratory Safety Knowledge

The pretest showed a significant difference in laboratory safety knowledge between conditions, RL (M = 55.9%, SD = .142) and VR (M = 72.1%, SD = .171), prior to any treatment at $t(38)=3.258$, $p=.001$. No difference was observed for basic safety questions at $t(38)=0.438$, $p=.664$, but a significant difference for inference questions at $t(38)=3.813$, $p=.0005$, where the VR condition scored significantly higher than the RL condition (see Table 4). There were no application questions.

The posttest showed a significant difference between the two conditions at $t(38)=2.334$, $p=.025$, where RL performed significantly lower (M = 62.1%, SD = .129) compared to VR (M = 73.2%, SD = .171). No difference for basic safety questions at $t(38)=0.461$, $p=.648$, but a significant difference for inference questions at $t(38)=2.648$, $p=.012$. Both RL and VR conditions improved from pretest to posttest, where RL showed an average of 14.2% increase and VR showed an average of 3.1% increase. There was no significant difference between the two conditions regarding overall lab safety knowledge learning at $t(38)=1.717$, $p=.094$.

A paired-sample t-test was performed on the pre-posttest lab safety knowledge scores for each condition. The RL condition significantly increased from pretest (M = 55.8%, SD = .142) to posttest (M = 62.1%, SD = .129) at $t(19)=2.761$, $p=.012$. There was no significant increase for basic questions at $t(19)=1.000$, $p=.330$, but a significant increase for inference questions at $t(19)=2.653$, $p=.0157$. The VR condition improved their score from pretest (M = 72.1%, SD = 0.171) to posttest (M = 73.2%, SD = 0.171), but not significantly at $t(19)=0.483$, $p=.635$. There was no significant increase for VR in basic questions at $t(19)=0.370$, $p=.716$ or inference questions at $t(19)=0.403$, $p=.691$.

Table 4 Comparison of assessment percent scores of VR ($n=20$) and real-life (RL) ($n=20$) conditions

Variable	Pretest		Posttest		*Diff*	*t*	*p*	*Effect size (Cohen's d)*
	M (SD)	*SE*	*M (SD)*	*SE*				
Lab safety knowledge-Total								
Real life	55.8 (14.2)	3.2	62.1 (12.9)	2.9	6.3	2.761	.012*	.617
Virtual reality	72.1 (17.1)	3.8	73.2 (17.1)	3.8	1.1	.483	.635	.108
Lab safety knowledge -Basic								
Real life	88.0 (13.6)	3.0	89.0 (12.1)	2.7	1.0	1.000	.330	.224
Virtual reality	90.0 (15.2)	3.4	91.0 (15.2)	3.4	1.0	.370	.716	.083
Lab safety knowledge -Inference								
Real life	42.5 (16.2)	3.6	50.8 (15.5)	3.5	8.3	2.653	.016*	.593
Virtual reality	64.6 (20.2)	4.5	65.8 (20.0)	4.5	1.2	.403	.691	.088

*This value is significant at the Bonferroni-adjusted value of $\alpha=.0167$

Laboratory Safety Behaviors

Video data recorded participants' laboratory safety behaviors during the two chemistry experiments. One participant was removed from the VR condition due to video recording failure, resulting in 19 participant data for the VR condition, and 20 participant data for the RL condition. After training on two videos, the interrater reliability between two raters was found to be within strong agreement with $\kappa = 0.884$ ($p < 0.001$), 95% CI (0.678, 1.0). Bonferroni corrections were also made for the seven individual aspects of the safety behaviors, rendering significance now at the more conservative level of $\alpha = .007$.

First, we analyzed the combined amount of safety behaviors observed for each condition. This consisted of both the *procedures and skills* and the *cleanup behaviors* categories. There was a significant difference between the two conditions on all observed safety behaviors during the two experiments at $t(37) = 5.669$, $p < .0001$. The participants in the RL condition (M = 68.6%, SD = .095) averaged significantly more safety behaviors during the two experiments than the VR condition (M = 48.8%, SD = .122). However, it was noticed that participants' safety behaviors decreased from the first to the second experiment for both RL (M = −12.2%, SD = .143) and VR (M = −16.4%, SD = .210). A paired-sample t-test was performed on each of the safety behavior categories comparing any differences from experiment 1 to 2. Both conditions showed significant decrease in their total safety behavior scores from the first to the second experiment (see Table 5).

Looking more closely at the two categories within safety behaviors, a Bonferroni-correction of $\alpha = .025$ was used to test significance. The averages for observed *procedures and skills* for RL were M = 66.2% (SD = .087) and VR was M = 55.1% (SD = .124). A significant difference was found between conditions in following *procedures and skills* at $t(37) = 3.232$, $p = .0026$. This included wearing safety equipment and conducting procedures such as safely placing the thermometer on a non-slip surface. Table 5 lists the statistics for each specific type of procedure or skill that contributed to the overall value for this category.

Table 5 Comparison of lab safety behaviors for VR ($n = 20$) and real-life ($n = 20$) conditions in percent

Variable	Exercise 1		Exercise 2		*Diff*	*t*	*p*
	M (SD)	*SE*	*M (SD)*	*SE*			
Lab safety behaviors—Total							
Real life	75.5 (11.5)	2.6	63.2 (12.1)	2.7	−12.3	3.839	.001*
Virtual reality	56.4 (22.4)	5.0	40.0 (12.1)	2.7	−16.4	3.492	.002*
Wore eye goggles[a]							
Real life	95.0 (22.4)	5.0	90.0 (30.8)	6.9	−5.0	1.000	.330
Virtual reality	85.0 (36.6)	8.2	90.0 (30.8)	6.9	5.0	−.567	.577
Wore safety gloves[a]							
Real-life	100 (0)	0	100 (0)	0	0	–	–
Virtual reality	90.0 (30.8)	6.9	90.0 (30.8)	6.9	0	–	–
Wore safety apron/coat[a]							
Real-life	90.0 (30.8)	6.9	90.0 (30.8)	6.9	0	0	1.000
Virtual reality	55.0 (51.0)	11.8	70.0 (47.0)	10.5	15.0	−1.371	.186
Thermometer safety[a]							
Real-life	70.0 (47.0)	10.5	35.0 (48.9)	10.9	−35.0	3.199	.005*
Virtual reality	53.0 (51.3)	11.8	26.0 (45.2)	10.4	−27.0	1.564	.135
Cleaning glassware[b]							
Real-life	100 (0.0)	.00	85.0 (36.6)	8.2	−15.0	1.831	.083
Virtual reality	89.0 (32.3)	7.6	28.0 (46.1)	10.9	−61.0	4.267	.001*
Dispose of solids in trash[b]							
Real life	15.0 (98.8)	22.1	35.0 (81.3)	18.2	20.0	−.748	.464
Virtual reality	37.0 (68.4)	15.7	0 (0)	0	−37.0	2.348	.031
Dispose of liquids in drain[b]							
Real life	100 (0)	0	85.0 (36.6)	8.2	−15.0	1.831	.083
Virtual reality	74.0 (45.2)	10.4	21.0 (41.9)	9.6	−53.0	3.293	.004*

*This value is significant at the Bonferroni-adjusted value of $\alpha = .007$

[a]Procedures and Skills subgroup

[b]Clean-up behaviors subgroup

Comparing changes in *procedures and skills*, the RL condition significantly decreased in this category in experiment 1 (M = 75.7%, SD = .114) to experiment 2 (M = 59.5%, SD = .105) at t(19) = 5.456, $p < .0001$. The VR condition did not significantly decrease from experiment 1 (M = 61.7%, SD = .191) to experiment 2 (M = 50.5%, SD = .151) at t(18) = 2.101, $p = .05$, but it averaged a lower amount of observed behaviors on both experiments. There was no significant difference between the two conditions in the amount of decrease at t(37) = 1.227, $p = .228$.

A significant difference between conditions for clean-up behaviors was observed. The participants in the RL condition (M = 73.8%, SD = .211) followed cleaning procedures in the first experiment significantly more than the VR condition (M = 35.5%, SD = 0.192, t(37) = 5.916, $p < .0001$). Table 5 depicts the specific cleaning behaviors that make up this category. Regarding any changes in clean-up behaviors demonstrated in the second experiment, VR decreased significantly more than RL at t(37) = 2.672, $p = .011$. Participants in the RL condition showed no difference between their clean-up behavior from experiment 1 (M = 75%, SD = .256) to experiment 2 (M = 72.5%, SD = .371) at t(19) = 0.233, $p = .818$. However, participants in the VR condition significantly decreased in their clean-up behavior from experiment 1 (M = 55.3%, SD = .318) to experiment 2 (M = 15.8%, SD = .208) at t(18) = 4.581, $p = .0002$.

Participant Anxiety Comments

Comments such as "It's ok if it goes really bad, right?", "It's been so long since I've done something like this", and "this is what happened in all of my science classes" fit into categories of "negative self-talk" that were used to exemplify anxious feelings (Bourne, 2015, p. 187). Anxiety comments were observed much more frequently in the RL condition than VR. Seven out of the 20 RL participants (35%) stated at least one comment at the beginning of the experiment, while only two out of the 20 VR participants (10%) expressed anxiety in this way. The totals of anxiety comments were compiled and analyzed using an independent samples t-test. RL averaged a higher number of comments (M = .45, SD = .826) than VR (M = .20, SD = .523). However, the difference between conditions was not significant at t(38) = 1.144, $p = .260$. Therefore, we cannot conclude that there was a significant difference between conditions for the participants' level of anxiety through the number of verbalized comments.

Discussion

General Chemistry Content

For learning general chemistry content, the experience in the VR chemistry lab seems comparable to the learning experience in a traditional RL chemistry laboratory, as there were no significant differences between the two conditions. This is consistent with other findings comparing virtual labs with hands-on experiments (Winkelmann et al., 2014). However, the VR condition did significantly improve in learning results whereas the control condition did not. A possible explanation for this difference could be the novelty of VR and increased student interest (Ferrell et al., 2019). Another possibility is that the lower level of environmental detail in VR, as opposed to a real laboratory, could contribute to less cognitive load and thus better potential for learning. Scheiter et al. (2009) observed similar results in their study of multimedia presentations of biology processes. Our question type analysis revealed significant pre-posttest increases for application knowledge but not basic and inference questions. This is partially supported by the findings of Wieman and Holmes (2015), where the learning of basic science content knowledge was found not to be heavily influenced by non-inquiry-based laboratory experiences.

Exercise Comprehension Assessments

For the Experiment Exercise Comprehension Assessment, the learning experience in the VR chemistry lab seems comparable to the traditional RL chemistry laboratory, as there were no significant differences between the two conditions for both the Single Mixture ECA and the Multiple Mixture ECA. For the Single Mixture ECA, a 10–20-min delay between assessments resulted in both conditions decreasing significantly in content recall. This decrease could have originated from participants having to move quickly to the next lab experiment and not having time to reflect on what they had just learned (Baddeley, 1983). The Multiple Mixture exercise was more complex than the first and could have required more cognitive resources, erasing the recently learned Single Mixture information. Johnstone's (1997) working memory research on chemistry students' lecture note–taking habits in relation to exam performance may provide some insight into the poor performance on the Single Mixture ECA posttest questions. Johnstone found that students who "elaborated" (i.e., made cross references, comments, etc.) on concepts in their notes did much better on exams than those who merely copied written and verbal notes from the class lecture (p. 266). One could argue that the absence of application questions on our ECA did not give students a chance to elaborate on the basic concepts encountered, nor did they have much time to consider those concepts since they took the assessment immediately after completing the lab experiment. The decrease in retained Single Mixture knowledge from the first experiment could be explained by the students' lack of elaboration.

The pre- to posttest scores on inference questions remained somewhat stable for the VR condition while the

RL condition showed significant decrease. It is possible that the VR environment allows participants to focus more on their thought processes, and why things are happening, while the RL environment may distract learners with hazards (e.g., breaking the beaker, spilling chemicals) that take their focus away from content learning. There is some evidence that shows VR environments can be designed to increase student focus and elaboration on specific content areas (Hamilton et al., 2020; Shin, 2017), but more research is needed to show that VR environments can provide an advantage for specifically developing inference-related knowledge.

Lab Safety Knowledge

One surprising finding was the significant difference between the two groups on the pretest of lab safety knowledge, where the VR condition scored higher than the RL condition. This may be due to the differences in the environments where participants took the pretest. The RL participants answered the pretest surrounded by laboratory equipment, which may have triggered anxiety. The VR participants answered the pretest questions in an office space. This may have led to less anxiety for VR participants, making it easier to concentrate. The VR condition demonstrated marginal gains from pre- to posttest in laboratory safety knowledge, while the RL condition had a significant increase. The marginal gains in the VR condition may be due to an already high pretest score, and the significant gains for RL could be due to relaxation of anxiety based on the completion of the experiments. It is also possible that VR's smaller improvement may be due to the lack of physical contact and less safety risks and consequences, confirming that VR environments may desensitize users over time in anxiety-inducing settings (Maples-Keller et al., 2017; North et al., 1997). While lower anxiety is a benefit in most circumstances, too little attention paid in some situations could result in unsafe behavior. In the case of a science laboratory, this could take the form of lowered attentiveness regarding safe laboratory procedures.

Lab Safety Behaviors

Overall, participants in the RL condition showed significantly more laboratory safety behaviors than participants in the VR condition. Both conditions showed a decrease over time (i.e., from the first to the second experiment) in safety behaviors, but a larger decrease in clean-up behavior was seen in the VR condition. RL participants continued to engage in clean-up behaviors at a rate closer to that of the first exercise.

Several possible reasons can be entertained. The most obvious is that VR participants understood that the VR lab environment could easily be cleaned by simply resetting the program, and thus did not abide by expected social or classroom norms. This is similar to other observations of aberrant student behavior in virtual world classroom settings (Wankel & Kingsley, 2009). Also, the VR laboratory environment is relatively awkward to maneuver within when compared to the normal bodily motions and gestures people can use in a physical lab space. As the second experiment required conducting many intricate processes, the participants in VR may have become weary of the unfamiliar gestures and did not want to expend any "extra" effort to clean. Another possibility was that VR participants may have foregone clean-up behaviors during the experiment since there was no real physical danger to leaving virtual chemicals exposed. This type of risk-taking behavior would corroborate with observations made in other virtual training or game environments, where participants felt comfortable exploring the relatively innocuous consequences of unsafe actions (de-Juan-Ripoll et al., 2018; Dickey, 2005). An interesting line of future research could observe the transfer possibility or longevity of these risky behaviors to real-life settings.

The comfort with which some VR participants conducted risky behaviors can be attributable to the ability of virtual environments to provide easily repeatable learning experiences (Heradio et al., 2016; Tatli & Ayas, 2013). This was observed during our study where VR participants "broke" lab equipment and simply restarted the same steps immediately with new digital equipment. Similar to the findings in game-based environments, the VR laboratory environment welcomed exploratory learning habits with immediate feedback, minor consequences, and the opportunity to easily repeat experiences (Berns et al., 2013; Lau & Lee, 2015). Given the opportunity to fail and learning from that failure is a powerful learning strategy (Edmondson, 2011; Straehler-Pohl & Pais, 2014). Easy access to this learning strategy is a distinct advantage of virtual learning environments like VR. Although not specifically measured for in this study, the use of failure-related behaviors in VR, such as repeating steps and exploring alternate problem-solving techniques, could exemplify effective learning strategies in VR .

Another interesting finding was that participants in the VR condition scored higher than the RL condition on laboratory safety *knowledge*, but they performed less *actual* laboratory safety behaviors than the RL condition. Research has found that virtual world users may attempt to apply some real-life social norms but also behave in inappropriate manners (Lau & Lee, 2015; Sherblom et al., 2009). This may be the case for our participants' high score in the laboratory safety knowledge and low score for observed behavior. Alternatively, since the setting of the initial video on laboratory safety knowledge was depicted in a real-life laboratory, participants may have had trouble identifying the same equipment in the VR lab environment.

Also surprising was the finding that the frequency of anxiety comments did not significantly differ between conditions. Although the greater percentage of RL participants who verbalized anxiety hinted at this tendency, it is likely that only measuring verbal comments is not an ideal method for measuring anxiety. Since those feeling anxiety could simply have not verbalized it, a stricter measurement of stress, such as the use of skin conductance measurements, could bring a clearer picture of this anticipated effect. For future research, mitigation of anxiety is still an important potential benefit of VR laboratory experiences. Science anxiety is a researched phenomenon that can cause students to perform poorly or even avoid science classes altogether (Udo et al., 2004, p. 435). As VR has been shown to reduce student anxiety due to its game-like environment, it stands that VR could provide an avenue into science for those who suffer from science anxiety (McLellan, 1994).

Limitations

We acknowledge that there are limitations to this study. First, the age range of the participants was not the ideal age group (11–17 years old) for the lab experience. The gender-skewed sample population also limits the generalizability of the findings of this study. Another limitation was that the effect of the immediate surroundings during the pretest was underestimated and may have caused the significant difference observed between conditions in laboratory safety knowledge pretest scores. Finally, the presence of the experimenter in the room may have adversely affected certain participant behaviors. Recent research by Gallup et al. (2019) discovered that the presence of other people physically nearby a person using VR will inhibit that user's social actions. Therefore, it is possible that our VR participants may have behaved differently if they were conducting the experiment alone.

Practical Implications

VR technology has advanced so far that it can now provide immersive learning experiences even without certain sensory information (i.e., no haptic response on any other body part but the hands, no olfactory, and a relatively limited visual view). Although deserving of a larger discussion, it seems like the technical specifications of technology like the *HTC Vive* (HTC, 2020) are equal to or greater than the level of visual and gestural embodiment needed to access typical chemistry content. As this study shows, similar learning gains can be achieved through the VR simulation of a real-life chemistry laboratory experience.

On the other hand, the observations of laboratory safety behaviors for the VR condition demonstrate that even as immersive as VR is, users will usually be aware that they are in a simulation. As a result, they may behave in a more reckless or risk-taking manner, which would generally be taboo if transferred to a real-life laboratory. VR training simulations often take advantage of this detail to enable users to safely explore the limits of machinery and safety procedures (Zhao & Lucas, 2015). This relative safety of a VR simulation could also contribute to a learner maintaining stronger focus on the scientific processes or content, thus leading to greater abilities in making inferences and applying knowledge. However, with unannounced assessments of certain skills, such as our study observing adherence to lab safety behaviors, this can lead to unintended results. These tradeoffs demonstrate that the effects from embodiment in VR are complex and that conclusions generated from the observations should be made with care.

The actions performed in the VR condition demonstrate that designing digital gestures to be congruent to real-life gestures is important for content learning. Even though the level of contextual detail in the VR laboratory is not nearly as rich as reality, the actions that each participant engaged in were designed to be similar across conditions. The two conditions required users to gather and weigh chemicals, mix substances, and measure temperatures. Hand and finger gestures were mapped on the VR controllers to be as similar as possible to the real motions. Since general context, actions, and physical movements in each condition were comparable, similar amounts of meaning-making could be anticipated. In other contexts, similar setups have been shown to elicit math insights and proofs using dynamic gestures (Nathan & Walkington, 2017) as well as improving physical movements by following a 3D virtual teacher (Patel et al., 2006).

From a design perspective, this places much more responsibility on the designer of the VR experience; attention to the quality and motion of virtual gestures is needed to create an effective learning experience. In this study, the virtual environment contained many motions (i.e., spooning small amounts of chemicals, using a thermometer), which required intricate design to mimic the realistic function of the actual tools. This was not only true for the objects in the lab but also the design of the lab environment itself. Specific artifacts (i.e., periodic table wall poster, traditional lab worktables) of a typical science laboratory were used to properly reproduce its cultural atmosphere. The lack of such details can ruin the illusion of presence and result in users not responding realistically to the virtual environment (Slater et al., 2009). There is much to explore in effective design of VR learning environments.

Our research has important implications in creating foundational knowledge of how learning occurs in immersive learning environments and identifying the type of knowledge (e.g., basic, inferential, applied) and behavior that can

be acquired in those environments. Identifying how learning may or may not differ between a VR and RL chemistry laboratory can inform educators about the pedagogical risks, settings, and approaches that may encourage certain kinds of learning. For instance, it demonstrates that using VR, like real-life hands-on laboratory experiences, can promote learner-centered active learning (Barnes, 1989) in situated learning environments (Lave & Wenger, 1990). These types of environments allow students to discover new knowledge and refine cognitive skills in relation to recognizable situational features that can then be applied to real-world situations. This study addresses an important question if VR environments are to be used as a potential alternative to the traditional real-world science laboratory; can they provide a comparable level of learning? We observe that on many levels, they can.

Conclusion

This study compared students' learning and behaviors between a virtual reality (VR) chemistry laboratory and a traditional real life (RL) chemistry laboratory. Overall, learning performance was comparable across the two conditions, except in applying knowledge, where only VR participants showed a significant increase in their scores. Safe procedural behaviors were performed significantly less often by the VR condition and clean-up behaviors were observed significantly less in the VR condition. There is also some evidence that VR environment may allow the learner to elaborate and reflect more on the general chemistry content and laboratory safety knowledge compared to the RL environment.

VR learning environments pose a unique opportunity to explore learning theory and design. They allow for the fine-tuning of specific details in learning contexts to demonstrate possible cognitive offloading and anxiety-relieving techniques that are unavailable in the real world. There are also undesirable behaviors that students may engage in when in VR, for example, ceasing to clean up, given the virtual environment. This was evidenced by the decline in certain lab behaviors in this study. These types of outcomes need to be thoroughly observed to determine if they are temporary side effects or hints at longer-term consequences. Given the unique global conditions we live in and the affordances of VR technology, it is essential that we explore this innovative technology to determine its place among learning tools.

Funding This material is based upon work partially supported by the Google Virtual Reality Research Unrestricted Gift Award.

Data Availability Data is available upon request to the corresponding author.

Code availability Code is original and produced by the corresponding author. Contact corresponding author to inquire about availability of code.

Declarations

Ethics Approval Conducted with IRB approval from the authors' institution.

Consent to Participate Informed consent was obtained from all individual participants included in the study.

Consent for Publication Participants signed informed consent regarding publishing their data and photographs.

Disclaimer Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of Google LLC.

Competing Interests The authors declare no competing interests.

References

AACT. (2020). Classroom resources: Observing a chemical reaction. Retrieved from https://teachchemistry.org/classroom-resources/observing-a-chemical-reaction

Akuma, F. V., & Callaghan, R. (2019). A systematic review characterizing and clarifying intrinsic teaching challenges linked to inquiry-based practical work. *Journal of Research in Science Teaching, 56*, 619–648. https://doi.org/10.1002/tea.21516

Anderman, E., Sinatra, G. M., & Gray, D. L. (2012). The challenges of teaching and learning about science in the 21st century: Exploring the abilities and constraints of adolescent learners. *Studies in Science Education, 48*(1), 89–117.

Baddeley, A. (1983). Working memory. *Philosophical Transactions of the Royal Society B: Biological Sciences, 302*(1110), 311–324. https://doi.org/10.4324/9781315782119-2

Bailey, J. O., Bailenson, J. N., & Casasanto, D. (2016). When does virtual embodiment change our minds? *Presence: Teleoperators & Virtual Environments, 25*(2), 222–234. https://doi.org/10.1162/PRES

Barab, S., Thomas, M., Dodge, T., Carteaux, R., & Tuzun, H. (2005). Making learning fun: Quest Atlantis, a game without guns. *Educational Technology Research and Development, 53*(1), 86–107. https://doi.org/10.1007/BF02504859

Baragona, M. (2009). *Multiple intelligences and alternative teaching strategies: The effects on student academic achievement, conceptual understanding, and attitude*. (AAI3358373) [Doctoral dissertation, The University of Mississippi.] ProQuest Dissertations and Theses.

Barnes, D. (1989). *Active learning*. Leeds University TVEI Support Project.

Berns, A., Gonzalez-Pardo, A., & Camacho, D. (2013). Game-like language learning in 3-D virtual environments. *Computers & Education, 60*(1), 210–220. https://doi.org/10.1016/j.compedu.2012.07.001

Bourne, E. (2015). *The Anxiety and Phobia Workbook* (6th ed.). New Harbinger Publications.
Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*(1), 288–305. https://doi.org/10.4324/9780203990247
Butt, A. L., Kardong-Edgren, S., & Ellertson, A. (2018). Using game-based virtual reality with haptics for skill acquisition. *Clinical Simulation in Nursing, 16*, 25–32. https://doi.org/10.1016/j.ecns.2017.09.010
Castelvecchi, D. (2016). Low-cost headsets boost virtual reality's lab appeal. *Nature, 533*, 153.
Chao, J., Chiu, J. L., DeJaegher, C. J., & Pan, E. A. (2016). Sensor-augmented virtual labs: Using physical interactions with science simulations to promote understanding of gas behavior. *Journal of Science Education and Technology, 25*(1), 16–33. https://doi.org/10.1007/s10956-015-9574-4
Crippen, K. J., Archambault, L. M., & Kern, C. L. (2013). The nature of laboratory learning experiences in secondary science online. *Research in Science Education, 43*(3), 1029–1050. https://doi.org/10.1007/s11165-012-9301-6
Crosier, J. K., Cobb, S. V., & Wilson, J. R. (2000). Experimental comparison of virtual reality with traditional teaching methods for teaching radioactivity. *Education and Information Technologies, 5*(4), 329–343. https://doi.org/10.1023/A:1012009725532
Dalgarno, B., & Lee, M. J. W. (2010). What are the learning affordances of 3-D virtual environments? *British Journal of Educational Technology, 41*(1), 10–32. https://doi.org/10.1111/j.1467-8535.2009.01038.x
de-Juan-Ripoll, C. M., Soler-Domínguez, J. L., Guixeres, J., Contero, M., Álvarez Gutiérrez, N., & Alcañiz, M. (2018). Virtual reality as a new approach for risk taking assessment. *Frontiers in Psychology, 9*(2532), 1–8. https://doi.org/10.3389/fpsyg.2018.02532
Dede, C. (2009). Immersive interfaces for engagement and learning. *Science, 323*(5910), 66–69. https://doi.org/10.1126/science.1167311
Dickey, M. D. (2005). Engaging by design: How engagement strategies in popular computer and video games can inform instructional design. *Educational Technology Research and Development, 53*(2), 67–83. https://doi.org/10.1007/BF02504866
Edmondson, A. C. (2011). Strategies for learning from failure. *Harvard Business Review*.
Ferrell, J. B., Campbell, J. P., McCarthy, D. R., McKay, K. T., Hensinger, M., Srinivasan, R., Zhao, X., Wurthmann, A., Li, J., & Schneebeli, S. T. (2019). Chemical exploration with virtual reality in organic teaching laboratories. *Journal of Chemical Education, 96*(9), 1961–1966. https://doi.org/10.1021/acs.jchemed.9b00036
Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1319030111
Gallup, A. C., Vasilyev, D., Anderson, N., & Kingstone, A. (2019). Contagious yawning in virtual reality is afected by actual, but not simulated, social presence. *Scientific Reports*, *9*(294). https://doi.org/10.1038/s41598-018-36570-2
Gibson, J. J. (1966). *The senses considered as perceptual systems*. Houghton Mifflin.
Goldin-Meadow, S. (2011). Learning through gesture. *Wiley Interdisciplinary Reviews: Cognitive Science, 2*(6), 595–607. https://doi.org/10.1002/wcs.132
Halim, L., Rahman, N. A., Wahab, N., & Mohtar, L. E. (2018). Factors influencing interest in STEM careers: an exploratory factor analysis. *Asia-Pacific Forum on Science Learning and Teaching, 19*(2).
Hamilton, D., McKechnie, J., Edgerton, E., & Wilson, C. (2020). Immersive virtual reality as a pedagogical tool in education: a systematic literature review of quantitative learning outcomes and experimental design. *Journal of Computers in Education*. https://doi.org/10.1007/s40692-020-00169-2
Heradio, R., De La Torre, L., Galan, D., Cabrerizo, F. J., Herrera-Viedma, E., & Dormido, S. (2016). Virtual and remote labs in education: a bibliometric analysis. *Computers & Education, 98,* 14–38. https://doi.org/10.1016/j.compedu.2016.03.010
Hinojo-Lucena, F. J., Aznar-Díaz, I., Cáceres-Reche, M. P., Trujillo-Torres, J. M., & Romero-Rodríguez, J. M. (2020). Virtual reality treatment for public speaking anxiety in students. Advancements and Results in Personalized Medicine. *Journal of Personalized Medicine, 10*(1), 14. https://doi.org/10.3390/jpm10010014
Hofstein, A. (2004). The laboratory in chemistry education: Thirty years of experience with developments, implementation, and research. *Chemical Education Research and Practice, 5*(3), 247–264. https://doi.org/10.1039/b4rp90027h
HTC. (2020). *Vive Series Specs & Details*. https://www.vive.com/eu/product/vive/#vive-spec
Itō, M. (2010). *Hanging out, messing around, and geeking out: Kids living and learning with new media*. MIT Press.
Jenson, C. E., & Forsyth, D. M. (2012). Virtual reality simulation: Using three-dimensional technology to teach nursing students. *Computers, Informatics, Nursing, 30*(6), 312–318. https://doi.org/10.1097/nxn.0b013e31824af6ae
Johnson-Glenberg, M. C., & Megowan-Romanowicz, C. (2017). Embodied science and mixed reality : How gesture and motion capture affect physics education. *Cognitive Research: Principles and Implications, 2*(1), 24. https://doi.org/10.1186/s41235-017-0060-9
Johnstone, A. H. (1997). Chemistry teaching-science or alchemy? *Journal of Chemical Education, 74*(3). https://pubs.acs.org/sharingguidelines
Jones, A. L., & Stapleton, M. K. (2017). 1.2 million kids and counting-mobile science laboratories drive student interest in STEM. *PLoS Biol, 15*(5), e2001692. https://doi.org/10.1371/journal.pbio.2001692
Keiner, L., & Graulich, N. (2021). Beyond the beaker: Students' use of a scaffold to connect observations with the particle level in the organic chemistry laboratory. *Chemistry Education Research and Practice, 22*(1), 146–163. https://doi.org/10.1039/d0rp00206b
Krathwohl, D. R. (2002). A taxonomy for learning, teaching, and assessing: a revision of Bloom's taxonomy of educational objectives, (Abridged Edition). *Theory Into Practice, 41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2
Lau, K., & Lee, P. (2015). The use of virtual reality for creating unusual environmental stimulation to motivate students to explore creative ideas. *Interactive Learning Environments, 23*(1), 3–18.
Lave, J., & Wenger, E. (1990). *Situated learning: Legitimate peripheral participation*. Cambridge University Press.
Lindner, P., Miloff, A., Hamilton, W., Reuterskiöld, L., Andersson, G., Powers, M. B., & Carlbring, P. (2017). Creating state of the art, next-generation virtual reality exposure therapies for anxiety disorders using consumer hardware platforms: Design considerations and future directions. *Cognitive Behaviour Therapy, 46*(5), 404–420. https://doi.org/10.1080/16506073.2017.1280843
Loke, S. K. (2015). How do virtual world experiences bring about learning? a critical review of theories. *Australasian Journal of Educational Technology, 31*(1), 31. https://doi.org/10.14742/ajet.2532
Maples-Keller, J. L., Bunnell, B. E., Kim, S., & Rothbaum, B. O. (2017). The use of virtual reality technology in the treatment of anxiety and other psychiatric disorders. *Harvard Review of Psychiatry, 25*(3), 103–113. https://doi.org/10.1097/HRP.0000000000000138
Maxwell, S. E., Delaney, H. D., & Kelley, K. (2017). *Designing experiments and analyzing data: A model comparison perspective*. Routledge.
McLellan, H. (1994). *Virtual realities. In The Handbook of Research for Educational Communications and Technology*. AECT.
Merchant, Z., Goetz, E. T., Keeney-Kennicutt, W., Cifuentes, L., Kwok, O., & Davis, T. J. (2013). Exploring 3-D virtual reality technology

for spatial ability and chemistry achievement. *Journal of Computer Assisted Learning, 29*(6), 579–590. https://doi.org/10.1111/jcal.12018

Morell, V. (1994). Novel course III: Undergrad labs "get real." *Science, 266*(5186), 870–872. Retrieved from http://www.jstor.org/stable/2885577

Nathan, M. J., & Walkington, C. (2017). Grounded and embodied mathematical cognition: Promoting mathematical insight and proof using action and language. *Cognitive Research: Principles and Implications, 2*(1). https://doi.org/10.1186/s41235-016-0040-5

National Research Council. (2006). *America's lab report: Investigations in high school science*. Washington, DC: The National Academies Press. https://doi.org/10.17226/11311

NGSS. (2013). *Topic arrangements of the next generation science standards. From A Framework for K-12 Science Education: Practices, Cross-Cutting Concepts, and Core Ideas*. Washington, D.C.: Achieve, Inc.

North, M. M., North, S. M., & Coble, J. R. (1997). Virtual reality therapy: an effective treatment for psychological disorders. *Stud Health Technol Inform., 44,* 59–70.

Oviatt, S. (2013). Designing integrated interfaces that stimulate activity. *The Design of Future Educational Interfaces* (pp. 211–234). Routledge.

Patel, K., Bailenson, J.N., Hack-Jung, S., Diankov, R., & Bajcsy , R. (2006). The effects of fully immersive virtual reality on the learning of physical tasks. *Proceedings of PRESENCE 2006: The 9th Annual International Workshop on Presence*. August 24 – 26, Cleveland, Ohio, USA.

PSI Chemistry. (2018). Observing chemical reactions lab. *New Jersey Center for Teaching & Learning*. Retrieved from http://content.njctl.org/courses/science/chemistry/atomic-origins/observing-chemical-reactions-lab/observing-chemical-reactions-lab-2015-08-14.pdf

Psotka, J. (1995). Immersive training systems: Virtual reality and education and training. *Instructional Science, 23*(9), 405–431.

Rubio-Tamayo, J. L., Barrio, M. G., & García, F. G. (2017). Immersive environments and virtual reality: Systematic review and advances in communication, interaction and simulation. *Multimodal Technologies and Interaction, 1*(4), 1–20. https://doi.org/10.3390/mti1040021

Salzman, M., Dede, C, Loftin, R., & Chen, J. (1999). A model for understanding how virtual reality aids complex conceptual learning. *Presence: Teleoperators and Virtual Environments*, *8*, 293–316.

Scheiter, K., Gerjets, P., Huk, T., Imhof, B., & Kammerer, Y. (2009). The effects of realism in learning with dynamic visualizations. *Learning and Instruction, 19*, 481–494. https://doi.org/10.1016/j.learninstruc.2008.08.001

Seymour, N. E., Gallagher, A. G., Roman, S. A., O'brien, M. K., Bansal, V. K., Andersen, D. K., & Satava, R. M. (2002). Virtual reality training improves operating room performance: Results of a randomized, double-blinded study. *Annals of Surgery, 236*(4), 458–464. https://doi.org/10.1097/01.SLA.0000028969.51489.B4.

Sherblom, J. C., Withers, L. A., & Leonard, L. G. (2009). Communication challenges and opportunities for educators using second life. In C. Wankel & J. Kingsley (Eds.), *Higher Education in Virtual Worlds* (pp. 29–46). Emerald Group Publishing.

Shin, D. H. (2017). The role of affordance in the experience of virtual reality learning: Technological and affective affordances in virtual reality. *Telematics and Informatics, 34*(8), 1826–1836.

Slater, M., Lotto, B., Arnold, M. M., & Sanchez-Vives, M. V. (2009). How we experience immersive virtual environments: The concept of presence and its measurement. *Anuario De Psicología, 40*(2), 193–210.

Straehler-Pohl, H., & Pais, A. (2014). Learning to fail and learning from failure - ideology at work in a mathematics classroom. *Pedagogy, Culture and Society, 22*(1), 79–96. https://doi.org/10.1080/14681366.2013.877207

Standen, P. J., & Brown, D. J. (2006). Virtual reality and its role in removing the barriers that turn cognitive impairments into intellectual disability. *Virtual Reality, 10*(3–4), 241–252. https://doi.org/10.1007/s10055-006-0042-6

Tatli, Z., & Ayas, A. (2013). Effect of a virtual chemistry laboratory on students' achievement. *Educational Technology & Society, 16*(1), 159–170. Retrieved from http://www.ifets.info/journals/16_1/14.pdf

Tsaparlis, G. (2009). Learning at the macro level: the role of practical work. In J. Gilbert, & D. F. Treagust (Eds.), *Multiple Representations in Chemical Education, Models and Modeling in Science Education* (Vol. 4, pp. 109–135). Springer Science + Business Media B.V.

Udo, M. K., Ramsey, G. P., & Mallow, J. V. (2004). Science anxiety and gender in students taking general education science courses. *Journal of Science Education and Technology, 13*(4), 435–446. https://doi.org/10.1007/s10956-004-1465-z

Varela, F. J., Thompson, E., & Rosch, E. (1991). The embodied mind: Cognitive science and human experience. In *The Embodied Mind: Cognitive Science and Human Experience* (1st ed.). MIT Press.

Vogel, J. J., Greenwood-Ericksen, A., Cannon-Bowers, J., & Bowers, C. A. (2006). Using virtual reality with and without gaming attributes for academic achievement. Journal of Research on Technology in Education (International Society for Technology in Education), 39(1), 105–118. https://doi.org/Article

Wankel, C., & Kingsley, J. (2009). *Higher education in virtual worlds: Teaching and learning in second life*. Emerald Group Publishing Limited.

Weisberg, S. M., & Newcombe, N. S. (2017). Embodied cognition and STEM learning: overview of a topical collection in CR:PI. *Cognitive Research: Principles and Implications, 2*(38). https://doi.org/10.1186/s41235-017-0071-6

Wickstrom, L. (1991). *Starting with safety* [Film]. American Chemical Society. https://www.youtube.com/watch?v=9o77QEeM-68

Wieman, C., & Holmes, N. G. (2015). Measuring the impact of an instructional laboratory on the learning of introductory physics. *American Journal of Physics, 83*(972). https://doi.org/10.1119/1.4931717

Winkelmann, K., Scott, M., & Wong, D. (2014). A study of high school students' performance of a chemistry experiment within the virtual world of second life. *Journal of Chemical Education, 91*(9), 1432–1438. https://doi.org/10.1021/ed500009e

Zacharia, Z. C., Manoli, C., Xenofontos, N., de Jong, T., Pedaste, M., van Riesen, S. A. N., et al. (2015). Identifying potential types of guidance for supporting student inquiry when using virtual and remote labs in science: a literature review. *Educational Technology Research and Development, 63*(2), 257–302. https://doi.org/10.1007/s11423-015-9370-0.

Zhao, D., & Lucas, J. (2015). Virtual reality simulation for construction safety promotion. *International Journal of Injury Control and Safety Promotion, 22*(1), 57–67. https://doi.org/10.1080/17457300.2013.861853