2021 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR) | 978-1-6654-3225-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/AIVR52153.2021.00060

ScienceVR: A Virtual Reality Framework for STEM Education, Simulation and Assessment

Hossain Samar Qorbani
School of Information *Technology*
Carleton University
Ottawa, Canada
hossain.samarqorbani@carleton.ca

Ali Arya
School of Information *Technology*
Carleton University
Ottawa, Canada
arya@carleton.ca

Nuke Nowlan
School of Information *Technology*
Carleton University
Ottawa, Canada
nuket.nowlan@carleton.ca

Maryam Abdinejad
Department of Physical and Environmental Sciences
University of Toronto
Toronto, Canada
maryam.abdinejad@mail.utoronto.ca

Abstract—This paper addresses the use of Virtual Reality (VR) in Science, Technology, Engineering, and Math (STEM) education. There are limited studies investigating the proper design and effectiveness of VR in STEM education, and current VR frameworks and applications lack explicit links to the established learning theories and assessment mechanisms to evaluate learning outcomes. We present *ScienceVR*, an educational virtual reality design framework, illustrated through a science laboratory prototype, to bridge some of the gaps identified in the design and development of a VR environment for learning. We established design guidelines and implemented an in-app data collection system to measure users' learning, performance, and task completion rate. Our evaluation using ANOVA and other non-parametric methods with 36 participants in three groups: immersive VR (IVR), desktop VR(DVR), and 2D indicated improved usability and learning outcomes for the IVR group. Task completion rate in the IVR group was higher (68% compared to DVR with 50%). For memorability, the IVR condition performed better than DVR while for learnability, IVR&DVR performed significantly better than 2D. IVR group has performed better and faster with more accuracy compared to the DVR group in completing the tasks.

Keywords—Virtual Reality, STEM Education, Immersion, Simulation, Interaction, Education and Technology

I. Introduction

Learning and teaching Science, Technology, Engineering, and Mathematics (STEM) has unique difficulties, in part, due to the abstract nature of concepts in these disciplines and also because of the range of issues such as cognitive load, required spatial thinking, and experiential and hands-on learning requirements [1][2][3]. Several scientific topics in STEM need a dedicated physical laboratory[4]. These labs often contain toxic and dangerous materials which can present a high level of danger if not maintained and handled carefully. Maintaining such physical facilities and the safety considerations around any science laboratory can be very costly.

Virtual Reality (VR) is increasingly used in games and training applications thanks to its immersion and engagement affordances [4][5]. The increasing application of VR in different training scenarios presents an opportunity to investigate the efficacy of this technology in STEM education where learners can manipulate 3D objects and gain a better understanding of complex subjects. However, current approaches to use VR in STEM education still suffer from a series of limitations. The lack of fidelity/authenticity, the absence of explicit and proper connection to the established learning theories, the lack of evaluation/assessment tools, and accessibility issues are frequently named as existing gaps in the literature related to the VR applications in education [6][7][8][9][10].

To address these shortcomings, we propose and investigate *ScienceVR*, a VR framework for STEM education that is authentic (for both content and interaction), customizable, learner-centred, and accessible. In this paper, we study Immersive VR (IVR) using Head-Mounted Displays (HMDs) and Desktop VR (DVR) experiences and compare them to the traditional two-dimensional (2D) method using text/video. We report on how a VR design based on established learning theories and an in-app data collection system can improve usability, assessment, and learning outcomes in ScienceVR compared to traditional 2D teaching methods. The accessibility in our framework is the subject of another study. Our specific research questions are as follow:

1. Can VR be more effective in teaching science lab procedures compared to the current teaching method (text-based /2D)?
2. How does immersive and desktop VR compare in usability, navigation, visualization, object manipulation, and efficiency in completing assigned tasks and sense of presence?
3. How to implement an in-app data collection system without relying on external/3rd party tools to measure user performance? What are the advantages of such a system?
4. What is the impact of immersion and sense of presence and on the engagement level and learning outcomes?

Focusing on a chemistry lab as an example, our study shows that VR can be effective in improving usability and learning and that immersive VR has the potential to outperform desktop VR. We also show that an in-app data collection system can provide process metrics that have potential in assessment and student guidance while further studies are required to measure the exact effect of these metrics. The framework can be enhanced with Artificial Intelligence (AI) tools for agents and data analysis and recommendation using the in-app metrics. In the following sections, we provide a review of related work, followed by design guidelines, experimental setup, results, some discussions, and concluding remarks.

978-1-6654-3225-2/21/$31.00 ©2021 IEEE
DOI 10.1109/AIVR52153.2021.00060

II. Related Work

A. Common characteristics of STEM disciplines

STEM education has specific characteristics that distinguish it from generic educational approaches. Proper design and utilization of educational technologies will not be possible without understanding these distinctive characteristics. Among them, the following can be highlighted [1][2][3][4][5]:

- **Abstract concepts:** Many STEM lessons deal with abstract concepts that require extra efforts and training by educators to provide various learning tools and technologies to convey such concepts to their students.
- **Spatial Skills:** To translate textbook representations of abstract concepts into understandable ideas, many students may need additional skills, particularly spatial thinking to learn these complex concepts.
- **Text-based content vs. 2D vs. 3D**: Many STEM disciplines such as Chemistry deal with complex abstract concepts that make them extremely challenging to comprehend, particularly if educators only rely on their textbooks and text-based content.
- **Laboratory setup and space:** Many scientific concepts in STEM education require a dedicated laboratory within an existing school building that is expensive to build and maintain.

  Other characteristics: In many STEM lessons, activities/experiments are hands-on inquiry and exploration. These activities can engage students both individually and through teamwork.

B. Applicable learning theories to VR environments

VR systems offer a series of specific features and affordances that make them particularly helpful for STEM education [10]. Among them, we can mention the followings whose consideration is necessary for the design of an effective VR solution in STEM education [8][14]:

- **Interactivity:** VR provides a different way of interacting with content compared to using a traditional keyboard and mouse.
- **Immersive experience:** Depending on the type of VR explained earlier, the user may feel various degrees of presence with different levels of immersion.
- **Multimedia and multisensory:** Similar to many other digital media platforms, VR displays can show graphic representations such as diagrams, videos, and animations, plus audio and haptic feedback.
- **User engagement and presence:** VR draws attention and can be very engaging. Research indicates that novelty, surprise, or uncertain events can attract students' attention and fascinate them [15][16].
- **Facilitating conceptual learning**: Creating 3D content and animations for scientific concepts and making them available on the faster, smaller mobile headsets are becoming relatively easier every day and facilitate design-based research. Studies suggest that students with difficulties understanding 2D representations of complex and abstract concepts can benefit from presenting data in a 3D virtual environment [17].

The assertion that Augmented and Virtual Reality can enhance learning experiences is grounded mainly in two interdependent theoretical frameworks: situated learning theory [18] and constructivist learning theory [7][19]. Situated learning states that learner efficiency increases by being in the relevant environments and that the level and even topic of learning are dependent on where and how the learner is situated. VR allows a sense of presence and immersion that can simulate various situations and, as such, improve learning. Constructivist learning and related theories such as experiential learning [20]. On the other hand, emphasize the importance of interaction, engagement in activities, and learning by doing. In the absence of physical environments, VR has the potential to provide an engaging alternative. In addition to these, other major learning theories such as Cognitive Load Theory [21], Scaffolding [22], and Signaling [23] have potential contributions in the design of educational VR applications. They suggest practices that can be effectively implemented in VR systems, as demonstrated in the design of our proposed framework.

C. VR for STEM education

As suggested by Biocca and Delaney [7], VR is "the sum of hardware and software system" that creates an all-inclusive, sensory illusion of being present in another environment. Three main characteristics of VR technology, as described by Ryan et al. [13] are immersion, presence, and interactivity. The effectiveness of using 2D and 3D content in multimedia applications for science learning has been the subject of many studies. However, from 2000 to 2013, the research on using new technologies such as VR in undergraduate studies has been modest. The main issues related to these studies were the limited scope and small sample size and the experimental nature of the work [24][25].

A literature review by Scavarelli et al. [14] into the related works for VR learning in 2019 identified the primary environments in which VR technology was used to enhance learning. Some of the examples provided in that review include the use of Google's VR ventures with Google cardboard headset and 360 degrees videos.

In 2016, a review was conducted by Gutierrez et al.[26] to examine "Virtual Technologies Trend in Education". It was intended to provide a "comprehensive understanding of AR/VR technologies" and discuss the possibility of using them in education. The authors refer to the amount of investment made by big companies to develop headsets and platforms while emphasizing the fact that creating and developing authentic experiences is necessary to make this technology popular.

A series of studies exploring the potential of using AR and VR in STEM from 2013 to 2019 were reviewed that reaffirm the need to create methods for assessment, and a more polished, practical immersive content (for STEM domains), that can be used by students [7][11][27][28][29]. As identified in these studies most of the impressive VR applications done thus far still need to run on the tethered VR Head-Mounted Displays

(HMDs) that are connected to a powerful computer which is seen as a limiting factor for the user studies). The other related areas of VR studies in STEM include “Atomic Structure” Virtual Lab [27] and “Water Cycle in Nature” [31] being among the small-scale pilot studies that use virtual laboratory concepts and make a comparison between Desktop VR and Text/2D-based content. The authors of these studies point out the difficulties teachers encounter to improve the motivation, engagement, and learning outcomes of students in STEM subjects. Lack of engagement is attributed to the perception that scientific subjects are difficult to learn. There is also a belief that interactive, engaging technology-based educational content can improve/increase learners' engagement [30][31].

Another experimental project by Parong and Mayer [32] was conducted to compare the instructional effectiveness of immersive VR against a desktop slideshow (PowerPoint) for teaching scientific knowledge, as well as examining the efficacy of adding a generative learning strategy to a VR lesson. The content of the experiment was a biology lesson about the human body that was delivered to a group of participants (55-57 college students) with two methods: Immersive VR (IVR) and 2D slides.

There is also a very limited set of studies on the accessibility of educational VR systems which is not within the scope of this paper and is the subject of our future study. Overall, existing research studies show that VR has a strong potential in the context of STEM education. However, several gaps are identified that include lack of authenticity, learning theories, assessment tools, and accessibility. Furthermore, existing works in this area are mostly limited to experimental development and usability testing with limited scope and scale.

III. ScienceVR Framework

A. *Characteristics of ScienceVR*

To address some of the existing gaps in educational VR, especially for STEM, we propose *ScienceVR*, a framework with the following main features as primary goals:

- Authenticity: providing content and interaction that resembles the physical experience and allows the closest simulation.
- Flexibility: customizing the experience for various purposes and users through reusable modules, intelligent tools, and scripts.
- Learning-centricity: basing the experience on educational theories and providing the ability to assess learning.
- Accessibility: offering software solutions that give equitable access to users with limited physical abilities (subject of another study).

Several design elements are used in building immersive VR experiences. The most important of these elements for educational applications are “basic interactions”, “realistic environment”, “immediate feedback” using haptic and visual feedback, “voice instructions”, “traveling/teleportation” (moving around), “virtual rewards” and “knowledge test” which are not often used effectively and together [7]. *ScienceVR* (illustrated in Figure 1) applies all the above design elements and a series of educational theories and pedagogical methods to achieve its four main features.

Fig. 1. Virtual chemistry lab. A prototye for ScienceVR

The authenticity of ScienceVR is primarily rooted in the implementation of content and interaction, and the use of major learning theories including “Situated learning, constructivism, and experiential learning”. According to “situated learning theory”, learners’ efficiency increases by being in the relevant environments; and based on “constructivism and experiential learning theories”, learning is an active, constructive, and goal-oriented process [7][18][19]. The virtual environment in ScienceVR is designed to replicate a science lab as exists in the real world. It uses photo-realistic textures, physics-based lighting, rendering, and reflections to create an authentic ambiance and equipment. This realistic surrounding enables learners to see themselves in an authentic-looking lab. Natural body movement such as the ability to walk around (within the safety limit of the headsets), and being able to reach/grasp (by moving arms and hands) created a natural flow for the users while interacting with virtual objects.

In addition, various forms of feedback such as voice messages and visual highlights offer a more engaging and informative user experience. Such feedback is in line with “Signaling” theory in learning which refers to the auditory or visual cues to assist learners to select and organize key information if a large pool of information is being conveyed [20]. To avoid overwhelming participants with visual stimuli that could potentially cause “cognitive load”, and inspired by scaffolding theories in education, the information was presented in a gradual, incremental method with tutorials and learning levels. Any extra information such as help tips was hidden and users had the option to make them appear when needed.

The in-app data collection system is a part of ScienceVR that makes it more feasible to collect process data and assess learning through observation of what the learner does throughout learning. This is more effective than relying on purely outcome-based measures such as tests or various external tools to collect data. The assessment tools and the essential use of educational theories in the design of the ScienceVR experiences are the foundation of learning-centricity in our proposed framework. It should be noted that the common practice for data collection in VR applications and games is to either rely on game engine tools (such as Unity Analytics) or 3rd party ones (such as Google Firebase). These are usually limited and often complicated, while many VR frameworks do not provide access to any public Application Programming Interface (API) for data collection.

ScienceVR follows a flexible design methodology. This virtual space can be further developed with modular components to make it even more customizable and simulate any science lab. Modularity can include adding customizable dimensions and floor maps to build a specific lab. By using a modern game engine to develop this framework we could export the output for multi-platform applications including different models and brands of HMDs and/or Desktop computers. Embedding specific accessibility features for users with accessibility needs (wheelchair-bound users for example) enables a wider range of audiences to use this framework.

B. *Implementation*

Our ScienceVR prototype was built for a Chemistry lab using Unity 3D which is a popular game engine to build 2D, 3D, and VR games and experiences accessible on desktop, mobile, and HMDs. For the current study, ScienceVR is only built for HMD and desktop:

- **HMD:** Due to the practicality and feasibility of the Android-based mobile headset, the Oculus Quest device was selected as the device of choice for immersive VR.
- **Desktop:** A Windows desktop version of the ScienceVR app was built for the desktop and could run on any laptop or desktop computer with Windows 10 OS.

From the production and development point of view, we kept the workflow simple to allow flexibility yet maintain the mood and feel of a science lab as realistic as possible. To minimize the production time a variety of ready-made assets were acquired from Unity and 3rd party asset stores. Other 3D objects were modelled using Autodesk 3Ds Max software. Figure 2 shows the virtual environments for our prototype.

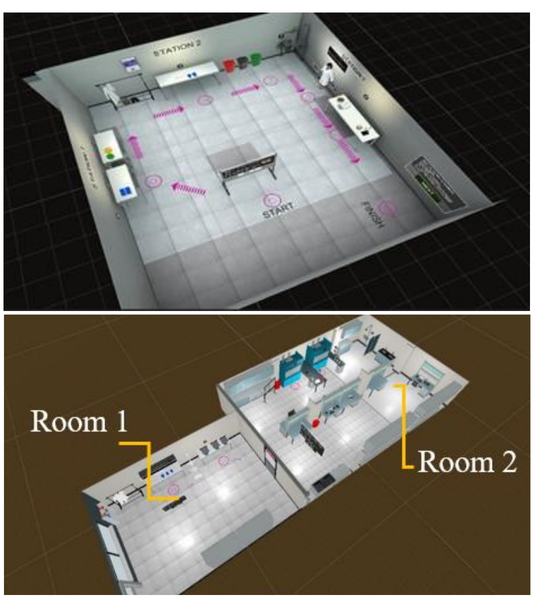

Fig. 2. (top) VR basic training area, (bottom) lab training, and testing areas

The virtual environment has three areas for training and testing purposes. Since our target audience had limited or no experience using VR, we included a basic training area to help them gain experience using touch controllers. This was an important design consideration based on scaffolding learning theory to help users acclimatize themselves and avoid potential motion sickness for some users.

- **Area one:** This area was designed to provide basic training activities such as picking up simple objects (cubes, sphere), which enabled participants to learn how to use touch controllers (Figure 2-top). Since a VR experience can overwhelm first-time users, participants were guided to interact with simple objects, travel, or teleport (locomotion), pick up objects and use help tips. This level was build based on the scaffolding principle in learning theories to help participants build skills and knowledge necessary to navigate and interact with objects.
- **Areas two and three:** While area one was for generic VR training, areas two and three offered specific Chemistry lab experience. Divided into two sections separated by a wall and a door (Figure 2-bottom), these two areas are the virtual Chemistry lab; area two is for more advanced interactions and safety training, including personal protective equipment (PPE) and safety questions, while area three is the actual lab involving the simulated chemistry experiment, equipment, and scientific experiment stations known as fume hoods.

IV. Evaluation

We conducted studies to answer four research questions: (1) Are VR systems more effective in teaching Science lab procedures compared to the current teaching method (text-based /2D)? (2) How do immersive and desktop VR compare in usability, navigation, visualization, object manipulation, and efficiency in completing assigned tasks and sense of presence. (3) How can we develop and apply an in-app data collection system without relying on a 3rd party tools to collect performance metrics? (4) Are the collected data a better indicator of learning compared to the knowledge test?

We used a chemistry lab safety procedure as an example for the educational context. In the traditional safety training format, students are given a standard safety manual based on Workplace Hazardous Materials Information System (WHMIS) for each lab experiment, and the safety process is explained to them by the instructor before entering the lab in a 10–15-minute class session. Students are then asked to sign a waiver confirming that they read and understood the safety procedure. We have simulated this procedure for three environments: Desktop VR (DVR), Immersive VR (IVR) for a head-mounted display, and 2D (Text/Video-based). DVR and IVR were the implementations of our proposed ScienceVR framework, while the 2D case was existing teaching material. This experiment was built in consultation with our partner instructor collaborating with us as a domain expert in Chemistry education. The study was approved by the institutional Ethics Board and conducted online and with downloaded programs and delivered equipment for IVR. A realistic virtual environment was built for this study, and a safety training procedure with two accident scenarios was added to this virtual environment to help us study users' interaction with the environment and record their reactions.

Participants were guided through a few signages and guiding tips to complete tasks that included finding/locating objects, touching/picking up objects, and reacting to accident scenarios. After completing the training in each area, participants were asked to complete the testing round without showing them any of the help tips or guides. They were informed that the accuracy and efficiency (speed) of their performance would be recorded. All participants completed a post-knowledge test and usability survey for their training environment at the end.

A. Study Design

The design of our experiment is "between subjects" with one independent variable (the environments to compare) and a series of lab tasks to perform (measured through dependent variables, which are our evaluation criteria). The environment has three conditions: 2D, IVR, and DVR. The experiment for each group has two parts:

- Training round using the selected environment. This was the main part of the research and included pre and post-knowledge tests to evaluate learning and knowledge improvement, in addition to a post-experience usability survey.
- Practical test using VR. This part should have been done within the physics lab, but due to COVID restrictions, we tested IVR and DVR groups using their environment.

The tasks for participants included common lab activities compiled by the domain expert (a Chemistry instructor and an actual lab procedure). The measurements are based on the primary goal (learning) plus common HCI metrics such as efficiency, accuracy, and general usability. Objective and subjective data were collected. The subjective usability evaluation included ease of use, memorability, learnability, pleasantness, clarity, visualization, and overall satisfaction. We also included a set of presence questions for the immersive VR.

TABLE I. LIST OF THE TASKS/PROCEDURES EXPECTED FOR EACH PARTICIPANT

Task	Subtask	Objects to interact with or identify			
Grasp	Select	PPE			
Verify	Check	PPE	Cont,lens		
Locate	Identify	Sink	Red bin	Green bin	Black bin
Travel	-	Sink	Red bin	Green bin	Black bin
Select	-	Sink	Red bin	Green bin	Black bin
Manipulate	-	Lab props and equipment			

For the detailed tasks outlined in Table 1, we collected objective and subjective data using usability questionnaires, in-game logged data, and knowledge tests. By collecting objective data, we can validate if users are completing the assigned tasks and what are the success and error rates. For objective data (effectiveness and efficiency), we used the logged data to track and record the interactions automatically during the experiment. This feature will eliminate errors in data collection and will not interrupt the process as participants will not experience any intermittent distraction. Considering the criteria mentioned above we have developed the following hypotheses:

Hypothesis 1: Learning outcomes will be positively impacted by the training environment. Comparing all three environments, we expect the IVR score to be the highest. Considering the multiple definitions for learning [33], in this study we refer to the learning that depends on "experience" as the source of the information to be learned.

Hypothesis 2: For the usability across three platforms, the 2D condition (in the form of 2D/ video) will rank higher than IVR and DVR for ease of use, and the IVR will be in second place. However, for other areas such as learnability, visualization, memorability, and overall satisfaction IVR and DVR will rank higher.

Hypothesis 3: Based on the data from the in-app data collection system, for the combination of efficiency and accuracy in task completion, we hypothesized that DVR would rank higher than IVR due to the familiarity of users with the interactions and desktop environment. We can not test for efficiency and accuracy in 2D as we have no access to physical labs during a pandemic.

Hypothesis 4: IVR environment will create a strong sense of presence and immersion, creates an engaging experience where participants stay longer in the simulation. It will positively impact learning.

B. Participants

We recruited 38 participants, students in Chemistry or any other Science or Engineering program at the undergraduate level. Internet accessibility and willingness to use a VR headset were the criteria for participation. Of 38 volunteers in this study, two participants were excluded from the results due to an incomplete experiment caused by motion sickness in IVR and one incomplete post-experiment test and survey.

C. Instruments

Participants in all three groups were asked to complete a survey asking demographic questions including age, gender, handedness, and the familiarity of VR. In addition, the following data collection instruments were used:

Knowledge Tests: Pre-Post knowledge tests were designed using questions from the WHMIS safety manual.

In-app data: A built-in data collection system of IVR and DVR was used to track and record users' interactions.

Usability Survey: A post-study usability survey was conducted based on a 7-point Liker scale (1 very negative and 7 very positive), on ease of use, memorability, learnability, pleasantness, clarity, visualization, and overall satisfaction, followed by one open-ended question to ask how they felt about the experience.

D. Procedure

Participants were asked to complete a series of tasks detailed in table 1 within the training and testing area respectively. Each task included one or two lower-level subtasks. For example, task 1 of grasping an object such as personal protective equipment (PPE) is composed of three subtasks in IVR: looking around, choosing an object (by pointing a ray cast, or extending the arm), and a “grasping” gesture. In DVR, on the other hand, the same tasks would be done by clicking on the object to add it to the inventory confirming that the object is “grasped”.

After completing the training round, participants were asked to do the testing round without showing them any of the help tips or guides to complete the experiment. They were informed that the accuracy and efficiency of their performance would be recorded.

E. Results

1) Demographic Information

Of 36 participants, 20 were males and 16 were females. The average age was 25 with a standard deviation of 6.45. All participants were university students in Chemistry or other science/engineering program. On average they had taken 5.10 chemistry courses with a standard deviation of 3.70. 94% had already completed WHMIS safety training. 66% had prior experience in immersive VR with a variety of games including “Beat Saber”, “Super-Hot demo”, 360 degrees experience e.g. roller coaster, flying in a plane, “Walking dead” etc. The remaining 34% were aware of VR but had never experienced it.

2) Learning Outcomes

The learning outcomes for three environments were analyzed based on pre and post-knowledge test results. Scoring was done based on a point system for a pre-post knowledge test out of 9 points. Pre and post-tests were at the same level but different, designed by the collaborating instructor.

The results for learning outcome were separated for pre, post and delta, (i.e., the difference between pre and post knowledge scores), with samples 1, 2, and 3 corresponding to 2D, IVR, and DVR. We performed three sets of one-way independent analysis of variances (ANOVA) with test scores across all three groups. ANOVA on the Pre-Knowledge Test showed no significant difference, indicating that all groups were at the same level of prior theoretical knowledge. However, for Post and Delta (Post minus Pre), ANOVA showed significant differences (Table 2), and further Tukey HSD post hoc test showed significant differences in the cases of IVR vs. 2D and DVR vs. 2D, with no significant difference for IVR vs. DVR. Table 3 shows the mean and standard deviation for all sets of score data. While the mean values are similar for the pre-test, VR participants scored higher at the post.

TABLE II. ANOVA SUMMARY -INDEPENDENT SAMPLES K=3

	PRE		POST		DELTA	
Source	Treatment	Error	T.	Err.	T.	Err.
SS	7.326	46.99	149.94	131.24	95.06	172.25
df	2	33	2	33	2	33
MS	3.66	1.42	74.97	3.97	47.53	5.21
F	2.57		18.58		9.11	
P	0.091770		0.0001		0.0007	

TABLE III. KNOWLEDGE TEST MEAN SCORE AND STANDARD DEVIATION FOR 2D, IVR, AND DVR

	2D	IVR	DVR
Pre	M = 6.06 SD = 1.48	M = 6.71 SD = 0.98	M = 7.15 SD = 1.06
Post	M = 3.00 SD = 1.26	M = 7.19 SD = 1.98	M = 7.46 SD = 2.53
Delta	M = -3.05 SD = 1.91	M = 0.48 SD = 2.56	M = 0.31 SD = 2.34

An analysis of covariance (ANCOVA) was conducted on the post-test scores of the three groups with gender types as covariates. With the value of $F\ (2,33) = 1.436,\ p < 0.240,$ the effect of gender on the difference between the pre and post-knowledge tests was not significant.

3) Usability

A set of Likert scale questions (1=lowest 7=highest) on the ease of use, memorability, learnability, pleasantness, clarity, visualization, and overall satisfaction for each environment was administered to determine the perceived levels of usability. Table 4 shows the mean and standard deviation for all usability data across three groups. Based on the mean values, participants found 2D easier to use, which is expected, but in other questions, IVR/DVR seemed preferred.

TABLE IV. USABILITY MEAN SCORE AND STANDARD DEVIATION FOR 2D, IVR, AND DVR

	2D	IVR	DVR
Q1 (Ease of use)	*M = 6* *SD = 0.6*	*M = 4.8* *SD = 1.1*	*M = 4.3* *SD = 1.7*
Q2 (Memorability)	*M = 4.7* *SD = 0.7*	*M = 5.8* *SD = 1.5*	*M = 5.7* *SD = 1*
Q3 (Learnability)	*M = 4.5* *SD = 0.7*	*M = 5.6* *SD = 1.6*	*M = 5.6* *SD = 0.9*
Q4 (Pleasantness)	*M = 4.4* *SD = 0.8*	*M = 4.7* *SD = 1.4*	*M = 5.3* *SD = 1.2*
Q5 (Clarity)	*M = 4.8* *SD = 0.6*	*M = 5.2* *SD = 1.6*	*M = 5* *SD = 1.3*
Q6 (Visualization)	*M = 4.3* *SD = 0.5*	*M = 5.6* *SD = 1.3*	*M = 5.9* *SD = 0.9*
Q7 (Overall Satisfaction)	*M = 4.8* *SD = 0.6*	*M = 4.8* *SD = 1.1*	*M = 5.3* *SD = 1.2*

Figure 3 shows the overall satisfaction percentage indicating lower overall satisfaction in the 2D group compared to the other two groups confirming our H2.

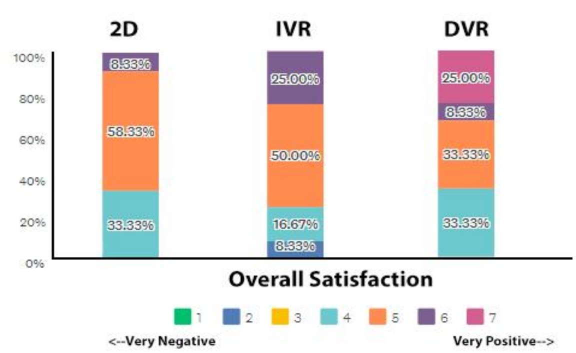

Fig. 3. Overall satisfaction for each group

IVR and DVR showed no significant difference in any of the seven questions based on their mean scores, while 2D seemed superior only in ease of use. For further usability analysis, a non-parametric Kruskal-Wallis test was conducted and the result showed significant differences among the three groups for ease of use, memorability, learnability, and visualization (Table 5). A

Mann-Whitney post hoc test was conducted for each pair of groups and the results (Table 6) indicate that:

- For Ease of Use, there was a significant difference between IVR/DVR and 2D. This was the only case where 2D outperformed VR.
- For Memorability, Kruskal-Wallis showed an overall difference and the paired confirmed that IVR was performing better than 2D, although IVR vs. DVR showed no significant difference.
- For Learnability, Kruskal-Wallis showed an overall difference and the paired test showed a significant difference between IVR/DVR and 2D.
- For Pleasantness and Clarity, there were no significant differences.
- For Visualization, there was a significant difference between IVR/DVR and 2D, as shown by Kruskal-Wallis and Mann-Whitney test. The difference between IVR and DVR was not significant.
- For Overall Satisfaction, Kruskal-Wallis did not show a significant difference.

TABLE V. KRUSKAL-WALLIS TEST FOR 2D, IVR, AND DVR.

	H	df	P
Q1 (Ease of use)	9.51	2	0.0086
Q2 (Memorability)	8.76	2	0.0125
Q3 (Learnability)	7.85	2	0.0197
Q4 (Pleasantness)	3.35	2	0.1873
Q5 (Clarity)	0.58	2	0.7483
Q6 (Visualization)	13.59	2	0.0011
Q7(Overall Satisfaction)	0.35	2	0.8395

*Level of Significance for the test was 0.05

TABLE VI. MANN-WHITNEY TEST FOR 2D, IVR, AND DVR. (SAMPLES M1, M2, AND M3 CORRESPOND TO 2D, IVR, AND DVR RESPECTIVELY)

	Mann-Whitney Post Hoc Test		
Q1 (Ease of use)	M1 vs M2	P=0.0051	U=23
	M1 vs M3	P= 0.0164	U=30
	M2 vs M3	P= 0.4179	U=57.5
Q2 (Memorability)	M1 vs M2	P=0.0078	U=25.5
	M1 vs M3	P= 0.0244	U=32.5
	M2 vs M3	P= 0.4533	U=58.5
Q3 (Learnability)	M1 vs M2	P=0.0404	U=36
	M1 vs M3	P= 0.0067	U=24.5
	M2 vs M3	P= 0.7718	U=66.5
Q4 (Pleasantness)	M1 vs M2	P=0.4009	U=57
	M1 vs M3	P= 0.0643	U=39.5
	M2 vs M3	P=0.4179	U=57.5
Q5 (Clarity)	M1 vs M2	P=0.3843	U=56.5
	M1 vs M3	P= 0.9283	U=70
	M2 vs M3	P= 0.7039	U=65
Q6 (Visualization)	M1 vs M2	P=0.0131	U=28.5
	M1 vs M3	P= 0.0002	U=7.5
	M2 vs M3	P= 0.749	U=65
Q7(Overall Satisfaction)	M1 vs M2	P=0.5222	U=60.5
	M1 vs M3	P= 0.749	U=66
	M2 vs M3	P=0.726	U=65.5

-Level of Significance for the test was 0.05
-Lower limit =42 Upper limit =102

An open-ended question revealed that the IVR group had a mostly positive experience. Participants felt that the IVR was "fun", "enjoyable learning experience", "realistic", "positive and very neat". Three IVR participants (25%) pointed out that they had "experienced slight motion sickness due to teleportation", and "having difficulty to move some objects". One participant mentioned being "very dizzy." All participants were asked to withdraw or take a break if experiencing any discomfort. Similarly, DVR participants had an overall positive experience commenting on how realistic the environment was. Other comments were related to some difficulties in "picking up the objects and rotating them" using keyboard/mouse combinations. Participants from the 2D group commented on how it was easy to run and watch the video but difficult to read and concentrate on the text version. We conclude that the students had a more positive experience and feelings about learning in immersive and desktop VR than learning from a 2D method.

4) Efficiency /Task Completion (In-app data)

Due to COVID-19 restrictions, we could not test in the lab. therefore, efficiency and task completion measurements were done only in VR. Looking at the data generated by the participants' interactions in IVR and DVR, we first made a list of questions to create an overview of how participants performed inside the IVR and DVR and how they interacted with the environment. These questions included: How many participants in each group conducted and completed the training and testing level (by gender)? How long did it take on average to complete the tasks? Type and frequency of interactions (Hover, Select) and average duration for the main interactions.

In each group, we had 12 participants, (7 M, 5 F). For the duration, four major time factors were measured: Total time/duration spent in each environment (in Minutes), the number of data points generated (via the interaction, i.e., hover or select), duration spent in training area1, duration spent in training area2 and duration spent in the testing area.

A one-way ANOVA was conducted to analyze the difference between IVR and DVR. On average participants spent about 25.50 minutes in IVR simulation compared to 20.36 minutes for DVR. The average number of data generated in IVR was 340 interactions vs. 165 for DVR. Although ANOVA does not show a significant difference in the two environments in the duration and the generated data points, further data analysis revealed that 60% of the IVR participants fully completed both training and testing areas/tasks (68% almost, with minor details missing), while only 33% of DVR participants completed both (50% almost completed). This justifies the fewer number of generated data points and the lower usability score in the DVR compared to the IVR. The completion percentage among female participants in IVR was slightly higher than males with 57%. In the DVR simulation, this percentage was 75% for male participants and 25% for female participants.

From the data review of the IVR and DVR participants, we can see that the IVR group is performing better than the DVR group in terms of conducting the task with more efficiency and accuracy. However, further analysis for this part is required and is in progress.

5) SUS (Presence) survey

A Slater Usoh Steed (SUS) questionnaire [34] was also administered to evaluate participants' sense of presence in IVR. SUS survey shows that nine participants (75%) reported a strong sense of "being there" in the immersive virtual lab. Over 50% of the participants reported that the IVR environment was a reality for them "almost all the time," suggesting that the IVR environment was close to an authentic science lab environment. Over 50% of the participants thought of the virtual space as a place/location they had visited, indicating the virtual space feeling "real" for half of the participants. Several comments by the participants also confirm this notion of how "realistic" it was. Over 58% of participants strongly felt the virtual environment was similar to other places they have been. The overall results indicate that there is a correlation between the feeling of "being" in the virtual environment and the similarity to the "real" world could be an attraction that encouraged participants to stay longer in the virtual environment and complete the tasks, which in turn reflected as the higher completion rate and higher post knowledge test.

V. Discussion and Conclusion

Overall, the evidence suggests that teaching science lab procedures in VR is more effective compared to the current 2D method. The results supported two of our hypotheses, while one was only partially supported, and another was not supported.

Hypothesis 1: Confirmed. The result shown in section two supported H1 that the IVR and DVR training has a positive effect on learning outcomes, as indicated by the higher score in the post-knowledge test. However, there was no significant difference between IVR and DVR in this area. It did not support our expectation that the IVR group will score the highest as the DVR group achieved a better post-test score. While the result shows the DVR group is spending less time in the simulation, it did not necessarily mean it was more efficient or accurate in completing the tasks. On the contrary, the rate of incomplete tasks is higher in DVR, which indicates that in an equal condition, more IVR users are completing the experiments successfully. For the difference in the results of pre and post-knowledge test scores and in-app data, we can deduce that IVR is more impactful in engaging participants and producing a better learning outcome as shown in the data analysis.

Hypothesis 2: Partially confirmed. The usability survey on seven areas of usability partially confirms that the 2D condition ranked higher only in the first area of "ease of use" compared to IVR and DVR, and scored lower in every other area, although the difference was not always significant. This indicates that the 2D group was not as satisfied with the content presentation/delivery method as the VR group. Participants in IVR and DVR groups frequently commented on the "realism" of the environment and how "fun", "engaging" and "game-like" the experience was which indicates the high level of engagement and attraction of the experience. No participant in the 2D group added a similar comment or answer to the open-ended question. As such we conclude that participants had a more positive experience and feelings about learning in immersive VR than the 2D method.

Hypothesis 3: Not confirmed. The results did not show a significant difference between IVR and DVR environments on the task efficiency (time to complete the tasks). It was particularly noticeable since several participants in the DVR group had commented about the difficulties to complete certain tasks such as picking up and rotating objects in the usability survey. Further investigation within the in-app data log revealed that only one-third of the DVR participants have completed both training and testing areas of the experiment which justifies the lower number of generated data points within a similar period for IVR. The accuracy of the task was mostly measured based on the collision detection records within the in-app data and a limited number of observations due to Covid-19 restrictions. From the observed and analyzed data we concluded that the IVR group has performed better and faster with more accuracy compared to the DVR group in completing the tasks supporting our third hypothesis. The results suggest that IVR is more impactful in engaging participants and producing a better learning outcome as shown in the data analysis.

Hypothesis 4: Confirmed. The SUS survey revealed that the majority of IVR participants reported a strong sense of "being there" in the virtual space. The overall results suggest a correlation between the feeling of "being" in the virtual environment and the similarity to the "real" world experience as an attraction and motivating factor to stay longer in the VR and complete the tasks (indicating a higher level of engagement). This feeling has positively reflected on the learning outcome as the higher task completion rate and higher post-knowledge test scores

Research Questions: Based on the results we can state that VR experiences are effective in improving both learning and usability. Developing in-app data collection based on various process metrics proved to be easy to implement and replicate. The In-app data collector was a valuable source of information that gave us a fuller picture of the user's journey within the VR environment through each interaction. The knowledge test may not provide such a comprehensive picture of the process.

Limitations: Comparing 2D, IVR, and DVR showed that 2D training in the form of text and even video format is easier to use, even though the learning outcome and other usability dimensions may be significantly worst compared to IVR and DVR. This illustrates the importance of ease of use, and familiarity which is relevant in DVR vs. IVR. This study was mostly conducted remotely due to the pandemic restrictions impacting our timeline parts of the evaluation. The practical test in this study should have been done in the physical lab. Replacing it with VR gave an advantage to IVR and DVR groups and as such made the results less insightful. The observation part of the study is partially missing since the research team could not be present while participants experimented (either in DVR or IVR) due to the social distancing measures. We used no AI tools within the framework but see potential in such use. Examples are agents to interact with the user, tracking activities and personal characteristics to provide personalized content and experience, and particularly, processing the in-app data for automated assessment, support, and recommendation purposes. We are evaluating various algorithms for automatic assessment which are the topic of another paper.

References

[1] M.E.Madden, M.Baxter, H.Beauchamp, K.Bouchard, D.Habermas, M.Huff, B.Ladd, J.Pearon, G.Plague, “Rethinking STEM Education: An Interdisciplinary STEAM Curriculum”, Procedia Computer Science, vol. 20, pp. 541-546, 2013.

[2] V.Potkonjak, "Virtual laboratories for education in science, technology, and engineering: A review." Computers & Education 95: pp.309-327, 2016.

[3] M. H.Towns , "A Review of Spatial Ability Literature, Its Connection to Chemistry, and Implications for Instruction", Journal of Chemical Education, 10.1021, 2010.

[4] B.M Alain, “To what degree does handling concrete molecular models promote the ability to translate and coordinate between 2D and 3D molecular structure representations? A case study with Algerian students”, The Royal Society of Chemistry, 17: pp. 862-877, 2016.

[5] M.Ibáñez,. and C. Delgado-Kloos, "Augmented reality for STEM learning: A systematic review." Computers & Education 123, pp.109-123, 2018.

[6] T.R. Kelley, J.G.Knowles, J.G. “A conceptual framework for integrated STEM education”, International Journal of STEM Education vol. 3, Issue 11, 2016.

[7] J. Radianti, T. A. Majchrzak, J. Fromm, and I. Wohlgenannt, “A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda”, sciencedirect, Computers & Education, vol.147, 2020.

[8] A.Saeed Alqahtani, L.Foaud Daghestani and L.Fattouh Ibrahi m, “Environments and System Types of Virtual Reality Technology in STEM: a Survey” International Journal of Advanced Computer Science and Applications (IJACSA), vol. 8, No.6, 2017.

[9] C. Anthes, R. J. García-Hernández, M. Wiedemann and D. Kranzlmüller. “State of the art of virtual reality technology,” IEEE Aerospace Conference, pp. 1-19, 2016.

[10] I. Jeelani, K. Han, and A. Albert, "Development of virtual reality and stereo-panoramic environments for construction safety training," Engineering, Construction and Architectural Management, Vol. 27 No. 8, pp. 1853-1876, 2020.

[11] H.A. El-Mounayri, C.R., Indiana, E. Fernandez, J. Connor Satterwhite, , “Assessment of STEM e-Learning in an Immersive Virtual Reality (VR) Environment,” Annual Conference & Exposition, New Orleans, 2016.

[12] N.Nowlan, P. Hartwick, and A.Arya, “Skill Assessment in Virtual Learning Environments,” IEEE International Conference (CIVEMSA), pp. 1-6, 2018.

[13] M.Mott, E.Cutrell, M.G. Franco, C. Holz, E. Ofek, R., Stoakley, and M.Ringel. “Accessible by Design: an Opportunity for Virtual Reality,” ISMAR Workshop on Mixed Reality and Accessibility, 2019.

[14] A.Scavarelli, A.Arya, and R.J. Teather, “Virtual Reality and Augmented Reality in Social Learning Spaces”: A Literature Review. Virtual Reality, vol. 25, pp.257-277, 2020.

[15] X. Chen, “STEM Attrition: College Students’ Paths Into and Out of STEM Fields”, National Center for Education Statistics, 2013.

[16] R.Raman, H. Vachharajani, and K. Achuthan, “Students motivation for adopting programming contests: Innovation-diffusion perspective”, Education and Information Technologies,. 23(5): pp.1919-1932, 2018.

[17] C.H.Chen, and C.H. Chiu, “Employing intergroup competition in multitouch design-based learning to foster student engagement, learning achievement, and creativity”. Computers & Education,. vol.103: pp. 99-113, 2016.

[18] J.S. Brown , A.Collins, P.Duguid, "Situated Cognition and the Culture of Learning", Educational Researcher. vol. 18, no.1, pp.32-42,1989.

[19] M.Dunleavy, C. Dede, and R. Mitchell, “Affordances and Limitations of Immersive Participatory Augmented Reality Simulations for Teaching and Learning”, Journal of Science Education and Technology, 18(1): pp. 7-22, 2008.

[20] A.Y Kolb, D.A. Kolb, “Experiential learning theory”. In Encyclopedia of the sciences of learning (pp. 1215–1219). Springer. (2012).

[21] A.Kolb, and D.Kolb,. "Eight important things to know about the experiential learning cycle", Australian Educational Leader, 40 (3), 2018.

[22] S.A. Ambrose, M. C.Lovett, M.DiPietro, M.K.Norman, “How Learning Works”, Jossey-Bass A Wiley Imprints. 2010.

[23] L.Long, G.Dabrowski and A.Grant "Using 360 Virtual Reality to Make Experiential Learning Accessible to All",Teaching & Learning , 2020.

[24] G.Korakakis, “3D visualization types in multimedia applications for science learning: A case study for 8th grade students in Greece”, Computers & Education, 52 (2): pp. 390-401, 2009.

[25] N.Hoyek, C. Collet, F.Di Rienzo, M.De Almeida,A.Guillot, "Effectiveness of three-dimensional digital animation in teaching human anatomy in an authentic classroom context", American Association for Anatomy, 2014.

[26] J.Martin-Gutiérrez, C.Efrén Mora,B.Añorbe-Díaz A.González-Marrero "Virtual Technologies Trends in Education" EURASIA Journal of Mathematics, Science & Technology Education, vol. 13 No.2, pp469-486, 2017.

[27] M. Akçayır and G. Akçayır, “Advantages and challenges associated with augmented reality for education: A systematic review of the literature,” Educational Research Review, vol. 20, pp. 1–11, 2017.

[28] A. Suh and J. Prophet, “The state of immersive technology research: A literature analysis,” Computers in Human Behavior, vol. 86, pp. 77-90, 2018.

[29] R.Al-Azawi, A. Albadi1, R.Moghaddas1,and J. Westlake2 "Exploring the Potential of Using Augmented Reality and Virtual Reality for STEM Education"-Springer Nature Switzerland. LTEC, CCIS 1011, pp.36-44,2019.

[30] I. Ghergulescu1, T. Lynch1, M. Bratu, A. Moldovan, C.H. Muntean, G.M. Muntean4 "STEM Education With Atomic Structure Virtual Lab For Learners With Special Education Needs", International Conference on Education and New Learning Technologies (EduLearn), 2018.

[31] Bogusevschi, D., Tal, I., Bratu, M., Gornea, B., Caraman, D., Ghergulescu, I., Muntean, G.-M. . "Water Cycle in Nature: Small-Scale STEM Education Pilot” EdMedia + Innovate Learning, Association for the Advancement of Computing in Education (AACE), pp. 1496–1505, 2018.

[32] J.Parong, R. E.Mayer, "Learning Science in Immersive Virtual Reality", Journal of Educational Psychology. Advance online publication Online: www.researchgate.net/publication/322708154_Learning_Science_in_Immersive_Virtual_Reality [accessed Jan 28 2021].

[33] A. B. Barron, E.A. Hebets, T. A. Cleland, C.L. Fitzpatrick, M. E. Hauber, J. R. Stevens, “Embracing multiple definitions of learning,”, sciencedirect: Trends in Neurosciences,vol38, No.7, pp. 405-407, 2015.

[34] M.Usoh, E.Catena, S.Arman, M.Slater, "Using Presence Questionnaires in Reality",MIT,Presence: Teleoperators and Virtual Environments vol.9 No.5, pp.497–503, 2000.