Investigating the Novelty Effect in Virtual Reality on STEM Learning

by

Wen Huang

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2020 by the
Graduate Supervisory Committee:

Rod Roscoe, Chair
Mina Johnson
Scotty Craig

ARIZONA STATE UNIVERSITY

May 2020

ABSTRACT

The science, technology, engineering, and math (STEM) education community is interested in using virtual reality (VR) to help students learn STEM knowledge. Prior research also provided evidence that VR learning can increase students' motivation and learning achievement. However, it was not clear whether the effect of VR on learning was partly from sensory novelty and whether the effectiveness was sustainable. This study was to satisfy the concern on the sustainability of VR STEM learning in instruction and address the research gaps in exploring the effect of VR on a STEM learning experience with a consideration of novelty.

The study used a mixed-methods experimental design and involved a three-session VR STEM learning intervention. The quantitative data was collected through the intervention by survey questionnaire, session quiz, and pre- and post-tests, while the interviews were taken after the intervention. The structural equation modeling method was used to explore the relationships among factors in the VR learning experience. Longitudinal quantitative comparisons were conducted with the multiple imputation method. Its purpose was to evaluate the changing magnitude of factors across sessions. After quantitative analysis, interview transcripts were analyzed. They were used to triangulate or provide context for understanding of quantitative results.

The results showed that motivation and engagement play a critical mediation role in an effective VR learning experience. While individuals' psychological responses and motivation may significantly increase in a VR learning experience for novelty, the novelty effect may not steeply decrease when individuals are becoming familiar with the novelty. This phenomenon is more observable in a VR condition having a high degree of

immersion and embodiment. In addition, novelty does not necessarily increase learning achievement. The increase of learning achievement is more dependent on a match between the learning content and the learning method. The embodied learning method is appropriate for instructing difficult knowledge and spatial knowledge. Reserving enough time for reflection is important to deep learning in a VR environment.

DEDICATION

I dedicate this work to my parents. Your encouragement and love gave me energy to complete the long journey.

ACKNOWLEDGMENTS

I would like to express my deep appreciation and gratitude to my advisor, Dr. Rod Roscoe, for the scholarly guidance and mentorship you provided to me during the last two years in my doctoral study. Especially, I would like to thank your support for my research direction and your patience in solving various issues I encountered during research. Without your support and help, this dissertation would not have been possible.

I am grateful to my committee members, Dr. Mina Johnson and Dr. Scotty Craig. I thank Dr. Mina Johnson for your guidance on the experimental design and sponsorship for the experimental implementation. Your research inspired me to think VR learning from the embodiment direction. I thank Dr. Scotty Craig for providing comments at each critical stage and the convenience for recruiting experimental participants in this study.

I would like to thank the students who showed warm supports for this research. Zhen Zhao, Elena Kalina, Vanessa Ly, Ricardo Nieland, Samuel Arnold, and Chelsea Johnson, I cannot forget your help in the pilot test, material preparation, data collection, and data analysis.

I would like to recognize the contribution from the experimental participants in this study. Most of you completed the task quite well.

In addition, I would like to thank all the faculties in the engineering education program, past and present. The courses and knowledge that I learned in the first two years have laid a solid foundation for my research.

This research is partly supported by the Graduate Research Support Program at Arizona States University. I would like to thank the Office of Knowledge Enterprise

Development, Graduate and Professional Student Association (GPSA), and the Graduate College for providing funding support.

TABLE OF CONTENTS

Page

LIST OF TABLES x

LIST OF FIGURES xiii

CHAPTER

1 INTRODUCTION 1

Research Problem 1

Prior Research on VR Learning 4

Purpose and Research Questions 6

2 LITERATURE REVIEW 8

VR and VR STEM Learning Research 8

Virtual Reality 8

Potential of VR in STEM Education 10

Models in VR Learning 11

Novelty Effect 14

Effectiveness of a VR Condition 16

A Model of an Effective VR Learning Experience 19

Conceptual Framework 19

Factors and Their Relationships 21

Research Model 31

Novelty in VR Learning 33

Novelty Effect in VR Learning 33

Attenuation of the Novelty Effect in VR Learning 34

CHAPTER Page

Summary 36

3 METHODS 39

Participants 40

Materials 41

Experimental Environment 42

Pre- and Post-Tests 43

Instrument Development 44

Session Quiz Development 45

Experimental Design 48

Interview Design 51

Procedure 52

Procedure of the Mental Rotation Test 52

Procedure of the Go-Version VR Training 53

Procedure of the Rift-Version VR Training 54

Procedure of the VR Learning Task 55

Procedure of Other Tasks 56

Data Analysis 56

Phase I: Grading Quizzes 58

Phase II: Structure Equation Modeling 62

Phase III: Data Comparison 65

Phase IV: Qualitative Data Analysis 74

4 RESULTS 77

CHAPTER Page

Grading Quizzes 77

Structural Equation Modeling 78

Exploratory Factor Analysis....78

Descriptive Statistics 84

Assessment of Measurement Model 85

Assessment of Structure Model 89

Data Comparison 92

Establishment of Predictor Model....93

Data Screening....96

Comparison by Category....99

Completer Analysis 111

Qualitative Results....128

5 DISCUSSION 144

Interpretations of Study Results 144

The Impact of a VR Condition on an Effective Learning Experience..144

The Magnitude of Novelty Effects in a VR Learning Experience....147

The Attenuated Effectiveness of a VR Condition 153

The Overall Trends of Assessed Factors 157

Contributions and Implications 161

Theoretical Aspects 161

Methods Aspects....163

Practical Aspects....164

CHAPTER Page

Study Limitations 166

Suggestions for Future Research 168

6 CONCLUSION 170

REFERENCES 172

APPENDIX

A IRB APPROVAL LETTER 190

B QUESTIONNAIRE ITEMS 193

C QUIZ QUESTIONS 196

D EXPERIMENTAL INSTRUCTIONS 207

E PREDICTOR MODEL 220

LIST OF TABLES

Table Page

1. A Comparison Between the Oculus Go and the Oculus Rift Systems 23

2. Experimental Participants 41

3. Relevant Instruments in the Literature 44

4. The Experimental Arrangement for Each Sessions in Each Group 51

5. The Statistics of Dropouts and Their Reasons 66

6. The Number of Participants for Each Session in Each Group 67

7. VR Learning Time 67

8. An Example of a Longitudinal Dataset 70

9. The Grading Results of Session Quizzes 78

10. The Pattern Matrix of the Initial EFA 79

11. The Pattern Matrix of the Second-Round EFA 81

12. The Cronbach's Alpha of Each Factor 81

13. Factor Loadings of Data in the First Session 82

14. Factor Loadings of Data in the Second Session 83

15. Factor Loadings of Data in the Third Session 84

16. Descriptive Statistics of Remaining Data 85

17. Evidence of Convergent Validity for the Measurement Model 87

18. Standardized Factor Correlation Matrix 88

19. Parameter Estimates for the Hypothesized Structural Model 89

20. Parameter Estimates of Indirect Effects for the Structural Model 91

21. Parameter Estimates for the Modified Structural Model 92

Table Page

22. Comparison of Dropout Rate Among Different Age Groups 94

23. Comparison of Dropout Rate Among Different Races 94

24. The Number of Predictors for Target Variables 94

25. Comparison of Participants Who Obtained a Score of Two or Three in the Pre-Tests Among Groups 97

26. The Distribution of Majors Among Groups 98

27. Comparison of Scores in the Pre-Tests Among Groups 99

28. Descriptive Statistics of Presence 100

29. Descriptive Statistics of Sense of Embodiment 102

30. Descriptive Statistics of Motivation 104

31. Descriptive Statistics of Cognitive Load 106

32. Descriptive Statistics of Engagement 108

33. Descriptive Statistics of Quiz Score 109

34. Descriptive Statistics of Pre- and Post-Test Scores 111

35. Comparison of Pre- and Post-Test Scores within Groups 111

36. Descriptive Statistics of Presence for Completers 113

37. Within-Subjects Contrasts of Presence for Completers 114

38. Descriptive Statistics of Sense of Embodiment for Completers 115

39. Comparison of Sense of Embodiment for Completers Between the Go Group and the Rift Group 116

40. Within-Subjects Contrasts of Sense of Embodiment for Completer 117

Table Page

41. Descriptive Statistics of Motivation for Completers 118

42. Within-Subjects Contrasts of Motivation for Completers 119

43. Descriptive Statistics of Cognitive Load for Completers 120

44. Within-Subjects Contrasts of Cognitive Load for Completers 121

45. Descriptive Statistics of Engagement for Completers 123

46. Within-Subjects Contrasts for Engagement for Completers 124

47. Descriptive Statistics of Quiz Score for Completers 125

48. Within-Subjects Contrasts of Quiz Score for Completers 126

49. Descriptive Statistics of Pre- and Post-Test Scores for Completers 127

50. Comparison of Pre- and Post-Test Scores in Each Completer Group 127

51. Comparison of Themes and Quantitative Results 129

52. Comparison of Effects from VR Condition to Presence and Motivation 148

53. Comparison of the Standard Deviation Between Presence and Motivation for the Go-Rift Group 149

54. The Change of Assessed Factors From the First Session to the Third Session for the Go Group and the Rift Group 154

55. Changes of the Mean of Item En_3 Across Sessions 156

LIST OF FIGURES

Figure Page

1. The Conceptual Framework of an Effective VR Learning Experience 20

2. The Hypothesized Model of an Effective VR Learning Experience.....31

3. The Influence of Novelty on Assessed Factors Based on Predictions 37

4. An Example of the Recognizing Questions 46

5. An Example of the Recalling Questions 47

6. An Example of the Understanding Questions.....48

7. An Example of the Evaluating Questions 48

8. The Measurement Model of an Effective VR Learning Experience.....88

9. The Structural Model of an Effective VR Learning Experience.....90

10. Presence for Each Session in Each Group 101

11. Sense of Embodiment for Each Session in Each Group.....103

12. Motivation for Each Session in Each Group.....105

13. Cognitive Load for Each Session in Each Group 106

14. Engagement for Each Session in Each Group 108

15. Quiz Score for Each Session in Each Group.....110

16. Presence for Each Session in Each Completer Group 113

17. Sense of Embodiment for Each Session in Each Completer Group.....116

18. Motivation for Each Session in Each Completer Group 119

19. Cognitive Load for Each Session in Each Completer Group.....121

20. Engagement for Each Session in Each Completer Group 123

21. Quiz Score for Each Session in Each Completer Group 125

Figure Page

22. Mean Scores in Session Quizzes for the Low-Level and the High-Level Cognitive Processing Questions 150

CHAPTER 1

INTRODUCTION

Research Problem

STEM is the acronym of science, technology, engineering, and mathematics. It represents four different subjects in our daily life. An integration of these four subjects offers students a chance to view the world as a whole rather than learn separated pieces of knowledge and skills (Dugger, 2010). In the 21st century, globalization and the knowledge-based economy enhance the importance of science and technological innovations (National Science Foundation, 2007). This puts forward higher requirements for students' capabilities in STEM than before. However, there is a popular belief that it is hard to learn STEM (Pew Research Center, 2018). Particularly, some content in STEM subjects is abstract or hard to reach in life. This is recognized as one of the main reasons causing students to feel frustrated and lose interest in STEM learning (Drew, 2011). Approximately 40 percent of STEM major undergraduate students either switch to non-STEM majors or drop out (Drew, 2011). Concerning this issue, researchers and educators have attempted to use a diversity of digital tools to enhance conceptual understanding, promote thinking skills, and motivate students to learn STEM (European Schoolnet, 2017; Microsoft Corporation, 2016).

Virtual reality (VR) is one of the fastest-growing emerging information technologies at present. Among various audiences (e.g., healthcare & medicine, retail, manufacturing), education has the largest demand for VR in a long term (Superdata, 2018). VR enables instruction and learning beyond the bound of real world. This

advantage makes VR useful in the STEM education field. Many STEM researchers and educators are enthusiastic about applying VR in the STEM education field.

Furthermore, the recently published *Charting a course for success: America's strategy for STEM education* showed that the country encourages expanding digital platforms for instruction and learning in STEM education (Committee of STEM Education, 2018). This strategy suggested that researchers and educators should realize the full benefits of these new tools in content and skills learning. One of the key federal actions is to expand research on the development of curricula using digital tools. Particularly, it mentioned the benefit of virtual environments in increasing the learner's engagement.

However, compared with the scene of prosperity in the academic and education field, there is a significantly different phenomenon in the industry. After the flourishing in 2016 and 2017, the public's interest in VR is gradually waning, and the VR industry is experiencing a harsh winter. While researchers and entrepreneurs predicted the total consumer VR revenue in the world would reach 4.5 billion in 2018, the final result was only 3.6 billion, 20% less than expectation (Superdata, 2018, 2019). A survey questioned 595 professionals from around the world at the end of 2018 revealed that approximately 42% of them selected the growth of VR for consumer is weaker than hoped (VR Intelligence & Superdata, 2018). Goeldi (2018), an active columnist in the Internet technology field, wrote on his blog that only 18 of his total 630 Facebook friends had an account on the Facebook social VR platform, and he had never seen any of them being online; while participants in a public VR chatroom was often only a few dozen to a couple hundred, Facebook commonly had more than 3 million units online. He wrote that

"if my early-adopter friends and I can't get sustainably excited about a new technology, there's nearly zero chance that it will fly in the mainstream" (para. 7).

This contrast is thought-provoking. Entertainment is the first and largest market of VR technology at present. Compared to the entertainment field, the education community's reaction to VR is still at the stage of active exploration. There is no sign of losing interest in VR instruction or VR learning reported from openly published publications. However, Rogers (2003) indicated in his *diffusion of innovation theory* that if there is no rich knowledge to support the use of an innovation in a social system, the diffusion of this innovation in this society will be restrained. In history, researchers have broadly recorded the phenomenon that students or teachers gradually lose interest in learning with various digital learning methods, such as Massive Open Online Courses (MOOCs), virtual community of practice, and mobile learning. Their research revealed that keeping individuals' engagement and persistence in learning with an assistant tool is a challenge in the education field; if no outside factor intervenes, the rate of participation will significantly decline in the learning process (Callum & Jeffrey, 2013; Cheung, C. & Lee, 2009; Higashi, Schunn, & Flot, 2017; Jung & Lee, 2018).

No one is willing to see this situation happening in VR STEM learning. VR has already shown several competitive advantages in supporting STEM education. It can present phenomenons and scenarios which are hard to observe with the naked eye or in the local area (Bell & Fogler, 1997; Borst, Ritter, & Chambers, 2016; Papachristos, Vrellis, & Mikropoulos, 2017); serve as a simulated operational environment, where students can learn how to manipulate the mechanical equipment with more practical opportunities but less financial cost (Crespo, García, & Quiroz, 2015; Mineev, 2017;

Quevedo et al., 2017); enhance the memory and understanding of abstract concepts through embodied motions and gestures (Johnson-Glenberg & Megowan-Romanowicz, 2017; Parmar, 2017).

Considering these significant benefits, the STEM education community should not wait until the public start to lose interest in VR STEM learning, but actively face this potential challenge. This involves an understanding of how VR influences students' STEM learning experience and whether this influence is durable. The understanding may help the community determines whether losing interest will be a problem in the VR STEM learning field in the future. If it is really a problem, the revealed mechanism may help the community more accurately find ideas to solve this problem.

Prior Research on VR Learning

Prior VR STEM learning research provided evidence that using VR in instruction has the potential to increase individuals' motivation and learning achievement in STEM, and students like using VR to learn the STEM knowledge and skills. In addition, researchers have broadly explored how and why VR can benefit learning through modeling. They found that this is because VR offers distinctive affordances (e.g., presence, motivation, and embodiment) to individuals in their learning experience. Nevertheless, it is still not quite clear how the characteristics of VR, or the affordances of VR, affect an actual learning experience and then enhance learning outcomes (e.g., knowledge and cognitive skills). Filling this research gap has the potential to help instructional designers more efficiently use VR to develop STEM courses and support students to successfully reach the expected learning objectives.

Furthermore, many VR STEM learning studies only used either a single or a short-term measurement to collect data. However, previous studies in the digital learning field showed that due to the novelty effect, this data collection method might produce biased estimates. For example, students may show a high motivation to learn STEM knowledge in a technology-enhanced learning environment at the beginning for being curious about this type of learning, but their motivation decreases over time when they become familiar with the surroundings. If data are only collected at the beginning, this study may overestimate the effect of the learning environment on students' motivation. Thus, to reveal the real effect of VR on individuals' learning, researchers should consider the impact of novelty on the effectiveness of VR in learning. If the impact is large, repeated measurements should be used to collect data. However, there is a lack of research on the possible novelty effect in the VR learning field at present, there are also few studies in the VR STEM learning field used repeated measurements to collect data in literature.

Third, minimal research attention has been directed toward instructional methods when the effect of VR on STEM learning is assessed. This is incongruent with the result of prior research on the relationship between media and instructional method in the multi-media learning field, which emphasizes the contribution of instructional method on learning. Different instructional methods may lead to differences in learning experience and learning achievement even when the type of media used is the same. Particularly, prior literature showed that the efficiency of VR learning is not always higher than learning with a desktop virtual environment (VE). Thus, it is important to find the most appropriate instructional method in VR learning and consider the possible differences

caused by instructional methods. In the recent years, researchers have paid attention to the high-end VR systems (e.g., Oculurs Rift and HTC Vive), which allow individuals to actively interact with virtual objects in the learning process. This embodied learning method has the potential to effectively enhance individuals' engagement and their learning achievement. As a new research direction, its application and strength need more explorations. These explorations have the potential to guide the design of VR applications and maximize the potential of different VR conditions in instruction.

Purpose and Research Questions

This study was to satisfy the concern on the sustainability of VR STEM learning in instruction and address the research gaps in exploring the effect of VR on a STEM learning experience with a consideration of novelty. A mixed-methods experimental design (Creswell & Clark, 2017) was used in this study. This is a type of design in which quantitative and qualitative data are embedded within a major intervention. In this study, the quantitative data were collected before, during, and after the intervention by survey questionnaire, session quiz, and pre- and post-tests. They were used to validate a VR learning experience model and longitudinally assess the possible changes of individuals' engagement and learning in different VR conditions. The qualitative data were collected after the intervention by interviews. They were used to triangulate or provide context to understand the quantitative results.

The research questions that lead this study are:

1. How does the VR condition impact an effective VR learning experience?
2. How much does novelty influence the VR learning experience?

3. How does the effectiveness of a VR condition change after the novelty effect happens?

CHAPTER 2

LITERATURE REVIEW

This chapter includes three sections. The first section is a review of prior research in VR STEM learning. It indicates three research gaps and provides a rationale for the study. The second section corresponds to the first research question in this study. It first proposes a conceptual framework, and then establishes a research model. The third section focuses on the second and third research questions. It includes a series of predictions which are related to the effect of novelty on STEM learning in different VR conditions.

VR and VR STEM Learning Research

Virtual reality. Virtual reality (VR) is a computer-generated fully-immersive 3-D multimedia environment that individuals can use special electronic equipment to interact with in a seemly real way (Aukstakalnis & Blatner, 1992; Milgram & Kishino, 1994; Onyesolu & Eze, 2011; Oxford, 2019). It incorporates 3-D technologies and creates a condition that simulates real or imagined life scenarios (Haag, Cummings, & Dawkins, 1998). The head mounted display (HMD) headset is the most common and important tool to implement VR STEM education in schools (Johnson-Glenberg, 2018). It provides a fully interactive 3-D environment to users with a head tracking design. Users can walk, adjust sight, and touch virtual objects in a 360-degree virtual environment with an affordable-price. Currently, there are mainly three types of VR headsets available to the public: (1) A simple headset that uses a smartphone without any other features or functionality, e.g., Google Cardboard; (2) A mid-range headset that either incorporates a

smartphone or be standalone, but offers additional functions, e.g., Samsung Gear VR, Oculus Rift Go; and (3) A high-end headset that can track individuals' position, e.g., Oculus Rift and HTC Vive (Mcmillan, Flood, & Glaeser, 2017). While most of the high-end headsets need to be connected to a high-performance personal computer, all-in-one headsets (e.g., Oculus Quest and Vive Cosmos) have been released since 2019. This type of headsets can provide the high-end VR service without a personal computer.

VR is different from desktop VE. Desktop VE is non-immersive. It allows individuals to view the 3-D virtual world presented on a standard high-resolution monitor and access the real world at the same time (Onyesolu & Eze, 2011). Besides this, VR supports various interactive methods with sensory facilities (e.g., HMD, glove, tracking systems), while individuals usually interact with a desktop VE through keyboard and mouse (Onyesolu & Eze, 2011).

Furthermore, there are also differences between VR and augmented reality (AR) and between VR and mixed reality (MR). AR places a computer-generated layer on an individual's view of the real-world (Blair, 2016; Oxford, 2019). The layer is at least semi-transparent and allows individuals to see the real-world around (Blair, 2016). This is different from VR, which fully immerses individuals into a virtual world. Relatively, researchers have not reached an agreement on the definition of MR. Milgram and Kishino (1994) referred MR as the technologies which involve the merging of real and virtual worlds and AR is a part of MR in a virtuality continuum. Blair (2016) viewed MR as a combination of VR and AR. He indicated that an AR environment allows digital content to be simply overlaid onto the real environment, while a MR environment integrates the digital content into the real world through an interactive way. An example of MR is that

an individual visited a virtual theater with an HMD headset as well walking on a physical omnidirectional floor (Schuster, Hoffmann, Bach, Richert, & Jeschke, 2014). The floor supported his or her orientation. Overall, despite differences existing in the definition of MR, the most distinguishing difference between VR and MR is whether real objects are involved in the environment.

Potential of VR in STEM education. Prior literature showed that VR has the potential to increase individuals' motivation and learning achievement in STEM education, most individuals take a positive attitude to VR STEM learning. Following are some cases.

Evidence for motivation. VR, as a novel technology, increases individuals' motivation to engage in STEM learning activities. Individuals were excited to learn in VR environments no matter they used a mid-range or a high-end VR system (Webster & Dues Jr., 2017). Angulo and de Velasco (2014) reported that immersive VR architectural models attracted audiences to spend more time reviewing and providing suggestions than physical models in an undergraduate architecture design showcase. In another case, K-12 students were motivated by the colorful interfaces of a virtual construction model and eliminated their initial concern on a lack of technological background at a civil engineering education workshop (Dinis, Guimaraes, Carvalho, & Martins, 2017a). In an immersive animated 3-D atlas experience activity, both students and lecturers were attracted by the interaction with the VR application and agreed that replacing traditional lecture methods with this VR application could enhance the education process (Hamrol, Górski, Grajewski, & Zawadzki, 2013).

Evidence for learning and skills. Parmar (2017) found that the VR activities more successfully helped students understand concepts in computer programming and increase their creativity than the desktop VE. Similarly, Alhalabi (2016) found that students who used VR equipment with 6 degrees of freedom exhibited higher achievement in engineering learning than those in the control groups. Additionally, VR also showed strength in instructing operational tasks. Bharathi and Tucker (2015) found that students spent less time to complete a virtual assembly task in a VR environment than they worked with a desktop VE.

Students' attitudes towards VR STEM learning. Most students reported positive attitudes towards their VR learning experience. Angulo and de Velasco (2014) observed that students reported higher ratings of affective appraisals for immersive architectural simulation models than physical ones. Oberhauser and Lecon (2017) found that students highly rated the method of programming with a VR application for the fun and intuition in understanding the structure of program code. Bell and Fogler (1995) found that most students perceived the VR learning experience helped them more accurately and completely understand chemical engineering concepts. Dinis et al. (2017a) found that most students felt their abilities of identifying building system structure increased after they learned in VR environment.

Models in VR learning. To understand how and why a VR environment benefits learning, researchers have proposed different models and continuously make progress in this field. *Affordance* is a concept frequently used in related studies. This word is made up by Gibson (Greeno, 1994). It is to describe what the utility of an object can afford to the others (Gibson, 2014). One object can provide either benefit or injury to one or many

objects, while different objects can afford the same benefit or injury to the same object. This flexibility rescues researchers from understanding objects only by their surface characteristics.

Perhaps the first VR learning model in history is from the study of Salzman, Dede, Loftin, and Chen (1999). Their model was used to understand how VR aided complex conceptual learning. They claimed that not only the VR features but also other factors (e.g., individual characteristics, concepts to be learned) affected the learning process and then learning outcomes. In their model, the instructed content partly determines the effectiveness of VR features, while the individual's characteristics moderate the effect of VR features (affordances) on learning by influencing the learning and interaction experience. Although this model has a detailed description of external factors that may influence learning, the node *learning process* is a black box without any introduction or analysis.

Dalgarno and Lee (2010) established a model to present the potential learning benefits of a 3-D desktop virtual learning environments (VLEs). In this model, representational fidelity and learner interaction are two unique characteristics of 3-D VLEs. These characteristics contribute to the experience of presence, identity construction, and co-presence in a multi-user virtual world. VLEs benefit learning with a series of affordances. This model has a detailed explanation of possible affordances but does not involve any expected *learning outcomes* in 3-D VLEs.

The study of Lee, Wong, and Fung (2010) and Makransky and Lilleholt (2018) proposed their model in VR / VE learning, respectively. The model proposed by Lee, Wong, and Fung showed how a desktop VE enhances learning outcomes. In this model,

desktop VE features offers a series of affordances in a learning experience, which include psychological factors (e.g., presence, motivation, and cognitive benefits) and reflective thinking. These affordances further affect individuals' learning outcomes. The model proposed by Makransky and Lilleholt investigated the emotional value of VR in education. From their model, two paths of affordances were identified by which VR features affect perceived learning outcomes: (1) the first path includes affordances in the affective domain, such as presence, motivation, enjoyment, and active learning, and (2) the second path includes affordances in the cognitive domain, such as cognitive benefit and reflective thinking. The structure of these two models are very similar. Both are from VR features to learning outcomes through a list of affordances. There is no mediator between affordances and learning outcomes in these models.

In recent years, there is a growing number of researchers who refer to *embodiment* as one affordance to understand individuals' interaction with a VR environment (Joo, Ahn, Bailenson, & Park, 2014). Shin (2017) utilized embodiment in his model to examine how a VR environment influenced learning. In that model, the usability of a VR application (comfortable usage) offers two educational affordances: embodiment and empathy. Both have a positive effect on the learnability of the VR application.

In summary, despite that differences exist, all these models show that learning experience mediates the effect of VR features on learning outcomes. Affordance is used to represent the potential benefits of VR features in the VR learning experience. Some researchers have started to use embodiment to understand the benefit of VR in recent years.

However, the characteristics of VR cannot cause but only affect individuals' meaningful learning (Roxana Moreno & Mayer, 2007). None of these models have considered how the characteristics of VR, or its affordances, affect an actual learning experience and then enhances tangible learning outcomes (e.g., knowledge and cognitive skills). This is an important issue. At present, VR STEM education is at the beginning stage. The available principles of instructional design are still very limited. Studying on how the VR condition influences an effective learning experience has the potential to support researchers to contribute to this field. These design principles will help instructional designers efficiently develop VR STEM courses, which will benefit students for successfully reaching the expected learning objectives. If knowledge in this field is insufficient, the design of VR STEM courses may be disoriented, and the actual learning outcomes are unpredictable.

Novelty effect. Some researchers were worried about whether the results of present VR education / learning research were conclusive, as many studies only collected research data through either a single or a short-term measurement. Huang and Roscoe (2019) reviewed 40 publications on VR in engineering education, primarily appearing from 2015 to 2017. They found that only two out of 25 experimental studies took repeated measures, and no article had taken in-depth longitudinal analysis. Similarly, Jensen and Konradsen (2017) did a review of 21 experimental studies of HMD VR in education and training since 2013. They indicated that all but one reviewed research focused on a very short-term use in an experimental setting. Researchers suggested to take a long-term view to assess the effectiveness of VR in learning to avoid the possible

influence of novelty (Hussein & Nätterdal, 2015; Jensen & Konradsen, 2017; Parmar, 2017).

In general, novelty refers to the experience and / or information that individuals feel unexpected, surprising, and new (Huang, 2003). It is the opposite of familiarity and can be created by innovative content or information technology. The novelty effect is an increased motivation or an increase in perceived usability of something due to its newness (Koch, Luck, Schwarzer, & Draheim, 2018). However, the increased motivation and perceived usability may gradually decrease when individuals are familiar with this innovation. The novelty effect may happen in two occasions: initially and reoccurring (Koch et al., 2018).

Clark (1983) and Clark and Craig (1992) indicated that the novelty effect is an uncontrolled effect which disturbs researchers to evaluate the effect of media on learning. Due to the novelty, the increased attention paid by students sometimes leads to an increased effort which influences their learning gains. These gains tend to diminish when students are familiar with the medium. Clark's viewpoint is supported by the results of many studies in the STEM education field. Following are some cases.

Krendl and Broihier (1992) examined students' (4-10 grade) perceptions about the instructional effectiveness of computer at approximately 30 years ago when personal computer was novel to the public. The whole study lasted over three years. They found that both students' preference of computer and perceived learning from computer were significantly declined after this period.

Tsaya and Kofinas (2017) investigated the effectiveness of serious games on students' engagement in a two-semester online learning course. They found that despite

that students more engaged in the gamified course than the non-gamified course, students' engagement gradually decreased in each semester, respectively.

Li and Ma (2010) did a meta-analysis of the effects of computer technology on k-12 students' mathematics learning from 46 studies involving more than 36 thousand students. They found that short technology interventions (6 months or less) were much more effective in promoting mathematics achievement than longer interventions (between 6 and 12 months).

Similarly, Merchant, Goetz, Cifuentes, Keeney-kennicutt, and Davis (2014) did a meta-analysis study of the effectiveness of desktop VE on individuals' learning. They found that there was a statistically significant inverse relationship between learning gains and the number of treatments for individuals who engaged in desktop VE serious games. In other words, students' learning gains diminished along with the playing time passing.

Considering the results of these studies, it is rational to predict that the novelty effect is prevalent in the multi-media learning field. Unjustified ignoring the novelty factor may negatively affect researchers evaluating the potential of multi-media technologies on learning. Therefore, a thorough evaluation of VR in STEM learning should involve a consideration of the possible novelty effect. Despite that, prior studies on the novelty effect in the VR field are rare.

Effectiveness of a VR condition. The research methods that many prior VR learning studies used is worth discussing. Common experimental designs in those studies are: (1) comparing individuals' feedback or learning outcomes between a desktop VE and a VR environment after an intervention, and (2) checking the differences through pre- and post-tests within a group of participants (Bharathi & Tucker, 2015; Kozhevnikov,

Gurlitt, & Kozhevnikov, 2013; Guido Makransky, Terkildsen, & Mayer, 2019; Mineev, 2017; Shu, Huang, Chang, & Chen, 2018). Both of these methods do not consider the possible differences among VR environments. However, prior research showed that the effect of different VR environments on learning are not always the same.

For example, although there are many studies showed that individuals in a VR environment had the potential to learn better than those learned with a desktop VE, there are also cases showed that a VR environment did not benefit learning at all when compared with desktop VEs. Mineev (2017) did not find any statistically significant difference between VR and desktop VE in training skills of maintaining computer systems. Similarly, the study of Timcenko et al. (2017) did not show any significant difference between students who used VR and desktop VE in their collaborative behaviors for a problem-solving game. Guido Makransky et al. (2019) even found that individuals learned less from a VR-version science simulation than they learned with the desktop-version application having the same content.

A possible explanation to the difference among these results is that the potential of media in learning is not only determined by its characteristics but also influenced by its appropriate use in learning activities, namely instructional method. Researchers in the multi-media field have proposed their views on the relationship between media and learning. In general, they agree that instructional method is an important factor that directly influences the efficiency of multi-media learning, a match among the technical affordances of a learning environment, instructional method, and learning objective has the potential to create a high cost-effective learning experience (Clark, 1994; Clark, 1983;

Fowler, 2015; Jonassen et al., 1994; Kozma, 1994; Mayer, 2003; Moreno & Mayer, 2002).

In addition, some studies found that manipulation and haptics sense can improve individuals' engagement and learning achievement in a technology-enhanced learning environment (Jang, Vitale, Jyung, & Black, 2017; Johnson-Glenberg & Megowan-Romanowicz, 2017; Lindgren & Johnson-Glenberg, 2013; Parmar, 2017; Schönborn, Bivall, & Tibell, 2011). Their explanation is from the embodied cognition theory. That is, actively interaction with the environment has the potential to deepen individuals' memory through offloading the cognitive work onto actions (Wilson, 2002). Specific to the VR field, research on this type of learning method, namely embodied learning, is very limited. Until now, embodied learning can only be implemented with the high-end VR systems (e.g., Oculurs Rift and HTC Vive) for their capability of a body-level tracking. There are many practical issues unsolved. For example, there is a lack of research on its best application scenes and researchers do not know whether its effectiveness is durable in a VR environment. Despite that, study on this direction is important for VR education. It has the potential to guide the design of VR applications and maximize the potential of VR in instruction.

Overall, scientifically measuring the effectiveness of VR in STEM education requires researchers to consider the contribution of instructional method on learning. Some studies have introduced the embodied learning method to increase individuals' engagement and learning achievement in a technology-enhanced learning environment. This method can be implemented with the high-end VR systems. As a new research direction, it is hopeful and needs deeper explorations.

A Model of an Effective VR Learning Experience

Due to the deficiency of prior research in understanding how individuals learn in a VR environment, a cross-sectional research model is proposed in this section. Its establishment includes three steps. First, a conceptual framework is set up based on theories in motivation and learning. Second, the main factors in a VR learning experience are described and their possible relationships are identified with evidence from prior literature. Finally, the research model is established based on the conceptual framework and the possible relationships among factors.

Conceptual framework. Malone and Lepper's theory of curiosity in intrinsically motivating instruction (1987) and the cognitive-affective theory of learning with media (CATLM; Moreno, 2005; Moreno & Mayer, 2007) are the starting points of understanding how a VR condition influences individuals' learning. Malone and Lepper indicated that curiosity is the most direct intrinsic motivation for learning, and it includes sensory curiosity and cognitive curiosity. Sensory curiosity is the changes of light, sound, manipulation, or other sensory stimuli, which may attract individuals' attention in an environment. Cognitive curiosity is evoked when an instructional environment can stimulate individuals to reflect their prior knowledge and make up deficiencies. Everyone has a cognitive drive to build a personal satisfied knowledge structure (Malone, 1981).

As a theory for instruction, Malone and Lepper's theory is situated into the VR learning field in this study. Presence and sense of embodiment represent individuals' feelings of themselves with the VR environment and their feelings with the virtual body, respectively (Kilteni, Groten, & Slater, 2012). Both may arouse individuals' sensory curiosity. By contrast, the completeness of individuals' prior knowledge will partly

weaken individuals' cognitive curiosity in learning (Johnson-Glenberg, 2018; Wade & Kidd, 2019). The enhanced or undermined curiosity will directly influence individuals' motivation.

Furthermore, CATLM is a theory that explores individuals' cognitive processing in a multimodal learning environment (Moreno & Mayer, 2007). According to the theory, for meaningful learning to occur, an individual needs to purposefully select useful information for further processing from the sensory memory to the working memory. Then, the individual will organize this new information into a coherent mental structure and integrate it with the prior knowledge. If the cognitive processing demands exceed the processing capacity of the cognitive system, cognitive overload will happen. When an individual lacks motivation, he or she may fail to engage in the learning process. In other words, prior knowledge and motivation are critical for a successful learning experience. Prior knowledge may lower a high cognitive load in cognitive processing, while motivation ensures individuals' engagement in the cognitive processing. Figure 1 illustrates the framework based on the two theories. This framework is the foundation for establishing a research model which is shown at the final step in the subsection.

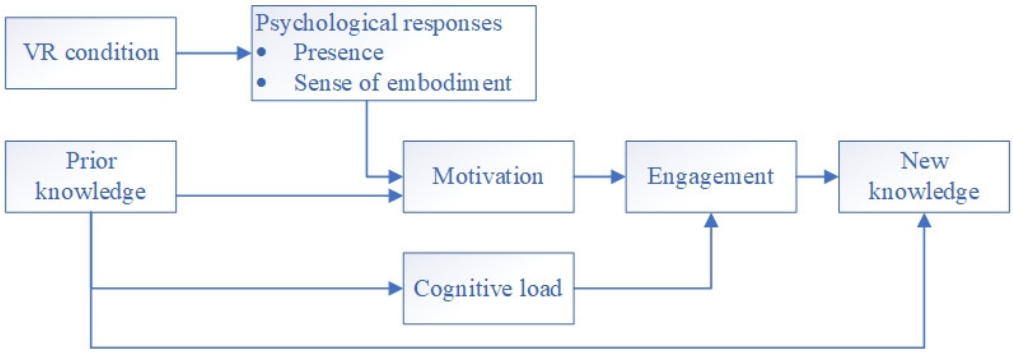


Figure 1. The conceptual framework of an effective VR learning experience

Factors and their relationships. This part mainly describes the main factors shown in the conceptual framework. It also includes their causal relationships which may be beyond the conceptual framework. The sequence of description is as follows: (1) VR condition, (2) presence, (3) sense of embodiment, (4) motivation, (5) cognitive load, (6) engagement, (7) prior knowledge, and (8) learning achievement.

VR condition. One of the most important characteristics of VR is full immersion. Immersion "describes the extent to which the computer displays are capable of delivering an [……] illusion of reality to the senses of a virtual reality participant" (Slater & Wilbur, 1997, p. 606). Its degree can be objectively measured by a series of technical indicators of a VR system, such as the field of view, color resolution, fidelity (e.g., whether or not present dynamically changing shadows), the types of sensory modalities, and the consistent degree between physical movement and the environment's response (e.g., the change of visual display, sound direction, and the time delay) (Slater, Usoh, & Chrysanthou, 1995; Slater & Wilbur, 1997). A VR system is more immersive than another VR system if it can more fully and / or realistically simulate individuals' perceptions as being in the real world (e.g., a more complete virtual representation of the participant's body) (Slater & Sanchez-Vives, 2016).

The differences in technical parameters not only lead to the different degree of immersion between VR systems, but also the differences in affording individuals the capability of manipulation and movement. Commonly, the high-end VR systems not only have a higher level of immersion but also affords individuals more freedom in action (i.e., embodiment) than the mid-range VR systems. This capability is attributed to another distinctive capability of VR: embodiment (Johnson-Glenberg, 2018). The degree of

embodiment in a VR educational environment is mainly determined by three aspects: (1) the degree of immersion, (2) the magnitude of gesture, and (3) congruence of gesture to content (Johnson-Glenberg, 2018; Johnson-Glenberg & Megowan-Romanowicz, 2017). A high-embodied VR condition allows free exploration and active learning. It has the potential to enhance learning achievement (Johnson-Glenberg, 2018).

The Oculus Go system and the Oculus Rift system are two types of VR systems released by the Facebook Oculus company (Oculus, 2020). They belonged to the mid-range VR system and the high-end VR system, respectively. Table 1 is a comparison between these two types of VR systems. Compared with the Oculus Go system, the Oculus Rift system offers more vivid and extensive virtual surrounding, which leads to a higher level of immersion. Its external "Constellation" IR camera tracking system can track individuals' body position and gestures (i.e., grasp and rotate). This benefits individuals to learn with actions in a VR environment, which belongs to a high-level embodiment (Johnson-Glenberg, 2018). In short, compared with the Oculus Go system, the Oculus Rift system has the potential to offer a higher-level VR condition with a higher degree of immersion and embodiment.

Table 1

A Comparison Between the Oculus Go and the Oculus Rift Systems

	Oculus Go	Oculus Rift
Technology*		
Display	1280 x 1440	2160x1200 (1080x1200 per eye) OLED panels
Refresh rate	60 Hz	90 Hz
Field of view	100 degrees	110 degrees
Headtracking	3DOF headtracking	6DOF tracking capabilities with external "Constellation" IR camera tracking system
Controller	Unimanual touch controller	Bimanual touch controllers
Audio	Headphones purchased separately	Integrated over-ear headphones with 3D directional audio support and built-in microphone
Function		
Auditory sense	Yes	Yes
Visual sense	Yes	Yes
Tactile sense	No	Yes
Proprioceptive sense	Yes (weak)	Yes (strong)
Virtual manipulation	No	Yes

Note. * Data was retrieved from https://arstechnica.com/gaming/2016/03/the-ars-review-oculus-rift-expands-pc-gaming-past-the-monitors-edge/ and https://www.niora.net/en/p/oculus_go

Presence. Presence is one of the most distinguishing feelings that individuals will experience when they engage in a VR environment (Johnson-Glenberg, 2018). Due to the lack of a clear and unified definition, researchers sometimes use it and *immersion* interchangeably. Relatively, the definition from Slater and Wilbur (1997) has a high degree of adoption. "Presence is a state of consciousness, the (psychological) sense of being in the virtual environment" (Slater & Wilbur, 1997, p. 607). It is the individual's potential responses to immersion in psychology and behavior. The responses may include unconscious physiological behaviors, conscious volitional behaviors, and cognitive

processing (Sanchez-Vives & Slater, 2005). In a narrow sense, presence only refers to the sense of "being there", namely place illusion (PI) (Slater, 2009). PI is humans' response to an immersive environment within its boundaries (Slater, 2009). The occurrence of PI requires the sensorimotor contingencies (SC) (Slater, Spanlang, & Corominas, 2010). That is, if the perceived changes of actions in perception is much in line with those occurred in the real world, the brain will adopt the perceived illusion as true.

A high level of presence requires the physical disappearance of display devices in individual's consciousness and realistic context (Held & Durlach, 1992; Hendrix & Barfield, 1996). Individuals' sense of presence is strongly influenced by the features of virtual environment (i.e., the fidelity of the representation, immediacy of control) in both VR and desktop VE conditions (Dalgarno & Lee, 2010; Lee et al., 2010; Makransky & Lilleholt, 2018).

Sense of Embodiment. *Presence* is a concept that reflects the influence of immersion on individuals' psychology and behaviors, while *sense of embodiment* corresponds to individuals' responses to the potential of embodiment in a VR condition. Kilteni, Groten, and Slater (2012) referred sense of embodiment in VR as "the ensemble of sensations that arise in conjunction with being inside, having, and controlling a body especially in relation to virtual reality applications" (p. 374). It includes one's perception between the self and the body representation in the VR environment, namely body ownership. For example, an individual felt a connection with the virtual body after the individual was virtually stroked (Bourdin, Barberia, Oliva, & Slater, 2017). An individual might have an illusion that a virtual arm was a part of his / her self when there were synchronous tactile stimuli on the real hand and virtual stimuli on the virtual arm (Slater,

Perez-marcos, Ehrsson, & Sanchez-Vives, 2008). Another factor of sense of embodiment is the perceived capability of naturally controlling the agency in the VR environment. It requires a match between the predicted and the actual sensory feedback (Kokkinara, Kilteni, Blom, & Slater, 2016; Sato & Yasuda, 2005). When this prerequisite was satisfied, individuals might have the illusion of walking, even when they just sat without movement (Kokkinara et al., 2016).

Prior research showed the importance of dynamic movements to presence. Slater, McCarthy, and Maringelli (1998) found that individuals who played more body movements, such as blending down and looking up, reported a higher degree of presence than those played less movements in a VR task. In another study, Slater et al. (1995) found that providing virtual objects with dynamic shadows to match individuals' manipulation enhanced individuals' presence in VR environments. They explained that this was due to individuals' identifying with their virtual body.

Motivation. Motivation is termed as "[an individual's] willingness or desire to engage in a task" (Garris & Driskell, 2002, p. 451). It is relevant to an individual's choice to engage in an activity and the energy involved to persist. In general, an individual's motivation includes intrinsic motivation and extrinsic motivation. Intrinsic motivation is defined as "the doing of an activity for its inherent satisfactions" (Ryan & Deci, 2000, p. 56). It is for the fun or challenge entailed rather than external rewards. On the contrary, extrinsic motivation arises from outside. It leads to individuals' activities or behavior for attaining separable outcomes (e.g., money, fame, grades), rather than personal enjoyment (Ryan & Deci, 2000). In the digital learning field, intrinsic motivation is more effective than extrinsic rewards (Tüzün, Yilmaz-Soylu, Karakuş, Inal, & Kizilkaya, 2009).

Furthermore, motivation is situated and contextualized (Lee et al., 2010; Linnenbrink & Pintrich, 2002). VR facilitates experiential learning tasks that would be impossible to undertake in the real world and allows individuals to make choices for personalized learning. These intrinsically motivate individuals to learn in a VR environment (Dalgarno & Lee, 2010).

Cognitive load. According to the cognitive load theory, there are three types of cognitive load: intrinsic cognitive load, extraneous cognitive load, and germane cognitive load (Sweller, 2010; Sweller, Merrienboer, & Paas, 1998). Intrinsic cognitive load is the cognitive load generated by the element interactivity in learning materials. It is related to the complexity of learning content and whether or not it is easy to be understood. Extraneous cognitive load concerns the instructional design. It reflects the effort required to process the designed instruction. Germane cognitive load is the available working memory resources to deal with the intrinsic cognitive load. It is the actual cognitive resources being able to use in meaningful learning activities.

In this study, cognitive load is a combination of intrinsic cognitive load and extraneous cognitive load. A combination of high intrinsic cognitive load and high extraneous cognitive load may cause cognitive overload, which undermines learning efficiency (Paas & Van Merriënboer, 1994). The level of element interactivity is determined by the interaction between the instructional materials and individuals' expertise (Sweller et al., 1998). In other words, the complexity of learning task and individuals' prior knowledge are critical for intrinsic cognitive load. Individuals can learn materials of low element interactivity with lower intrinsic cognitive load than those of higher element interactivity due to the less reference among elements (Sweller, 2010).

For example, it is easier to remember several words together than to remember them with sequence. This is because the later task requires additional cognitive load to remember their relative position in the list.

A high-embodied condition benefits individuals' learning in STEM (Johnson-Glenberg, 2018). A possible explanation is that meaningfulness of gestures and movements can deepen individuals' memory by lightening cognitive load (Goldin-meadow, Nusbaum, Kelly, & Wagner, 2001; Ping & Goldin-meadow, 2010). Paas and Sweller's evolutionary view of human cognitive architecture (2012) provides a more detailed explanation. They indicated that using movement as a type of knowledge that humans are familiar with can facilitate the acquisition of new knowledge. This method requires less cognitive resources but can construct higher-quality cognitive schemas. It breaks the constraint of working memory on obtaining novel information.

Engagement. Engagement is commonly defined as the behaviors that individuals involve themselves physically, cognitively, and emotionally in a task (Bouvier, Lavoué, & Sehaba, 2014; Kahn, 2013). In game or gamified course studies, the definition of engagement is a little different. It is referred as a concept concerned with individual's in-the-moment experience during play (Boyle, Connolly, Hainey, & Boyle, 2012; Shernoff & Strati, 2011). This engagement includes intensity of arousal, attention, deeply involvement, and persistence (Cordova & Lepper, 1996; Parker & Lepper, 1992; Garris & Driskell, 2002).

Flow theory is commonly used to explain engagement from the psychological aspect (Csikszentmihalyi, 1997). Flow represents an optimal experience where individuals can focus on their personal goals or interest and screen out unwanted stimuli

(Csikszentmihalyi, 1990). Being in the state of flow is regarded as a similar state of being highly engaged in an activity (Whitton, 2011).

Prior literature showed that motivation influences engagement. Perceived challenging, interesting, and enjoyable promotes individuals' intrinsic motivation and keeps them engage in tasks (Keller, Ringelhan, & Blomann, 2011). The study of Schüler (2007) showed that motivation moderated the relationship between challenge-skill balance and flow experience. Weibel and Wissmath (2011) found that motivation positively affected flow, and flow influenced performance in computer games.

In addition, presence has the potential to impact flow / engagement. The study of Takatalo (2002) recognized presence as a prerequisite for flow / engagement. Their research showed that presence was correlated with the dimensions defined by their model of flow experience in VE. Similarly, Cheng, Chieng, and Chieng (2014) found that presence positively and significantly affected flow after they examined the mechanism of forming the flow state in a 3-D computer simulated interactive environment.

Besides presence, sense of embodiment may affect engagement. The natural user interface was regarded as the key to facilitate participants' movement in a MR or a VR environment, where participants can act as being in the real world (Brondi, Avveduto, & Carrozzino, 2015; Lindgren, Tscholl, Wang, & Johnson, 2014). It enhanced participants' sense of embodiment (Bianchi-Berthouze, 2013). Furthermore, individuals who were in a high-embodied condition had a higher degree of engagement than those in lower-embodied conditions (Johnson-Glenberg & Megowan-Romanowicz, 2017; Lindgren et al., 2014). Thus, engagement is related to sense of embodiment.

Last, cognitive load also has an effect on engagement. Flow experience represents a match between individuals' ability and the challenge of task (Csikszentmihalyi, 1999). If the task is too difficult, individuals may be overwhelmed by frustration, while the task is too easy, individuals will feel bored. Both these two situations will undermine individuals' engagement in the learning experience (Przybylski, Rigby, & Ryan, 2010; Sharek & Wiebe, 2014). In other words, a moderate cognitive load benefits individuals' engagement in learning. This study only focused on the situation that cognitive load negatively affects engagement.

While prior research showed the positive effects of motivation and presence on engagement, there are also evidence and argument to support the opposite relationships. Flow represents a match between personal skills and demands. The skills-demands compatibility enhanced intrinsic motivation in working or learning tasks (Csikszentmihalyi & Lefevre, 1989; Keller et al., 2011). Thus, flow / engagement has the potential to influence intrinsic motivation. Moreover, some studies found that engagement influences presence. This is because they defined engagement as a concept related to involvement. Bouvier, Lavoué, and Sehaba (2014) indicated that as presence requires a shift in consciousness from the real world to the activity world, presence is a state that occur after engagement. Witmer and Singer (1998) regarded engagement (or involvement) as a necessary condition for presence. In their study, all items measuring *involvement* were significantly correlated with the *presence* scale.

Prior knowledge. Prior knowledge is critical in an effective learning experience. The conceptual framework has already covered its possible effects on motivation, cognitive load, and learning achievement based on existing theories. Besides that, there is

practical evidence to support the influence of prior knowledge on constructing new knowledge in a learning process. The study of Desanctis et al. (1994) and King (1994) showed that prior knowledge helped individuals' reading comprehension. The study of Kesteren, Ruiter, and Henson (2012) found that individuals were able to better memorize new information than normal through building connections between new information and prior ones. However, the study of Shaffer et al. (2016) indicated that the prior knowledge could improve learning efficiency only when it had been covered in depth before.

Learning achievement. Learning achievement is defined as the indicator of learning outcomes. Prior research showed that individuals' learning achievements in a VR environment can be assessed from different aspects: the score in a post-test which examined individuals' memory or comprehension of instructed knowledge and skills (e.g., Alhalabi, 2016; Bell & Fogler, 1995; Bharathi & Tucker, 2015; Dinis et al., 2017b; Hussein & Nätterdal, 2015; Mineev, 2017; Oberhauser & Lecon, 2017; Oberhauser et al., 2016), the efficiency of implementing assigned tasks in a VR environment (e.g., Alhalabi, 2016; Bharathi & Tucker, 2015; Mineev, 2017; Oberhauser & Lecon, 2017; Oberhauser et al., 2016; Ragan et al., 2015), and the capability of transferring instructed knowledge into the real-world practice (e.g., Dinis et al., 2017c; Ragan et al., 2015; Waller & Knapp, 1998).

Prior research showed evidence that motivation may influence individuals' learning. Motivation directed individuals' attention to the learning task and helped them persist when they encountered difficulties (Svinicki, 2004). A positive relationship was found between intrinsic motivation and learning performance in classroom teaching (Keller, 2010; Lee et al., 2010; Lepper & Iyengar, 2005; Pintrich, 2003).

Furthermore, practical cases supported a causal relationship from engagement to learning achievement. Hou and Li (2014) found that flow was an important indicator predicting desirable learning outcomes in a digital problem-solving-based adventure game. Similarly, the study of Hsu, Tsai, and Wang (2012) showed that students in the high engagement group outperformed those in the low engagement group in a digital game-based learning activity.

Research model. Based on the conceptual framework and evidence of relationships among factors, a research model that explores how a VR environment influences a STEM learning experience is proposed (see Figure 2). This model has two external variables (i.e., VR condition and prior knowledge), one output factor (i.e., learning achievement), and five internal factors (i.e., presence, sense of embodiment, motivation, cognitive load, and engagement). The internal factors are influenced by the external variables and impact the output factor.

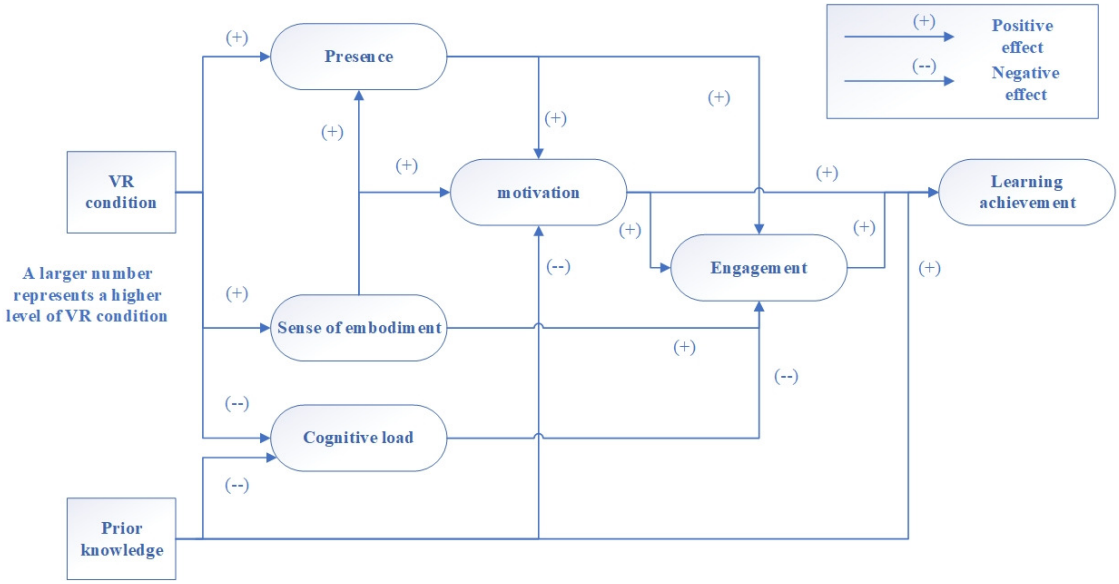


Figure 2. The hypothesized model of an effective VR learning experience

Related hypotheses are:

H1: VR condition has a directly positive effect on presence.

H2: VR condition has a directly positive effect on sense of embodiment.

H3: Sense of embodiment has a directly positive effect on presence.

H4: Presence has a directly positive effect on motivation.

H5: Sense of embodiment has a directly positive effect on motivation.

H6: Prior knowledge has a directly negative effect on motivation.

H7: VR condition has a directly negative effect on cognitive load.

H8: Prior knowledge has a directly positive effect on cognitive load.

H9: Motivation has a directly positive effect on engagement.

H10: Presence has a directly positive effect on engagement.

H11: Sense of embodiment has a directly positive effect on engagement.

H12: Cognitive load has a directly negative effect on engagement.

H13: Prior knowledge has a directly positive effect on learning achievement.

H14: Motivation has a directly positive effect on learning achievement.

H15: Engagement has a directly positive effect on learning achievement.

It is noteworthy that the effects of engagement on presence and motivation are not labeled. This is because the purpose of this model is to understand an effective learning experience. The measurement of an effective learning experience is satisfactory learning achievement. Relatively, the effects of presence and motivation on engagement need more attention, but not the reversed effects.

Novelty in VR Learning

This section mainly explores the possible influence of novelty on a VR learning experience. It includes two subsections. The first subsection investigates the possible change of assessed factors when individuals come to a high-level VR condition from a low-level VR condition. The second subsection evaluates the changing tendency of assessed factors after the novelty effect happens.

Novelty effect in VR learning. The degree of novelty is determined by the incongruency between a learner's past experience and the present one (Berlyne, Craw, Salapatek, & Lewis, 1963; Kesteren et al., 2012; Pearson, 1998). There are four types of novel experiencing: external sensation, internal sensation, external cognition, and internal cognition (Pearson, 1998). External means that the novelty is from the environment or activity, while internal means that the novelty is self-generated (e.g., daydream). This study specially focuses on external sensory novelty.

A high-level VR condition provides more flexibility in learning than a low-level VR condition. It allows individuals to learn with movement and manipulation. Thus, the difference between learning in a high-level VR condition and traditional learning with text or picture is more significant than the difference between a low-level VR condition and traditional learning. In other words, learning in a high-level VR condition is more novel than learning in a low-level VR condition. According to the definition of the novelty effect, it is predicted that individuals' motivation will increase in a high-level VR condition when they come from a low-level VR condition.

Furthermore, motivation influences individuals' conscious and unconscious psychological behaviors by adjusting attention resources. It has the capability to enhance

the sensibility of selective attention (Engelmann & Pessoa, 2007) and maintain individuals' focused attention (Robinson et al., 2012; Simon, 1994; Tomporowski & Tinsley, 1996). At the same time, the attention system directly controls the whole perception process (Wicken, 1992). Prior research showed that sufficient, involuntary focused attention is the pre-condition of generating spatial presence in a VR environment (Bystrom, Barfield, & Hendrix, 1999; Wirth et al., 2007; Witmer & Singer, 1998). When individuals were engaging in an environment with little blatant disturbance, motivation positively influenced individuals' unconscious embodied perceptions (Zestcott, 2017).

Moreover, motivation influences engagement and learning achievement. Both CATLM and prior case studies asserted that motivation positively influences individuals' engagement and engagement is critical for constructing new knowledge (Hou & Li, 2014; Hsu et al., 2012; Keller et al., 2011; Moreno & Mayer, 2007; Schüler, 2007). There are also studies that showed the positive effect of motivation on learning (Keller, 2010; Lee et al., 2010; Lepper & Iyengar, 2005; Pintrich, 2003; Svinicki, 2004). Therefore, when individuals are in a high-level VR condition, their engagement and learning achievement may also increase.

Attenuation of the novelty effect in VR learning. Berlyne (1970) studied the relationships among novelty, complexity, and hedonic value. He found that individuals' interest in complex novel stimulus was more durable than simple novel ones. In his study, complexity referred in particular to visual complexity, which covered several different properties, such as "the number of distinguishable elements, dissimilarity between elements, and the degree to which several elements were responded to as a unit" (Berlyne & Parham, 1968, p. 377). He explained that this was because "rewarding effects of

perceptual processing came into play" (Berlyne, 1970, p. 285). A positive habituation was the pre-condition that a stimulus condition was pleasant and rewarding. Compared with complex stimulus, individuals would be more easily preponderated by the tedium factor after repeatedly encountered simple stimulus.

A typical high-level VR condition has more powerful features to facilitate individuals operating the avatar in the virtual environment. For example, an individual has two controllers and can tactilely perceive virtual objects in a high-level VR condition. Therefore, the stimuli in a high-level VR condition are more diversified and complex than a low-level VR condition in most occasions. According to Berlyne's explanation of the relationship among novelty, complexity, and hedonic value (Berlyne, 1970), individuals' motivation should decrease more slowly in a high-level VR condition after the novelty effect happens than in a low-level VR condition.

Furthermore, as it was reviewed in the last subsection, motivation may positively influence presence, sense of embodiment, and engagement. Thus, it is predicted that individuals' presence, sense of embodiment, and engagement decrease more slowly in a high-level VR condition after the novelty effect happens than in a low-level VR condition, respectively.

Finally, besides the positive effectiveness of motivation, a high-level VR condition is embedded the capability of embodied learning. According to the embodied cognition and the embodied learning theories, learning with manipulation off-loads cognitive work onto the environment and body action aids learning (Goldin-Meadow, 2011; Johnson-Glenberg, 2018; Wilson, 2002). Therefore, individuals' learning in a high-level VR condition has the potential to reach a higher rate of knowledge-retention than in

a low-level VR condition. Since prior knowledge is critical for constructing new knowledge, embodied learning should benefit the individual to learn relevant content in the future. Thus, due to the positive effectiveness of motivation and the potential of embodied learning in a high-level VR condition, it is predicted that individuals' learning achievement decreases more slowly in a high-level VR condition after the novelty effect happens than in a low-level VR condition.

Summary. In conclusion, a high-level VR condition is more novel and complex than a low-level VR condition. It is predicted that when individuals come to the high-level VR condition from the low-level VR condition, their motivation will increase. Due to the possible influence of motivation, their presence, sense of embodiment, engagement, and learning achievement may also increase at the same time. After reaching peak values, their values will decrease. It is predicted that their values will have a slower decrease in a high-level VR condition than in a low-level VR condition. This slower decrease may be attributed to the rewarding effect of perception and the strength of embodied learning in knowledge retention. Figure 3 is a diagram that describes the influence of novelty on assessed factors based on predictions in this section. It is appliable to the possible changes of presence, sense of embodiment, motivation, engagement, and learning achievement over time.

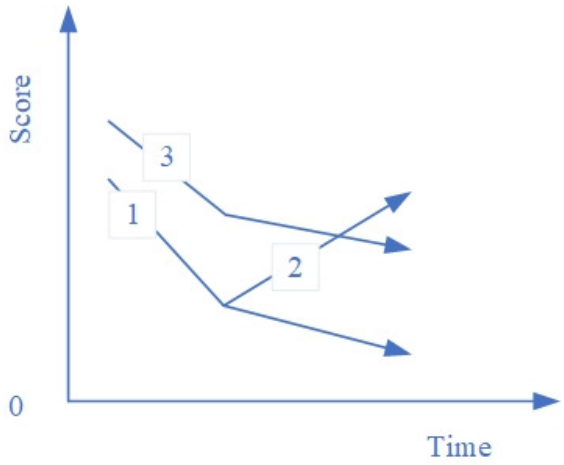


Figure 3. The influence of novelty on the assessed factors based on predictions. "1" represents the decrease in a high-level VR condition. "2" represents the increase from a low-level VR condition to a high-level one. "3" represents the decrease in a low-level VR condition.

Related predictions are:

P1: Presence will increase in a high-level VR condition.

P2: Sense of embodiment will increase in a high-level VR condition.

P3: Motivation will increase in a high-level VR condition.

P4: Engagement will increase in a high-level VR condition.

P5: Learning achievement will increase in a high-level VR condition.

P6: Presence will decrease more slowly in a high-level VR condition than in a low-level VR condition after the novelty effect happens.

P7: Sense of embodiment will decrease more slowly in a high-level VR condition than in a low-level VR condition after the novelty effect happens.

P8: Motivation will decrease more slowly in a high-level VR condition than in a low-level VR condition after the novelty effect happens.

P9: Engagement will decrease more slowly in a high-level VR condition than in a low-level VR condition after the novelty effect happens.

P10: Learning achievement will decrease more slowly in a high-level VR condition than in a low-level VR condition after the novelty effect happens.

CHAPTER 3

METHODS

The purpose of this study was to explore and understand the impact of novelty on an effective VR STEM learning experience. It included three research questions: (1) How does the VR condition impact an effective VR learning experience, (2) How much can the novelty effect influence the VR learning experience, and (3) How does the effectiveness of a VR condition change after the novelty effect happens.

In order to answer these questions, the study employed a multi-session mixed-methods experimental design (Creswell & Clark, 2017). Specifically, the research was conducted based on an experiment which included three repeated but independent sessions. In addition, participants were randomly assigned into three groups and treated differently through sessions. Participants' scores in pre- and post-tests, psychological scales, and session quizzes were collected. Their responses in interviews were audio recorded.

The quantitative strand took the primary role in this study. In order to answer the first question, the structural equation modeling (SEM) method was used to validate the hypothesized model proposed in the literature review chapter and establish the causal relationships among factors. SEM integrates the measurement and structural approaches in a generalized analytical framework (Wang & Wang, 2012). It allows researchers to directly identify the quality of a proposed model and the causal relationships among factors in the model with directly statistical data. Thus, SEM is appropriate, both for validating the proposed model that describes how VR affects an effective learning experience and for establishing the causal relationships among the factors in the model.

In addition, due to dropout, the proportion of incomplete cases was larger than 15% in the study (27 out of 77). The quantitative data analysis to answer the second and the third questions was conducted under the framework of the multiple imputation method. Compared to traditional methods for handling missing data (e.g., listwise, simple regression imputation, mean imputation), the multiple imputation method shows its strength in minimizing bias for estimates when the proportion of missing data is large (Rubin, 2004; Buuren, 2018). Therefore, the multiple imputation method was chosen to handle incomplete data in this study.

The specific tasks of the qualitative strand were to triangulate and explain the quantitative results in this study. This is because due to the relatively high cost of a VR experiment in money and time, this VR learning study was hard to reach the minimum sample size recommended by the priori power analysis for the quantitative data comparison. Besides that, the qualitative method can reveal "the complexity of human life [……] with many moving parts" (Creswell, 2016, p 32). It helps researchers understand and answer questions like "how" and "why" that are associated with complex phenomena. This is what quantitative methods cannot do.

The following sections describes the participants, materials, and procedure in the experiment. Then it introduces the specific phases in data analysis.

Participants

The experimental recruitment information was posted on one SONA system in the Polytechnic campus and one SONA system in the Tempe campus at Arizona State University in Fall 2019 after the approvement of the university human subjects

institutional review board (IRB). Exclusion criteria were: (1) a serious cybersickness history and (2) the age was younger than 18. A total of 77 undergraduate students were finally recruited in the experiment. Table 2 shows their demographics. The incentives were course credits and Amazon eGift Cards. Furthermore, as a multiple session experiment, dropout happened during the experiment. The second session had 64 participants, while the third session only had 50 participants.

Table 2

Experimental Participants

		Number	Percentage
Gender	Male	52	67.53
	Female	25	32.47
Major	Engineering	45	58.44
	Science	14	18.18
	Business and Humanity	18	23.38
Age	18	41	53.25
	19	17	22.08
	20-21	7	9.09
	>21	12	15.58
Race	American Indian, Alaska Native, and other races	7	9.09
	Asian or Asian American	12	15.58
	Black or African American	6	7.79
	Hispanic or Latino	16	20.78
	White or Caucasian	36	46.75

Materials

The VR experimental environment, questions for the pre- and post-tests, the instrument for a survey questionnaire, and questions for session quizzes were prepared before the experiment. The preparation of the VR experimental environment involved VR systems and the VR application. Questions for the pre- and post-tests included questions for a mental rotation test and questions for an astronomy knowledge test. The instrument

was developed based on existing instruments in literature. After content validity and face validity were obtained, the instrument was deployed on an online system as a survey questionnaire in this experiment. The quiz questions were developed based on the revised Bloom's taxonomy in cognitive domain (Krathwohl & Anderson, 2009). It was deployed on another online system after obtained content validity and face validity.

Experimental environment. ***Hardware***. This experiment employed one suite of the Oculus Rift system, one suite of the Oculus Go system, and a laptop with the NVIDIA GTX 1060 graphics card. These facilities were used to establish two different learning environments: a VR condition with a high degree of immersion and embodiment, namely a high-level VR condition, and another VR condition with a low degree of immersion and embodiment, namely a low-level VR condition.

Software. *Titan of Space* was selected to be used in this experiment. This is a VR STEM education application developed by DrashVR, LLC (http://www.drashvr.com/). It allows an individual to engage in a virtual tour through the solar system. The individual can learn the basic information of celestial bodies in the solar system (i.e., size, distance, and components) through text reading, 3-D observation, and manipulation. The learning difficulty is stable across scenarios. In this experiment, its contents were evenly separated to four sections with a consideration of the integrity of learning content and the time spent: (1) Earth, Mercury, Venus, Mars and their moons; (2) Ceres, Jupiter and its moons; (3) Saturn and its moons; (4) Chariklo, Uranus, Neptune, Pluto, and their moons. In the self-paced guided tour mode, an individual needed approximately 15-20 minutes to complete the learning task in each section which includes all reading and manipulation work. The corresponding manuscript for instruction were prepared at the same time.

Furthermore, *Titan of Space* has versions for both the Oculus Rift and the Oculus Go systems. *Titans of Space Plus 2.5.7* and *Titans of Space (mobile)* were installed on the Oculus Rift and the Oculus Go systems respectively in this experiment. A significant advantage of *Titans of Space Plus 2.5.7*, namely the Rift version, is that the individual can fly out of the virtual spaceship, grab, rotate, and look around a celestial body from any angle. The Go version does not afford these functions.

Pre- and post-tests. Pre- and post-tests were designed to examine participants' spatial capability and astronomy knowledge. The tests included a mental rotation test and an astronomy knowledge test. The mental rotation test used the first 12 items in Peters and his colleagues' redrawn Vanderg and Kuse mental rotations test (Peters et al., 1995). Items were printed on two separate pages with another two introduction pages attached in the test booklet. Each item has a target figure shown on the left and four figures on the right as options. Only two of the options are the rotated versions of the target figure. The scoring method in this experiment was consistent with its original design. That is one point per item, in which a participant should correctly identified the two options that match the target figure. The maximum score was 12. Due to the request of Dr. Peters, the mental rotation test only took the form of paper-and-pen. The purpose of using mental rotation test in this experiment was to examine the possible effect of VR learning on individuals' capability of mental rotations. The astronomy knowledge test involved ranking the eight planets from smallest to biggest and from closest to furthest away from the sun. The scoring method was one point for one correct ranking place, and the total score was 16. The astronomy knowledge test was deployed on an online survey system named SurveyMonkey [link: https://www.surveymonkey.com/]. The score of the pre-

astronomy knowledge test was used as the indicator of participants' prior knowledge, while the difference between the scores in the pre- and the post- astronomy knowledge tests represented the effectiveness of a VR condition in instructing spatial knowledge.

Instrument development. Following the instrument development process introduced by McCoach, Gable, and Madura (2013), the instrument was developed in four steps: (1) literature reading and existing similar instruments search, (2) item writing or revision, (3) content validity assessment, and (4) face validity assessment. Table 3 is the instruments referred to in this study. Most referred instruments are either directly from or supported by the creators of relevant theories. The instrument of sense of embodiment is from a compilation of related items in prior literature. Items in these instruments had been purposefully selected and situated into the VR learning context.

Table 3

Relevant Instruments in the Literature

Author(s)	Number of Items Used	Factor
Slater, Khanna, Mortensen, and Yu (2009)	4	Presence
Gonzalez-franco and Peck (2018)	7	Sense of embodiment
Deci and Ryan (1994)	4	Intrinsic motivation
Cierniak, Scheiter, & Gerjets (2009)	3	Cognitive load
Jackson, Eklund, and Martin (2004)	7	Engagement

Furthermore, a 0- to 100-point slider interfaces instead of Likert scales was used in this instrument in this study. This is because granular measurement tools (e.g., 5-point Likert scales) are regarded as not sensitive to within-person changes over time, and fine-grained measures (e.g., 0- to 100-point slider interface) are recommended (Grimm, Ram, & Estabrook, 2016).

Both content validity and face validity were obtained before the formal test started. A professor with expertise in VR STEM education evaluated the accuracy of the items. Five undergraduate and graduate students who were interested in VR education judged whether the items' content could be understood by undergraduate students and whether this instrument could be used to measure students' learning experience in a VR experiment. After content validity and face validity were obtained, a total of 25 items were remained in the final instrument. The final instrument was then deployed onto the SurveyMonkey website [link: https://www.surveymonkey.com/] as a survey questionnaire.

Session quiz development. Corresponding to the prepared learning content for each section, four sets of quiz questions were developed before the experiment. They were deployed on an online quiz system named ClassMarker [link: https://www.classmarker.com/]. The development of quiz questions was based on the revised Bloom's taxonomy in cognitive domain (Krathwohl & Anderson, 2009). The revised Bloom's taxonomy defines six dimensions of cognitive process based on their complexity. The quiz in this experiment included four types of questions and investigated participants' learning in three out of these six dimensions. The three dimensions are remember, understand, and evaluate.

The dimension of remember. Remembering involves retrieving knowledge from long-term memory (Krathwohl & Anderson, 2009). There was one recognizing question and two recalling questions to investigate participants' learning in this dimension per quiz.

The recognizing question. Participants needed to connect the images of four different celestial bodies to their names per quiz. These four celestial bodies and their images had been shown in that day's VR learning content. Figure 4 is an example of the recognizing question.

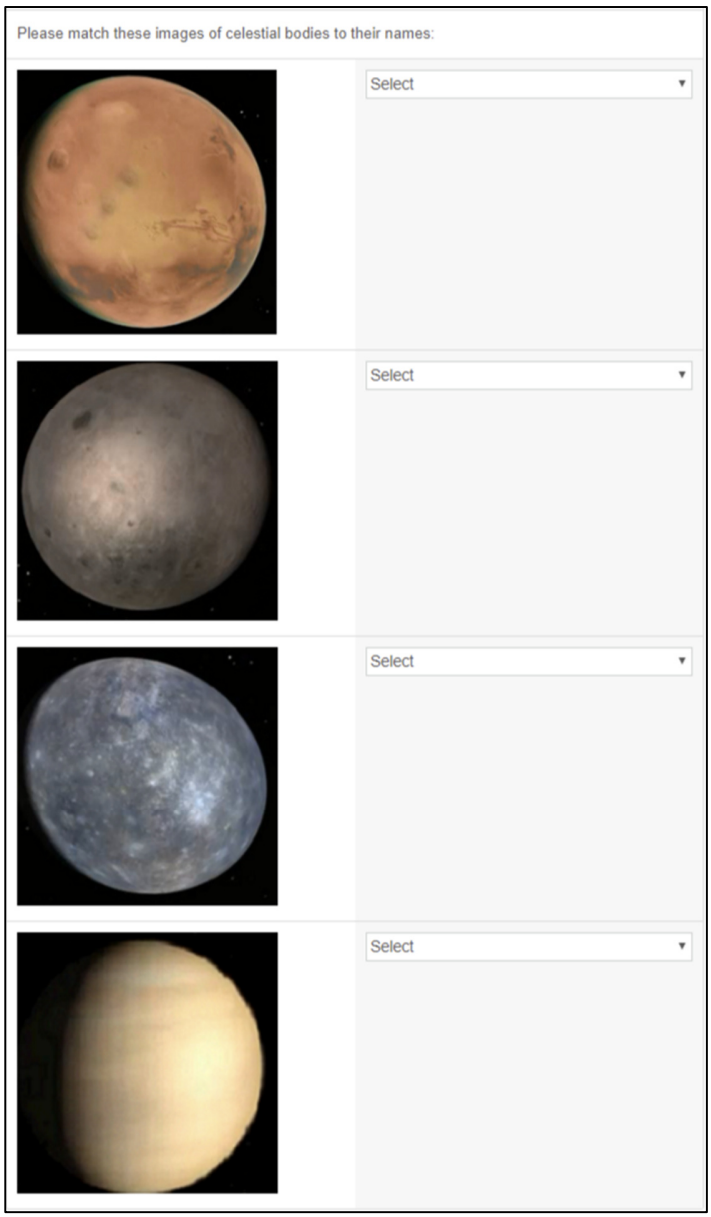


Figure 4. An example of the recognizing questions

The recalling question. There were two recalling questions per quiz. Its purpose was to investigate whether participants could clearly remember the characteristics of celestial bodies that they had learned on that day. The common question design was that participants were imagined to build a settlement on a celestial body, they needed to point out the geological and / or climatic and / or orbit characteristics on the celestial body based on what they had learned in the VR tour. Figure 5 is an example of the recalling questions.

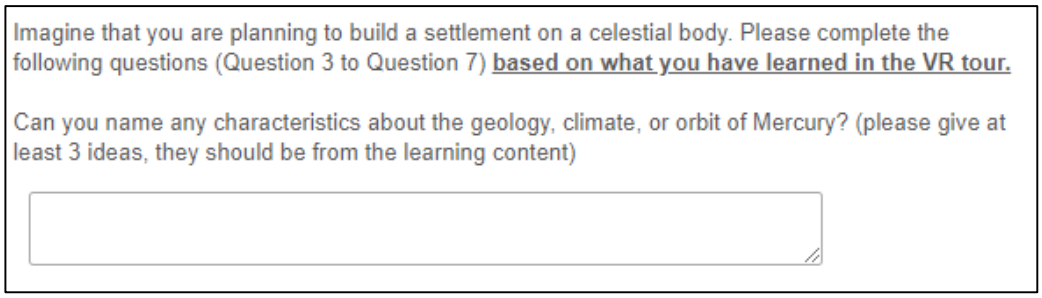


Figure 5. An example of the recalling questions

The dimension of understand. Understanding is related to construct new information with an organizational structure, interpret the knowledge with the learner's own words, and transfer prior knowledge to a new context (Krathwohl & Anderson, 2009). There were two questions per quiz to investigate participants' learning in this dimension. These questions followed participants' answers to the recalling questions. The common question design was that participants needed to point out what life on the celestial body would be like. Figure 6 is an example of the understanding questions.

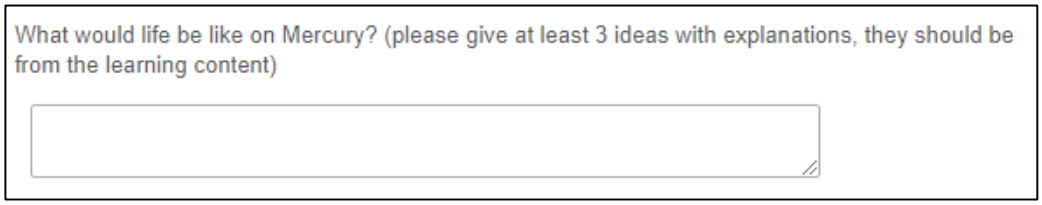


Figure 6. An example of the understanding questions

The dimension of evaluate. Evaluate is to put scattered information together to form a whole and make judgement (Krathwohl & Anderson, 2009). After completed the questions that tested the cognitive process of understanding, participants would need to answer which celestial body had greater potential for human to live. Figure 7 is an example of the evaluating questions.

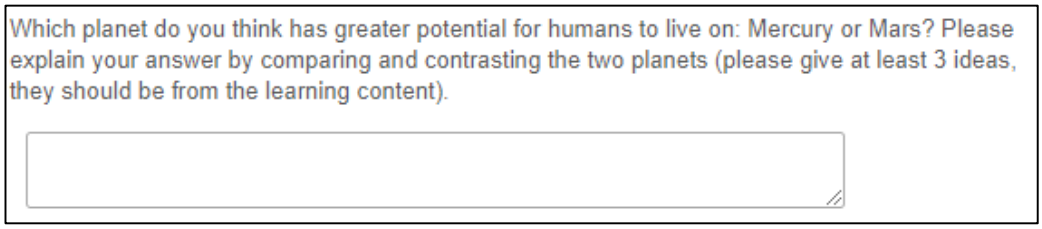


Figure 7. An example of the valuating questions

Experimental Design

The experiment had three sessions with three groups of participants. The materials of VR training and learning in the three sessions were from the content of the VR application *Titans of Space*. Specifically, the content of training was its first section, which included Earth, Mercury, Venus, Mars and their moons; the content of session 1 learning task was its second section, which included Ceres, Jupiter and its moons; the content of session 2 learning task was its third section, which included Saturn and its

moons; the content of session 3 learning task was its fourth section, which included Chariklo, Uranus, Neptune, Pluto, and their moons.

The three groups were named as the Go group, the Go-Rift group, and the Rift group. Participants in the Go group used the Go-version VR application to learn across sessions. It was to simulate a continuous learning experience in a low-level VR condition. Participants in the Go-Rift group used the Go-version VR application in the first two sessions and then used the Rift-version VR application in the third session. This was to simulate a change of the learning environment, which was from a low-level to a high-level VR condition, which is namely a novel condition. Participants in the Rift group used the Rift-version VR application to learn across sessions. This was to simulate a continuous learning experience in a high-level VR condition. After participants completed the VR learning task, they needed to fill in the survey questionnaire and answer the quiz immediately in each session.

If a session was the first time that a participant engaged in one of the two VR conditions, a short VR training (less than 10 minutes) was provided before the participant started the formal VR learning. Otherwise, the participant was only required to read the *instruction* and the *sample quiz* (questions for the content of training) before learning, which helped the participant recall the key points in the learning task.

Pre-test and post-test were assigned at the beginning of the first session and the end of the third session, respectively. The pre-test and the post-test had the same questions. Both included an online astronomy knowledge test and a three-minute paper-based mental rotation test. There were two purposes for designing the pre- and post-tests in this experiment. One was to ensure that the cognitive level of the participants in these

three groups was generally similar, so that the data collected during and after the experiment were comparable among groups. The other was to check whether participants' spatial skills and astronomy knowledge increased after this VR learning activity.

Table 4 shows the experimental arrangement for each session in each group. The tasks that participants took were different among sessions within a group and the tasks that participants took at the same session were also different among groups. For example, when a participant in the Go group started the first session experiment, this participant would (1) take a pre-test, (2) read the *sample quiz*, (3) take a Oculus Go system training, (4) complete the session 1 VR learning task, (5) finish the VR experience questionnaire and the session 1 quiz in sequence; when this participant engaged in the second session experiment, this participant only needed to (1) read the *instruction* and the *sample quiz*; (2) complete the session 2 VR learning task, and (3) finish the VR experience questionnaire and the session 2 quiz.

Table 4

The Experimental Arrangement for Each Session in Each Group

Group		Session 1	Session 2	Session 3
Go	VR System	Oculus Go	Oculus Go	Oculus Go
	Tasks before learning	Pre-test Sample quiz Go training	Instruction Sample quiz	Instruction Sample quiz
	Learning task	Session 1 content	Session 2 content	Session 3 content
	Tasks after learning	Questionnaire Session 1 quiz	Questionnaire Session 2 quiz	Questionnaire Session 3 quiz Post-test
Go-Rift	VR System	Oculus Go	Oculus Go	Oculus Rift
	Tasks before learning	Pre-test Sample quiz Go training	Instruction Sample quiz	Sample quiz Rift training
	Learning task	Session 1 content	Session 2 content	Session 3 content
	Tasks after learning	Questionnaire Session 1 quiz	Questionnaire Session 2 quiz	Questionnaire Session 3 quiz Post-test
Rift	VR System	Oculus Rift	Oculus Rift	Oculus Rift
	Tasks before learning	Pre-test Sample quiz Rift training	Instruction Sample quiz	Instruction Sample quiz
	Learning content	Session 1 content	Session 2 content	Session 3 content
	Tasks after learning	Questionnaire Session 1 quiz	Questionnaire Session 2 quiz	Questionnaire Session 3 quiz Post-test

Interview Design

Participants were randomly selected to be interviewed immediately after they finished other tasks in the third session, respectively. The form of interview was one-to-one. The questions were semi-structured and related to the assessed factors in this study (i.e., presence, sense of embodiment, motivation, cognitive load, engagement, and perceived learning effectiveness). In general, there were two types of questions for each assessed factor. The first type was asking interviewees to describe their general feelings

with explanations. For example, the question “Did you have a sense of being in the universe when you were in the virtual tour? Or a sense of being in the real-world?” was to assess the degree of presence that an interviewee felt. “Did you feel that you could control the virtual body as if it was your own?” was to evaluate interviewees’ sense of embodiment in the VR learning experience. The second type of questions were asking interviewees to provide a ranking of their feelings across sessions. For example, a follow-up question of “Did you have a sense of being in the universe when you were in the virtual tour? Or a sense of being in the real-world?” was “Did this sense change across sessions? If there were changes, how and why? Could you give me an order from most to least sense of being in the universe?”.

Procedure

Participants were randomly assigned into the three groups and engaged in this study thrice, respectively. The interval between two neighboring sessions was 4-7 days. The first 40 participants were equally assigned into the three groups. However, the Rift group showed a higher attrition rate than the other two groups. Thus, more participants were assigned into the Rift group than the other two groups in the later 37 ones, but the assignment was still random. The purpose was to guarantee that there were roughly the same number of participants who completed the three sessions among groups.

Procedure of the mental rotation test. The questions in the mental rotation test were the same in the pre-test and post-test stages. The questions were offered by Dr. Michael Peters directly. Based on Dr. Peters’ request of keeping the test questions private, this test was paper based. The experimenters first offered the introduction page of the

mental rotations test to participants to read, so that they could understand what mental rotation is and how to answer questions in the test. When participant felt that they were ready for the test, they would inform the experimenter. Then the experimenter simply introduced the test requirement to the participant, “There are 12 items in the test paper. However, the test time is only 3 minutes. Therefore, you do not have to finish all these questions, but try to make sure that your answers are correct.” After it, participants would start their test. Three minutes later, the experimenter would say: “Time is up.” The participant would stop the test and submit the test paper to the experimenter.

Procedure of the Go-version VR training. The Go-version VR training required approximately 6 minutes. The experimenters offered two documents named *Go-version training instruction* (this document is a part of the document *experimental instructions*) and *sample quiz* to participants to read first. After showing these two documents, the experimenter said “Please read these two documents first. After that, you will wear the headset, engage in the VR environment, and do the VR training task introduced in the training task document by steps. Questions in the *sample quiz* are similar to the questions in the quiz that you need to answer after the VR learning task. Reading this sample helps you understand the learning goal in the VR experience”. Then the participant would start reading these two documents. When the participant felt ready, they informed the experimenter. The experimenter summarized the training task and introduced the Go VR system to the participant. The experimenter indicated that the trigger button on the controller was the only useful button in this study. Next, the participant would wear the Go headset and a pair of headphones, hold the controller with the dominant hand, and start to implement the training task by him or herself. After approximately 3.5 minutes of

experience, the experimenter asked the participant whether he or she could freely control the VR application. If the response was yes, the experimenter would remove the participant's headset. If no, the experimenter would communicate with the participant, find and solve possible problems.

Procedure of the Rift-version VR training. The Rift-version VR training required 8-10 minutes. The experimenters offered two documents named *Rift-version training instruction* and *sample quiz* to participants to read first. After showing these two documents, the experimenter said "Please read these two documents first. After that, you will wear the headset, engage in the VR environment, and do the training task introduced in the document by steps. Questions in the *sample quiz* are similar to the questions in the quiz that you need to answer after the VR learning task. Reading this sample helps you understand the learning goal in the VR experience". Then the participant would start reading these two documents. When the participant felt ready, they informed the experimenter. The experimenter summarized the training task and introduced the Rift VR system to the participant. Particularly, the experimenter held the trigger button on the controller, and said that the participant could freely rotate a celestial body or grab the surface of a celestial body when the first-person avatar touched the surface of the celestial body in the VR environment. Next, the participant would wear the Rift headset, hold the controllers on both hands, and start the training task by him or herself. There were four steps or named skills that the participant needed to master in the training process: (1) reading text information on the dashboard by pages, (2) clicking the buttons on the two sides of the dashboard, (3) holding the trigger button and freely rotating the celestial body, and (4) flying out of the spaceship and grab the surface of the celestial body like

climbing a mountain. When both the participant and the experimenter felt that the participant had mastered the skills and was ready for the formal VR learning, the experimenter would remove the participant's headset.

Procedure of the VR learning task. If there was a VR training before the formal learning task (i.e., the first session in the three groups, and the third session in the Go-Rift group), the experimenter would introduce the learning task to participants after they finished their training tasks. Then these participants started their formal VR learning experience. If there was not a VR training (i.e., the second and third sessions in the Go and Rift groups respectively, the second session in the Go-Rift group), the experimenters asked participants to read the documents of *Go-version training instruction* or *Rift-version training instruction* and *sample quiz* first. The experimenter also explained to the participant that reading these documents helped them recall what they had learned in the previous session(s). When the participant felt ready, they informed the experimenter. The experimenter showed the participant the corresponding learning instruction document and summarized the learning task. Then the participant started the VR learning experience. Participants assigned with the Go VR system wore the Go headset and a pair of headphones, held the controller with the dominant hand. Participants assigned with the Rift VR system wore the Rift headset and held the controllers on both hands.

To minimize the difference in time spent across groups and sessions, the following rules of learning duration were taken:

1. Participants who finished the first-round learning in 15 minutes could learn again based on their willingness. The maximum learning time was 20 minutes. When 20 minutes was reached in the middle of the second-round learning, the

experimenter removed the participant's headset. If 20 minutes was still not reached after the second-round learning, the experimenter also removed the participant's headset directly to avoid participants' perceived boredom.

2. If participants spent more than 15 minutes to finish the first-round learning, they did not have the opportunity to learn again. The experimenter removed the participant's headset directly after they finished learning.

Procedure of other tasks. Participants sat in front of a laptop and answered the questions in the astronomy knowledge test, questionnaire, or quizzes. There was no time limitation. A total of 20 participants were randomly selected to be interviewed after they finished other tasks in the third session. The interview was face-to-face, and audio recorded after the permission of interviewees. Most interviews were completed in 20 minutes.

Data Analysis

Data analysis was conducted in four phases. They were grading quizzes, structure equation modeling (SEM), quantitative data comparison, and qualitative data analysis. Phase I (grading quizzes) was the starting point of data analysis. It focused on providing unbiased grades for participants' responses to the quizzes in each session. The grading rubric was developed after all items were graded by the first researcher. Then a second researcher was involved to improve the rubric and establish the interrater reliability of the grading results. When the actual data exceeded the recommended threshold value of inter-rater reliability, the grading results were used as the measurement of participants'

learning achievement. This phase was indispensable for establishing the learning model in phase II and comparing participants' learning achievement in phase III.

Phase II (SEM) aimed to validate the hypothesized model proposed in the literature review chapter. It included a priori power analysis, an exploratory factor analysis (EFA), a confirmatory factor analysis (CFA), and an assessment of the structural model. A priori power analysis provided evidence whether the sample size was enough for SEM with anticipated effect size. EFA checked the unidimensionality of all factors in the model and confirmed that the model could be analyzed as a standard confirmatory model at the next step. CFA assessed the confirmatory model, namely the measurement model, with multiple goodness-of-fit indices and provided evidence for convergent validity and discriminant validity. The assessment of structure model was to estimate and interpret the magnitude of causal effect between assessed variables based on the measurement model. It also reported the good-of-fit indices for the structural model. This procedure was consistent with a two-step SEM process (Anderson & Gerbing, 1988). That is, a well-rounded measurement model is the foundation of the final structural model.

Phase III (data comparison) focused on data comparison. The primary analysis was performed under the framework of the multiple imputation method, while secondary analysis was performed using data from participants who completed the three sessions in the experiment. The multiple imputation method included three steps: imputation, analysis, and pooling. Following the strategies recommended by Buuren (2018), a predictor model was established in the imputation step. This predictor model helped the researcher impute missing data and then conduct data analysis. The specific data analysis

included assessing the changes of each factor from the first session to the third session for both the Go group and the Rift group, and comparing the value of each factor between the second session and the third session for the Go-Rift group. A comparison of pre- and post-test scores among groups was also a part of the analysis work. The secondary analysis did the same comparison as the primary analysis. Its purpose was to validate the analysis results after multiple imputation.

Phase IV (qualitative analysis) interpreted interviewees' perceptions and experience in the VR learning experiment. It included a two-round process. The first round was conducted based on a series of questions in longitudinal qualitative inquiry (Saldaña, 2002). Its purpose was to explore the possible changes among sessions within each group and differences among groups in the same session. The second-round analysis reorganized the first-round results. It directly focused on triangulating and understanding the significant effects or changes found in the quantitative results.

Phase I: Grading quizzes. There were four types of questions in the quiz taken in each session: recognizing question, recalling question, understanding question, and evaluating question. The grading of recognizing questions was completed automatedly by the online quiz system. Its scoring method was one point for one correct match between the image of a celestial body and its name. A participant could get four points if this participant correctly answered all the four recognizing questions in a quiz. Manual grading was required for the remaining questions in each quiz. The revised Bloom's taxonomy was used to guide the manual grading process. After the first-round grading, an initial version of the manual grading rubric was developed. Then the second researcher was invited to grade a total of 75 items (25 per quiz) with this rubric.

When the second researcher completed his grading, the two researchers met together to discuss their grading results and possible issues in the rubric. Then the first researcher revised the rubric, and updated the scores for all the 191 items, while the second researcher graded another 30 items (10 per quiz) with the revised rubric.

After the second-round grading, Pearson's correlation coefficient (Hunt & Dentistry, 1986) was used to measure the agreement and consistency between the two researchers for their grading on the 30 items. The reason of using Pearson's correlation coefficient to measure the interrater reliability was because the quiz score was a continuous variable, and the grading results showed that it was normally distributed. The threshold value was $r = 0.8$, as $r \geq 0.8$ represents a very strong correlation (Evans, 1996).

Manual grading rubric. The rubric was congruent with the design principle of session quizzes in the experiment. That is, participants' VR learning achievement should be well presented from multiple dimensions. The specific dimensions assessed included remember, understand, and evaluate. A participant's quiz score was all points that the participant obtained in a quiz in that session. There was no limit on how many scores a participant could have in a quiz. The principles of assessing participants' knowledge were as follows:

The dimension of remember. Only the question that assessed participants' recalling of knowledge needed to be manually graded. Recall means retrieving the relevant knowledge directly from long-term memory, such as recollecting a friend's birthday (Krathwohl & Anderson, 2009). The question that assessed participants' capability of recalling knowledge was "Imagine that you are planning to build a settlement on a celestial body. [……] Can you name any characteristics about the geology, climate, or

orbit of [the name of a celestial body]? (please give at least 3 ideas, they should be from the learning content)".

The grading rule was *one point per correct fact.* This fact could be either from text information in the VR environment, such as "Io has 400 active volcanos" or from visual observation, such as "Enceladus has a dark side". The criterion were: (1) this fact had been introduced in the VR application, (2) it was relevant to the question, (3) and the fact was not common knowledge that everyone knows, such as "Jupiter is a planet".

The dimension of understand. Understand means individuals already know how to construct meaning from learning materials and build connections between new pieces of knowledges and between new knowledge and prior knowledge (Krathwohl & Anderson, 2009). Classifying, summarizing, and explaining are three sub-cognitive processes in this dimension (Krathwohl & Anderson, 2009). The design of questions for the dimension of understand in quizzes was "What the life would be like on [the name of a celestial body]?". To answer this type of question, participants first classified all facts that they had learned in the VR experience by whether the fact might influence humans' life on the target celestial body. Then they were instructed to present their ideas with those facts which might influence human's life. Some participants generalized their ideas to a theme at the beginning or summarize their ideas as a single statement at the end, or both. It had two grading rules:

1. One point per idea. "Idea" meant something new, but not an explanation of a fact which was not directly related to life. This idea should be supported by explanations or facts. For example, "the volcanoes could cause life to be dangerous". Another example was, "IO has huge amount of radiation [, which] will harm the live thing." Independent

ideas without explanation or facts to support could not get points. For example, “It would be a struggle to travel, difficult to find something nice to look at, difficult to stay for long.” Moreover, common knowledge, such as “humans cannot breathe due to lacking oxygen”, also could not get any point.

2. The theme and summary generalized from idea(s) also could get points. One point per theme or summary. The relationship between a theme / summary and ideas should be understandable.

Example 1:

Life on Io would be almost impossible for humans at the very least. With the 400+ volcanoes located on the celestial body, the threat of lava would always be prevalent. The sulfur deposits would also make building on top much more difficult.	A theme, 1 point An idea, 1 point An idea, 1 point

Example 2:

Life on Enceladus would be […...] would probably sustain life pretty well, with all the water one needs to live readily available, which could be converted to fuel machinery and breathe.	A theme, 1 point An idea, 1 point An idea, 1 point

The dimension of evaluate. Evaluate is defined as making judgements based on standards (Krathwohl & Anderson, 2009). The standards can be quality, effectiveness, efficiency and any other factors that are regarded as important by either students or teachers (Krathwohl & Anderson, 2009). In this study, participants were required to use their standards to decide which celestial body had greater potential for human to live on. This standard should be presented through comparing the two celestial bodies listed in the question. The design of questions for the dimension of evaluate in quizzes was

“Which celestial body do you think has greater potential for humans to live on: [the first celestial body name] or [the second celestial body name]?”.

The grading rule of responses to the evaluating questions was o*ne point per criterion*. Some responses did not include any structured comparison or contrast but showed very persuasive advantages or weaknesses of one celestial body which might influence the potential that humans live on the celestial body. This type of answer still could get points. For example, “Europe itself has water which is the source for all living things.” Furthermore, some participants answered that human could not live on both celestial bodies and provided reasons. If their reasons made sense, their responses could get point(s). For example, “Neither if an option, the length of time to solely reach these two bodies would outweigh the positives of living there.” In addition, an uncertain but imaginative reason was acceptable. For example, “While they’re both relatively devoid of craters, [Triton] having Neptune as an orbital neighbor seems handier for avoiding asteroids than having another dwarf planet.” However, a list of abstract factors without any analysis could not get a point. For example, “Europa has a greater potential because of the proximity, chemical make-up, and temperature.”

Phase II: Structure equation modeling (SEM). ***The priori power analysis.*** In general, there is still no consensus about the minimum sample size in a study where SEM is used. In this study, the minimum sample size was calculated by the power analysis framework proposed by Westland (2010). With anticipated effect size 0.30, desired statistical power level 0.80, number of latent variables 6, number of observed variables 25, and p value 0.05 entered into the web program provided by Soper (2020), the program displayed that the minimum sample size to detect effect was 161, the minimum

sample size for model structure was 94, and recommended minimum sample size was 161. Therefore, the 191 cases collected in this experiment was enough for SEM.

EFA. There were three steps in the SEM phase. First, an EFA was conducted on the collected cases with SPSS 25. Its purpose was to check the unidimensionality of all factors in the model and to ensure that the item loadings were in line with the factors used in prior research (Cheng et al., 2014). The Kaiser–Meyer–Olkin measure of sampling adequacy (0.755) and statistically significant Bartlett's test of sphericity ($p < 0.001$) indicated that EFA was applicable to the dataset. A combination of the scree plot (Cattell, 1966) and parallel analysis (Horn, 1965; O'Connor, 2000) was used to estimate the probable number of factors to retain. The factor extraction method was principal axis factoring (PAF). Since oblique rotation techniques can be used whenever the factors are correlated (McCoach, Gable, & Madura, 2013), an oblique rotation method (i.e., Promax method) was chosen to improve the interpretation of factors. The criteria for retaining items were: (1) all final items load > 0.40 on one factor and < 0.30 on all other factors and (2) if a factor had at least three items, its Cronbach's $\alpha > 0.70$.

CFA. As the second step of SEM, CFA was conducted with Mplus 7.4. First, the nature of collected data was checked to decide the type of estimator. This is because Maximum Likelihood (ML) estimation, the default estimator in Mplus, assumes that data are multivariate normally distributed (Enders, 2010a). Using this estimator may cause bias for parameter estimate on multivariate non-normal data. In this step, potential outliers were identified as well. Outliers are data that reside further than 1.5 times the interquartile range (IQR) from the median number. Retaining outliers in analysis may undermine the results of normality analysis (Hair, Black, & Babin, 2009). Therefore,

researchers need to investigate the possible outliers and determine whether these data can be retained. In addition, the recommended acceptable range of univariate skewness in CFA is between -2 and 2, the range of kurtosis is between -7 and 7 (Finney & Distefano, 2006; West, Finch, & Curran, 1995), and the value of multivariate kurtosis index (Mardia's normalized multivariate kurtosis) is between -3 and 3 (Bentler, 2006).

After the estimator was determined and identified outliers were removed, the goodness-of-fit of the measurement model was assessed. Evidence that a model fits data well in CFA included (1) Root Mean Square Error of Approximation (RMSEA) < 0.06 (Hu & Bentler, 1999); (2) Comparative Fit Index (CFI) / Tucker-Lewis Index (TLI) > 0.90 (Bentler & Bonett, 1980); (3) Standardized Root Mean Square Residual (SRMR) < 0.08 (Hu & Bentler, 1999).

In addition, convergent validity and discriminant validity were assessed for the measurement model, respectively. Convergent validity requires items of a specific factor converge or share a high proportion of variance in common (Hair et al., 2009). To assess convergent validity, three measurements were used: (1) factor loadings, (2) average variance extracted (AVE), and (3) composite reliability (CR).

To establish convergent validity, there should be high loadings for all items on a factor. The threshold value of standardized factor loading is 0.50 (Hair et al., 2009).

AVE is the mean variance of the item loadings on a factor (Hair et al., 2009). It can be calculated by the following equation:

$$\text{AVE} = \sum_{k=1}^{n} L_k^2 / \text{n}$$

The L_k is the standardized factor loading. The k is the number of items. An AVE higher than 0.50 is required for adequate convergence of the items on a factor.

CR indicates the internal consistency of items on a factor. It can be calculated by the following equation:

$$CR = (\sum_{k=1}^{n} \lambda i)^2 / [(\sum_{k=1}^{n} \lambda i)^2 + (\sum_{k=1}^{n} \varepsilon i)]$$

The λ_i is the standardized factor loading. The ε_i is the error variance. The k is the number of items. CR is suggested to be higher than 0.70 (Hair et al., 2009).

Discriminant validity provides evidence that a factor is distinct from other factors. Its value is determined by the AVE values of any two factors and the squared correlation between these two factors. A ratio higher than one provides evidence for a high discriminant validity.

Structural model. The final step of SEM was to assess the structural model, which includes the causal relationships among factors. These relationships are represented by the model's parameter estimate (e.g., path coefficient, p-value). A significant path coefficient would support its corresponding hypothesis proposed in the literature review chapter. The indices of model-data fit were used again to determine whether these relationships fitted well to the data.

In addition, when the structural model reached the criteria of good model-data fit, a furthermore model modification would be conducted. Specifically, it was possible that *session* and its interaction with *VR condition* would significantly influence participants' perception and psychological responses. Therefore, extra paths from *session* and from its interaction with *VR condition* to other factors were established and assessed.

Phase III: Data comparison. ***The priori power analysis.*** The priori power analysis showed that it was hard for this VR learning study to reach the minimum sample size for a desirable power in the quantitative data comparison. Specifically, it needed at

least 159 samples in a one-way ANOVA omnibus test (f = 0.25, p = 0.05, power = 0.80, number of groups = 3), 28 samples per group in a one-way repeated measures ANOVA omnibus test (f = 0.25, p = 0.05, power = 0.80, number of measurements = 3), 128 samples in an independent samples T test (Cohen's d = 0.50, p = 0.05, power = 0.80), and 34 samples in a paired-samples T test (Cohen's d = 0.50, p = 0.05, power = 0.80).

Missing data statistics. A total of 77 students participated in the first session of this experiment. Due to dropout, the number of participants in the second session was 64 and the number in the third session was 50. The main reasons of their dropout were: (1) personal reasons, such as cybersickness, cold, family emergency, and did not need more course credits; (2) objective reasons, such as the VR equipment could not work or was not available and student did not obey the experimental rules; (3) unidentified reasons, experimenters had sent emails to participants who did not show up on time, unidentified means that experimenters did not receive any reply from the participant.

Table 5 is the statistics of dropouts and their reasons. Table 6 shows the number of participants in each session for each group. The Rift group (41.94%) had a higher attrition rate than the Go group (30.43%) and the Go-Rift group (30.43%). This might be caused by its more complicated learning task and longer experimental time than the other two groups in the first two sessions (see Table 7).

Table 5

The Statistics of Dropouts and Their Reasons

Session	Reasons		
	Personal	Objective	Unidentified
Second	3	3	7
Third	4	2	8

Table 6

The Number of Participants for Each Session in Each Group

Group	Session 1	Session 2	Session 3	Sum
Go	23	19	16	58
Go-Rift	23	20	16	59
Rift	31	25	18	74
Sum	77	64	50	191

Table 7

VR Learning Time

Group	Session 1	Session 2	Session 3
Go	710.34 (SD=276.52)	933.00 (SD=255.94)	854.63 (SD=225.80)
Go-Rift	648.64 (SD=256.59)	951.58 (SD=227.05)	1059.88 (SD=349.77)
Rift	1003.35 (SD=276.52)	1057.04 (SD=204.66)	1092.17 (SD=281.05)

Note. Unit: second

Furthermore, conducting a comparison between the completer and non-completer groups is another important part of missing data analysis. This work was included in the process of establishing the predictor model for multiple imputation. It will be introduced in the following subsections.

Multiple imputation. The number of participants in the third session was 50. This meant only 50 out of 77 cases were complete in this study. The rate of incomplete data was approximately 35.06%. Compared to traditional methods for handling missing data (e.g., listwise, simple regression imputation, mean imputation), the multiple imputation method shows its strength in minimizing bias for estimates when the proportion of missing data is large (Rubin, 2004; Buuren, 2018). Therefore, the multiple imputation method was chosen to handle incomplete data in this study. In general, the multiple imputation method includes three phases: imputation, analysis, and pooling (Enders, 2010d). In the imputation phase, several complete-case datasets are created with plausible

data values for the missing data in the original dataset. After this, the statistical analysis is performed on these complete-case datasets independently in the analysis phase. Finally, the parameter estimates from different datasets are pooled into one estimate. In this study, the imputation phase was implemented on the R platform, the analysis and pooling phases were carried out on SPSS 25. Particularly, SPSS automatically finished the pooling work after data were analyzed in each complete-case dataset.

Multivariate Imputation by Chained Equations (MICE). This study used the implementation package of multivariate imputation by chained equations (MICE; Buuren & Groothuis-Oudshoorn, 2011) on the R platform to create complete-case datasets in the imputation phase. The MICE algorithm is an implementation of Gibbs sampling method on the missing data imputation. Specifically, this algorithm involves a process that repeatedly iterates all missing data and performs univariate regression imputation for variables in sequence after having filled in all missing values for the variables in the imputation model (Carlin, 2014; Raghunathan, Lepkowski, Hoewyk, & Solenberger, 2001; Buuren & Groothuis-Oudshoorn, 2011). Its imputation relies on the identified univariate regression models for each variable. The algorithm and its implementation package allow researchers to precisely select the imputation predictors for each variable. This characteristic is particularly useful for imputing missing data in small-scale longitudinal studies. In this study, participants filled in the same survey questionnaire thrice and there were approximately 25 items in the survey questionnaire. If all variables were included in the imputation model, the number of predictors for each assessed variable would be at least 75. Due to multicollinearity and computational problems, using

a large number of variables as predictors in a small-scale dataset may not be feasible (Enders, 2010c; Buuren, 2018).

Furthermore, both White and Wood (2011) and Buuren (2018) suggested that the number of imputations should be at least equal to the percentage of incomplete cases in a multiple imputation using chained equations. Since the percentage of missing data was 17.32% in this study (40 out of 231 items, 3 * 77 = 231), the imputation number was set as 20.

Due to the advantage of incorporating noise into the multiply imputed data, the MICE algorithm can much speed up the convergence speed over other approaches (Buuren, 2014). The number of literation is usually set between 5 to 20 (Brand, 1999; Buuren, Boshuizen, & Knook, 1999). Thus, the number of iterations was set as 20 in this study.

Predictor model. The number of predictors for each variable is recommended to be between 15 and 25 in a regression-based multiple imputation model (Buuren, 2018). Buuren et al. (1999; 2018) provided the specific steps for selecting predictor variables in MICE: (1) including all variables expected to be in the complete data model, which would be applied to the data analysis after imputation, (2) including the variables or factors that are related to the nonresponses, (3) including variables that correlate with the target variables, and (4) removing from the variables selected in steps 2 and 3 which have too many missing values within the incomplete cases. When longitudinal data are being imputed and the number of predictors is needed to be reduced, the cross-lagged predictors are recommended to leave out (Buuren, 2018). Table 8 shows an example of a longitudinal dataset. In this example, the values of a variable set (X, Y, Z) are recorded

thrice. The columns X_1, X_2, X_3 are the records of variable X at three different time points in sequence, while columns X_1, Y_1, Z_1 are the records of the variable set (X, Y, Z) at the first time point. The cross-lagged predictors of X_1 are X_2, X_3, Y_1, and Z_1.

Table 8

An Example of a Longitudinal Dataset

Id	X_1	X_2	X_3	Y_1	Y_2	Y_3	Z_1	Z_2	Z_3
1	98	96	89	65	87	98	98	76	77
2	94	77	64	66	89	54	99	44	87
3		…..							

The establishment of predictor model in this study was in line with the recommendation and strategies provided by Buuren and his colleagues. First, all variables collected in this experiment were screened and some were identified as the candidate predictors. Sense of embodiment, presence, motivation, engagement, cognitive load, and learning achievement were the factors to be assessed after imputation. They were measured by 18 variables (items) thrice. Therefore, these 18 variables, which added up to 54 variables, were recognized as the candidate predictors. Furthermore, participants' scores in the pre- and post-tests also would be analyzed after imputation. This included four variables (i.e., scores in the pre-mental rotation test, pre-astronomy test, post-mental rotation test, and post-astronomy test). In addition, demographic information (i.e., age, major, gender, race), group number, and the type of VR system used in each session (i.e., the Oculus Go and the Oculus Rift) were possible variables that influenced the occurrence of missing data and / or correlated with the value of target variables. Thus, these eight variables were also categorized as the candidate predictors.

Although the reasons of dropout had a general classification in this experiment, the proportion of unidentified reasons was large (15 in 27 cases, approximately 55.56%).

In addition, it was possible that dropouts in the same category might have significantly different performance in the experiment. For example, the dropouts for having enough course credits should have better VR learning experience than those left for cybersickness. However, a more detailed division might lead to a very small number of cases in each category. This would lose the value of prediction. Therefore, the reason of dropout was not in the predictor model. Similarly, the VR learning time in each session was not in the model. It had the same missing pattern as the assessed variables in the study. In other words, its contribution to imputation was small. Since the number of candidate predictors for each variable had already far more than 15, the VR learning time in each session was not included in the predictor model.

A total of 66 variables were selected as candidate predictor variables for the predictor model. Then these variables were identified to be the predictors for each target variable (the variable that had missing data) one by one. First of all, group number was identified as a predictor for all target variables in the final predictor model due to the requirement of the further comparison. Second, the value distribution of the remaining 65 variables between the completer and non-completer groups were checked one by one. The statistics methods were independent samples T test and χ^2 test. If there was a significant difference between the two groups, the variable would be in the final predictor model for all target variables.

When the previous step was completed, independent samples T tests and / or one-way ANOVA tests were conducted to determine whether there were significant differences for target variables under different dimensional categories (i.e., age, race, gender, major, VR system). A significant difference would mark the categorical variable

as a predictor for the target quantitative variable in the final model. For example, if gender was not selected as the predictor for all target variables in the last step, an independent samples T test would be conducted to measure the difference between male and female participants for their scores in different target variables. If there was a significant mean difference found from a target quantitative variable, gender would be a predictor of this target quantitative variable.

Furthermore, Pearson correlation analysis was conducted to the remaining quantitative variables in the three sessions, respectively. If the magnitude of the correlation between one variable and the target variable was in the top 10 for the target variable in that session, this variable would be categorized as a predictor for the target variable. The four variables from pre-test and post-test were included in the Pearson correlation analysis for all three sessions. Their predictors determined in this step were the variables which had the top 10 magnitude correlation in the three sessions.

After the three steps, the predictor model was established. The predictors of one target variable included (1) its group number, (2) variables that showed a significant difference distribution between the non-completer and completer groups, (3) categorical variables that the target variable had significant different mean values between or among their categories, (4) 10 quantitative variables that had the top 10 magnitude correlation with this target variable, and (5) two variables which had the same item content as the target variable but were measured in different sessions.

Quantitative data comparison. The assessed factors in the quantitative data comparison included presence, sense of embodiment, motivation, engagement, cognitive load, learning achievement, and scores in the pre- and post-tests. The value of each

assessed factor was the mean of its items in the measurement model, respectively. For example, the value of motivation was the mean of items Mo_1, Mo_3, and Mo_4, while the value of learning achievement was the quiz score. Before conducting the further data analysis, data collected from participants who received full mark or zero score in either the pre-mental rotation test or the pre-astronomy knowledge test were removed to avoid floor and ceiling effects. Then χ^2 test and one-way ANOVA test were conducted to ensure that there was no significant difference among the three groups in both their demographic information and pre-test scores. This provided evidence that the values of assessed factors were comparable among the three groups.

After the preliminary work, data in the first session were compared between the low-level VR condition (the Go and the Go-Rift groups) and the high-level VR condition (the Rift group). The purpose was to know the possible influence of novelty on assessed factors at the beginning and establish a baseline. After that, analysis was conducted to assess the possible changes of these factors from the second session to the third session for the Go-Rift group. The purpose was to find the possible influence of novelty on participants' VR learning experience in a recurring situation. This is because the high-level VR condition was a novel condition for the participants in the Go-Rift group after they had experienced the low-level VR condition twice. Paired-samples T tests were conducted in this analysis. Specifically, the values of assessed factors in the third session were compared with their values in the second session for the Go-Rift group. Then the values of assessed factors in the third session were compared among the three groups. The goal of this step was to find whether the change of the assessed factor for the Go-Rift group would lead to its value significantly different from the other two groups. Since the

pooling result of one-way ANOVA analysis was still not available in SPSS, the step of omnibus test was skipped. Independent-samples T test was used to pair-wisely compare the values between groups directly.

In addition, the changes of assessed factors from the first session to the third session were compared between the Go group and the Rift group. The purpose was to answer the third question in this study. That is whether the assessed factors decrease slower in a high-level VR condition. Paired-samples T test was used in this analysis.

Finally, paired-samples T tests were conducted to check the possible score differences between the pre-test and the post-test in the mental rotation test and the astronomy knowledge test in each group. The purpose was to investigate whether participants' spatial ability and astronomy knowledge had a significant increase after this experiment and whether there were differences among groups.

Completer analysis. To validate the comparation results after multiple imputation, a completer analysis was conducted. While the assessed factors were the same as the comparison after multiple imputation, the dataset was from all the participants who completed the three sessions.

The steps were almost the same as the data comparison analysis after multiple imputation. There were two differences. The first was the repeated measures ANOVA method was used to evaluate the possible difference among sessions for each group. The second was the polynomial effect was assessed to determine whether there was a definite changing trend for an assessed factor through sessions in each group.

Phase IV: Qualitative data analysis. The reasons of involving a qualitative strand were to triangulate quantitative findings and to explain why some hypotheses and

predictions were rejected by the quantitative results. After the whole experiment was completed, the audio recording of interviews taken in the third session was transcribed. A total of 20 transcripts (six from the Go group, eight from the Go-Rift group, and six from the Rift group) were used in the data analysis.

The qualitative data analysis included two rounds. The first-round analysis focused on comparison. The analysis was framed by a series of questions that explored the possible changes in longitudinal data (Saldaña, 2002). These questions included:

1. What was different from one group of data to the next?
2. What increased or emerged through sessions?
3. What decreased or ceases through sessions?
4. What remained constant or consistent through sessions?
5. What context appeared to influence and affect participant change(s)?
6. Which changes interrelated through sessions?

In this study, comparison was conducted based on six dimensions: presence, sense of embodiment, motivation, cognitive load, engagement, and perceived learning effectiveness. The specific comparisons applied in the analysis included:

1. Comparison between the first session and the latter two sessions from interviewees in the Go group and the Rift group, respectively.
2. Comparison between the first two sessions and the third session from interviewees within the Go-Rift group.
3. Interviewees' general impression about their VR experience between the Go group and the Rift group.

To facilitate the comparison and the discovery of themes, a table was created. Its columns contained interviewee's ID, demographic attributes (e.g., gender, race, prior VR and video game experience), and the six dimensions, while each row contained the information and the responses from one interviewee. Through comparison, ranking patterns and observations related to these dimensions were found.

The second-round analysis was driven by the hypotheses proposed in the literature review chapter. It particularly focused on the significant findings from quantitative analysis. The specific method was establishing themes through extracting information from each observation. These themes were closely related to the hypotheses and quantitative findings.

"Deviating cases" are cases that do not confirm to the common patterns in any sub-analyzed units (Hermanowicz, 2013). Answering the questions of why and how they depart from the mode allows researchers to strengthen assertions and to qualify conclusions of patterns. Therefore, this type of cases was also checked in this study.

CHAPTER 4

RESULTS

The purpose of this study was to assess the possible impact of novelty on the effectiveness of VR in STEM learning. A hypothesized model was established to describe how the VR condition affects an effective learning experience. Predictions were also made on the possible changes of individuals' engagement and learning in this learning experience. To validate the model and predictions, an experiment was designed and implemented. Both quantitative and qualitative data were collected in this experiment and then they were analyzed in four phases: grading quizzes, SEM, quantitative data comparison, and qualitative data analysis.

This chapter presents the results of the data analysis. It includes four subsections. These subsections use the same titles as the four phases in the data analysis. They are grading quizzes, structural equation modeling, data comparison, and qualitative results.

Grading Quizzes

This section presents the grading results. Pearson's correlation coefficient between the two researchers' rating for the 30 items in the second-round grading was 0.905 ($p < 0.001$). Since the result was much higher than the threshold value ($r = 0.80$), the reliability of grading was established. The first researcher's grading results were accepted as participants' quiz scores in this study. Table 9 shows the statistics of grading results for the whole 191 items. They were from quizzes in the three sessions of the experiment. Items in the session 1 quiz had the highest means in the recognizing

question, the understanding question, the evaluating question, and the total score. Items in the session 2 quiz had the highest mean in the recalling question.

Table 9

The Grading Results of Session Quizzes

		Remember		Understand	Evaluate	Sum
		Recognize	Recall			
	Range	0-4	0-12	0-8	0-3	1-26
Mean	Session1	3.36	4.13	2.34	1.42	11.25
	Session2	2.82	4.48	2.05	0.77	10.13
	Session3	3.14	4.24	2.10	1.02	10.5

Structural Equation Modeling (SEM)

This section includes the results in Phase II, SEM. Specifically, it had three parts: EFA, CFA, and the assessment of structural model. CFA was completed in two steps: the nature of the data was analyzed first to determine the method for model estimation, then the quality of measurement model was assessed by criteria.

Exploratory factor analysis (EFA). In the EFA step, parallel analysis indicated a six-factor solution while the scree plot indicated a five-factor solution. Compared to the fix-factor model, the six-factor solution could be better interpreted and retaining items. Therefore, the six-factor model was accepted in the first-round EFA. Table 10 is the pattern matrix of this round EFA.

Table 10

The Pattern Matrix of the Initial EFA

Item	Factor					
	Presence	Sense of Embodiment	Motivation	Engagement	Cognitive Load	---
Pr_1	**0.66**	-0.05	0.13	0.13	0.05	-0.04
Pr_2	**0.76**	0.08	-0.01	-0.01	-0.02	0.14
Pr_3	**0.69**	-0.14	-0.09	0.03	-0.09	0.09
Pr_4	**0.60**	0.10	0.08	-0.19	-0.04	0.01
Em_1	0.29	**0.79**	-0.25	0.08	-0.01	0.00
Em_2	-0.05	**0.84**	0.01	-0.05	0.01	-0.29
Em_3	-0.12	0.00	0.09	0.09	0.01	-0.44
Em_4	-0.22	**0.81**	0.14	0.10	-0.00	0.09
Em_5	0.01	**0.48**	0.06	-0.13	0.04	0.22
Em_6	0.04	0.22	0.12	-0.05	-0.05	-0.40
Em_7	-0.03	0.01	0.20	0.09	0.04	0.40
Mo_1	0.09	-0.02	**0.85**	-0.03	-0.04	0.12
Mo_2	0.04	-0.03	0.32	0.08	-0.10	-0.12
Mo_3	-0.01	0.02	**0.83**	-0.02	0.04	-0.00
Mo_4	-0.01	0.02	**0.97**	-0.04	0.06	-0.01
En_1	-0.03	0.05	0.02	**0.59**	-0.19	0.10
En_2	0.11	-0.04	0.00	**0.82**	0.13	-0.02
En_3	-0.18	0.05	-0.11	**0.78**	0.00	0.15
En_4	0.03	-0.03	-0.03	**0.72**	0.11	-0.11
En_5	0.11	0.12	0.15	**0.41**	-0.05	-0.03
En_6	0.35	-0.08	0.10	0.04	0.19	-0.21
En_7	0.10	0.07	0.16	0.08	0.00	0.39
Cl_1	0.06	-0.02	-0.10	-0.10	**0.76**	0.06
Cl_2	-0.08	0.03	0.09	0.10	**0.84**	-0.01
Cl_3	-0.04	0.13	-0.10	-0.32	0.27	0.17

Note. Extraction Method: Principal Axis Factoring.
Rotation Method: Promax with Kaiser Normalization.
a. Rotation converged in 7 iterations.

Item Em_1 (I felt as if the virtual body or body part was my own body or body part.) belonging to *sense of embodiment* showed cross loading with *presence*. In addition, there were six items under the threshold value of factor loading ($R^2 = 0.40$). These six items were Em_3 (I felt as if I might have more than one body.), Em_6 (I felt as if the

virtual body was moving by itself.), Em_7 (When a virtual celestial body was in front of me, I felt that my body was very small.), M_2 (This activity did not hold my attention at all.), En_6 (I was not worried about what others might be thinking of me.), En_7 (The way time passed seemed to be different from normal.), and C_3 (It was difficult to learn in this VR environment.).

After these seven items were removed, the second-round EFA was conducted. Both parallel analysis and the scree plot indicated a five-factor solution in this time. The EFA result showed a clean factor structure and sufficient reliabilities for each factor. Table 11 is the pattern matrix of the second-round EFA. Table 12 is the Cronbach's Alpha of each factor. This five-factor structure also showed acceptable factor loadings for data in each session (see Table 13, Table 14, and Table 15). There were only two small deviations: item Pr_1 (Please rate your sense of being in the universe.) in session one and item En_5 (I had a feeling of total control over what I was doing.) in session three.

Table 11

The Pattern Matrix of the Second-Round EFA

Item	Factor				
	Presence	Sense of Embodiment	Motivation	Engagement	Cognitive Load
Pr_1	**0.63**	-0.04	0.11	0.16	0.06
Pr_2	**0.81**	0.12	-0.07	-0.00	-0.02
Pr_3	**0.67**	-0.12	-0.08	0.03	-0.06
Pr_4	**0.62**	0.10	0.06	-0.17	0.00
Em_2	0.06	**0.60**	-0.05	0.00	-0.00
Em_4	-0.13	**0.92**	0.03	0.10	0.00
Em_5	0.12	**0.56**	-0.01	-0.16	0.05
Mo_1	0.07	-0.03	**0.88**	-0.03	-0.05
Mo_3	-0.04	-0.01	**0.86**	-0.01	0.02
Mo_4	-0.04	0.00	**0.96**	-0.02	0.04
En_1	-0.00	0.01	0.04	**0.60**	-0.22
En_2	0.11	-0.07	-0.00	**0.84**	0.12
En_3	-0.13	0.05	-0.10	**0.69**	-0.03
En_4	-0.01	-0.07	-0.02	**0.71**	0.10
En_5	0.14	0.13	0.11	**0.42**	-0.06
Cl_1	-0.16	-0.01	-0.09	-0.01	**0.68**
Cl_2	0.12	0.04	0.07	0.04	**0.91**

Note. Extraction Method: Principal Axis Factoring.
Rotation Method: Promax with Kaiser Normalization.
Rotation converged in 5 iterations.

Table 12

The Cronbach's Alpha of Each Factor

Factor	Cronbach's Alpha
Presence	0.767
Sense of Embodiment	0.707
Motivation	0.913
Engagement	0.793

Table 13

Factor Loadings of Data in the First Session

Item	Factor				
	Presence	Sense of Embodiment	Motivation	Engagement	Cognitive Load
Pr_1	0.24	-0.03	0.37	0.24	0.08
Pr_2	**0.63**	0.22	0.04	-0.03	-0.03
Pr_3	**0.65**	-0.07	-0.20	0.20	-0.15
Pr_4	**0.64**	0.03	0.17	-0.23	0.03
Em_2	0.17	**0.70**	-0.26	0.17	0.05
Em_4	-0.19	**0.83**	0.19	-0.00	-0.05
Em_5	0.23	**0.62**	0.12	-0.15	0.07
Mo_1	0.08	-0.02	**0.85**	0.10	-0.01
Mo_3	-0.05	0.03	**0.91**	0.07	-0.02
Mo_4	-0.02	-0.01	**0.94**	-0.03	-0.04
En_1	0.17	-0.10	0.12	**0.70**	0.06
En_2	-0.03	0.01	0.03	**0.90**	0.05
En_3	-0.17	0.19	-0.16	**0.71**	-0.14
En_4	-0.03	-0.08	0.11	**0.61**	0.09
En_5	-0.04	0.05	0.15	**0.58**	-0.04
Cl_1	-0.01	0.06	-0.20	-0.05	**0.61**
Cl_2	-0.10	0.02	0.07	0.07	**0.71**

Note. Extraction Method: Principal Axis Factoring.
Rotation Method: Promax with Kaiser Normalization.
Rotation converged in 6 iterations.

Table 14

Factor Loadings of Data in the Second Session

Item	Factor				
	Presence	Sense of Embodiment	Motivation	Engagement	Cognitive Load
Pr_1	**0.92**	-0.11	-0.10	0.15	-0.10
Pr_2	**0.86**	0.11	-0.06	-0.07	-0.10
Pr_3	**0.63**	-0.03	-0.01	0.03	0.00
Pr_4	**0.59**	0.17	0.22	-0.20	0.20
Em_2	0.12	**0.51**	0.07	-0.08	-0.12
Em_4	-0.09	**0.91**	-0.04	0.13	-0.02
Em_5	0.06	**0.62**	-0.10	-0.09	0.12
Mo_1	0.13	0.01	**0.74**	-0.04	-0.12
Mo_3	-0.08	-0.09	**0.81**	-0.06	-0.03
Mo_4	-0.07	0.01	**0.95**	0.10	0.11
En_1	-0.01	0.16	0.10	**0.44**	-0.39
En_2	0.28	-0.20	0.07	**0.69**	0.04
En_3	-0.16	0.04	-0.16	**0.79**	-0.01
En_4	-0.04	-0.01	0.09	**0.68**	0.11
En_5	0.27	0.13	0.00	**0.49**	0.15
Cl_1	0.11	-0.02	-0.10	-0.08	**0.86**
Cl_2	-0.16	0.04	0.07	0.19	**0.89**

Note. Extraction Method: Principal Axis Factoring.
Rotation Method: Promax with Kaiser Normalization.
Rotation converged in 5 iterations.

Table 15

Factor Loadings of Data in the Third Session

Item	Factor				
	Presence	Sense of Embodiment	Motivation	Engagement	Cognitive Load
Pr_1	**0.66**	0.08	0.06	0.10	0.18
Pr_2	**0.79**	0.08	-0.02	0.04	-0.02
Pr_3	**0.78**	-0.26	0.06	-0.10	0.02
Pr_4	**0.56**	0.19	-0.09	-0.03	-0.17
Em_2	-0.19	**0.91**	0.04	-0.16	0.11
Em_4	0.19	**0.70**	-0.02	0.16	0.03
Em_5	0.20	**0.34**	-0.13	-0.14	-0.04
Mo_1	0.01	-0.08	**0.97**	-0.06	-0.07
Mo_3	0.03	-0.01	**0.96**	0.03	0.08
Mo_4	-0.04	0.04	**0.98**	-0.02	0.05
En_1	-0.15	0.00	-0.06	**0.62**	-0.29
En_2	0.05	0.04	-0.06	**0.79**	0.12
En_3	0.10	-0.11	0.06	**0.54**	-0.20
En_4	-0.04	-0.15	0.01	**0.94**	0.17
En_5	0.08	0.24	0.32	0.21	-0.36
Cl_1	0.12	-0.07	-0.01	-0.11	**0.78**
Cl_2	-0.07	0.16	0.05	0.14	**0.93**

Note. Extraction Method: Principal Axis Factoring.
Rotation Method: Promax with Kaiser Normalization.
Rotation converged in 5 iterations.

Descriptive statistics. Before conducting confirmatory factor analysis (CFA), data were screened and tested whether they met the criteria for conducting a multivariate analysis. This included identifying outliers and testing for normal distribution. A total of 17 cases were found to have observed outliers on one or several variable(s). Four of them were further deleted from the dataset for having more than two outliers on their variables. The remaining 13 were retained in the further analysis. This is because one or two observed variables in a case are identified as outliers will not seriously impact the overall analysis (Hair et al., 2009). Table 16 shows the descriptive statistics of remaining data. No observed variable exceeded the boundaries of univariate skewness (±2) and kurtosis

(±7). However, the Mardia's multivariate kurtosis statistic was 27.77 with $p < 0.01$. This meant that data were not multivariate normally distributed. Therefore, the robust maximum likelihood method (MLR) were chosen as the estimation method for further analysis.

Table 16

Descriptive Statistics of Remaining Data

Item	Range	Mean	SD	Univariate Skewness	Univariate Kurtosis
Pr_1	95	84.46	19.47	-1.97	4.28
Pr_2	100	70.42	26.62	-1.07	0.40
Pr_3	100	68.69	26.52	-1.02	0.45
Pr_4	100	61.80	29.39	-0.56	-0.69
Em_2	100	54.43	29.92	-0.16	-0.91
Em_4	100	64.76	30.53	-0.70	-0.57
Em_5	100	47.13	34.10	0.05	-1.31
Mo_1	69	89.98	13.81	-1.64	2.41
Mo_3	100	89.50	14.77	-1.08	6.85
Mo_4	70	91.30	12.39	-1.86	3.90
En_1	90	82.53	18.23	-1.25	1.42
En_2	75	83.20	16.94	-1.07	0.49
En_3	100	74.35	19.53	-0.67	0.23
En_4	84	81.57	18.25	-1.28	1.63
En_5	100	81.33	19.03	-1.24	1.66
Quiz	25	10.59	5.00	0.44	-0.32
Pre_AST*	16	7.10	4.13	0.18	-0.97

Note. Pre_AST represents the score in the pre-astronomy knowledge test

Assessment of measurement model. The remaining 187 cases were used in CFA to assess the quality of the measurement model. The assessment included (1) model fit indices, (2) the convergent validity, and (3) discriminant validity. Particularly, the validity assessments were only used for factors with at least three items (i.e., presence, sense of embodiment, motivation, and engagement).

A standard CFA model with more than one factor can be identified when it has more than one indicator per factor (Kline, 2011). In this study, learning achievement is a factor with a single item *quiz score*. For a single-item factor, a fixed 1.00 loading and a fixed measurement error variance should be specified, so that this only indicator can statistically identify its factor (Hayduk & Littvay, 2012). The value of measurement error variance is typically set as the indicator's variance multiplies its un-reliability (one minus the reliability). In this study, its factor loading was specified at one and the measurement error variance was specified at zero. The factor *cognitive load* only had two items after EFA. In this situation, an equal factor loading (1.00) was imposed on its two indicators (Kline, 2011).

The assessment results of the measurement model fitness using the MLR estimate method are reported below:

- RMSEA = 0.046.
- CFI = 0.948.
- TLI = 0.935.
- SRMR = 0.067.

All the values of these indices were in good model fit. Therefore, the goodness-of-fit statistics supported the measurement model.

Furthermore, the measurement results of convergent validity and discriminant validity also supported to use this model for further analysis. Table 17 shows evidence of convergent validity for the measurement model. The standardized factor loadings ranged from 0.55 to 0.92. They were higher than the threshold value of standardized factor loading (0.50). The AVE values ranged from 0.45 to 0.79. Particularly, the AVE values of presence (0.47), sense of embodiment (0.48), and engagement (0.45) were slightly below

the recommended rule of thumb (0.50). The CR values ranged from 0.73 to 0.92. All of them were higher than the low bound (0.70).

Table 17

Evidence of Convergent Validity for the Measurement Model

Factor	Item	R^2	Standardized Factor Loading	AVE	CR
Presence	Pr_1	0.53	0.73	0.47	0.78
	Pr_2	0.65	0.81		
	Pr_3	0.34	0.58		
	Pr_4	0.38	0.62		
Sense of Embodiment	Em_2	0.38	0.62	0.48	0.73
	Em_4	0.76	0.87		
	Em_5	0.30	0.55		
Motivation	Mo_1	0.80	0.90	0.79	0.92
	Mo_3	0.71	0.84		
	Mo_4	0.85	0.92		
Engagement	En_1	0.52	0.72	0.45	0.80
	En_2	0.62	0.78		
	En_3	0.38	0.61		
	En_4	0.42	0.65		
	En_5	0.32	0.57		

Note. R^2 = the square of a standard factor loading score.

Discriminant validity was established through a comparison between the AVE estimates for each factor with the squared inter-factor correlation associated with that factor. Results showed that all AVE estimates in Table 17 are greater than the squared correlation data above the diagonal of Table 18. This evidence supported the discriminant validity of the measurement model.

Table 18

Standardized Factor Correlation Matrix

Factor	1	2	3	4	5
1. Presence	1	0.12	0.20	0.09	< 0.01
2. Sense of Embodiment	0.34**	1	0.13	0.05	0.01
3. Motivation	0.45***	0.36***	1	0.21	0.04
4. Engagement	0.30**	0.23*	0.46***	1	0.16
5. Cognitive Load	-0.03	-0.090	-0.20**	-0.40***	1

*Note. Significant level: *=0.05, **=0.01, ***=0.001.* Value below the diagonal are correlation estimates among factors, diagonal elements are factor variances, and values above the diagonal are squared correlations.

In sum, based on the model fit indices, evidence of convergent validity, and evidence of discriminant validity, the five-factor model was appropriate to explain how the VR condition influences an effective learning experience. Figure 8 shows the final measurement model.

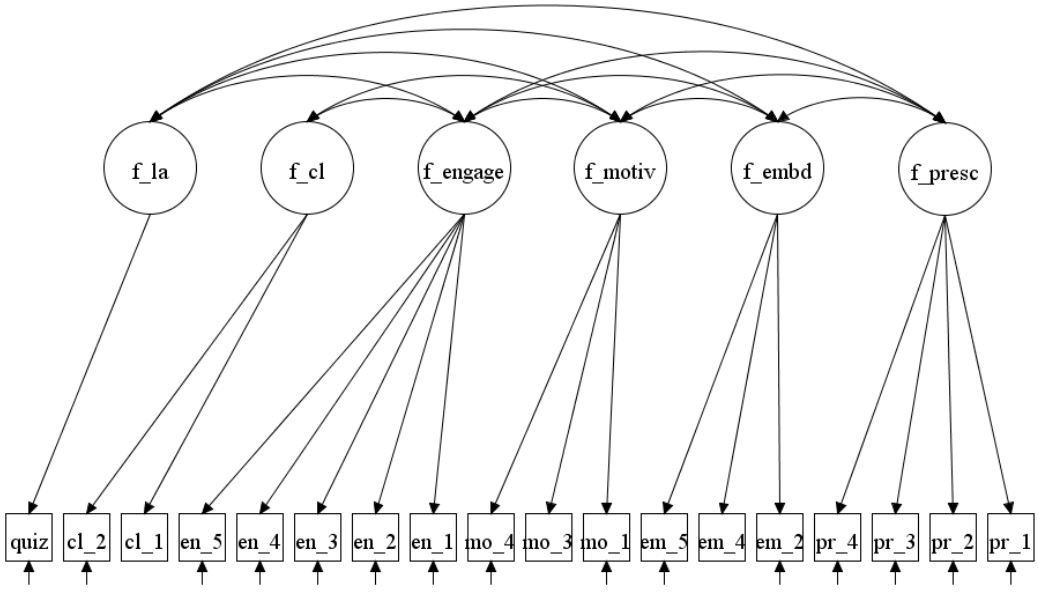


Figure 8. The measurement model of an effective VR learning experience

Assessment of structure model. The last step in phase II (SEM) was to assess the structure model which represents the structural relationship among factors and external variables. The results of model-fitness test showed that the hypothesized model adequately fit the data (RMSEA = 0.052, CFI = 0.922, TLI = 0.906, SRMR = 0.072). Due to the differences in range among scales, structural relationships were assessed by standardized path coefficient. If a standardized path coefficient was statistically significant, its source factor or external variable had a direct effect on the target factor. Table 19 shows a total of 15 direct paths evaluated in this model. These paths represented the hypotheses proposed in the literature review chapter.

Table 19

Parameter Estimates for the Hypothesized Structural Model

Structural Relationships	Std. Coefficient	Std. Error	p	Support Hypothesis
H1: VR condition -> Presence	-0.07	0.12	0.574	No
H2: VR condition -> Sense of embodiment	0.63	0.05	<0.001	Yes
H3: Sense of embodiment -> Presence	0.42	0.15	0.005	Yes
H4: Presence -> Motivation	0.37	0.11	0.001	Yes
H5: Sense of embodiment -> Motivation	0.20	0.10	0.045	Yes
H6: Prior knowledge -> Motivation	0.04	0.07	0.595	No
H7: VR condition -> Cognitive load	-0.02	0.07	0.749	No
H8: Prior knowledge -> Cognitive load	-0.16	0.06	0.010	Yes
H9: Motivation -> Engagement	0.31	0.10	0.002	Yes
H10: Presence -> Engagement	0.15	0.09	0.101	No
H11: Sense of embodiment -> Engagement	<0.01	0.09	0.971	No
H12: Cognitive load -> Engagement	-0.37	0.09	<0.001	Yes
H13: Prior knowledge -> Learning achievement	0.82	0.10	<0.001	Yes
H14: Motivation -> Learning achievement	<0.01	0.23	0.984	No
H15: Engagement -> Learning achievement	0.53	0.17	0.002	Yes

Direct effects among factors. As shown in Table 19, the level of a VR condition significantly influenced the sense of embodiment (β = 0.63). The sense of embodiment had a significant effect on presence (β = 0.42). Both sense of embodiment (β = 0.20) and presence (β = 0.37) contributed to motivation. Both motivation (β = 0.31) and cognitive load (β = -0.37) had a cause-effect relationship with engagement. Not only engagement (β = 0.54) but also prior knowledge (β = 0.81) significantly influenced learning achievement. Figure 9 illustrates the structural parameter estimates for the hypothesized structural model.

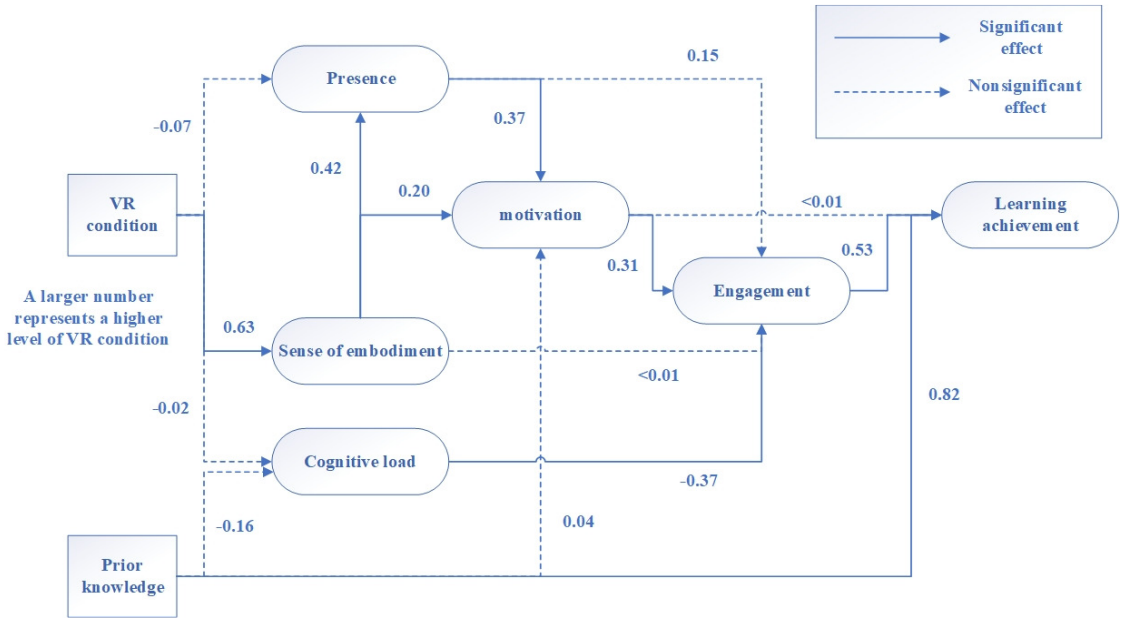


Figure 9. The structural model of an effective VR learning experience

Indirect effects. In addition to identifying the direct relationships among factors, several indirect effects were investigated as to know how some factors (e.g., motivation and engagement) delivered the effect of one factor onto another (see Table 20). Results showed that there were significant indirect effects from motivation to learning achievement through engagement (β = 0.17, p = 0.038) and from presence to engagement

through motivation (β = 0.12, p = 0.019). The sum of indirect effects from sense of embodiment to engagement was also significant (β = 0.13, p = 0.013).

Table 20

Parameter Estimates of Indirect Effects for the Structural Model

Structural Relationships	Mediator	Std. Coefficient	p
Sense of embodiment -> Engagement	Sum	0.13	0.013
	Presence	0.06	0.167
	Motivation	0.06	0.093
Presence -> Engagement	Motivation	0.12	0.019
Motivation -> Learning achievement	Engagement	0.17	0.038

Model modification. After extra paths from external variable *session* and its interaction with *VR condition* were added in this model, the model fitted the data better (RMSEA = 0.048, CFI = 0.924, TLI = 0.909, SRMR = 0.069). However, no other new significant finding was found. Table 21 shows the parameter estimates for the modified structural model.

Table 21

Parameter Estimates for the Modified Structural Model

Structural Relationships	Std. Coefficient	Std. Error	p	Support Hypothesis
H1: VR condition -> Presence	-0.05	0.13	0.709	No
H2: VR condition -> Sense of embodiment	0.63	0.06	<0.001	Yes
H3: Sense of embodiment -> Presence	0.41	0.15	0.006	Yes
H4: Presence -> Motivation	0.37	0.11	0.001	Yes
H5: Sense of embodiment -> Motivation	0.20	0.10	0.045	Yes
H6: Prior knowledge -> Motivation	0.04	0.07	0.588	No
H7: VR condition -> Cognitive load	-0.05	0.07	0.471	No
H8: Prior knowledge -> Cognitive load	-0.17	0.06	0.005	Yes
H9: Motivation -> Engagement	0.31	0.10	0.002	Yes
H10: Presence -> Engagement	0.16	0.10	0.103	No
H11: Sense of embodiment -> Engagement	0.01	0.09	0.959	No
H12: Cognitive load -> Engagement	-0.38	0.09	<0.001	Yes
H13: Prior knowledge -> Learning achievement	0.82	0.10	<0.001	Yes
H14: Motivation -> Learning achievement	0.01	0.23	0.983	No
H15: Engagement -> Learning achievement	0.54	0.17	0.002	Yes
N1: Session->Sense of embodiment	-0.03	0.08	0.748	No
N2: VR condition * Session-> Sense of embodiment	0.04	0.08	0.596	No
N3: Session->Presence	-0.09	0.08	0.251	No
N4: VR condition * Session->Presence	0.07	0.08	0.364	No
N5: Session->Cognitive load	0.13	0.07	0.071	No
N6: VR condition * Session->Cognitive load	-0.08	0.07	0.277	No

Data Comparison

This section includes the results of establishing a predictor model, data screening, comparison by categories, and completer analysis. The established predictor model was used to support the generation of 20 datasets with complete cases for the further data analysis. The data screening was to ensure the data were comparable among the three groups. Then data comparison was conducted by category based on the 20 new datasets

under the framework of the multiple imputation method. Finally, the completer analysis was carried out to validate the analysis results after multiple imputation in the previous step.

Establishment of predictor model. The results are reported based on the steps of establishing a predictor model for multiple impution of longitudinal missing data, which was introduced in the methods chapter. The first step was to compare the possible differences between the non-completer and completer groups. χ^2 test results indicated that there was no significant difference between the non-completer and completer groups in gender, $\chi^2(1) = 0.14$, $p = 0.905$, and major, $\chi^2(2) = 3.53$, $p = 0.171$. Since the pre-condition of χ^2 test could not be satisfied (the count of cells that had less than five cases was larger than the minimum expected count), χ^2 test was not used to compare the non-completer and completer groups in age and race. However, significant differences were found on dropout rate among categories in these two dimensions based on observed values. The dropout rate of the age 19 group was 70.59%, while the dropout rate of the participants who were older than 21 was 25.00% (see Table 22). The dropout rate of American Indian, Alaska Native, and other races was 57.14%, while all Black or African American participants completed the three sessions (see Table 23). Furthermore, there was a marginally significant difference found from the mean of variable Cl 2.1 (The learning content was difficult.) between the non-completer and completer groups in the independent samples T test, $t(66) = 1.76$, $p = 0.083$. Therefore, age, race, and variable Cl 2.1 were marked as the predictors for all target variables in the final model.

Table 22

Comparison of Dropout Rate Among Different Age Groups

Category	Age	Total Number	Dropout Number	Dropout Rate (%)	Standard Deviation
1	18	41	10	24.39	0.44
2	19	17	12	70.59	0.47
3	20-21	7	2	28.57	0.48
4	>21	12	3	25.00	0.44

Table 23

Comparison of Dropout Rate Among Different Races

Category	Race	Total Number	Dropout Number	Dropout Rate (%)	Standard Deviation
1	American Indian, Alaska Native, and other races	7	4	57.14	0.54
2	Asian or Asian American	12	4	33.33	0.49
3	Black or African American	6	0	0.00	0.00
4	Hispanic or Latino	16	4	25.00	0.45
5	White or Caucasian	36	15	41.67	0.50

Furthermore, independent samples T tests and / or one-way ANOVA tests were conducted to determine whether there were significant differences for target variables under different dimensional categories. Results showed that male participants had significantly higher scores than female participants in item Pr_3 (Do you think of the universe more as images that you saw? Or more as somewhere that you visited?), $t(94) = 2.25$, $p = 0.026$, and the post-mental rotation test, $t(79) = 3.35$, $p = 0.001$. The Oculus Rift system had significantly higher scores than the Oculus Go system: item Pr_1 (Please rate your sense of being in the universe.), $t(185) = 2.51$, $p = 0.013$; item Pr_2 (To what extent were there times during the experience when the universe was the 'reality' for you?),

$t(188) = 3.22$, $p = 0.002$; item Pr_4 (During the experience, did you often think to yourself that you were just in a room? Or did the universe overwhelm you?), $t(188) = 1.98$, $p = 0.049$; item Em_2 (It felt as if the virtual body or body part I saw was someone else's.), $t(188) = 6.66$, $p < 0.001$; item Em_4 (It felt like I could control the virtual body or body part as if it was my own.), $t(178) = 8.25$, $p < 0.001$, and Em_5 (I felt as if the movements of the virtual body were influencing my own movements.), $t(172) = 5.32$, $p < 0.001$. Therefore, gender was the predictor for all the three variable columns of item Pr_3 and the score of the post mental rotation test, while the type of VR system was the predictor for the variables Pr_1, Pr_2, Pr_4, Em_2, Em_4, and Em_5 in their corresponding sessions.

At the next step, quantitative variables that had the top 10 magnitude correlation value with the target variable were identified through Pearson correlation analysis and added to the predictor model. Table 24 shows the number of predictors for target variables in the final predictor model.

Table 24

The Number of Predictors for Target Variables

Target Variables	Number of Predictors	Target Variables	Number of Predictors	Target Variables	Number of Predictors
Em_2.1	18	Pr_4.2	18	En_1.3	17
Em_2.2	18	Pr_4.3	18	En_2.1	17
Em_2.3	18	Mo_1.1	17	En_2.2	17
Em_4.1	18	Mo_1.2	17	En_2.3	17
Em_4.2	18	Mo_1.3	17	En_3.1	17
Em_4.3	18	Mo_3.1	17	En_3.2	17
Em_5.1	18	Mo_3.2	17	En_3.3	17
Em_5.2	18	Mo_3.3	17	En_4.1	17
Em_5.3	18	Mo_4.1	17	En_4.2	17
Pr_1.1	18	Mo_4.2	17	En_4.3	17
Pr_1.2	18	Mo_4.3	17	En_5.1	17
Pr_1.3	18	Cl_1.1	17	En_5.2	17
Pr_2.1	18	Cl_1.2	17	En_5.3	17
Pr_2.2	18	Cl_1.3	17	Quiz_1	0
Pr_2.3	18	Cl_2.1	16	Quiz_2	17
Pr_3.1	18	Cl_2.2	16	Quiz_3	17
Pr_3.2	18	Cl_2.3	16	Post_MRT	16
Pr_3.3	18	En_1.1	17	Post_AST	15
Pr_4.1	18	En_1.2	17		

Note. Post_MRT represents score in the post-mental rotation test. Post_AST represents score in the post-astronomy knowledge test.

Data screening. The purpose of data screening was to ensure that the data were effective for the comparison between the pre-test and the post-test and the data were also comparable among groups. First, to avoid a floor effect and a ceiling effect in the comparison between the pre-test and and the post-test, data from four participants were removed. Specifically, one participant obtained zero score in the pre-astronomy knowledge test, two participants obtained full marks in the pre-astronomy knowledge test, and one participant obtained full marks in the pre-mental rotation test. After that, χ^2 tests and one-way ANOVA omnibus tests were conducted to check whether data among

the three group were comparable in the remaining 73 cases. χ^2 tests results showed that there was no significant difference among groups in gender, χ^2 (2) = 1.81, p = 0.404, and major, $\chi^2(4) = 5.30$, p = 0.258. One-way ANOVA omnibus tests did no find any significant difference among groups in the pre-mental rotation test, F(2, 70) = 2.27, p = 0.111, and the pre-astronomy knowledge test, F(2, 70) = 1.89, p = 0.159. However, the post-hoc Fishers Least Significant Difference (LSD) tests showed that the scores were significant difference between the Go-Rift group and the Rift group in the pre-mental rotation test (p = 0.049) and marginally significant difference between the Go group and the Rift group in the pre-astronomy knowledge test (p = 0.057).

To explore this issue and find a solution, a follow-up frequency analysis of the score distribution within each group was conducted. Results showed that the proportion of participants in the Rift group who obtained a score of two or three in the pre-tests was much higher than the other two groups (see Table 25).

Table 25.

Comparison of Participants Who Obtained a Score of Two or Three in the Pre-Tests Among Groups

Test Type	Group	Total Number	Number of Two Points	Number of Three Points
Pre-Mental Rotation Test	Go	21	2 (9.52%)	2 (9.52%)
	Go-Rift	22	2 (9.10%)	3 (13.60%)
	Rift	30	8 (26.70%)	7 (23.30%)
Pre-Astronomy Knowledge Test	Go	21	2 (9.52%)	1 (4.76%)
	Go-Rift	22	2 (9.10%)	1 (4.50%)
	Rift	30	7 (23.30%)	5 (16.70%)

To ensure that the score distribution was similar among groups, a total of 10 cases in the Rift group were removed from the comparison. These cases matched one of the

following two criteria: (1) the scores in the two tests were either two or three points; (2) the participant only did the first session experiment.

Then, χ^2 tests and one-way ANOVA omnibus tests were conducted to check the remaining 63 cases. χ^2 tests results showed that there was no significant difference among the three groups in gender, $\chi^2(2) = 0.93$, $p = 0.630$. Since the pre-condition of χ^2 test could not be satisfied (the count of cells that had less than five cases was larger than the minimum expected count), χ^2 test was not used to compare the possible differences in the distribution of majors among the three groups. Despite that, the descriptive statistics showed that the distribution of majors among groups was very similar (see Table 26). Moreover, one-way ANOVA omnibus tests showed that there was no significant difference among groups in both the pre-mental rotation test, $F(2, 60) = 0.40$, $p = 0.673$, and the pre-astronomy knowledge test, $F(2, 60) = 0.28$, $p = 0.758$. The results of the post-hoc Fishers Least Significant Difference (LSD) tests also did not show any significant difference between groups in the pre-tests (see Table 27).

Table 26

The Distribution of Majors Among Groups

Group	Engineering	Science	Business and Humanity	Total Number
Go	12	4	5	21
Go-Rift	12	7	3	22
Rift	11	3	6	20

Table 27

Comparison of Scores in the Pre-Tests Among Groups

Test Type	Group	Cases	Mean	Standard Deviation
Pre-Mental Rotation Test	Go	21	4.57	2.33
	Go-Rift	22	5.18	2.61
	Rift	20	4.65	2.35
Pre-Astronomy Knowledge Test	Go	21	7.67	3.90
	Go-Rift	22	6.82	3.88
	Rift	20	7.05	3.72

Comparison by category. A total of 20 datasets were created after the initial incomplete dataset was imputed with assigned parameters on the R platform. Each dataset included 77 complete cases after imputation. Then these data were analyzed on SPSS 25. Comparison results are shown by category in the following subsections. Most tables in this section include a column named fraction of missing information (FMI). FMI is a measurement that assesses the missing data's influence on the sampling variance of a parameter estimate (Enders, 2010b). If the assessed variables are uncorrelated, FMI equals the rate of missing data. In this study, due to the correlation among variables, its value was typically less than the missing data rate.

Presence. Table 28 and Figure 10 show the mean of presence for each session in each group. There was no significant difference of presence between the low-level VR condition (the Go and the Go-Rift group) and the high-level VR condition (the Rift group) in the first session, $t(61) = 1.05$, $p = 0.299$.

While the mean of presence for the Go-Rift group increased a lot from the second to the third session, there was no significant difference between the second session and the third session in the Go-Rift group, $t(860) = 1.39$, $p = 0.165$, nor between the Go-Rift

group and the Go group in the third session, t(287) = 1.60, p = 0.111. This is different from the prediction that individuals' presence will increase in a high-level VR condition.

The mean of presence did not have a significant decrease from the first session to the third session for both the Go group and the Rift group. Relatively, the decrease for the Rift group (-4.63) was larger than the Go group (-3.57). By percentage, the decrease for the Rift group (4.63 / 75.16 * 100% = 6.16%) was still larger than the Go group (3.57 / 69.58 * 100% = 5.14%). This is different from the prediction that presence decreases more slowly in a high-level VR condition after the novelty effect happens.

Table 28

Descriptive Statistics of Presence

Group	Session	Mean	Std. Error of Mean	FMI
Go	1	69.58	3.63	0.03
	2	68.46	4.71	0.10
	3	66.00	4.50	0.24
Go-Rift	1	69.48	5.14	0.00
	2	66.81	5.93	0.08
	3	74.98	3.39	0.17
Rift	1	75.16	4.41	0.00
	2	69.25	5.00	0.00
	3	70.54	5.60	0.16

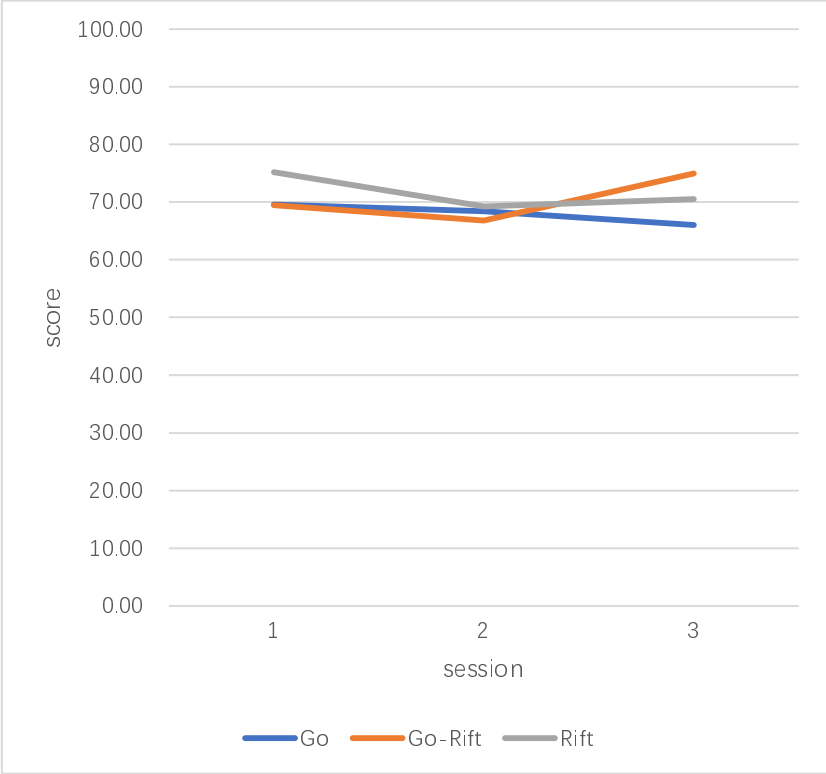


Figure 10. Presence for each session in each group

Sense of embodiment. Table 29 and Figure 11 show the mean of sense of embodiment for each session in each group. There was a significant difference of sense of embodiment between the low-level VR condition (the Go and the Go-Rift group) and the high-level VR condition (the Rift group) in the first session, $t(61) = 5.12$, $p < 0.001$, Cohen's $d = 1.540$.

The mean of sense of embodiment for the Go-Rift group had a significant increase from the second session to the third session, $t(1287) = 3.86$, $p < 0.001$, Cohen's $d = 1.118$. Its value was also significantly higher than the Go group in the third session, $t(333) = 2.91$, $p = 0.004$, Cohen's $d = 0.993$. This is consistent with the prediction that individuals' sense of embodiment will increase in a high-level VR condition.

The mean of sense of embodiment did not have a significant decrease from the first session to the third session for both the Go group and the Rift group. In addition, the decrease for the Rift group (-5.26) was smaller than the Go group (-6.89). By percentage, the decrease for the Rift group (5.26 / 74.37 * 100% = 7.08%) was much smaller than the Go group (6.89 / 50.51 * 100% = 13.64%). This is consistent with the prediction that sense of embodiment decreases more slowly in a high-level VR condition after the novelty effect happens.

Table 29

Descriptive Statistics of Sense of Embodiment

Group	Session	Mean	Std. Error of Mean	FMI
Go	1	50.51	4.87	0.07
	2	47.34	4.69	0.22
	3	43.62	5.10	0.24
Go-Rift	1	38.70	5.09	0.00
	2	38.69	5.37	0.06
	3	63.66	4.91	0.24
Rift	1	74.37	3.89	0.00
	2	66.35	4.21	0.00
	3	69.10	4.77	0.29

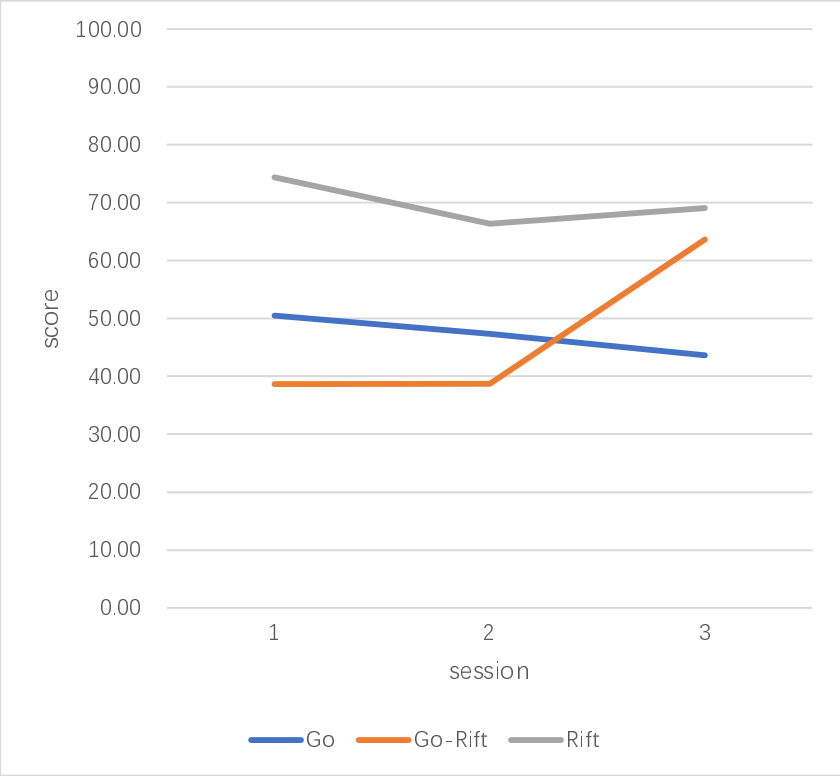


Figure 11. Sense of embodiment for each session in each group

Motivation. Table 30 and Figure 12 show the mean of motivation for each session in each group. There was no significant difference of motivation between the low-level VR condition (the Go and the Go-Rift group) and the high-level VR condition (the Rift group) in the first session, $t(61) = 1.06$, $p = 0.295$.

The mean of motivation for the Go-Rift group had a significant increase from the second session to the third session, $t(996) = 2.01$, $p = 0.045$, Cohen's $d = 0.621$. Its value was significantly higher than the Go group in the third session, $t(815) = 2.34$, $p = 0.019$, Cohen's $d = 0.778$. This is consistent with the prediction that individuals' motivation will increase in a high-level VR condition.

The mean of motivation did not have a significant decrease from the first session to the third session for both the Go group and the Rift group. In addition, the decrease for

the Rift group (-2.05) was smaller than the Go group (-3.70). By percentage, the decrease for the Rift group (2.05 / 93.93 * 100% = 2.18%) was still smaller than the Go group (3.70 / 92.61 * 100% = 4.00%). This is consistent with the prediction that motivation decreases more slowly in a high-level VR condition after the novelty effect happens.

Table 30

Descriptive Statistics of Motivation

Group	Session	Mean	Std. Error of Mean	FMI
Go	1	92.61	2.35	0.06
	2	90.57	2.29	0.08
	3	88.91	2.33	0.19
Go-Rift	1	88.77	2.99	0.00
	2	88.36	3.01	0.04
	3	94.96	1.49	0.30
Rift	1	93.93	2.23	0.00
	2	90.88	2.73	0.00
	3	91.88	2.96	0.09

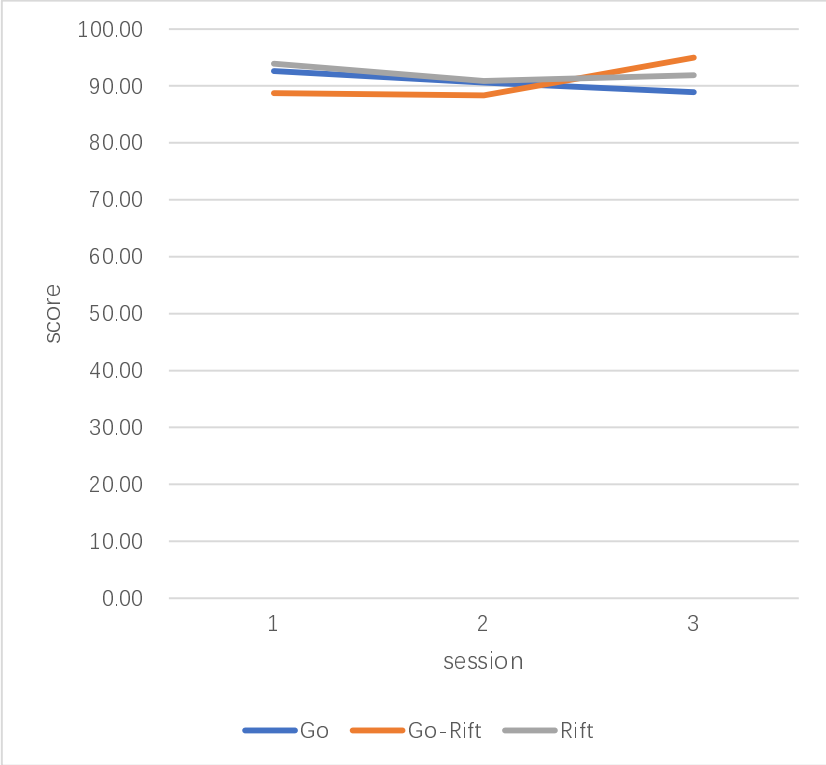


Figure 12. Motivation for each session in each group

Cognitive load. Table 31 and Figure 13 show the mean of cognitive load for each session in each group. Overall, the mean of cognitive load was stable across sessions and was similar among groups in the same session. There was a slight increase from the first session to the second session for all groups.

Table 31

Descriptive Statistics of Cognitive Load

Group	Session	Mean	Std. Error of Mean	FMI
Go	1	29.13	4.93	0.03
	2	34.43	4.73	0.22
	3	35.19	6.63	0.24
Go-Rift	1	33.64	3.97	0.00
	2	41.60	5.89	0.10
	3	41.22	5.24	0.12
Rift	1	32.88	3.56	0.00
	2	38.30	5.80	0.00
	3	39.54	6.06	0.07

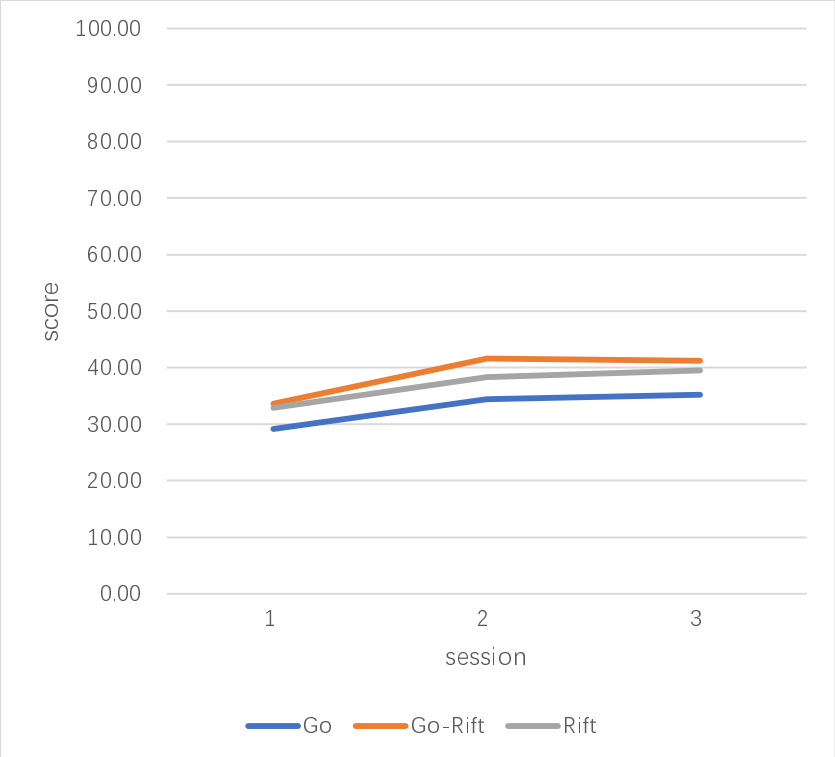


Figure 13. Cognitive load for each session in each group

Engagement. Table 32 and Figure 14 show the mean of engagement for each session in each group. There was no significant difference of engagement between the

low-level VR condition (the Go and the Go-Rift group) and the high-level VR condition (the Rift group) in the first session, $t(61) = -0.03$, $p = 0.980$.

All the three groups had the highest value in the second session. Then the Go group had a significant decrease from the second session to the third session, $t(1246) = -2.48$, $p = 0.013$, Cohen's $d = 0.628$, while the decrease for the Go-Rift group and the Rift group was less visible. There was no significant difference found between the Go group and the Go-Rift group in the third session, $t(1124) = 0.95$, $p = 0.342$, nor between the Go group and the Rift group in the third session, $t(2303) = 1.63$, $p = 0.104$. This is different from the prediction that individuals' engagement will increase in a high-level VR condition.

The mean of engagement had a significant decrease for the Go group from the first session to the third session, $t(1408) = -2.07$, $p = 0.039$, Cohen's $d = 0.539$. In addition, the mean of engagement for the Rift group increased (2.61), while the Go group decreased (-7.86). This is consistent with the prediction that engagement decreases more slowly in a high-level VR condition after the novelty effect happens.

Table 32

Descriptive Statistics of Engagement

Group	Session	Mean	Std. Error of Mean	FMI
Go	1	82.73	2.81	0.03
	2	84.36	3.04	0.06
	3	74.87	3.96	0.10
Go-Rift	1	77.47	2.90	0.00
	2	81.82	2.74	0.08
	3	79.57	3.11	0.18
Rift	1	80.09	3.61	0.00
	2	82.86	2.93	0.00
	3	82.70	2.84	0.15

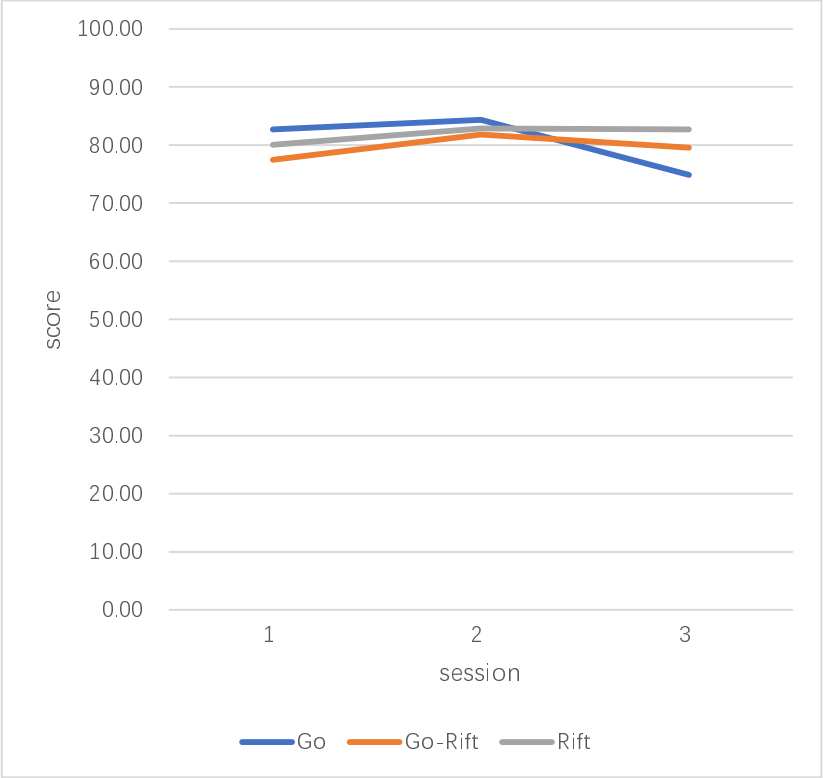


Figure 14. Engagement for each session in each group

Quiz score. Table 33 and Figure 15 show the mean of quiz score for each session in each group. There was no significant difference of quiz score between the low-level

VR condition (the Go and the Go-Rift group) and the high-level VR condition (the Rift group) in the first session, $t(61) = -1.92$, $p = 0.059$.

Moreover, the increase of quiz score from the second session to the third session for the Go-Rift group was not significant, $t(188) = 0.68$, $p = 0.500$. Its differences with the other two groups were also not significant in the third session. This is different from the prediction that individuals' learning achievement will increase in a high-level VR condition.

The mean of quiz score did not have a significant decrease from the first session to the third session for both the Go group and the Rift group. In addition, the decrease for the Rift group (-0.40) was smaller than the Go group (-0.85). By percentage, the decrease for the Rift group (0.40 / 9.90 * 100% = 4.04%) was still smaller than the Go group (0.85 / 13.05 * 100% = 6.51%). This is consistent with the prediction that learning achievement decreases more slowly in a high-level VR condition after the novelty effect happens. Despite that, the mean for the Rift group was lowest among groups across all the three sessions.

Table 33

Descriptive Statistics of Quiz Score

Group	Session	Mean	Std. Error of Mean	FMI
Go	1	13.05	1.49	0.00
	2	11.23	1.37	0.11
	3	12.20	1.35	0.22
Go-Rift	1	12.41	0.98	0.00
	2	10.37	1.13	0.05
	3	11.10	1.17	0.18
Rift	1	9.90	1.03	0.00
	2	10.20	0.99	0.00
	3	9.50	1.12	0.23

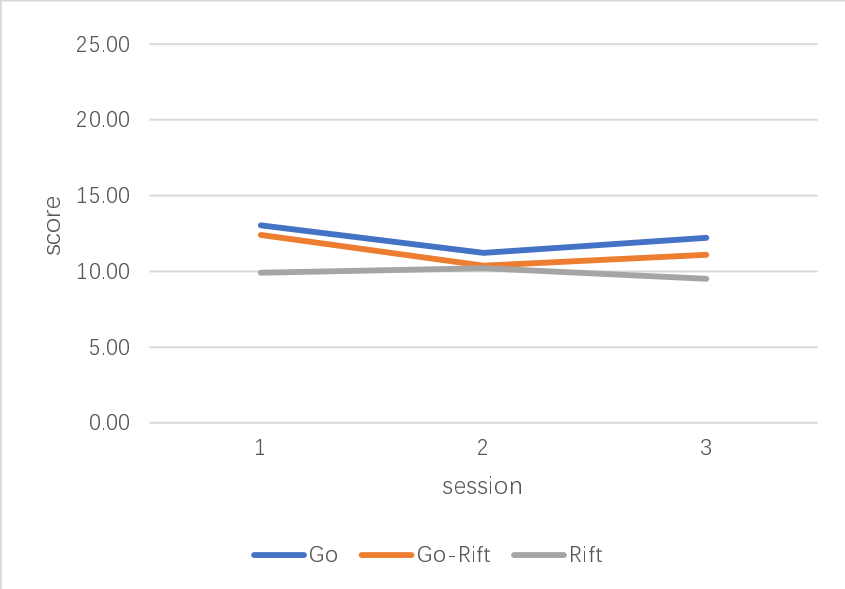


Figure 15. Quiz score for each session in each group

Pre- and post-tests. Compared with scores in the pre-tests, all the three groups had some increase in both the post-mental rotation test and the post-astronomy knowledge test (see Table 34 and Table 35). Both the Go group and the Go-Rift group had a significant increase in the post-mental rotation test: Go group, $t(164) = 2.80$, $p = 0.006$, Cohen's $d = 1.192$; Go-Rift group, $t(146) = 3.59$, $p < 0.001$, Cohen's $d = 1.411$, while no significant difference was found when comparing their pre- and post-test scores in the astronomy knowledge test, respectively.

The Rift group showed the best performance in both the post-mental rotation test and the post-astronomy knowledge test. It had the highest scores in these two tests. In addition, it had a significant increase in both the mental rotation test, $t(240) = 5.09$, $p < 0.001$, Cohen's $d = 1.228$, and the astronomy knowledge test, $t(240) = 1.94$, $p = 0.044$, Cohen's $d = 0.422$.

Table 34

Descriptive Statistics of Pre- and Post-Test Scores

Test Type		Group	Mean	Std. Error of Mean	FMI
Mental Rotation Test	Pre-Test	Go	4.57	0.51	0.00
		Go-Rift	5.18	0.56	0.00
		Rift	4.65	0.53	0.00
	Post-Test	Go	6.31	0.7	0.23
		Go-Rift	7.12	0.74	0.16
		Rift	7.58	0.66	0.18
Astronomy Knowledge Test	Pre-Test	Go	7.67	0.85	0.00
		Go-Rift	6.82	0.83	0.00
		Rift	7.05	0.83	0.00
	Post-Test	Go	8.51	0.93	0.18
		Go-Rift	7.62	1.01	0.23
		Rift	8.71	0.9	0.23

Table 35

Comparison of Pre- and Post-Test Scores in Each Group

Test Type	Group	Mean Difference	Std. Error Mean	t	df	p	FMI
Mental Rotation Test	Go	1.74	0.62	2.80	164	**0.006**	0.29
	Go-Rift	1.93	0.54	3.59	146	**<0.001**	0.31
	Rift	2.93	0.58	5.09	240	**<0.001**	0.24
Astronomy Knowledge Test	Go	0.84	1.03	0.82	624	0.415	0.14
	Go-Rift	0.80	0.74	1.08	78	0.283	0.43
	Rift	1.66	0.86	1.94	199	**0.044**	0.26

Note. Pre-test scores are the reference

Completer analysis. Among 77 participants, 16 in the Go group, 16 in the Go-Rift group, and 18 in the Rift group completed the three sessions. These completers' data were analyzed on SPSS 25. One-way ANOVA omni tests were conducted to check whether the mean scores were equal among groups in the pre-mental rotation test and the pre-astronomy knowledge test, respectively. Results showed that there was no significant difference among groups in both the pre-mental rotation test, $F(2, 47) = 1.99$, $p = 0.147$,

and the pre-astronomy knowledge test, $F(2, 47) = 2.21$, $p = 0.121$. However, the post-hoc Fishers Least Significant Difference (LSD) tests showed that there were marginally significant differences between the Go-Rift group and the Rift group in the pre-mental rotation test ($p = 0.078$), between the Go group and the Go-Rift group in the pre-astronomy knowledge test ($p = 0.084$), and between the Go group and the Rift in the pre-astronomy knowledge test ($p = 0.065$).

Presence. Table 36 and Figure 16 show the mean of presence for each session in each completer group. There was no significant difference of presence between the low-level VR condition (the Go and the Go-Rift group) and the high-level VR condition (the Rift group) in the first session, $t(48) = 1.16$, $p = 0.253$.

In addition, there was no significant difference among sessions for the Go-Rift group ($F = 2.76$, $p = 0.080$) nor among the three groups in the third session, $F(2,47) = 1.75$, $p = 0.184$. Both are consistent with the analysis results after multiple imputation and do not support the prediction that individuals' presence will increase in a high-level VR condition.

There was also no significant difference among sessions for both the Go group ($F = 1.22$, $p = 0.299$) and the Rift group ($F = 0.47$, $p = 0.580$). Different from the analysis result after multiple imputation, the Rift group had a smaller decrease (-2.47) than the Go group (-6.33) from the first session to the third session. By percentage, the decrease for the Rift group ($2.47 / 76.47 * 100\% = 3.23\%$) was still smaller than the Go group ($6.33 / 69.98 * 100\% = 9.05\%$). This is consistent with the prediction that presence decreases more slowly in a high-level VR condition after the novelty effect happens.

The trend analysis results showed that none of these three groups had a significant linear or quadratic contrast (see Table 37). In other words, they did not show any definite changing trend through sessions.

Table 36

Descriptive Statistics of Presence for Completers

Group	Session	Mean	Std. Error of Mean	Std. Deviation
Go	1	69.98	4.18	16.71
	2	66.27	5.33	21.33
	3	63.65	4.56	18.23
Go-Rift	1	70.25	5.17	20.67
	2	62.38	6.99	27.98
	3	75.53	3.64	14.54
Rift	1	76.47	4.46	18.91
	2	73.57	4.84	20.53
	3	74.00	5.71	24.23

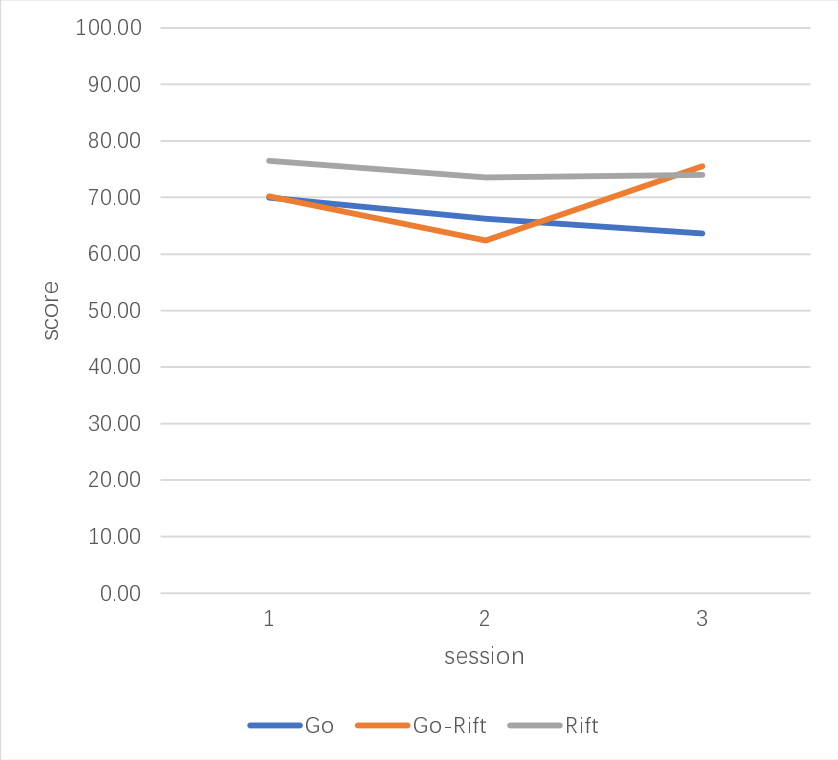


Figure 16. Presence for each session in each completer group

Table 37

Within-Subjects Contrasts of Presence for Completers

Group	Effect	F	p	η^2
Go	Linear	1.755	0.205	0.105
	Quadratic	0.038	0.847	0.003
Go-Rift	Linear	0.968	0.341	0.061
	Quadratic	4.243	0.057	0.221
Rift	Linear	0.373	0.550	0.021
	Quadratic	0.797	0.385	0.045

Sense of embodiment. Table 38 and Figure 17 show the mean of sense of embodiment for each session in each completer group. There was a significant difference of sense of embodiment between the low-level VR condition (the Go and the Go-Rift group) and the high-level VR condition (the Rift group) in the first session, $t(48) = 4.98$, $p < 0.001$, Cohen's $d = 1.468$.

Besides that, there was a significant difference among sessions for the Go-Rift group ($F = 11.40$, $p = 0.002$, $\eta^2 = 0.432$). The post-hoc test result showed that the Go-Rift group had a significant increase from the second session to the third session ($p = 0.008$, Cohen's $d = 0.859$). In addition, there was a significant difference among groups in the third session, $F(2, 47) = 14.01$, $p < 0.001$, $\eta^2 = 0.374$. Specifically, the Go-Rift group was significantly higher than the Go group in the third session ($p = 0.002$, Cohen's $d = 1.181$). Both are consistent with the analysis results after multiple imputation and the prediction that individuals' sense of embodiment will increase in a high-level VR condition.

There was no significant difference among sessions for both the Go group ($F = 1.28$, $p = 0.292$) and the Rift group ($F = 1.62$, $p = 0.213$). In addition, the Rift group had a smaller decrease (-2.13) than the Go group (-8.67). By percentage, the decrease for the

Rift group (2.13 / 76.67 * 100% = 2.78%) was much smaller than the Go group (8.67 / 48.44 * 100% = 17.90%). This is consistent with the analysis result after multiple imputation and the prediction that sense of embodiment decreases more slowly in a high-level VR condition after the novelty effect happens. Besides this, the Rift group had a significant higher sense of embodiment than the Go group across sessions (see Table 39).

The trend analysis results showed that the Go-Rift group had significant linear and quadratic contrasts (see Table 40). This means that the changes of the Go-Rift group had a definite curve, as well as an overall upward trend.

Table 38

Descriptive Statistics of Sense of Embodiment for Completers

Group	Session	Mean	Std. Error of Mean	Std. Deviation
Go	1	48.44	6.57	26.30
	2	45.69	5.49	21.96
	3	39.77	5.66	22.65
Go-Rift	1	36.42	5.76	23.04
	2	38.63	6.39	25.56
	3	65.75	5.33	21.31
Rift	1	76.67	4.65	19.72
	2	68.59	4.83	20.50
	3	74.54	3.48	14.76

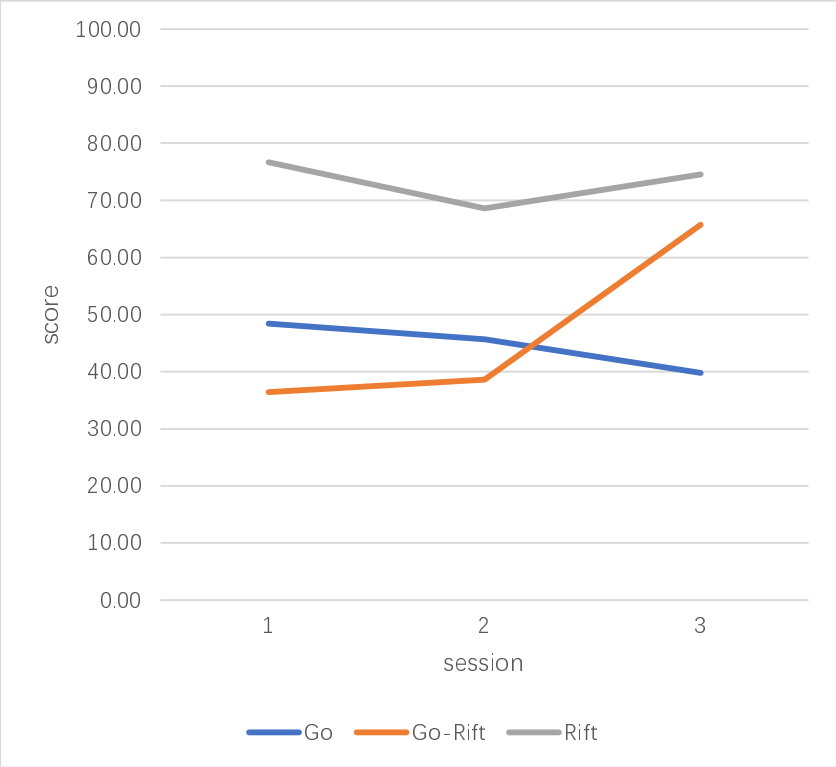


Figure 17. Sense of embodiment for each session in each completer group

Table 39

Comparison of Sense of Embodiment for Completers Between the Go Group and the Rift Group

Session	Mean Difference	t	df	p	Cohen's d
1	28.23	3.57	32	**0.001**	1.225
2	22.91	3.15	32	**0.004**	1.081
3	34.77	5.23	32	**<0.001**	1.842

Note. The Go completer group is the reference group

Table 40

Within-Subjects Contrasts of Sense of Embodiment for Completers

Group	Effect	F	p	η^2
Go	Linear	1.62	0.222	0.097
	Quadratic	0.23	0.641	0.015
Go-Rift	Linear	12.93	**0.003**	0.463
	Quadratic	7.63	**0.015**	0.337
Rift	Linear	0.25	0.625	0.014
	Quadratic	2.63	0.123	0.134

Motivation. Table 41 and Figure 18 show the mean of motivation for each session in each completer group. There was no significant difference of motivation between the low-level VR condition (the Go and the Go-Rift group) and the high-level VR condition (the Rift group) in the first session, $t(48) = 1.12$, $p = 0.267$.

Despite that, there was a significant difference among sessions for the Go-Rift group ($F = 5.55$, $p = 0.009$, $\eta^2 = 0.270$). The post-hoc test result showed that the Go-Rift group had a significant increase from the second session to the third session ($p = 0.035$, Cohen's $d = 0.719$). In addition, there was a significant difference among groups in the third session, $F(2, 47) = 3.32$, $p = 0.045$, $\eta^2 = 0.124$. Particularly, the post-hoc test result showed that the Go-Rift group was significantly higher than the Go group in the third session ($p = 0.041$, Cohen's $d = 1.182$). Both are consistent with the analysis results after multiple imputation and the prediction that individuals' sense of embodiment will increase in a high-level VR condition.

There was no significant difference among sessions for both the Go group ($F = 0.31$, $p = 0.739$) and the Rift group ($F = 0.08$, $p = 0.927$). In addition, the Rift group had a smaller decrease (-0.43) than the Go group (-2.42) from the first session to the third

session. By percentage, the decrease for the Rift group (0.43 / 92.89 * 100% = 0.46%) was also smaller than the Go group (2.42 / 89.46 * 100% = 2.71%). This is consistent with the analysis result after multiple imputation and the prediction that motivation decreases more slowly in a high-level VR condition after the novelty effect happens.

The trend analysis results showed that the Go-Rift group had significant linear and quadratic contrasts (see Table 42). This means that the changes of Go-Rift group had a definite curve, as well as an overall upward trend.

Table 41

Descriptive Statistics of Motivation for Completers

Group	Session	Mean	Std. Error of Mean	Std. Deviation
Go	1	89.46	3.12	12.48
	2	88.21	2.65	10.59
	3	87.02	2.58	10.33
Go-Rift	1	86.87	3.90	15.58
	2	85.46	3.94	15.77
	3	96.33	1.04	4.17
Rift	1	92.89	3.49	14.82
	2	91.80	2.77	11.74
	3	92.46	3.18	13.51

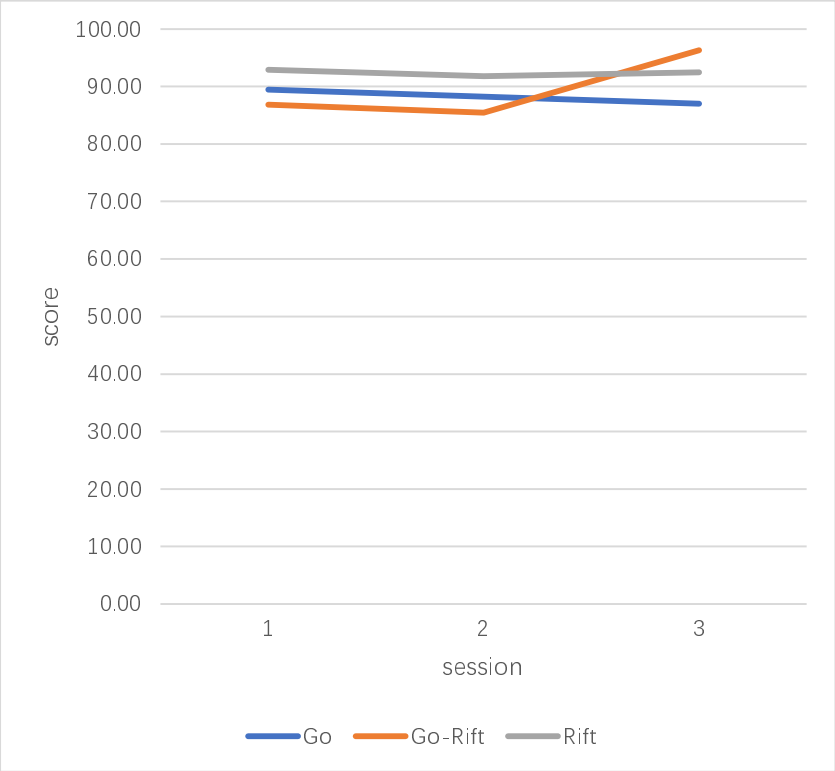


Figure 18. Motivation for each session in each completer group

Table 42

Within-Subjects Contrasts of Motivation for Completers

Group	Effect	F	p	η^2
Go	Linear	0.388	0.543	0.025
	Quadratic	0.000	0.986	0.000
Go-Rift	Linear	5.432	**0.034**	0.266
	Quadratic	5.773	**0.030**	0.278
Rift	Linear	0.023	0.881	0.001
	Quadratic	0.125	0.728	0.007

Cognitive load. Table 43 and Figure 19 show the mean of cognitive load for each session in each completer group. There was no significant difference of cognitive load between the low-level VR condition (the Go and the Go-Rift group) and the high-level

VR condition (the Rift group) in the first session, $t(48) = 0.71$, $p = 0.483$. Furthermore, the mean of cognitive load was stable across sessions and was similar among groups in the same session. There was a slight increase from the first session to the second session for all groups.

The trend analysis results showed that none of the three groups had a significant linear or quadratic contrast (see Table 44). In other words, they did not show any definite changing trend through sessions.

Table 43

Descriptive Statistics of Cognitive Load for Completers

Group	Session	Mean	Std. Error of Mean	Std. Deviation
Go	1	26.41	5.56	22.24
	2	37.06	5.84	23.37
	3	39.59	7.62	30.50
Go-Rift	1	29.69	4.77	19.08
	2	40.19	6.35	25.41
	3	39.63	5.91	23.62
Rift	1	32.22	4.55	19.30
	2	34.61	5.58	23.67
	3	35.69	5.84	24.79

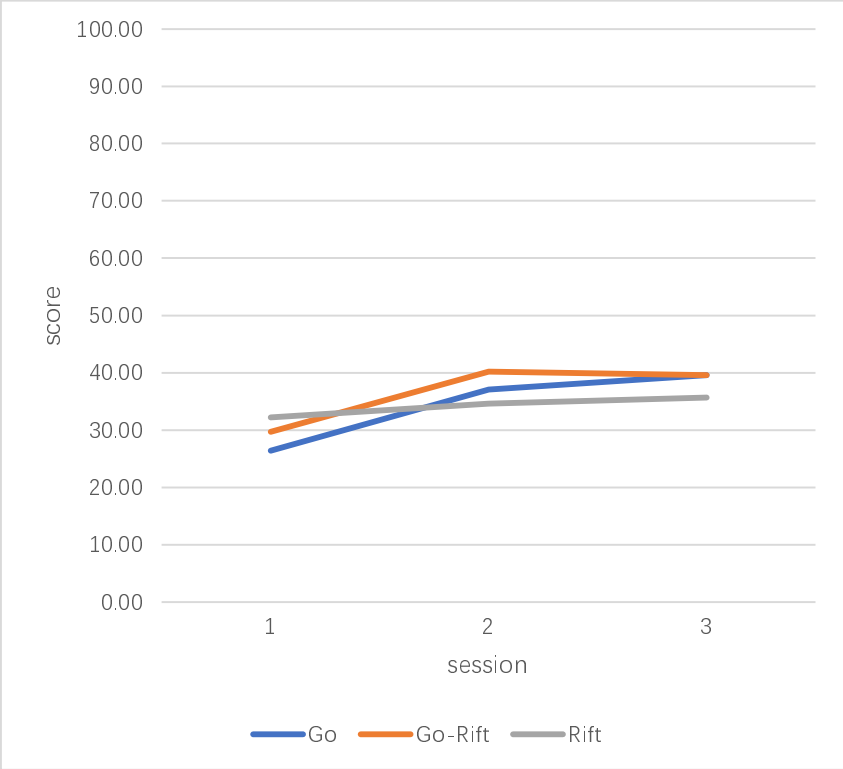


Figure 19. Cognitive load for each session in each completer group

Table 44

Within-Subjects Contrasts of Cognitive Load for Completers

Group	Effect	F	p	η^2
Go	Linear	2.717	0.120	0.153
	Quadratic	0.637	0.437	0.041
Go-Rift	Linear	4.119	0.061	0.215
	Quadratic	1.168	0.297	0.072
Rift	Linear	0.477	0.499	0.027
	Quadratic	0.030	0.865	0.002

Engagement. Table 45 and Figure 20 show the mean of engagement for each session in each completer group. There was no significant difference of engagement

between the low-level VR condition (the Go and the Go-Rift group) and the high-level VR condition (the Rift group) in the first session, $t(48) = 0.29$, $p = 0.775$.

The Go-Rift group had a small decrease but not increase from the second session to the third session. This is consistent with the analysis result after multiple imputation but different from the prediction that individuals' engagement will increase in a high-level VR condition.

Only the Go group had a significant difference among sessions ($F = 3.41$, $p = 0.046$, $\eta^2 = 0.185$). However, the post-hoc test results showed that there was no significant difference between the first session and the third session ($p = 0.134$) nor between the second session and the third session ($p = 0.239$). Furthermore, there was a significant difference among groups in the third session, $F(2, 47) = 3.65$, $p = 0.034$, $\eta^2 = 0.134$. The post-hoc test result showed that the Rift group was significantly higher than the Go group in the third session ($p = 0.034$, Cohen's $d = 0.887$). In addition, the mean of engagement for the Rift group increased (2.42) from the first session to the third session, while the Go group decreased (-9.94). This is consistent with the analysis result after multiple imputation and the prediction that engagement decreases more slowly in a high-level VR condition after the novelty effect happens.

The trend analysis results showed that the Go group had a significant linear contrast (see Table 46). This means that the changes of the Go group had a definite overall downward trend.

Table 45

Descriptive Statistics of Engagement for Completers

Group	Session	Mean	Std. Error of Mean	Std. Deviation
Go	1	81.09	3.42	13.67
	2	79.91	3.59	14.35
	3	71.15	4.46	17.82
Go-Rift	1	79.99	3.38	13.53
	2	82.59	3.58	14.30
	3	80.56	3.42	13.68
Rift	1	81.77	3.84	16.30
	2	84.77	2.66	11.27
	3	84.19	2.65	11.24

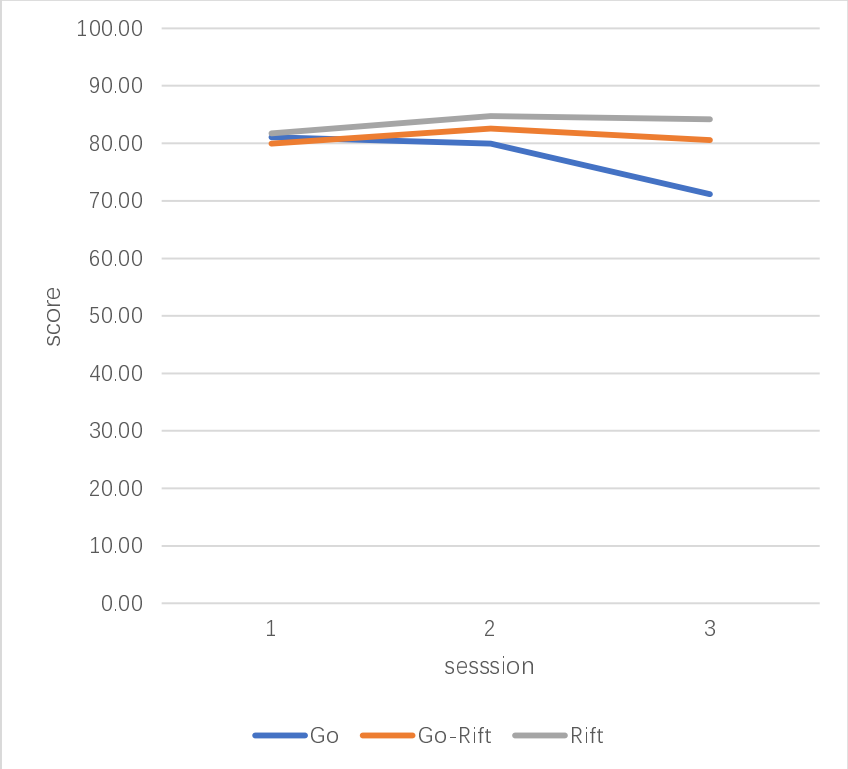


Figure 20. Engagement for each session in each completer group

Table 46

Within-Subjects Contrasts of Engagement for Completers

Group	Effect	F	p	η^2
Go	Linear	4.79	**0.045**	0.242
	Quadratic	1.37	0.260	0.084
Go-Rift	Linear	0.04	0.839	0.003
	Quadratic	1.10	0.310	0.069
Rift	Linear	0.04	0.839	0.003
	Quadratic	1.10	0.310	0.069

Quiz score. Table 47 and Figure 21 show the mean of quiz score for each session in each completer group. There was no significant difference of quiz score between the low-level VR condition (the Go and the Go-Rift group) and the high-level VR condition (the Rift group) in the first session, $t(48) = 1.51$, $p = 0.137$.

Moreover, there was no significant difference among sessions in the Go-Rift group ($F = 1.86$, $p = 0.174$) nor among the three groups in the third session, $F(2, 47) = 1.21$, $p = 0.307$. Both are consistent with the analysis results after multiple imputation and do not support the prediction that individuals' learning achievement will increase in a high-level VR condition.

There was no significant difference among sessions for both the Go group ($F = 0.53$, $p = 0.597$) and the Rift group ($F = 0.22$, $p = 0.808$). Moreover, the means for the Rift group were lowest among groups across all the three sessions. Despite this, the decrease for the Rift group (-0.61) was larger than the Go group (-0.37). By percentage, the decrease for the Rift group ($0.61 / 9.78 * 100\% = 6.24\%$) was still larger than the Go group ($0.85 / 12.00 * 100\% = 3.08\%$). Both are different from the analysis results after

multiple imputation and the prediction that learning achievement decreases more slowly in a high-level VR condition after the novelty effect happens.

The trend analysis results showed that none of the three groups had a significant linear or quadratic contrast (see Table 48). In other words, they did not show any definite changing trend through sessions.

Table 47

Descriptive Statistics of Quiz Score for Completers

Group	Session	Mean	Std. Error of Mean	Std. Deviation
Go	1	12.00	1.57	6.28
	2	10.81	1.54	6.16
	3	11.63	1.23	4.92
Go-Rift	1	12.00	1.11	4.43
	2	10.19	1.29	5.18
	3	10.88	1.24	4.96
Rift	1	9.78	1.00	4.24
	2	9.33	0.97	4.10
	3	9.17	1.03	4.37

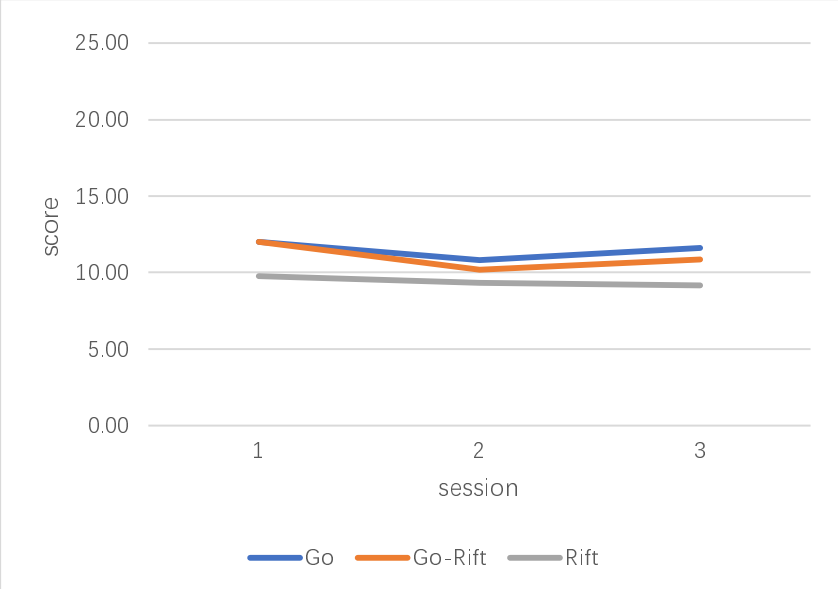


Figure 21. Quiz score for each session in each completer group

Table 48

Within-Subjects Contrasts of Quiz Score for Completers

Group	Effect	F	p	η^2
Go	Linear	0.13	0.720	0.009
	Quadratic	0.76	0.396	0.048
Go-Rift	Linear	1.19	0.293	0.073
	Quadratic	2.82	0.114	0.158
Rift	Linear	0.41	0.532	0.023
	Quadratic	0.03	0.871	0.002

Pre- and post-tests. Compared with the pre-tests, all the three groups had some increase in both the post-mental rotation test and the post-astronomy knowledge test (see Table 49 and Table 50). Aside from the Go group in the post-astronomy knowledge test, all groups had a significant increase in both the post-mental rotation test and the post-astronomy knowledge test.

The increase of the Rift group had the largest effect size among groups in both these two tests. This is consistent with the analysis results after multiple imputation. Specifically, the Go group had a significant increase in the post-mental rotation test, $t(15) = 2.73$, $p = 0.016$, Cohen's $d = 0.685$; the Go-Rift group had a significant increase in both the post-mental rotation test, $t(15) = 3.44$, $p = 0.004$, Cohen's $d = 0.862$, and the post-astronomy knowledge test, $t(15) = 2.22$, $p = 0.042$, Cohen's $d = 0.555$; the Go-Rift group had a significant increase in both the post-mental rotation test, $t(17) = 6.69$, $p < 0.001$, Cohen's $d = 1.577$, and the post-astronomy knowledge test, $t(17) = 2.43$, $p = 0.026$, Cohen's $d = 0.573$.

Table 49

Descriptive Statistics of Pre- and Post-Test Scores for Completers

Test Type		Group	Mean	Std. Error of Mean	Std. Deviation
Mental Rotation Test	Pre-Test	Go	5.44	0.79	3.14
		Go-Rift	5.63	0.80	3.20
		Rift	3.89	0.47	2.00
	Post-Test	Go	7.06	0.78	3.11
		Go-Rift	7.06	0.90	3.60
		Rift	6.33	1.01	4.28
Astronomy Knowledge Test	Pre-Test	Go	9.00	1.05	4.20
		Go-Rift	6.44	0.95	3.81
		Rift	7.06	0.56	2.36
	Post-Test	Go	9.06	1.02	4.09
		Go-Rift	7.50	1.09	4.35
		Rift	7.94	0.86	3.64

Table 50

Comparison of Pre- and Post-Test Scores in Each Completer Group

Test Type	Group	Mean Difference	Std. Error Mean	Std. Deviation	t	df	p
Mental Rotation Test	Go	1.63	0.60	2.38	2.73	15	**0.016**
	Go-Rift	1.44	0.42	1.67	3.44	15	**0.004**
	Rift	3.17	0.47	2.01	6.69	17	**<0.001**
Astronomy Knowledge Test	Go	0.06	1.03	4.11	0.06	15	0.952
	Go-Rift	1.06	0.48	1.91	2.22	15	**0.042**
	Rift	1.61	0.66	2.81	2.43	17	**0.026**

Summary. The results of completer analysis are very similar to the analysis results after multiple imputation. There are only two differences directly related to initial predictions and one minor difference in data comparison which does not influence final conclusions. The first is that the result of completer analysis showed that the Rift group had a smaller decrease from the first session to the third session, while the analysis result after multiple imputation showed that the Rift group had a larger decrease from the first

session to the third session. The second is that that the completer analysis result showed that the Rift group had a larger decrease from the first session to the third session, while the analysis result after multiple imputation showed that the Rift group had a smaller decrease from the first session to the third session. The minor difference in data comparison is that the Go group did not show the highest engagement among groups in the first session and the highest engagement in the second session among sessions within its group.

Qualitative Results

A total of 12 themes were found after two rounds of qualitative analysis. Eight of them provided evidence to support findings in the quantitative strand. Three offered the context to help researchers understand why some hypotheses and predictions were not supported by the quantitative results. One showed a deviation when it was compared with the quantitative result. Table 51 is a comparison of these themes and corresponding quantitative results.

Table 51

Comparison of Themes and Quantitative Results

Themes	Quantitative Results
Support significant findings in quantitative research	
1. VR condition influenced sense of embodiment	VR condition had a directly positive effect on sense of embodiment
2. VR condition / Sense of embodiment influenced presence	Sense of embodiment had a directly positive effect on presence
3. Sense of embodiment influenced motivation	Sense of embodiment had a directly positive effect on motivation
4. Presence influenced motivation	Presence had a directly positive effect on motivation
5. Interviewees' sense of embodiment increased when they came into the high-level VR condition from the low-level VR condition	Sense of embodiment increased significantly in a high-level condition
6. Interviewees' motivation increased when they came into the high-level VR condition from the low-level VR condition	Motivation increased significantly in a high-level condition
7. The sense of embodiment had more potential to increase in a high-level VR condition than in a low-level VR condition after the first session.	Sense of embodiment decreased more slowly in a high-level VR condition than in a low-level VR condition after the novelty effect happened
8. Complexity delayed the decrease of motivation for the Rift group	Motivation decreased more slowly in a high-level VR condition than in a low-level VR condition after the novelty effect happened
Provide context to understand rejected hypotheses or predictions	
1. Learning material was easily understood	VR condition did not have a directly negative effect on cognitive load
2. The process of becoming familiar with operation more significantly delayed the decrease of motivation than the decrease of presence in the high-level VR condition	Presence did not decrease more slowly in a high-level VR condition than in a low-level VR condition after the novelty effect happened

3. Manipulation distracted text reading	a. Motivation did not have a directly positive effect on learning achievement b. Engagement did not increase significantly in a high-level VR condition c. Learning achievement did not increase significantly in a high-level VR condition
Deviation	
1. Interviewees' presence increased when they came into the high-level VR condition from the low-level VR condition.	Presence did not increase significantly in a high-level VR condition.

The following are the theme passages in the order as Table 51. The first eight are those which supported quantitative results.

1. ***VR condition influenced sense of embodiment*** (Quantitative result: VR condition had a directly positive effect on sense of embodiment.)

Due to the differences between VR applications, interviewees in the Go group had a lower sense of embodiment than those in the Rift group. Interviewees in the Go group pointed out the design deficiencies of their application: the insufficient capability in control and the rigid virtual body. Their feeling of body ownership and free control was very low or even did not exist through sessions. For example, T_29, an interviewee in the Go group, said, "I just, I don't know. I just didn't feel connected to it [the virtual body]. It was really rigid. It never moved." The response from P_21, another interviewee in the Go group, is similar:

> I mean it was sitting the entire time and I was standing so I didn't really feel like that body was mine. I mean, if it was one to one then that'd be something else, but really, I just had a virtual model of a pointer and it just controlled what I did.

By contrast, five out of six interviewees in the Rift group felt that they could control the virtual body as their own, and the left one felt the sense of embodiment was little for her unfamiliarity with operating the VR application. The five interviewees explained that this was because the virtual hands and fingers could be manipulated and moved as real ones. P_39 was an interviewee in the Rift group. He said, "Because I had the controller and it had my hand on it and so, when I brought the planet over and I was manipulating it, I felt like I was manipulating it with my hand." Overall, interviewees in the Rift group had a higher sense of embodiment than those in the Go group.

2. ***VR condition / Sense of embodiment influenced presence*** (Quantitative result: Sense of embodiment had a directly positive effect on presence.)

In general, interviewees in the Go group agreed that they had a sense of being in the universe. Particularly, when the spaceship was in the state of auto pilot and they looked around the universe, they had the highest sense of being in the universe. This feeling was more often evoked in the second and the third sessions, when they became familiar with the application. T_36 described his experience, "In the second session, it felt kind of real, everything looked so in detail. When I was traveling, you could actually see all the planets, like going- It felt like the autopilot was fast." T_11 recalled his experience in the third session:

> What's interesting is when I clicked past Pluto and it started taking me toward the sun, I think it sounded like there was wind and stuff going and it sounded more realistic as opposed to everything else. Cause there was just like music playing throughout it.

Nevertheless, a lack of control over the avatar might undermine this sense and made an interviewee feel that she was in a room. P_31, an interviewee in the Go group, said:

> There were times I knew I was in the room, if I moved or something just moved my arm or something like that. But most of the time, I felt like I was in the spaceship and the virtual body felt like I was inside there. So it was cool with the layout of the VR, it felt like I was actually in the space shuttle going to each body and learning about it.

By contrast, besides being in the state of auto pilot, interviewees in the Rift group said that when they were in front of the celestial bodies, they felt small. This feeling also made them have a sense of being in the universe across sessions. Feeling small is individuals' response to visual stimuli, which is a form of perceived embodiment (Gonzalez-Franco & Peck, 2018). P_30, an interviewee in the Rift group, said, "[I had the

highest sense of being in the universe in the first session.] During the first one, I definitely got lost in the virtual reality, where I was just so small in comparison to everything else."

In summary, interviewees' sense of presence was influenced by the capability of flexible control and body ownership. Interviewees in the Rift group had more opportunities to enjoy the sense of being in the universe in this VR learning experience than those in the Go group.

3. ***Sense of embodiment influenced motivation*** (Quantitative result: Sense of embodiment had a directly positive effect on motivation.)

Interviewees in the Rift group felt that their VR learning experience was interesting. When talking about why they were interested in this VR application or this VR learning experience, interviewees in the Rift group referred their interest to the interaction with the virtual environment. Their responses to whether they were interested in this experience were mostly very confirmative. Interviewee P_30 said:

> Very, extremely, a hundred times. [……] It was a completely different activity than I've ever done in my life. The way that I learned wasn't just listen to a lecture or taking notes. It was way more hands on, way more detailed, and I felt as though I could really envelop the material at hand versus just reading it on a piece of paper.

Interviewee P_25 added:

> I'd say it's more than interesting. I'd definitely say it's groundbreaking when it comes to learning because rather than just say do a lecture you're interacting with the planets, you're getting to see all the features, you get to twist around the planet. One of the things I did with some of the bigger planets is that I would climb up let's say Neptune I would climb up the middle and then I'd fly up above it so I could see it from a top down view. That was definitely something that you can't really do when you're sitting in a lecture just scribbling away at notes.

Interviewee P_37 did a comparison:

> I think it's an experience that a lot of people don't have, being in space, and flying around, and seeing these things, and looking around freely. [……] The most interesting thing was being able to grab the planets, and move them around, and look at them from different angles. You can't do that with a picture or a video. You can't look at the top and then twist it around and look at the bottom.

4. ***Presence influenced motivation*** (Quantitative result: Presence had a directly positive effect on motivation)

Interviewees in the Go group also felt that their VR learning experience was interesting. When talking about why they were interested in this VR application or this experience, they responded that this was because of visual learning. P_29 said, "Well, yeah. […...] I just got to see them. I felt like saw them in person. You know I didn't, I know I didn't. I just felt like I was there in person to see all the planets. So I thought that was really interesting how I got to visualize that." P_31 recalled her experience:

> I thought it was pretty cool. If it was an application, I would probably play it all the time. So I thought it was pretty interesting just to learn about all the celestial bodies and see the images that are taken by actual cameras in the satellites and everything. So you can get a glimpse of the surface of the celestial body you were looking at. So it felt more real than the virtual body itself. Like, "Oh this is what it actually looks like.

5. ***Interviewees' sense of embodiment increased when they came into the high-level VR condition from the low-level VR condition*** (Quantitative result: Sense of embodiment increased significantly in a high-level VR condition.)

Interviewees in the Go-Rift group had the highest sense of embodiment in their third session. Six interviewees provided effective responses. Five of them ranked the third session as the one in which they had the highest sense of embodiment. They attributed this to the VR application (Rift version), in which they had visible virtual hands and fingers, and more flexible control than before. P_23 did a comparison of her experience between these two different conditions:

> I feel like I could, because I could see what my hands are doing. I can move my thumbs and it would move the thumbs too in the game. So it was like, okay that's kind of cool. So like whenever I would move my thumb I would see the thumb like come up and it was, it was really nice. But yeah, it's pretty good. […...] Whenever I would grab the planets, it did exactly what I was doing. It felt more real to me than if I was just like here grabbing nothing.[……] for the first part, the first two sections, I feel like it wasn't that as well an experience because I only had one hand here. I could feel this hand just on my side. You know? The first two weren't as good as this last one. This last one I actually turned, I moved, and I got scared and you know.

T_22 was the only exception. He ranked his feeling to be equal across sessions. His argument was that the visible movement of the virtual body was still very limited in the third session. He proposed that further improvement could include adding the visible leg movement and the capability of sitting down in the spaceship.

6. ***Individuals' motivation increased when they came into the high-level VR condition from the low-level VR condition*** (Quantitative result: Motivation increased significantly in a high-level VR condition.)

Most interviewees in the Go-Rift group had the highest interest and motivation to continue to play in their third session. Eight interviewees provided effective responses. Seven of them ranked the third session as the one that they had the highest interest and / or motivation to continue to play. Three interviewees referred fun as the main reason. P_37 evaluated this experience:

> I think it's an experience that a lot of people don't have, being in space, and flying around, and seeing these things, and looking around freely. [……] The most interesting thing was being able to grab the planets, and move them around, and look at them from different angles. You can't do that with a picture or a video. You can't look at the top and then twist it around and look at the bottom.

T_22 also attributed his high interest in the third session to the various tools and functions. This tools and functions made him more engaged in the environment, and then feel more interest. T_37 was the only one that did not rank the third session as the highest

one. He ranked the first session as the highest for newness. He explained that his interest and motivation was constrained by playing time; he was very busy every day. The Rift-version application, which was used in the third session for the Go-Rift group, was more complicated than the Go-version. On average, participants spent longer time to play in the Rift environment than the Go environment. This partly supported why T_37 did not choose the third session as the highest one:

> Just kind of in general my interest to keep playing was kind of based off of how I felt like how long I felt I was in it just because I know that there's like I have to answer questions, I need to do a little survey and stuff. I didn't want to take up too much time. So my interest in keep playing was mainly based off of that. So it's like if I felt like I was taking a long time I'd probably say oh, no I read the information, I generally understand it. I think I'm good. Whereas one of the other times I think I went back for just a little bit. I felt like, okay, I haven't been in here for that long and I would like to go over this information a little bit more. So.

7. ***The sense of embodiment had more potential to increase in a high-level VR condition than a low-level VR condition after the first session.*** (Quantitative result: Sense of embodiment decreased more slowly in a high-level VR condition than in a low-level VR condition after the novelty effect happened.)

Five interviewees in the Rift group provided effective responses. Four of them ranked the first session as the one that they felt least control and body ownership due to unfamiliarity. P_25 was one of them. He explained:

> I feel like session two was also probably the best where you could control it the best. Session one is probably the least because you're getting used to it and then session three is moderate because you're already used to it, but I see session two or three being either the most interesting or moderate. I see either one of them.

The fifth interviewee in the Rift group felt that his capability of control and feeling of body ownership was equal across sessions for his own intuition.

By contrast, the sense of embodiment for interviewees in the Go group was almost the same across sessions. Six interviewees in the Go group provided effective

ranking. Three of them responded that their feeling was equal across sessions and they did not feel that the virtual body was their own. P_21 explained that this was because of the VR application, “Pretty much the same across the three sessions. [……]. I mean, realistically we're using the same program and it was the same body that I was as in and so, yeah I mean nothing really changed that often within the body, so yeah.” The other three interviewees in the Go group ranked the first session as the highest or the lowest for different reasons: one ranked it as the highest for newness, one ranked it as the lowest for being unfamiliar with the environment, the last one ranked it as the lowest for he could not point to a target precisely with the virtual controller in that session.

Overall, being unfamiliar with operation might influence interviewees’ sense of embodiment for both the Rift group and the Go group. Relatively, its influence on the Rift group was larger. This led to a lower decrease of sense of embodiment for the Rift group across sessions.

8. ***Complexity delayed the decrease of motivation for the Rift group*** (Quantitative result: Motivation decreased more slowly in a high-level VR condition after the novelty effect happened.)

Three out of five interviewees in the Rift group mentioned that although they were motivated in the first session for newness, they liked the second or the third session more than the first session. This was because they were more used to the environment and the learning task at that time. The response from P_30 is a typical example:

> One being the least, 10 being the most interested. [The first session], I would say probably about an eight. I would say an eight because I was interested, but I was also very nervous. In the second one I was about an eight and a half, because although I felt as though it was more difficult in that session, I still was determined to fulfill it and complete this part of the research. In this one I'm about a 10. […...] Just because I feel more comfortable and it was intriguing about the moons and stuff.

By contrast, newness was a main factor that influenced interviewees' interest and intrinsic motivation to play for the Go group. Three out of five interviewees in the Go group ranked the first session as the one that they had the highest interest and intrinsic motivation. They attributed that to newness. For example, P_29 said, "After the end of the first session I was really interested because it was just new to me, and I was excited, and then I guess my interest kind of declined a little bit over time. […...] It just felt like the same thing just repeated." P_31 was supper excited after the first session:

> I would say actually the first session would be the highest [motivation to play] because I started out with it and it was an astronomy VR and astronomy is interesting to me. So after I finished with the first session, it was like, "Oh, can't wait for the second session." So it would probably just go in order for the first would be the highest, and then the second is the middle.

The following three themes are those which provided context to help researchers understand why and how the initial hypotheses or predictions were rejected by quantitative results.

1. ***Learning itself was easily understood.*** (Quantitative result: VR condition did not have a directly positive effect on cognitive load.)

All but one interviewee said that the learning material was not difficult. This was because it was only factual information, neither complex nor abstract. P_39 said, "It was very clear, and it wasn't confusing at all, and it was straight forward." It was possible for a novice in the astronomy field to learn easily. P_23 said, "Well it was really new to me, because I didn't really know much about solar system. I felt like, because I kind of knew what I was looking for, it wasn't as hard as I thought it was going to be." T_37 took the same attitude, "I didn't feel like it was difficult though there's a lot of information that I've never really researched about. I don't really know much about planets in general, but it's a

lot of information that's easy to understand and comprehend." T_28 was the only exception. She attributed the perceived difficulty to her lack of knowledge in the astronomy field. Furthermore, despite that most interviewees did not feel that the learning content was difficult, six of them pointed out that memorizing all the facts in a session was sometimes hard. P_29 said:

> The information individually wouldn't be difficult. I can memorize all the facts about Pluto, but then to memorize all the facts about... Or try and learn all the facts about Pluto, and then Neptune and then Uranus, and then all the moons in between those. That was like a bit much. There was a lot of information to try and learn at once.

Similarly, P_26 said, "I did not think the learning task was very difficult at all. Like I said, it's just me myself, I have a hard time remembering certain little details. I'll remember one or two here but remembering all of them is a little hard."

2. ***The process of becoming familiar with operation more significantly delayed the decrease of motivation than the decrease of presence in the high-level VR condition.*** (Quantitative result: Presence did not decrease more slowly in a high-level VR condition after the novelty effect happened.)

Five out of six interviewees in the Rift group provided effective ranking responses for their interest and motivation across sessions. Three of them did not select the first session as the session that they had the highest interest and motivation. They mentioned that although they were motivated in first session for newness, they liked the second or the third session more than the first session. This was because they were more used to the environment and the learning task at that time. The response from P_30 is a typical example:

> One being the least, 10 being the most interested. [The first session], I would say probably about an eight. I would say an eight because I was interested, but I was also very nervous. In the second one I was about an eight and a half, because although I felt as though it was more difficult in that session, I still was

> determined to fulfill it and complete this part of the research. In this one I'm about a 10. […...] Just because I feel more comfortable and it was intriguing about the moons and stuff.

By contrast, four of them ranked the first session as the highest session that they had the sense of being in the universe than in a room. They attributed this to the environment's newness but realistic. This sense was decreased when they were familiar with the environment and found issues in fidelity. For example, T_28 said:

> In the first session, I was very curious and felt very realistic. This sense of reality was slightly dropped in the following sessions. It was mainly reflected in the universe, which should be active, while it was still slightly like a mirror in VR. Of course, it was more realistic than 2-D pictures.

The response from P_30 was similar:

> During the first one, I definitely got lost in the VR reality, where I was just so small in comparison to everything else. The second one, I was a lot more in a sense that I was in a room. In the third one I was just like, I was a lot more comfortable. I was like, I'm in a room and I'm in the solar system.

P_25 was the only interviewee in the Rift group that did not choose the first session as the one affording the highest sense of being in the universe. He said that there were two reasons. One was that he could not comfortably control the application in the first session. The other was that celestial bodies in the first session were less interesting than those in other sessions.

In summary, the process of becoming familiar with operation influenced both the decrease of motivation and the decrease of presence for interviewees in the Rift group. Relatively, its influence on the decrease of motivation was perceived by more interviewees. Besides the influence of unfamiliarity with operation, the decrease of presence was also influenced by the fidelity of the virtual environment.

3. ***Manipulation distracted text reading*** (Quantitative results: a. Motivation did not have a directly positive effect on learning achievement; b. Engagement did not increase significantly in a high-level VR condition; c. Learning achievement did not increase significantly in a high-level VR condition.)

Most interviewees in the Go-Rift group felt that manipulation was distracting for memorizing text information in the third session. Seven interviewees provided effective responses. Six of them agreed that rotating, zooming in and out, and grabbing was distracting in memorizing text information. For example, P_37 said, "In this [third] session, I would read the text boxes, and then I would spend time flying towards the body and spinning around looking at it. By the time I was done with that, I'd kind of forgotten what I had read." He felt that observing 3-D models was more helpful. "In the first one, there was just enough to see, that it didn't distract from reading the text and/or retaining the information." P_32 took a similar point of view as P_37. He did not feel that interacting with the celestial bodies was connected very well with the actual information. He felt that observing models was "a little straighter forward and there's less to think about in the ideas of controls and everything".

P_23 was the only exception. She said that she had the highest concentration in the third session. Rotating, zooming in and out, and grabbing was interesting but not distracting. She felt that this learning method was more effective than observing 3-D models. She said:

> For the first and second [sessions], I had to only read, and just click one button. I was just concentrating on reading and I feel like it confused me. I had to read so much. […In this session,] I was able to grab it and move [a celestial body] in front of me, above me, beside me, and stuff like that. I felt like it was a lot better than having to just look at it from far away.

P_24 took another point of view. She admitted that learning in the first two sessions were easier with less distracted grabbing and clicking buttons. She also thought that interaction enhanced it memorable, “The third one is if you wanted to go back and you remember it a little more because it was a fun experience for you.”

Following is the theme that showed a deviation with the corresponding quantitative result.

1. ***Interviewees’ presence increased when they came into the high-level VR condition from the low-level VR condition*** (Quantitative result: Presence did not increase significantly in a high-level VR condition.)

While this prediction was rejected by the quantitative result, it was supported by the responses from most interviewees in the Go-Rift group. Specifically, six out of seven interviewees ranked the third session as the one that they had the highest degree of being in the universe. They explained that this was because of more interaction with the surrounding and more free control on the virtual body. For example, P_32 said:

> […...] with this [third session] it happened a lot more often [that it was like in the universe but not in a room] because that it had the more options to fly out, to interact and just in general have more interactivity with the actual celestial body and the information. So I feel like this one was a little more out -of-body I guess.

Furthermore, T_22 mentioned that due to more options, he could be “more engaged in terms of looking at the universe and looking at the planets and understanding everything”. This increased his feeling of being in the universe in the third session. However, T_29 ranked the third session as the lowest and the first session as the highest. She felt that some functions in the third session was not realistic:

> I think changing the map on the celestial body just made it seem like it was more of a virtual reality, as opposed to the first one. [………Ranking the first session as

the highest one] probably because it was my first time doing virtual reality. So it was new.

CHAPTER 5

DISCUSSION

The purpose of this study was to assess the impact of a VR condition on an effective STEM learning experience with a consideration of novelty. Based on prior literature, a model that described how VR influenced a learning experience and a series of predictions that related to novelty in VR learning were established. To validate the model and these predictions, a multi-session experiment was designed and implemented. Quantitative data were collected before, during, and after the experiment, and then analyzed. Interview were taken after the experiment. Its purpose included triangulating and explaining the quantitative results. This chapter focuses on interpretations of study results, its contributions and implications, limitations, and recommendations for future research.

Interpretations of Study Results

The interpretations of study results were conducted based on the research questions in this study. It included four parts: (1) the impact of a VR condition on an effective learning experience, (2) the magnitude of novelty effects in a VR learning experience, (3) the attenuated effectiveness of a VR condition after the novelty effect happens, and (4) the overall trends of assessed factors.

The impact of a VR condition on an effective learning experience. The first research question in this study is how a VR condition impacts an effective VR learning experience. The established cross-sectional model implies an uncertainty effect of

embodiment on VR learning and the important mediation role of motivation and engagement in the VR learning process. This conclusion is based on the following facts:

The result that VR condition did not show a significant effect on cognitive load is against most prior research that advocated embodied learning (Goldin-meadow et al., 2001; Hu, Ginns, & Bobis, 2015; Paas & Sweller, 2012). The limitation of modality effects in the traditional multimedia learning field may explain the weak relationship between VR condition and cognitive load in this study. Modality effects refers to a phenomenon that individuals have a higher learning achievement for a reduced cognitive load when both visual and audio channels are involved in the learning process (Sweller et al., 1998). Modality effect is recognized as invalid in a low element interactivity environment, in which information being processed is not constrained by the capability of working memory (Sweller et al., 1998; Tindall-ford, Chandler, Sweller, & Bartlett, 1997). The explanation may be applicable to this study, which initially predicted that a high-level VR condition could reduce cognitive load in learning for its capability of allowing manipulation. Text information was the main source of learning content in this VR learning experience. Evidence from the mean of cognitive load and responses from interviewees showed that text information was easily understood. Manipulation has the potential to lower cognitive load through creating embodied representations (Glenberg, 1997; Ping & Goldin-meadow, 2010). However, when the learning difficulty is small or moderate, this effectiveness may not be obvious.

In addition, the results showed that the both sense of embodiment and presence did not have a significantly direct effect on engagement, while the indirect effect of presence on engagement through motivation was significant. This finding highlights the

mediation role of motivation for enhancing participants' engagement. If a participant didn't have an interest in using VR to learn astronomy knowledge, this participant would not engage in the learning activity even in a high-level VR condition. This finding is in line with the claim that high presence is an antecedent of engagement but not a requisite one in a virtual or video environment (Hoffman & Nadelson, 2010; Hoffman & Novak, 2009). Most prior studies that found a significant direct effect from presence and engagement did not involve a consideration of mediators in their research (Animesh et al., 2011; Nijs et al., 2012; Zaman, Ananda, & Dai, 2010).

Only the sum of indirect effects of sense of embodiment on engagement was statistically significant. This means that the two mediators, motivation and presence, are dispensable for delivering the effect of sense of embodiment onto engagement. It implies that a high-level engagement also requires the participant to be conscious of surrounding environment besides personal interest. This consciousness might help the participant reach a match between personal capabilities and the demands of assigned tasks, and then more deeply involved in the learning process. This learning process included not only remembering text information, but also visually exploring the universe with or without manipulation.

Research result also showed that the effect of motivation on learning achievement was mediated by engagement and there was no necessary direct relationship from motivation to learning achievement. This is consistent with the study of Makransky, Terkildsen, and Mayer (2019). They interpreted their results as meaning that motivation caused by hedonism creates extraneous processing, which may distract the process of knowledge construction. Their explanation is also applicable to this study. Interviewees in

the Go-Rift group said that although the capability of manipulation made them feel fun with a higher interest in the third session, manipulation distracted their learning at the same time. In other words, the source of interest or motivation determines whether the enhanced motivation can effectively increase the learning achievement. Furthermore, participants in this study used independent controllers to interact with the virtual environment, which is more natural than the touchpad used in the study of Makransky, Terkildsen, and Mayer. This study provides evidence to support generalizing their idea to broader application scenarios. Moreover, this result is also congruent with the assumption in CATLM that motivation influences learning through increasing or decreasing cognitive engagement (Roxana Moreno & Mayer, 2007; Pintrich, 2003). On the contrary, this result is different from the study of Lee, Wong, and Fung (2010) and Salzman, Dede, Loftin, and Chen (1999), which claimed that motivation influenced learning outcomes. Both did not involve engagement as a mediator between motivation and learning outcomes in their models. Specially, while Salzman et al. predicted there was an additional factor beyond motivation that caused the learning differences among participants in their VR experiment, they did not point out what the additional factor was.

The magnitude of novelty effects in a VR learning experience. The second research question in this study is how much novelty can influence a VR learning experience. Both quantitative and qualitative results confirmed that when novelty happens, sense of embodiment and motivation will increase significantly. However, no significant quantitative increase was found when presence, engagement, learning achievement were assessed. The deviation of presence is explained by the discrepancy

among individuals in understanding this concept, while the deviation of engagement and learning achievement may be from the application scenarios of embodied learning.

Although the qualitative results showed that most interviewees felt a higher degree of presence in the high-level VR condition than the low-level VR condition, no significant difference was found between the second session and the third session for the Go-Rift group. An explanation is that there was a relatively large difference among participants when they assessed their perceived presence. The follow-up analysis showed that while the total effect of a VR condition on presence and on motivation were almost the same (see Table 52), the standard deviation of presence was much larger than motivation (see Table 53). Thus, when the difference between the second session and the third session was assessed, presence had a larger p-value in the paired-samples T test than motivation. This finding may be explained by the study of Slater (2004). Slater said that presence is similar to individuals' perceptions of a colorful experience, as both of them cannot be identified by any mental activities; the establishment of presence depends on individuals' interpretations of cues in an environment. In other words, individuals are relatively easier to reach a consensus about their perceived fun and interest than their interpretations of sensory cues, when they are learning in a VR environment.

Table 52

Comparison of Effects from VR Condition to Presence and Motivation

	Direct Effect	Indirect Effect (Mediator)	Total Effect
VR -> Presence	-0.07	0.26 (Sense of embodiment)	0.19
VR -> Motivation	0.00	0.19 (Sum) -0.03 (Presence) 0.12 (Sense of embodiment) 0.10 (Sense of embodiment and Presence)	0.19

Table 53

Comparison of the Standard Deviation between Presence and Motivation for the Go-Rift Group

Factor	Session 1	Session 2	Session 3
Presence	24.11	27.13	14.95
Motivation	14.01	14.45	4.23

Moreover, the hypotheses that engagement and learning achievement will increase in a high-level VR condition were not supported by the quantitative results. The quantitative results showed that the mean of engagement for the Go-Rift group decreased from the second session to the third session and the mean of learning achievement for the Go-Rift group had a slightly increase from the second session to the third session. The responses from interviewees in the Go-Rift group provide explanation to these two unexpected results. Most interviewees agreed that the Rift-version VR application was more novelty; although they had higher motivation, they needed more efforts and time to become familiar with the application, which distracted their learning of text information. This is consistent with the conclusion in the study of Jeno, Vandvik, Eliassen, and Grytnes (2019). Jeno et al. claimed that need-satisfaction is a moderator between novelty and learning achievement. It is different from the study of Tsaya and Kofinas (2017), in which students' engagement decreased over time after novelty happened. This difference is possibly caused by the different definitions of engagement. In this study, engagement referred to individuals' feeling in real time. Data was collected from participants' responses to the questionnaire immediately after the VR learning activity. In the study of Tsaya and Kofinas, engagement was related to students' involvement in a course. They used the number of activities that students completed on an online learning system per week as the measurement of engagement. Since differences exist in definition and

measurement approach, it is understandable that divergent results were obtained in these two studies.

To understand why the Rift group had the lowest mean scores in the session quizzes among groups, a follow-up comparison was conducted. This comparison separated all the questions in a session quiz into a low level and a high level based on the cognitive process dimension defined in the revised Bloom's taxonomy. Specifically, questions that assessed participants' low-level cognitive process included recognizing questions and recalling questions, while questions that assessed participants' high-level cognitive process included understanding questions and the evaluating question.

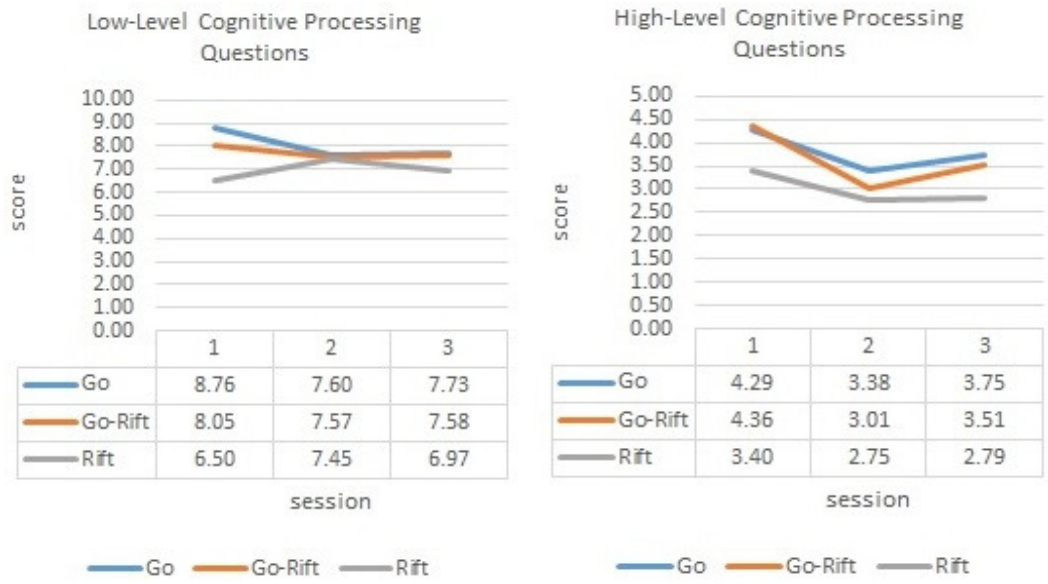


Figure 22. Mean scores in session quizzes for the low-level and the high-level cognitive processing questions

Figure 22 shows the changes of participants' scores categorized by the cognitive process dimension for each session in each group. The data analysis method was multiple imputation. The data were from the 63 participants after data screening in the data comparison phase. Participants in the low-level VR condition had a significantly higher

score for the low-level cognitive processing questions than the high-level VR condition in the first session ($t = 2.30$, $p = 0.022$, Cohen's $d = 2.131$), while the difference in the third session was small. It means that when participants in the high-level VR condition were familiar with the VR application, their capability of the low-level cognitive processing had the potential to be the same as those in the low-level VR condition. In addition, the difference between the Go group and the Rift group for the high-level cognitive processing questions was large through sessions. Particularly, the Go group had a significantly higher score than the Rift group in the third session ($t = 2.366$, $p = 0.018$, Cohen's $d = 0.226$). The difference between the Go group and the Go-Rift group was small in the third session. This is because there was an outlier in the Go-Rift group. A participant in the Go-Rift group spent an unusual long time (approximately 40 minutes) to finish the evaluating question in the session 3 quiz for being extremely interested in this study, and her total score for the high-level cognitive processing questions in the third session was 10. This was the highest score among all participants in the session 3 quiz. If this score were removed, the mean of the Go-Rift group would decrease from 3.51 to 3.07, which was also much lower than the Go group. These results imply that participants' capability of the high-level cognitive processing was undermined in the high-level VR condition, and this capability was not regained even when they were familiar with the VR application. A possible explanation is that the experimental design for the high-level VR condition was much more complicated than the low-level VR condition and participants in the high-level VR condition did not have sufficient learning time. Memorizing the text information in the VR application was the priority in the VR learning experience, as its content was directly related to the questions in session quizzes.

Participants continuously improved their learning strategies to memorize the text information and became familiar with the VR application through sessions. Therefore, the difference in the low-level cognitive processing was gradually narrowed between participants in these two VR conditions. However, besides memorizing text information, participants in the high-level VR condition needed to spend extra time on other assigned tasks, such as flying towards a celestial body and grabing a celestial body. They did not have enough to reflect what they had learned from the text information and make connections in the learning process as those in the low-level VR condition. Thus, their capability of the high-level cognitive processing was lower than those in the low-level VR condition across sessions.

Although the mean of learning achievement did not have a significant increase from the second session to the third sessions for the Go-Rift group, participants in the Rift group showed the best performance in both the post-mental rotation test and the post-astronomy knowledge test. Their scores in these tests also had the largest increase from the pre-test to the post-test among groups. The explanation for their performance in the mental rotation test is their repeated practice of manipulation and grab during the learning process. Performance on the mental rotation test is sensitive to practice and experience (Baenninger & Newcombe, 1989; Kail, 1986; Peters et al., 1995). This study provided evidence that practice in a VR environment is also effective for improving the performance in the mental rotation test. Furthermore, the fact that the Rift group had the best performance in the post astronomy knowledge test might be attributed to embodied learning. There was no text information in the VR application showed any celestial body's size and its distance from the sun. Thus, participants needed to obtain this

information from other methods. Compared with the other two groups, participants in the Rift group had more opportunities to freely explore the virtual solar system with manipulation. This process helped participants establish a representation of the solar system in their memory. The larger a planet's size was, a larger magnitude gesture was required when individuals grabbed on its surface. According to the taxonomy of embodiment for education (Johnson-Glenberg, 2018), grab provided deeper memory for participants to compare the sizes among planets than static observation. As for the ranking of the eight planets from closest to furthest away from the sun, the capability of flexible movement supported participants in the Rift group to actively observe those planets from different angles. This free exploration involved more feedback channels than static observation. It strengthened participants' memory for the relatively location of each planet and added retrieval cues. Thus, participants in the Rift group could have the highest score in ranking the sequence of these planets from closest to furthest away from the sun. This result is in line with prior similar studies (Kozhevnikov et al., 2013; Shu et al., 2018).

The attenuated effectiveness of a VR condition. The third research question in this study is how the effectiveness of a VR condition changes after the novelty effect happens. Results confirmed that most assessed factors decreased slowly after the novelty effect happened. Particularly, the decrease was more slowly in a high-level condition than a low-level condition. Presence was the only exception. Table 54 shows the change of all assessed factors from the first session to the third session for the Go group and the Rift group.

Table 54

The Change of Assessed Factors from the First Session to the Third Session for the Go Group and the Rift Group

Group	Presence	Sense of Embodiment	Motivation	Engagement	Learning Achievement
Go	-3.57 (-5.14%)	-6.89 (-13.64%)	-3.70 (-4.00%)	-7.86 (-9.50%)	-0.85 (-6.51%)
Rift	-4.63 (-6.16%)	-5.26 (-7.08%)	-2.05 (-2.18%)	2.61 (3.26%)	-0.40 (-4.04%)

The quantitative results showed that motivation and sense of embodiment decreased more slowly for the Rift group than the Go group from the first session to the third session. This is congruent with the study of Berlyne (1970), as the Rift-version VR application is more complicated than the Go-version VR application. Qualitative results supported this explanation. The Rift group had more interviewees mentioned the process that they became familiar with operating the VR application than the Go group. In other words, the process of becoming familiar with the VR application was more significant for the Rift group. Therefore, due to the complexity of the Rift-version VR application, motivation and sense of embodiment decreased more slowly for participants in the Rift group than the Go group.

The prediction that learning achievement decreases more slowly in a high-level VR condition after the novelty effect happens was made from two ideas. One is the rewarding effect of the perception process (Berlyne, 1970). The other is the embodied learning method. Based on the discussion in the last subsection, the effect of embodied learning on the quiz score was minimum in this study for knowledge assessed in the session quiz were easily understood. In addition, while the analysis result after multiple imputation supported this prediction, the result of completer analysis did not support the

prediction. This means the rewarding effect of the perception process on learning achievement is also weak when the learning method does not match the learning content, even the learning task is complicated.

The prediction that engagement decreases more slowly in a high-level VR condition after the novelty effect happens was validated by the quantitative result. An interesting finding is that there was a large decrease of engagement from the second session to the third session for the Go group, despite that this decrease did not lead to a significant difference among groups in the third session. There were five items used to measure engagement in this study. These items are independent and represent different dimensions of engagement. Thus, paired-samples T tests were conducted to check the changes of each items in the engagement scale across sessions for each group (see Table 55). Result showed that only item En_3 (I had a good idea about how well I was doing while I was involved in the activity.) had significant changes during the experiment. Specifically, the mean of item En_3 had a significant increase for both the Go group and the Go-Rift group from the first session to the second session, then the mean of item En_3 for the Go group significantly decreased from the second session to the third session.

Table 55

Changes of Item En_3 Across Sessions

Group	Session Comparison	Mean Difference	Std. Error Mean	t	df	p	FMI
Go	Δ2-1	12.19	5.33	2.29	642	**0.023**	0.14
	Δ3-2	-9.90	4.52	-2.19	92	**0.031**	0.40
Go-Rift	Δ2-1	10.41	4.37	2.38	1327	**0.017**	0.10
	Δ3-2	-3.34	3.82	-0.88	149	0.383	0.31
Rift	Δ2-1	8.15	6.47	1.26	--	0.208	0.00
	Δ3-2	0.34	4.89	0.07	1117	0.945	0.11

Note. Δ2-1 is the paired-samples T test for the mean in the second session and the mean in the first session; Δ3-2 is the paired-samples T test for the mean in the third session and the mean in the second session.

According to the authors of this engagement scale, item En_3 corresponds to the dimension of unambiguous feedback in a flow state (Jackson & Eklund, 2004). That is, "the performer receives clear, unambiguous information that he or she processes effortlessly, keeping performance heading in the right direction" (p. 8). The responses from interviewees showed that they became more familiar with the application's operation and the learning task, and then felt more comfortable in the virtual environment in the second and third sessions than the first session. Therefore, their engagement increased from the first session to the second session. Moreover, the significant decrease of engagement for the Go group might be explained from the little feedback that they could receive in a Go-version VR environment. Learning with the Go-version VR application only involved static reading and observation. In other words, participants could only receive visual and audio feedback in that environment. By contrast, participants in the Rift-version VR environment could fly out of the spaceship and grab the celestial body in the learning process. This involved not only visual and auditory, but also proprioceptive and tactile feedback. Specifically, visual and proprioceptive feedback

helped participants adjust the flying direction and speed, so that they could reach the target celestial body; visual feedback also supported participants to compare and contrast their observation with text information when they closely manipulated the celestial body. Vibration, as a type of tactile feedback, reminded participants that they had touched the surface of a celestial body and could start to grab its surface. In sum, various channels of feedback deepened these participants' perceptions of the surrounding environment and helped them implement the assigned learning task. These diverse types of feedback together delayed the decrease of engagement in a high-level VR condition.

While the completer analysis supported the prediction that presence decreases more slowly in a high-level VR condition after the novelty effect happens, both the quantitative and qualitative results did not support it. This may be explained as the fidelity of the environment itself also influences individuals' perception of presence, or whether the environment is complex is not the only critical factor that determines individuals' perception of presence. This phenomenon was particularly remarkable for participants in the Rift group. Only one out of five interviewees in the Rift group mentioned that the unfamiliarity with the task undermined the perception of being in the universe in the first session, while the number was three out of five when they assessed their motivation in the first session. One interviewee in the Rift group indicated that she gradually found issues of fidelity from the virtual solar system during the learning process after feeling very realistic at the beginning, these issues decreased her sense of being in the universe over time.

The overall trends of assessed factors. This study compared the changes of five factors through sessions in different VR conditions. Based on the consistent degree

between the experimental results and the initial predictions from highest to lowest, the sequence of these factors is sense of embodiment, motivation, presence, engagement, and learning achievement.

The assessment results of sense of embodiment are consistent with all initial predictions of sense of embodiment. Specifically, participants in the high-level VR condition showed a significantly higher sense of embodiment than the low-level VR condition through sessions; when participants came to the high-level VR condition from the low-level VR condition, their sense of embodiment was significantly increased; participants in the high-level VR condition showed a smaller decrease of sense of embodiment than a low-level VR condition from the first session to the third session. Besides these, the changes of sense of embodiment from the low-level VR condition to the high-level VR condition for the completer cases showed a definite curve, as well as an overall upward trend. All these results reveal that sense of embodiment is a very sensitive factor to novelty when the novelty is related to manipulation.

The assessment results of motivation are consistent with most initial predictions of motivation. The only difference is that participants in the high-level VR condition (the Rift group) did not show a significantly higher motivation than the low-level VR condition (the Go group across the three sessions and the Go-Rift group in the first two sessions), despite that the mean of motivation in the high-level VR condition was higher than the low-level VR condition across sessions. In addition, while no significant different was found between the Go group and the Rift group in the third session, a significant difference was found between the second session and the third session for the

Go-Rift group. All these imply that although motivation is sensitive to novelty, its degree of sensitivity is lower than sense of embodiment.

The assessment results of presence have some deviations with initial predictions. Despite that the mean of presence in the high-level VR condition was higher than the low-level VR condition in the comparison between the Go group and the Rift group and the comparison within the Go-Rift group, no significant difference was found in any comparison. In other words, the sensitivity degree of presence to novelty is moderate. As it was explained in the last subsection, this may be due to a relatively large difference among participants when they assessed their perceived presence. In addition, there is a difference between the analysis result after multiple imputation and the result of completer analysis on whether the high-level VR condition has a slower decrease of presence than the low-level VR condition through sessions. According to the explanation in the last subsection, this may be because not only the complexity but also the fidelity of the environment influences individuals' perception of presence.

The assessment results of engagement showed some unexpected phenomenons. The most impressive phenomenon is the straight decrease from the second session to the third session for the Go group, which was interpreted as a lack of real-time feedback in the last subsection. Another interesting phenomenon is that the mean in the second session did not decrease when compared with the first session for all groups. A possible explanation is that engagement or flow required a match between participants' skills and the demands of the VR learning task. Participants were more familiar with the VR application and the learning task in the second session than the first session. Thus, the degree of engagement was stable from the first session to the second session.

Furthermore, participants in the low-level VR condition did not show a lower degree of engagement than the high-level VR condition in the first session. Particularly, the analysis results after multiple imputation showed that the mean of the low-level VR condition was even higher than the high-level VR condition in the first session. This situation is explained by the complexity of the VR application. That is, the Rift-version VR application is more complicated than the Go-version VR application. Participants in the high-level VR condition needed more time and energy to become familiar with it. This familiarization process decreased their engagement at the beginning.

Learning achievement is at the end of the sequence. Most initial predictions of learning achievement were not validated by both the analysis results after multiple imputation and the results of completer analysis. As it was discussed in the previous subsection, this might be caused by a mismatch between the learning method and the assessment method in the session quiz. Specifically, participants could obtain most of the knowledge for the session quiz from the text information in the VR application; these information were easily understood; due to the limited experimental time but relatively complicated learning tasks, participants in the high-level VR condition did not have enough time to reflect and have deep learning. Thus, both the strength of embodied learning in knowledge retention and the rewarding effect of perceptual processing were not shown in the session quiz.

In summary, motivation, as a critical indicator of the novelty effect, influences the changes of other assessed factors in a novel VR condition. Despite this, their changing magnitude and the time of change for novelty are also influenced by other elements. In this study, since the high-level VR condition was much more flexible than the low-level

VR condition in manipulation, the novelty effect on sense of embodiment was strengthened and more obvious than motivation. By contrast, the novelty effect on presence was partly undermined by the relatively large difference among participants on their perceived presence, and its decrease was speeded up by the fidelity of the condition. Due to the influence of the skill demand for operating the VR application, the peak value of engagement was delayed to the second session. In addition, switching to the high-level VR condition helped participants avoid a large decrease of engagement as the low-level VR condition from the second session to the third session for a lack of real-time feedback. Finally, the learning achievement, namely the score of the session quiz, did not show a obvious change for novelty, this is attributed to a mismatch between the learning method and the assessment method.

Contributions and Implications

Contributions and implications of this study include theoretical aspects, method aspects, and practical aspects.

Theoretical aspects. This study investigated the novelty effect in VR on STEM learning. It contributes to the research in the fields of VR STEM education, novelty study, and embodied learning, respectively. First, this study explored the pairwise relationships between key factors in a fully immersive VR learning experience. The established cross-sectional model is possibly the first complete model that explores how a fully immersive VR condition influences an effective learning process. This model involves motivation and engagement as mediators in an actual VR learning process. Although a VR condition can offer many affordances to benefit learning, if the individual does not have an interest

or does not have required knowledge or skills to engage in the learning process, this VR learning experience cannot output satisfied learning achievement. In other words, affordances are not equal to satisfied learning achievement, motivation and engagement mediate the effects of affordances on the final learning achievement. This model fills the research gap on how VR affects a specific learning experience and provides a more comprehensive interpretation of a VR learning experience than prior models.

Second, this study longitudinally evaluated the effect of a VR condition on a STEM learning experience. It may be the first study that explored the novelty effect in a VR environment. It demonstrated that the novelty effect also exists in the VR STEM learning field. In this study, most assessed factors (i.e., presence, sense of embodiment, motivation, and engagement) more or less increased and then decreased for novelty. However, their changing magnitude and the time that their peak values appeared were also influenced by other elements. The changes of engagement and learning achievement in a VR environment showed their distinctive characteristics, which are different from the results in traditional media- and computer-supported learning (Clark, 1983; Li & Ma, 2010; Merchant, Goetz, Cifuentes, Keeney-kennicutt, & Davis, 2014; Tsaya & Kofinas, 2017). An important conclusion is that novelty does not necessarily increase individuals' learning achievement.

Third, this study identified the potential of embodied learning in some typical application scenarios. The advantage of embodied learning is helping individuals to memorize and understand knowledge through establishing concrete representations (Glenberg, 1997; Ping & Goldin-meadow, 2010). Its strength was shown in learning spatial knowledge. However, when the learning content is easily understood and the

learning time is limited, the benefit of manipulation may be small, and manipulation may even undermine deep learning. The Rift group had the lowest scores in the session quizzes among groups supports this conclusion. This is a new discovery in the embodied learning research field. In addition, using the revised Bloom's taxonomy to explore the potential of embodied learning is a new research direction. With further extension, it may help researchers more clearly define the application scenarios of embodied learning than ever. This is different from a discussion on whether a taxonomy is correct (Skulmowski & Rey, 2018) and the possible influence of individuals' characteristics on the effectiveness of embodied learning (Agostinho et al., 2015).

Method aspects. This study innovatively measured individuals' engagement and learning thrice through the intervention and then compared the values across sessions. This is different from the evaluation methods used in most prior VR learning studies, which only involved a single or a short-term measurement. Study results showed that although the attenuation of the novelty effect was slow for most assessed factors, some unexpected large decrease might happen during a continuous VR learning experience. For example, the mean of engagement for the Go group had a large decrease from the second to the third sessions in this study. If a study only includes either a single or a short-term measurement, bias may be produced in some specific context.

Furthermore, this study did a comparison of individuals' psychological responses, motivation, engagement, and learning achievement between a low-level VR condition and a high-level VR condition. When the VR condition was switched from a low level to a high level, participants had a significant increase of sense of embodiment and motivation. The differences between the two conditions were large for presence,

engagement, quiz scores, and scores in the post-test as well. These results suggested that *a fully immersive VR environment* is a very ambiguous concept in an academic study. Thus, when researchers present their results of research in the VR learning field, it is recommended to clearly state the VR system or the level of a VR condition (i.e., immersion and embodiment) they used not only in the methods section, but also in the discussion section.

Moreover, this study used the multiple imputation method to handle missing data. The multiple imputation method is commonly used for the longitudinal studies in the social science and the medical science fields. This study extended its application to the VR learning study field. It benefits data analysis and experimental implementation. Specifically, this method has the potential to effectively use cases with missing data to do the data analysis with smaller bias than traditional methods (e.g., listwise deletion, pairwise deletion, and simple regression-based imputation) when the proportion of cases with missing data is large (>15%). From the practical aspect, VR learning experiments are often low cost-effective for the relatively limited equipment resources and the large space occupied. Effectively using cases with missing data to the data analysis means that experimenters can more efficiently collect data than before. It either saves the total experimental time or increases the number of effective cases. Both of them are positive for most VR learning studies.

Practical aspects. First, and most important, this study demonstrated the sustainability of VR STEM learning in instruction. Results showed that a VR environment, especially the high-level VR condition, has the potential to sustain individuals' motivation and engagement for novelty, while the decrease of the novelty

effect is relatively slow for most assessed factors (less than 10% in a three-session VR STEM learning activity). In other words, a sustainable VR STEM learning in instruction is highly possible if the learning content is continuously updated. This eliminates the STEM education community's concern on a steep decrease of students' interest in VR STEM learning over time. Policymakers can be cautiously optimistic about their VR education plan and promote the application of VR in schools' STEM activities.

Furthermore, this study has the potential to serve as a reference for researchers to evaluate the real effect of novelty on their own studies. Many researchers have mentioned that the optimistic results of VR learning experiments might be partly caused by individuals' perception of novelty (Hussein & Nätterdal, 2015; Jensen & Konradsen, 2017; Parmar, 2017). This issue is addressed from two aspects. On one hand, since the some assessed factors had a significant increase for novelty in this study, it should be cautious when a researcher wants to refer to study results published one or two years ago, as the public are more familiar with VR than that time; on the other hand, the relatively slow attenuation of the novelty effect means that if the result of a study is only generalized to a limited extent, the influence of novelty may be very small. Thus, if no special requirement and operation, the novelty effect can be ignored. A typical example is to generalize the results of a one-session study to a three-session study.

Besides the possible benefits to policymakers and researchers, the findings of this study may guide instructional designers when they develop a VR learning class. A novel VR condition may significantly increase students' motivation and delay the decrease of engagement. Therefore, introducing changes and novelty is recommended to prevent students from feeling tedium in the VR learning process. In addition, both motivation and

engagement play a mediation role in an effective VR learning experience. In other words, trying methods to increase students' motivation and engagement can enlarge the benefits of a VR condition on the learning achievement. According to the definitions of intrinsic motivation and flow, these effective methods include a match between students' personal interest and instructional content and a match between students' knowledge / skills and the demand of a learning task. Thus, fully understanding the preference of target population and providing a training section before a formal VR learning activity are valuable practices for designing a successful VR STEM learning course.

Finally, independent software vendors are hoped to be inspired for developing more cost-effective VR learning applications with the guidance from researchers. Both the low-level and the high-level VR conditions have their own strength and weakness in supporting STEM instruction. Developing an application that runs in a high-level VR condition typically needs more money and time than developing an application that runs in a low-level VR condition. If the instructed knowledge is easily understood and the difficulty is moderate, it may not be necessary to use a high-level VR condition to develop this VR application.

Study Limitations

While the current study aids in understanding how novelty influences an effective VR STEM learning experience, it has several limitations that should be considered. First, as a longitudinal study, the number of participants, the number of repeated measurements, and the duration of the intervention are relatively insufficient. The experiment had 77 participants, three sessions, and the interval between neighboring sessions were 4-7 days.

This insufficiency may undermine the generalization of the quantitative results. Particularly, when the decrease of assessed variables from the first session to the third session were compared between the low-level VR condition and the high-level VR condition for answering the third research question. Most of these decreases were small (less than 10%). If the experiment had four or more sessions, and more participants, the results might be more significant, persuasive, or reveal new issues.

Second, different content among sessions may lead to bias in assessment. Some participants' interest to the learning content itself was not equal among sessions. For example, several participants said that they liked the second session because there were more big celestial bodies. In addition, there was no limitation on the score that an individual could obtain in a quiz. This might also cause bias when the mean of quiz score was compared among sessions.

Third, the rate of dropout was relatively large in this study. Particularly, the Rift group had a higher attrition rate than the other two groups for its relatively complicated learning task. The multiple imputation method was used to handle missing data and a completer analysis was conducted to validate the analysis results after multiple imputation. Although their results were almost the same, as the assumption of multiple imputation is data missing at random (MAR), imputed data cannot fully represent real data. The completer analysis has the drawback that some marginally significant differences existed among groups in the pre-tests.

Last, participants in this study were from two fundamental-level undergraduate courses. Although participants were diverse in gender, race, major, and school year, a large percentage of them were freshmen whose major were in the STEM field. In

addition, the data were collected only within one semester at a large public university in the Southwest of United States. Therefore, due to the limitation of the sampling method, possible deviations may exist when the results are generalized to a broader population.

Suggestions for Future Research

This study provides initial findings about the effect of VR on a STEM learning experience with a consideration of novelty. Much work remains to be done in this area and future studies might be based on this one with expansion.

This study used a three-session laboratory experiment to simulate a VR STEM learning activity. Laboratory experimentation has some limitations to conduct further research in this direction. In this study, participants came to the room thrice. Many participants had shown their exhaustion in the third session. For example, some interviewees said that they would like to finish the experiment fast when they were in the third session. If the experiment was extended to four sessions, the rate of dropout might be very high. Furthermore, the experiment was implemented person by person. This method is low efficiency when compared with experiments which only need computers or mobile phones. Since all-in-one VR systems have been released from 2019, the high-level VR condition can be created by cheaper hardware than before. Considering the weaknesses of laboratory experiment and the technology development, future longitudinal VR learning research is recommended to be integrated into classroom instruction.

Moreover, the study results did not support the prediction that learning achievement in a high-level VR condition is higher than a low-level VR condition from

the aspect of quiz score. One possible reason is that questions in the seesion quizzes were mostly based on knowledge that were easily understood. It may not be the fields that the embodied learning method can show its strength. This issue led to a lack of knowledge on the possible changes of the effectiveness of embodied learning over time in a scenario where its advantage can be shown to be significant. The fact that participants in the Rift group, namely those learned in the high-level VR condition, had the best performance among groups in the post-test only demonstrated that the cumulative effect of embodied learning on learning spatial knowledge is significant in a VR environment. The research gap may be filled by a study that repeated measures participants' achievement in learning complicated knowledge or spatial knowledge with sufficient reflection time in a high-level VR condition.

Finally, the novelty effect is possibly a ubiquitous phenomenon in the human society. It is hoped that more studies on the novelty effect will emerge in the future. Although many researchers claimed that one of the limitations in their studies was lacking a consideration of the novelty effect, systematic studies on the novelty effect are rare at present. This is a pity in the modern society, where user-centered is one of the most important design principles for products and services. Understanding the novelty effect in a specific field and actively applying it in practice have the potential to help the industry win customers, the government serve the public, and finally increase citizens' happiness.

CHAPTER 6

CONCLUSION

The purpose of this study was to assess the impact of a VR condition on STEM learning with a consideration of novelty. It involved a series of exploration on a VR STEM learning experience: how the VR condition impacts an effective VR learning experience, how much novelty influences the VR learning experience, and how the effectiveness of the VR condition changes after the novelty effect happens.

Findings of the study showed that both motivation and engagement played an important mediation role in a VR learning experience. It means trying methods to increase students' motivation and engagement may enlarge the benefits of VR on learning achievement. In addition, results also showed that while the novelty effect might be significant, the attenuation of the novelty effect was slow in the VR STEM learning experience. Relatively, this slow attenuation was more significant in a high-level VR condition. This implies that it is possible to conduct a multiple-session VR STEM learning practice with expected learning achievement. Last, novelty does not necessarily increase learning achievement. The increase of learning achievement is more dependent on a match between the learning content and the learning method. The embodied learning method showed its strength in instructing spatial knowledge through establishing representations but uncertain effectiveness in instructing easily understood knowledge within a limited time. Thus, it is recommended to use the embodied learning method to instruct difficult knowledge or spatial knowledge with sufficient time for reflection in VR STEM education.

It is expected that the results enable policymakers to more firmly promote the use of VR in STEM education. It also has the potential to help researchers evaluate the possible influence of novelty on their studies. Instructional designers may use the findings to guide their VR STEM courses. Independent software vendors are hoped to be inspired for developing cost-effective VR learning applications with the guidance from researchers.

The future research directions can be a classroom-based longitudinal VR learning study, a more detailed longitudinal evaluation of embodied learning in a VR environment, and a broadly exploration of novelty effects in the human society.

REFERENCES

Agostinho, S., Tindall-ford, S., Ginns, P., Howard, S. J., Leahy, W., & Paas, F. (2015). Giving learning a helping hand: Finger tracing of temperature graphs on an iPad. *Educational Psychology Review*, *27*(3), 427–443. https://doi.org/10.1007/s10648-015-9315-5

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

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103*(3), 411–423.

Angulo, A., & de Velasco, G. V. (2014). Immersive simulation of architectural spatial experiences. In *Proceedings of the XVII Conference of the Iberoamerican Society of Digital Graphics (SIGraDi)* (pp. 495–499). Valparaiso: Blucher.

Animesh, A., Pinsonneault, A., Yang, S., Oh, W., Animesh, A., & Pinsonneauit, A. (2011). An odyssey into virtual worlds: Exploring the impacts of technological and spatial environments on intention to purchase virtual products. *Mis Quarterly*, *35*(3), 789–810.

Aukstakalnis, S., & Blatner, D. (1992). *Silicon mirage: The art and science of virtual reality*. Peach Pit Press.

Baenninger, M., & Newcombe, N. (1989). The role of experience in spatial test performance: A meta-analysis. *Sex Roles*, *20*(5–6), 327–344.

Bell, J. T., & Fogler, H. S. (1995). The investigation and application of virtual reality as an educational tool. In *the Proceedings of the American Society for Engineering Education 1995 Annual Conference*.

Bentler, P. M. (2006). *EQS 6 structural equations program manual*. Encino, CA: Multivariate Software, Inc.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588–606.

Berlyne, D. E. (1970). Novelty, complexity, and hedonic value. *Perception & Psychophysics*, *8*(5), 279–286. https://doi.org/10.3758/BF03212593

Berlyne, D. E., Craw, M. A., Salapatek, P. H., & Lewis, J. L. (1963). Novelty, complexity, incongruity, extrinsic motivation, and the GSR. *Journal of Experimental Psychology*, *66*(6), 560–567.

Berlyne, D. E., & Parham, L. C. C. (1968). The dimensionality of visual complexity, interestingness, and pleasingness. *Canadian Journal of Psychology*, *176*(5), 376–387.

Bharathi, A. K. B. G., & Tucker, C. S. (2015). Investigating the impact of interactive immersive virtual reality environments in enhancing task performance in online engineering design activities. In *Proceedings of the ASME 2015 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference* (pp. 1–11). Boston, MA.

Bianchi-Berthouze, N. (2013). Understanding the role of body movement in player engagement. *Human–Computer Interaction*, *28*(1), 40–75. https://doi.org/10.1080/07370024.2012.688468

Blair, L. (2016). Understanding the differences between virtual reality, augmented reality and mixed reality. *Network World*, 1–4.

Bloom, B. S., Mesia, B., & Krathwohl, D. (1964). *Bloom's taxonomy*. New York, NY: David McKay Publications. Retrieved from http://www.csus.edu/indiv/e/estiokom/bloomstaxonomy.pdf

Bourdin, P., Barberia, I., Oliva, R., & Slater, M. (2017). A virtual out-of-body experience reduces fear of death. *PloS One*, *12*(1), 1–19. https://doi.org/10.1371/journal.pone.0169343

Bouvier, P., Lavoué, E., & Sehaba, K. (2014). Defining engagement and characterizing in digital gaming. *Simulation & Gaming*, *45*(4–5), 491– 507. https://doi.org/10.1177/1046878114553571

Boyle, E. A., Connolly, T. M., Hainey, T., & Boyle, J. M. (2012). Engagement in digital entertainment games: A systematic review. *Computers in Human Behavior*, *28*(3), 771–780. https://doi.org/10.1016/j.chb.2011.11.020

Brand, J. P. L. (1999). *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets*. Erasmus University.

Brondi, R., Avveduto, G., & Carrozzino, M. (2015). Immersive technologies and natural interaction to improve serious games engagement. In *International Conference on Games and Learning Alliance*. https://doi.org/10.1007/978-3-319-40216-1

Buuren, S. van. (2014). Fully conditional specification. In G. Molenberghs, G. Fitzmaurice, M. G. Kenward, A. Tsiatis, & G. Verbeke (Eds.), *Handbook of missing data methodology* (pp. 267–290). Chapman and Hall/CRC.

Bystrom, K. E., Barfield, W., & Hendrix, C. (1999). A conceptual model of the sense of presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, *8*(2), 241–244.

Callum, K. Mac, & Jeffrey, L. (2013). The influence of students' ICT skills and their adoption of mobile learning. *Australasian Journal of Educational Technology*, *29*(3), 303–314. https://doi.org/10.1234/ajet.v29i3.298

Carlin, J. B. (2014). Multiple imputation: A perspective and historical overview. In G. Molenberghs, G. Fitzmaurice, M. G. Kenward, A. Tsiatis, & G. Verbeke (Eds.), *Handbook of missing data methodology* (pp. 239–263). Chapman and Hall/CRC.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245–276.

Cheng, L.-K., Chieng, M.-H., & Chieng, W.-H. (2014). Measuring virtual experience in a three-dimensional virtual reality interactive simulator environment: A structural equation modeling approach. *Virtual Reality*, *18*(3), 173–188. https://doi.org/10.1007/s10055-014-0244-2

Cheung, C. & Lee, M. (2009). Understanding the sustainability of a virtual community: model development and empirical test. *Journal of Information Science*, *35*(3), 279–298. https://doi.org/10.1177/0165551508099088

Cierniak, G., Scheiter, K., & Gerjets, P. (2009). Explaining the split-attention effect: Is the reduction of extraneous cognitive load accompanied by an increase in germane cognitive load ? *Computers in Human Behavior*, *25*(2), 315–324. https://doi.org/10.1016/j.chb.2008.12.020

Clark, R. E. (1983). Reconsidering research on learning from media. *Review of Educational Research*, *53*(4), 445–459.

Clark, R. E. (1994). Media will never influence learning. *Educational Technology Research and Development*, *42*(2), 21–29.

Clark, R. E., & Craig, T. G. (1992). Research and theory on multi-media learning effects. *Interactive Multimedia Learning Environments*, 19–30. https://doi.org/10.1007/978-3-642-77705-9

Committee of STEM Education. (2018). *Charting a course for success: America's strategy for STEM education*. National Science & Technology Council. Retrieved from http://www.whitehouse.gov/ostp.

Cordova, D. I., & Lepper, M. R. (1996). Intrinsic motivation and the process of learning: Beneficial effects of contextualization , personalization , and choice. *Journal of Educational Psychology*, *88*(4), 715–730.

Creswell, J. W. (2016). *30 essential skills for the qualitative researcher*. Thousand Oaks, CA: SAGE Publications Ltd.

Creswell, J. W., & Clark, V. L. P. (2011). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oak: CA: Sage Pubications.

Csikszentmihalyi, M. (1990). Flow: The psychology of optimal experience. In *Flow and the foundations of positive psychology* (pp. 209–226). Dordrecht: Springer.

Csikszentmihalyi, M. (1997). *Finding flow: The psychology of engagement in everyday life*. New York, NY: Basic Books.

Csikszentmihalyi, M. (1999). If we are so rich, why aren't we happy? *American Psychologist*, *54*(10), 821–827. https://doi.org/10.1037/0003-066x.54.10.821

Csikszentmihalyi, M., & Lefevre, J. (1989). Optimal experience in work and leisure. *Journal of Personality and Social Psychology*, *56*(5), 815–822.

Dalgarno, B., & Lee, M. J. W. (2010). What are the learning affordances of 3-D virtual environments? *British Journal of Educational Technology*, *41*(1), 10–33. https://doi.org/10.1111/j.1467-8535.2009.01038.x

Deci, E. L., & Ryan, R. M. (1994). Intrinsic Motivation Inventory (IMI). Retrieved February 11, 2019, from http://selfdeterminationtheory.org/intrinsic-motivation-inventory/

Desanctis, G., Poole, M. S., Science, S. O., May, N., Desanctis, G., & Scott, M. (1994). Capturing the complexity in advanced technology use: Adaptive structuration theory. *Organization Science*, *5*(2), 121–147.

Dinis, F. M., Guimaraes, A. S., Carvalho, B. R., & Martins, J. P. P. (2017a). Development of virtual reality game-based interfaces for civil engineering education. In *IEEE Global Engineering Education (EDUCON)* (pp. 1195–1202). Athens, Greece: IEEE.

Dinis, F. M., Guimaraes, A. S., Carvalho, B. R., & Martins, J. P. P. (2017b). Virtual and augmented reality game-based applications to civil engineering education. In *2017*

IEEE Global Engineering Education Conference (EDUCON) (pp. 1683–1688). https://doi.org/10.1109/EDUCON.2017.7943075

Drew, C. (2011). Why science majors change their minds (It's just so darn hard). Retrieved February 5, 2019, from https://www.nytimes.com/2011/11/06/education/edlife/why-science-majors-change-their-mind-its-just-so-darn-hard.html?pagewanted=all

Dugger, W. E. (2010). Evolution of STEM in the United States. In *6Th Biennial International Conference on Technology Education Research* (pp. 1–8). https://doi.org/10.1.1.476.5804

Enders, C. K. (2010a). An introduction to maximum likelihood estimation. In *Applied missing data analysis* (pp. 56–85). Guilford press.

Enders, C. K. (2010b). Analysis and pooling phases of multiple imputation. In *Applied missing data analysis*. Guilford press.

Enders, C. K. (2010c). Practical issues in multiple imputation. In *Applied missing data analysis* (pp. 254–286). New York, NY: Guilford press.

Enders, C. K. (2010d). The imputation phase of multiple imputation. In *Applied missing data analysis*. Guilford press.

Engelmann, J. B., & Pessoa, L. (2007). Motivation sharpens exogenous spatial attention. *Emotion*, *7*(3), 668–674. https://doi.org/10.1037/1528-3542.7.3.668

European Schoolnet. (2017). *ICT in STEM Education - Impacts and Challenges: On Students. A STEM Alliance Literature Review*. Brussels, Belgium. Retrieved from http://www.stemalliance.eu/ict-paper-4-on-schools

Evans, J. D. (1996). *Straightforward statistics for the behavioral sciences*. Thomson Brooks/Cole Publishing Co.

Finney, S. J., & Distefano, C. (2006). Non-normal and categorial data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 269–314). Greenwich, CT: IAP.

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

Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, motivation, and learning: A research and practice model. *Simulation & Gaming*, *33*(4), 441–467. https://doi.org/10.1177/1046878102238607

Gibson, J. J. (2014). *The ecological approach to visual perception: classic edition*. Psychology Press.

Glenberg, A. M. (1997). What memory is for. *Behavioral and Brain Sciences*, *20*, 1–55.

Goeldi, A. (2018). Is the current generation of VR already dead? Retrieved May 4, 2019, from https://innospective.net/is-the-current-generation-of-vr-already-dead/

Goldin-Meadow, S. (2011). Learning through gesture. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(6), 595–607. https://doi.org/10.1002/wcs.132

Goldin-meadow, S., Nusbaum, H., Kelly, S. D., & Wagner, S. (2001). Explaining math: Gesturing lightens the load. *Psychological Science*, *12*(6), 516–522.

Gonzalez-Franco, M., & Peck, T. C. (2018). Avatar embodiment: Towards a standardized questionnaire. *Frontiers in Robotics and AI*, *5*(June), 1–9. https://doi.org/10.3389/frobt.2018.00074

Greeno, J. G. (1994). Gibson's affordances. *Psychological Review*, *101*(2), 336–342. https://doi.org/10.1037/0033-295X.101.2.336

Grimm, K. J., Ram, N., & Estabrook, R. (2016). *Growth modeling: Structural equation and multilevel modeling approaches*. Guilford Publications.

Haag, S., Cummings, M., & Dawkins, J. (1998). *Management information systems for the information age*. New York: Irwin/McGraw Hill.

Hair, J. F., Black, W. C., & Babin, B. J. (2009). Confirmatory factor analysis. In *Multivariate data analysis* (pp. 661–699). Upper Saddle River, New Jersey: Prentice Hall, Inc.

Hamrol, A., Górski, F., Grajewski, D., & Zawadzki, P. (2013). Virtual 3D atlas of a human body - Development of an educational medical software application. *Procedia Computer Science*, *25*, 302–314. https://doi.org/10.1016/j.procs.2013.11.036

Hayduk, L. A., & Littvay, L. (2012). Should researchers use single indicators, best indicators, or multiple indicators in structural equation models? *BMC Medical Research Methodology*, *12*, 159.

Held, R. M., & Durlach, N. I. (1992). Telepresence. *Teleoperators and Virtual Environments*, *1*, 109–112.

Hendrix, C., & Barfield, W. (1996). The sense of presence within auditory virtual environments. *Presence: Teleoperators and Virtual Environments*, *5*(3), 290–301.

Hermanowicz, J. C. (2013). The longitudinal qualitative interview. *Qualitative Sociology*, *36*(2), 189–208.

Higashi, R. M., Schunn, C. D., & Flot, J. B. (2017). Different underlying motivations and abilities predict student versus teacher persistence in an online course. *Educational Technology Research and Development*, *65*(6), 1471–1493. https://doi.org/10.1007/s11423-017-9528-z

Hoffman, B., & Nadelson, L. (2010). Motivational engagement and video gaming: A mixed methods study. *Educational Technology Research and Development*, *58*(3), 245–270. https://doi.org/10.1007/s11423-009-9134-9

Hoffman, D. L., & Novak, T. P. (2009). Flow online: Lessons learned and future prospects. *Journal of Interactive Marketing*, *23*(1), 23–34. https://doi.org/10.1016/j.intmar.2008.10.003

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179–185.

Hou, H., & Li, M. (2014). Evaluating multiple aspects of a digital educational problem-solving-based adventure game. *Computers in Human Behavior*, *30*, 29–38. https://doi.org/10.1016/j.chb.2013.07.052

Hsu, C. Y., Tsai, C. C., & Wang, H. Y. (2012). Facilitating third graders' acquisition of scientific concepts through digital game-based learning: The effects of self-explanation principles. *The Asia-Pacific Education Researcher*, *21*(1), 71–82.

Hu, F., Ginns, P., & Bobis, J. (2015). Getting the point: Tracing worked examples enhances learning. *Learning and Instruction*, *35*, 85–93. https://doi.org/10.1016/j.learninstruc.2014.10.002

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 37–41. https://doi.org/10.1080/10705519909540118

Huang, M. (2003). Designing website attributes to induce experiential encounters. *Computers in Human Behavior*, *19*(4), 425–442.

Huang, W. & Roscoe, R. (2019). Head mounted display-based virtual reality in engineering education: A review of recent research. Unpublished manuscript.

Hunt, R. J., & Dentistry, C. (1986). Percent agreement, Pearson's correlation, and kappa as measures of inter-examiner reliability. *Journal of Dental Research*, *65*(2), 128–130.

Hussein, M., & Nätterdal, C. (2015). *The benefits of virtual reality in education: A comparison study*. University of Gothenburg.

Jackson, S., & Eklund, B. (2004). *The Flow Scales Manual*. Fitness Information Technology.

Jang, S., Vitale, J. M., Jyung, R. W., & Black, J. B. (2017). Direct manipulation is better than passive viewing for learning anatomy in a three-dimensional virtual reality environment. *Computers & Education*, *106*, 150–165. https://doi.org/10.1016/j.compedu.2016.12.009

Jeno, L. M., Vandvik, V., Eliassen, S., & Grytnes, J. A. (2019). Testing the novelty effect of an m-learning tool on internalization and achievement: A self-determination theory approach. *Computers and Education*, *128*(October 2018), 398–413. https://doi.org/10.1016/j.compedu.2018.10.008

Jensen, L., & Konradsen, F. (2017). A review of the use of virtual reality head-mounted displays in education and training. *Education and Information Technologies*, *23*(4), 1515–1529. https://doi.org/10.1007/s10639-017-9676-0

Johnson-Glenberg, M. C. (2018). Immersive VR and education: Embodied design principles that include gesture and hand controls. *Frontiers in Robotics and AI*, *5*(81), 1–19. https://doi.org/10.3389/frobt.2018.00081

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*(24). https://doi.org/10.1186/s41235-017-0060-9

Jonassen, D. H., Campbell, J. P., Davidson, M. E., David, H., Campbell, J. P., & Davidson, M. E. (1994). Learning with media: Restructuring the debate. *Educational Technology Research and Development*, *42*(2), 31–39.

Joo, S., Ahn, G., Bailenson, J. N., & Park, D. (2014). Short- and long-term effects of embodied experiences in immersive virtual environments on environmental locus of control and behavior. *Computers in Human Behavior*, *39*, 235–245. https://doi.org/10.1016/j.chb.2014.07.025

Jung, Y., & Lee, J. (2018). Learning engagement and persistence in massive open online courses (MOOCS). *Computers and Education*, *122*, 9–22. https://doi.org/10.1016/j.compedu.2018.02.013

Kahn, W. A. (2013). Psychological conditions of personal engagement and disengagement at work. *Academy of Management Journal*, *33*(4), 692–724.

Kail, R. (1986). The impact of extended parctice on rate of mental rotation. *Journal of Experimental Child Psychology*, *42*(3), 378–391.

Keller, J. M. (2010). Generating and sustaining attention forethought. In *Motivational design for learning and performance: The ARCS model approach* (pp. 75–96). Boston: MA: Springer US. https://doi.org/10.1007/978-1-4419-1250-3

Keller, J., Ringelhan, S., & Blomann, F. (2011). Does skills-demands compatibility result in intrinsic motivation? Experimental test of a basic notion proposed in the theory of flow-experiences. *Journal of Positive Psychology*, *6*(5), 408–417. https://doi.org/10.1080/17439760.2011.604041

Kesteren, M. T. R. Van, Ruiter, D. J., & Henson, R. N. (2012). How schema and novelty augment memory formation. *Trends in Neurosciences*, *35*(4), 211–219. https://doi.org/10.1016/j.tins.2012.02.001

Kilteni, K., Groten, R., & Slater, M. (2012). The sense of embodiment in virtual reality. *Presence: Teleoperators and Virtual Environments*, *21*(4), 103–105. https://doi.org/10.1162/PRES_a_00124

King, A. (1994). Guiding knowledge construction in the classroom: Effects of teaching children how to question and how to explain. *American Educational Research Journal*, *31*(2), 338–368. https://doi.org/10.3102/00028312031002338

Kline, R. B. (2011). Identification. In *Principles and practice of Structural Equation Modeling* (3rd ed., pp. 124–153). New York, NY: Guilford Press.

Koch, M., Luck, K. Von, Schwarzer, J., & Draheim, S. (2018). The novelty effect in large display deployments – Experiences and lessons-Learned for evaluating prototypes. *IEEE Access*, *6*(99), 3140–3148. https://doi.org/10.18420/ecscw2018

Kokkinara, E., Kilteni, K., Blom, K. J., & Slater, M. (2016). First person perspective of seated participants over a walking virtual body leads to illusory agency over the walking. *Scientific Reports*, *6*, 1–11. https://doi.org/10.1038/srep28879

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

Kozma, R. (1994). Will media influence learning? Reframing the debate. *Educational Technology Research and Development*, *42*(2), 7–19.

Krathwohl, D. R., & Anderson, L. W. (2009). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.

Krendl, K. A., & Broihier, M. (1992). Student responses to computers: A longitudinal study. *Journal of Educational Computing Research*, *8*(2), 215–227.

Lee, E. A.-L., Wong, K. W., & Fung, C. C. (2010). How does desktop virtual reality enhance learning outcomes? A structural equation modeling approach. *Computers & Education*, *55*(4), 1424–1442. https://doi.org/10.1016/j.compedu.2010.06.006

Lepper, M. R., & Iyengar, S. S. (2005). Intrinsic and extrinsic motivational orientations in the classroom: Age differences and academic correlates. *Journal of Educational Psychology*, *97*(2), 184–196. https://doi.org/10.1037/0022-0663.97.2.184

Li, Q., & Ma, X. (2010). A meta-analysis of the effects of computer technology on school students' mathematics learning. *Educational Psychology Review*, *22*(3), 215–243. https://doi.org/10.1007/s10648-010-9125-8

Lindgren, R., & Johnson-Glenberg, M. (2013). Emboldened by embodiment: Six precepts for research on embodied learning and mixed reality. *Educational Researcher*, *42*(8). https://doi.org/10.3102/0013189X13511661

Lindgren, R., Tscholl, M., Wang, S., & Johnson, E. (2014). Enhancing learning and engagement through embodied interaction within a mixed reality simulation. *Computers & Education*, *95*(March 2016), 174–187. https://doi.org/10.1016/j.compedu.2016.01.001

Linnenbrink, E. A., & Pintrich, P. R. (2002). Motivation as an enabler for academic success. *School Psychology Review*, *31*(3), 313–327.

Makransky, G., & Lilleholt, L. (2018). A structural equation modeling investigation of the emotional value of immersive virtual reality in education. *Educational Technology Research and Development*, 1–24.

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

Malone, T. W. (1981). Toward a theory of intrinsically motivating instruction. *Cognitive Science*, *5*(4), 333–369.

Mayer, R. E. (2003). The promise of multimedia learning: Using the same instructional design methods across different media. *Learning and Instruction*, *13*, 125–139.

McCoach, D. B., Gable, R. K., & Madura, J. P. (2013a). Evidence based on the internal structure of the instrument: Factor analysis. In B. McCoach, R. K. Gable, & J. P. Madura (Eds.), *Instrument development in the affective domain* (3rd ed., pp. 87–120). New York, NY: Springer Science & Business Media. https://doi.org/10.1007/978-1-4614-7135-6

McCoach, D. B., Gable, R. K., & Madura, J. P. (2013b). *Instrument development in the affective domain* (3rd ed.). New York, NY: Springer Science & Business Media. https://doi.org/10.1007/978-1-4614-7135-6

Mcmillan, K., Flood, K., & Glaeser, R. (2017). Virtual reality, augmented reality, mixed reality, and the marine conservation movement. *Aquatic Conservation: Marine and Freshwater Ecosystems*, *27*, 162–168. https://doi.org/10.1002/aqc.2820

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

Microsoft Corporation. (2016). ICT tools for STEM teaching and learning. Retrieved May 4, 2019, from http://www.stemalliance.eu/documents/99712/104016/STEM_A_and_MS_ICT_Tools_in_Edu_paper_v06_Final.pdf

Milgram, P., & Kishino, F. (1994). A taxonomy of mixed reality visual displays. *IEICE Transactions on Information Systems*, *E77–D*(12), 1–15.

Mineev, G. (2017). *The impact of immersive virtual reality on effectiveness of educational games*. Utrecht University.

Moreno, R. (2005). Instructional technology: Promise and pitfalls. In L. PytlikZillig, M. Bodvarsson, & R. Bruning (Eds.), *Technology-based education: Bringing researchers and practitioners together* (pp. 1–19). Greenwich, CT: Information Age Publishing.

Moreno, R., & Mayer, R. (2007). Interactive multimodal learning environments. *Educational Psychology Review*, *19*(3), 309–326. https://doi.org/10.1007/s10648-007-9047-2

Moreno, R., & Mayer, R. E. (2002). Learning science in virtual reality multimedia environments: Role of methods and media. *Journal of Educational Psychology*, *94*(3), 598–610. https://doi.org/10.1037//0022-0663.94.3.598

National Science Foundation. (2007). *National action plan for 21st century science, technology, engineering and math education*. *NSF Publications*. Arlington, VA. Retrieved from http://www.nsf.gov/nsb/documents/2007/stem_action.pdf

Nijs, L., Coussement, P., Moens, B., Amelinck, D., Lesaffre, M., & Leman, M. (2012). Interacting with the music paint machine: Relating the constructs of flow experience and presence. *Interacting With Computers*, *24*(4), 237–250. https://doi.org/10.1016/j.intcom.2012.05.002

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers*, *32*, 396–402.

Oberhauser, R., & Lecon, C. (2017). Gamified virtual reality for program code structure comprehension. *The International Journal of Virtual Reality*, *17*(2), 79–88.

Oberhauser, R., Silfang, C., & Lecon, C. (2016). Code structure visualization using 3D-flythrough. In *ICCSE 2016 - 11th International Conference on Computer Science and Education* (pp. 365–370). https://doi.org/10.1109/ICCSE.2016.7581608

Oculus. (2020). Oculus. Retrieved March 24, 2020, from https://www.oculus.com/

Onyesolu, M. O., & Eze, F. U. (2011). Understanding virtual reality technology: Advances and applications. In *Advances in Computer Science and Engineering*. InTech.

Oxford. (2019). Oxford Dictionaries. Retrieved May 31, 2019, from https://www.oxforddictionaries.com/

Paas, F. G., & Van Merriënboer, J. J. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational Psychology Review*, *6*(4), 351–371.

Paas, F., & Sweller, J. (2012). An evolutionary upgrade of cognitive load theory: Using the human motor system and collaboration to support the learning of complex cognitive tasks. *Educational Psychology Review*, *24*(1), 27–45. https://doi.org/10.1007/s10648-011-9179-2

Parker, L. E., & Lepper, M. R. (1992). Effects of fantasy context on children's learning and motivation: Making learning more fun. *Journal of Personality and Social Psychology*, *62*, 625–633.

Parmar, D. (2017). *Evaluating the effects of immersive embodied interaction on cognition in virtual reality*. Clemson University.

Pearson, P. H. (1998). Relationships between global and specified measures of novelty-seeking. *Journal of Consulting and Clinical Psychology*, *34*(2), 199–204.

Peters, M., Laeng, B., Latham, K., Jackson, M., Zaiyouna, R., & Richardson, C. (1995). A redrawn Vandenberg and Kuse mental rotations test-different versions and factors that affect performance. *Brain and Cognition*, *28*(1), 39–58.

Pew Research Center. (2018). Women and men in STEM often at odds over workplace equity. Retrieved May 4, 2019, from http://www.pewsocialtrends.org/2018/01/09/blacks-in-stem-jobs-are-especially-concerned-about-diversity-and-discrimination-in-the-workplace/

Ping, R., & Goldin-meadow, S. (2010). Gesturing saves cognitive resources when talking about nonpresent objects. *Cognitive Science*, *34*(4), 602–619. https://doi.org/10.1111/j.1551-6709.2010.01102.x

Pintrich, P. R. (2003). Motivation and classroom learning. In W. M. Reynolds & G. E. Miller (Eds.), *Handbook of psychology: Educational psychology* (pp. 103–122). New York: Wiley.

Przybylski, A. K., Rigby, C. S., & Ryan, R. M. (2010). A motivational model of video game engagement. *Review of General Psychology*, *14*(2), 154–166. https://doi.org/10.1037/a0019440

Ragan, E. D., Bowman, D. A., Kopper, R., Stinson, C., Scerbo, S., & Mcmahan, R. P. (2015). Effects of field of view and visual complexity on virtual reality training effectiveness for a visual scanning task. *IEEE Transactions on Visualization and Computer Graphics*, *21*(7), 794–807.

Raghunathan, T. E., Lepkowski, J. M., Hoewyk, J. Van, & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, *27*(1), 85–95.

Robinson, L. J., Stevens, L. H., Threapleton, C. J., Vainiute, J., McAllister-Williams, R. H., & Gallagher, P. (2012). Effects of intrinsic and extrinsic motivation on attention and memory. *Acta Psychologica*, *141*(2), 243–249. https://doi.org/10.1016/j.actpsy.2012.05.012

Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). New York, NY: Free Press.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*(1), 54–67. https://doi.org/10.1006/ceps.1999.1020

Saldaña, J. (2002). Analyzing change in longitudinal qualitative data. *Youth Theatre Journal*, *16*(1), 1–17. https://doi.org/10.1080/08929092.2002.10012536

Salzman, M. C., Dede, C., Loftin, R. B., & Chen, J. (1999). A model for understanding how virtual reality aids complex conceptual learning. *Presence: Teleoperators and Virtual Environments*, *8*, 293–316. https://doi.org/10.1162/105474699566242

Sanchez-Vives, M. V, & Slater, M. (2005). From presence to consciousness through virtual reality. *Nature Reviews Neuroscience*, *6*(4), 332–339.

Sato, A., & Yasuda, A. (2005). Illusion of sense of self-agency: Discrepancy between the predicted and actual sensory consequences of actions modulates the sense of self-agency, but not the sense of self-ownership. *Cognition*, *94*(3), 241–255. https://doi.org/10.1016/j.cognition.2004.04.003

Schönborn, K. J., Bivall, P., & Tibell, L. A. E. (2011). Exploring relationships between students' interaction and learning with a haptic virtual biomolecular model. *Computers & Education*, *57*(3), 2095–2105. https://doi.org/10.1016/j.compedu.2011.05.013

Schüler, J. (2007). Arousal of flow experience in a learning setting and its effects on exam performance and affect. *Zeitschrift Für Pädagogische Psychologie*, *21*(3/4), 217–227. https://doi.org/10.1024/1010-0652.21.3.217

Schuster, K., Hoffmann, M., Bach, U., Richert, A., & Jeschke, S. (2014). Diving in? How users experience virtual environments using the virtual theatre. In *Proceedings of the 3rd International Conference on Design, User Experience, and Usability (DUXU 2014)* (pp. 636–646). Springer. https://doi.org/10.1007/978-3-319-07626-3_60

Shaffer, J. F., Dang, J. V., Lee, A. K., Dacanay, S. J., Alam, U., Wong, H. Y., … Sato, B. K. (2016). A Familiar(ity) problem: Assessing the impact of prerequisites and content familiarity on student learning. *PloS One*, *11*(1), 1–13. https://doi.org/10.1371/journal.pone.0148051

Sharek, D., & Wiebe, E. (2014). Measuring video game engagement through the cognitive and affective dimensions. *Simulation & Gaming*, *45*(4–5), 569– 592. https://doi.org/10.1177/1046878114554176

Shernoff, D. J., & Strati, A. D. (2011). Advances in engineering education measuring engagement as students learn dynamic systems and control with a video game department of mechanical engineering. *Advances in Engineering Education*, 1–32.

Shin, D. (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. https://doi.org/10.1016/j.tele.2017.05.013

Shu, Y., Huang, Y.-Z., Chang, S.-H., & Chen, M.-Y. (2018). Do virtual reality head-mounted displays make a difference? A comparison of presence and self‐efficacy between head‐mounted displays and desktop computer‐facilitated virtual environments. *Virtual Reality*, 1–10.

Simon, H. A. (1994). The bottleneck of attention: Connecting thought with motivation. *Integrative Views of Motivation, Cognition, and Emotion*, *41*, 1–21.

Skulmowski, A., & Rey, G. D. (2018). Embodied learning: Introducing a taxonomy based on bodily engagement and task integration. *Cognitive Research: Principles and Implications*, *3*(1), 6.

Slater, M. (2004). How colorful was your day? Why questionnaires cannot assess presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, *13*(4), 484–494.

Slater, M. (2009). Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*, 3549–3557. https://doi.org/10.1098/rstb.2009.0138

Slater, M., Khanna, P., Mortensen, J., & Yu, I. (2009). Visual realism enhances realistic response in an immersive virtual environment. *IEEE Computer Graphics and Applications*, *29*(3), 76–84.

Slater, M., McCarthy, J., & Maringelli, F. (1998). The influence of body movement on subjective presence in virtual environments. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *40*(3), 469–477.

Slater, M., Perez-marcos, D., Ehrsson, H. H., & Sanchez-Vives, M. V. (2008). Towards a digital body: The virtual arm illusion. *Frontiers in Human Neuroscience*, *2*(6), 1–8. https://doi.org/10.3389/neuro.09.006.2008

Slater, M., & Sanchez-Vives, M. V. (2016). Enhancing our lives with immersive virtual reality. *Frontiers in Robotics and AI*, *3*(December), 1–47. https://doi.org/10.3389/frobt.2016.00074

Slater, M., Spanlang, B., & Corominas, D. (2010). Simulating virtual environments within virtual environments as the basis for a psychophysics of presence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *29*(4), 1–9. https://doi.org/10.1145/1778765.1778829

Slater, M., Usoh, M., & Chrysanthou, Y. (1995). The influence of dynamic shadows on presence in immersive virtual environments. In *Virtual Environments' 95* (pp. 8–21). Vienna, Austria: Springer. https://doi.org/10.1007/978-3-7091-9433-1

Slater, M., & Wilbur, S. (1997). A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, *6*(6), 603–616. https://doi.org/10.1007/s10750-008-9541-7

Soper, D. S. (2020). A-priori sample size calculator for structural equation models. Retrieved from http://www.danielsoper.com/statcalc

Superdata. (2018). Virtual reality market and consumers. Retrieved May 14, 2018, from https://www.superdataresearch.com/market-data/virtual-reality-industry-report/

Superdata. (2019). Oculus expected to sell 1.3M Quest units in 2019; XR revenue reached $6.6B in 2018 and is projected to increase 442% by 2022. Retrieved May 4, 2019, from https://www.superdataresearch.com/blog/

Svinicki, M. D. (2004). Motivating students to learn. In *Learning and motivation in the postsecondary classroom* (pp. 141–176). San Francisco, CA: Jossey-Bass.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational Psychology Review*, *22*(2), 123–138. https://doi.org/10.1007/s10648-010-9128-5

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

Takatalo, J. (2002). *Presence and flow in virtual environments: An explorative study*. University of Helsinki.

Timcenko, O., Kofoed, L. B., Schoenau-fog, H., & Reng, L. (2017). Purposive game production in educational setup investigating team collaboration in virtual reality. In *International Conference on Human-Computer Interaction* (Vol. 714, pp. 184–191). Cham: Springer. https://doi.org/10.1007/978-3-319-58753-0

Tindall-ford, S., Chandler, P., Sweller, J., & Bartlett, S. (1997). When two sensory modes are better than one. *Journal of Experimental Psychology: Applied*, *3*(4), 257–287.

Tomporowski, P. D., & Tinsley, V. F. (1996). Effects of memory demand andmotivation on sustained attention in young and older adults. *The American Journal of Psychology*, *109*(2), 187–204.

Tsaya, C. H.-H., & Kofinas, A. (2017). Student engagement and the novelty effect in a technology-mediated gamified course. In *Society for Research Into Higher Education* (pp. 1–8).

Tüzün, H., Yilmaz-Soylu, M., Karakuş, T., Inal, Y., & Kizilkaya, G. (2009). The effects of computer games on primary school students' achievement and motivation in geography learning. *Computers and Education*, *52*(1), 68–77. https://doi.org/10.1016/j.compedu.2008.06.008

Van Buuren, S. (2018). *Flexible imputation of missing data*. Chapman and Hall/CRC.

van Buuren, S., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, *18*, 681–694.

Van Buuren, S., & Groothuis-Oudshoorn, C. G. M. (1999). *Flexible multivariate imputation by MICE* (No. Technical Report PG/VGZ/99.054). Leiden.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*, 1–67.

VR Intelligence, & Superdata. (2018). XR industry survey 2018. Retrieved May 4, 2019, from https://img03.en25.com/Web/FCBusinessIntelligenceLtd/%7Bbf739f98-151f-4691-bcd1-d7ef276df5a3%7D_4947_VRX_Survey_PDF_07AUG18.pdf?elqTrackId=A61CEB3D05BE6D8389D1ACDF17DBCCB2&elqaid=21568&elqat=2

Wade, S., & Kidd, C. (2019). The role of prior knowledge and curiosity in learning. *Psychonomic Bulletin & Review*, 1–11.

Waller, D., & Knapp, D. (1998). The transfer of spatial knowledge in virtual environment training. *Presence*, *7*(2), 129–143.

Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using Mplus*. John Wiley & Sons Inc.

Webster, R., & Dues Jr., J. F. (2017). System usability scale (SUS): Oculus Rift DK2 and Samsung Gear VR. In *ASEE Annual Conference & Exposition*.

Weibel, D., & Wissmath, B. (2011). Immersion in computer games: The role of spatial presence and flow. *Journal of Computer Games Technology*, *2011*, 1–6. https://doi.org/10.1155/2011/282345

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and applications* (pp. 56–75). Newbury Park, CA: Sage Publication.

Westland, J. C. (2010). Lower bounds on sample size in structural equation modeling. *Electronic Commerce Research and Applications*, *9*(6), 476–487. https://doi.org/10.1016/j.elerap.2010.07.003

White, I. R., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Tutorial in Biostatistics*, *30*(4), 377–399. https://doi.org/10.1002/sim.4067

Whitton, N. (2011). Game engagement theory and adult learning. *Simulation & Gaming*, *42*(5), 596–609. https://doi.org/10.1177/1046878110378587

Wicken, C. D. (1992). Introduction to engineering psychology and human performance. In C. D. Wicken (Ed.), *Engineering psychology and human performance* (pp. 1–23).

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*(4), 625–636.

Wirth, W., Hartmann, T., Böcking, S., Vorderer, P., Klimmt, C., Schramm, H., & Biocca, F. (2007). A process model of the formation of spatial presence experiences. *Media Psychology*, *9*(3), 493–525. https://doi.org/10.1080/15213260701283079

Witmer, B. G., & Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence*, *7*(3), 225–241.

Zaman, M., Ananda, M., & Dai, Q. (2010). Experiencing flow with instant messaging and its facilitating role on creative behaviors. *Computers in Human Behavior*, *26*(5), 1009–1018. https://doi.org/10.1016/j.chb.2010.03.001

Zestcott, C. A. (2017). *The role of conscious attention in embodiment: Initial evidence of a dual process model of embodied cognition*. The university of arizona.

APPENDIX A

IRB APPROVAL LETTER

ASU Knowledge Enterprise Development

APPROVAL: EXPEDITED REVIEW

Rod Roscoe
IAFSE-PS: Human Systems Engineering (HSE)
480/727-2760
Rod.Roscoe@asu.edu

Dear Rod Roscoe:

On 7/24/2019 the ASU IRB reviewed the following protocol:

Type of Review:	Initial Study
Title:	Investigating the Novelty Effect in Virtual Reality (VR) on STEM Learning
Investigator:	Rod Roscoe
IRB ID:	STUDY00010384
Category of review:	(6) Voice, video, digital, or image recordings, (7)(b) Social science methods
Funding:	None
Grant Title:	None
Grant ID:	None
Documents Reviewed:	• consent form_new.pdf, Category: Consent Form; • all questions.pdf, Category: Measures (Survey questions/Interview questions /interview guides/focus group questions); • recruitment information_new.pdf, Category: Recruitment Materials; • IRB Protocol (3).docx, Category: IRB Protocol; • The VR application user manual.pdf, Category: Participant materials (specific directions for them);

The IRB approved the protocol from 7/24/2019 to 7/23/2025 inclusive. Three weeks before 7/23/2025 you are to submit a completed Continuing Review application and required attachments to request continuing approval or closure.

If continuing review approval is not granted before the expiration date of 7/23/2025 approval of this protocol expires on that date. When consent is appropriate, you must use final, watermarked versions available under the "Documents" tab in ERA-IRB.

In conducting this protocol you are required to follow the requirements listed in the INVESTIGATOR MANUAL (HRP-103).

Sincerely,

IRB Administrator

cc: wen huang
wen huang
Rod Roscoe

APPENDIX B

QUESTIONNAIRE ITEMS

Table B1

Questionnaire Items

Item	
Pr_1	Please rate your sense of being in the universe. I had a sense of being in the universe ..... (from “At no time” to “Almost all the time”)
Pr_2	To what extent were there times during the experience when the universe was the “reality” for you? There were times during the experience when the universe was the reality for me ..... (from “At no time” to “Almost all the time”)
Pr_3	Do you think of the universe more as images that you saw, or more as somewhere that you visited? The universe seemed to be more like ..... (from “Images that I saw” to “Somewhere I visited”)
Pr_4	During the experience, did you often think to yourself that you were just in a room, or did the universe overwhelm you? During the experience I was thinking that I was really in the room .... (from “Most of the time” to “Rarely”)
Em_1	I felt as if the virtual body or body part was my own body or body part. (from “Strongly disagree” to “Strongly agree”)
Em_2	It felt as if the virtual body or body part I saw was someone else's. (from “Strongly disagree” to “Strongly agree”)
Em_3	I felt as if I might have more than one body (from “Strongly disagree” to “Strongly agree”)
Em_4	It felt like I could control the virtual body or body part as if it was my own. (from “Strongly disagree” to “Strongly agree”)
Em_5	I felt as if the movements of the virtual body were influencing my own movements. (from “Strongly disagree” to “Strongly agree”)
Em_6	I felt as if the virtual body was moving by itself. (from “Strongly disagree” to “Strongly agree”)
Em_7	When a virtual celestial body was in front of me, I felt that my body was very small. (from “Strongly disagree” to “Strongly agree”)
Mo_1	I enjoyed doing this activity very much. (from "Not at all true" to "Very true")
Mo_2	This activity did not hold my attention at all. (from "Not at all true" to "Very true")
Mo_3	This activity is fun to do. (from "Not at all true" to "Very true")
Mo_4	I would describe this activity as very interesting. (from "Not at all true" to "Very true")

En_1	I felt that I was competent enough to meet the demands of the learning task. (from "Never" to "Always")
En_2	I had a strong sense of what I wanted to do. (from "Never" to "Always")
En_3	I had a good idea about how well I was doing while I was involved in the activity. (from "Never" to "Always")
En_4	I was completely focused on the learning task at hand. (from "Never" to "Always")
En_5	I had a feeling of total control over what I was doing. (from "Never" to "Always")
En_6	I was not worried about what others (i.e., physical watchers) might be thinking of me. (from "Never" to "Always")
En_7	The way time passed seemed to be different from normal. (from "Never" to "Always")
Cl_1	The learning content was difficult. (from "Not at all the case" to "Completely the case")
Cl_2	The learning content was complex. (from "Not at all the case" to "Completely the case")
Cl_3	It was difficult to learn in this VR environment. (from "Not at all the case" to "Completely the case")

APPENDIX C

QUIZ QUESTIONS

Questions in the Sample Quiz

Sample quiz

Question 1 of 7

Your assigned ID in this experiment: _____.

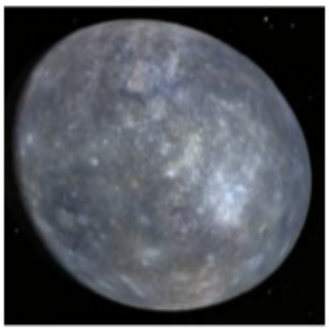
Show word count

Question 2 of 7

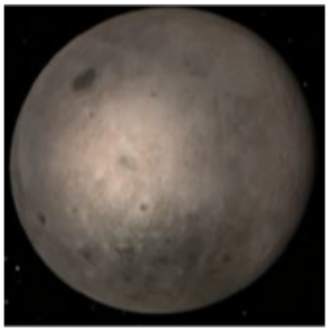
Please match these images of celestial bodies to their names:



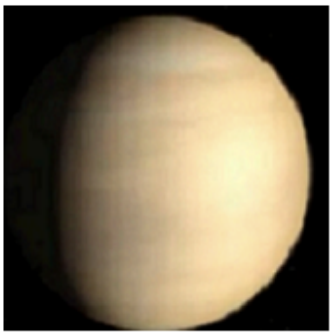
Select



Select



Select



Select

Question 3 of 7

Imagine that you are planning to build a settlement on a celestial body. Please complete the following questions (Question 3 to Question 7) **based on what you have learned in the VR tour.**

Can you name any characteristics about the geology, climate, or orbit of Mercury? (please give at least 3 ideas, they should be from the learning content)

Show word count

Question 4 of 7

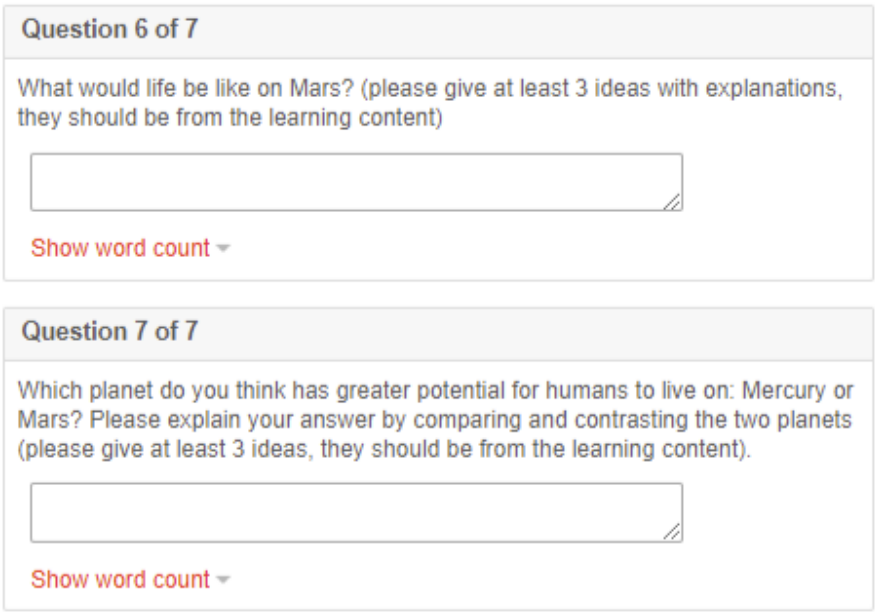
Can you name any characteristics about the geology, climate, or orbit of Mars? (please give at least 3 ideas, they should be from the learning content)

Show word count

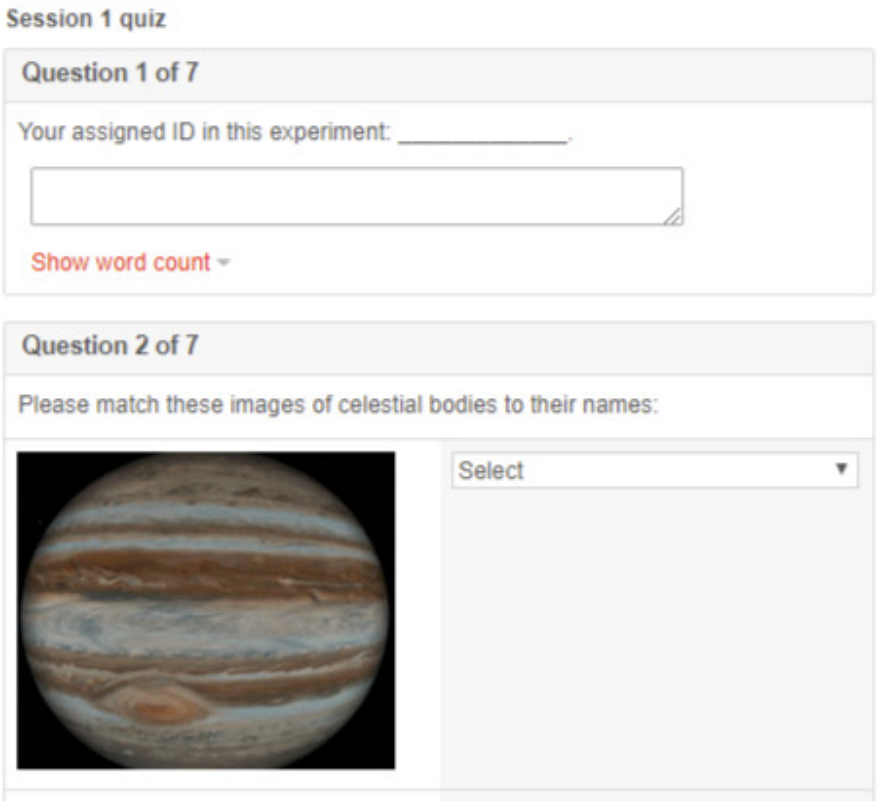
Question 5 of 7

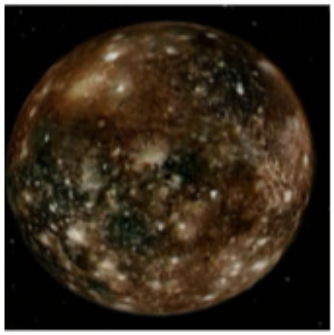
What would life be like on Mercury? (please give at least 3 ideas with explanations, they should be from the learning content)

Show word count

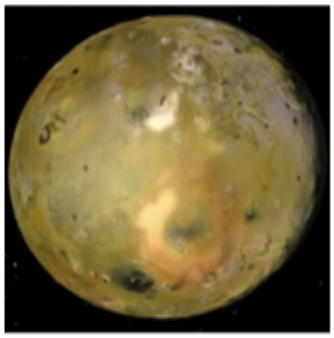


Questions in the Session 1 Quiz

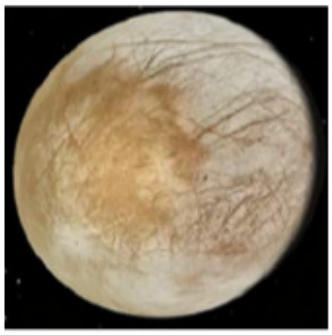




Select



Select



Select

Question 3 of 7

Imagine that you are planning to build a settlement on a celestial body. Please complete the following questions (Question 3 to Question 7) based on what you have learned in the VR tour.

Can you name any characteristics about the geology, climate, or orbit of Io? (please give at least 3 ideas, they should be from the learning content)

Show word count

Question 4 of 7

Can you name any characteristics about the geology, climate, or orbit of Europa? (please give at least 3 ideas, they should be from the learning content)

Show word count

Question 5 of 7

What would life be like on Io? (at least 3 ideas with explanations, they should be from the learning content)

Show word count

Question 6 of 7

What would life be like on Europa? (please give at least 3 ideas with explanations, they should be from the learning content)

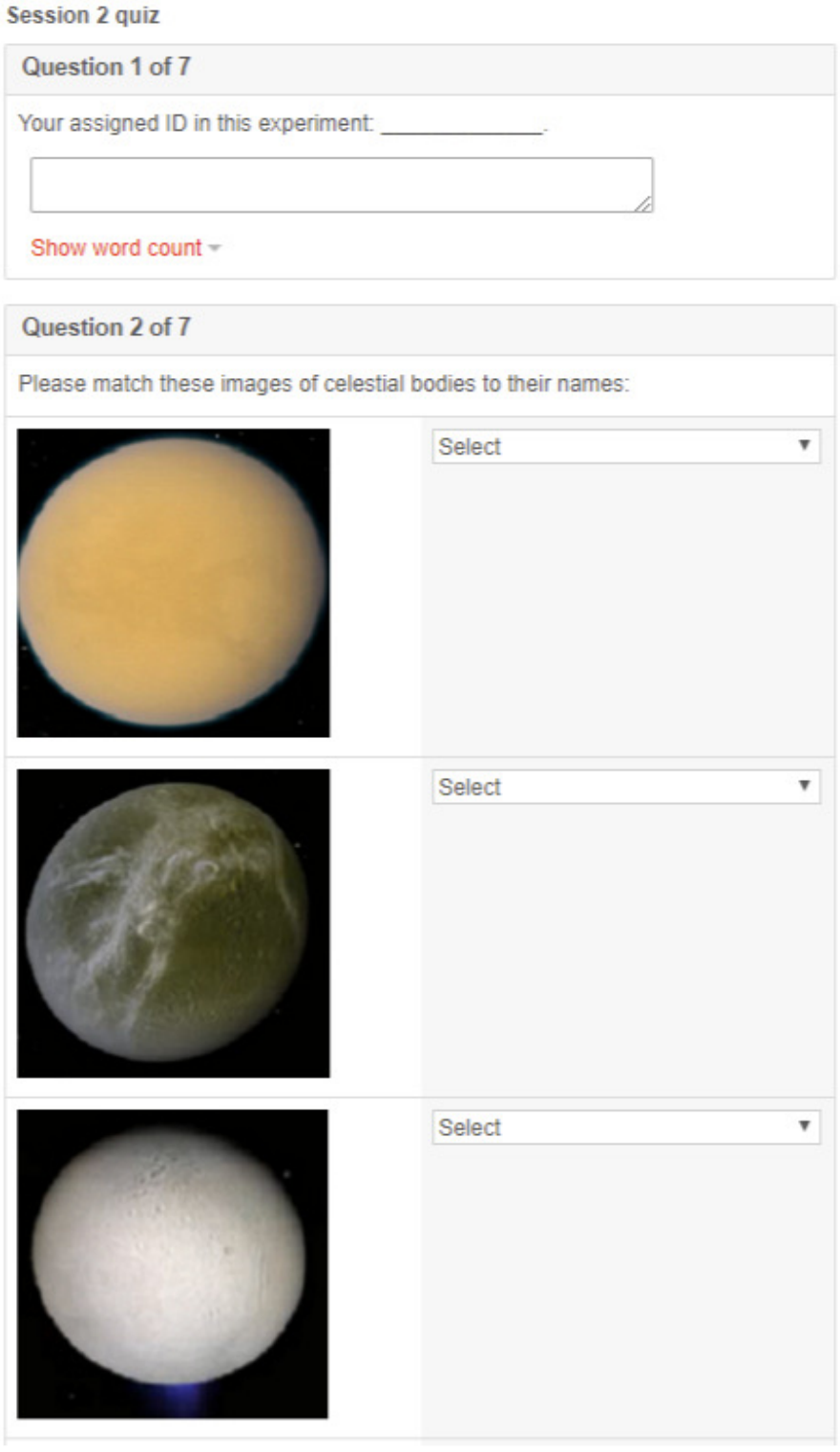
Show word count

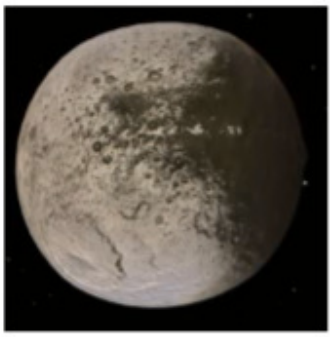
Question 7 of 7

Which celestial body do you think has greater potential for humans to live: Io or Europa? Please explain your answer by comparing and contrasting the two celestial bodies (please give at least 3 ideas, they should be from the learning content). You ideas will be useful in the later sessions.

Show word count

Questions in the Session 2 Quiz





Select

Question 3 of 7

Imagine that you are planning to build a settlement on a celestial body. Please complete the following questions (Question 3 to Question 7) based on **what you have learned in the VR tour**.

Can you name any characteristics about the geology, climate, or orbit characteristics of Enceladus? (please give at least 3 ideas, they should be from the learning content)

Show word count

Question 4 of 7

Can you name any characteristics about the geology, climate, or orbit characteristics of Lapetus? (please give at least 3 ideas, they should be from the learning content)

Show word count

Question 5 of 7

What would life be like on Enceladus? (please give at least 3 ideas with explanations, they should be from the learning content)

Show word count

Question 6 of 7

What would life be like on Lapetus? (please give at least 3 ideas with explanations, they should be from the learning content)

Show word count

Question 7 of 7

Which celestial body do you think has greater potential for humans to live: Enceladus or Lapetus? Please explain your answer by comparing and contrasting the two celestial bodies (please give at least 3 ideas).

Show word count

Questions in the Session 3 Quiz

Session 3 quiz

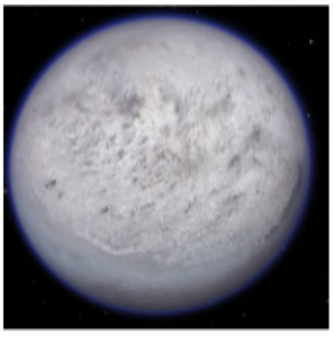
Question 1 of 7

Your assigned ID in this experiment: _____.

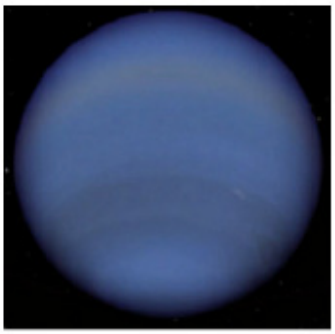
Show word count

Question 2 of 7

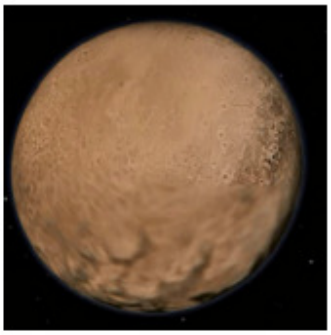
Please match these images of celestial bodies to their names:



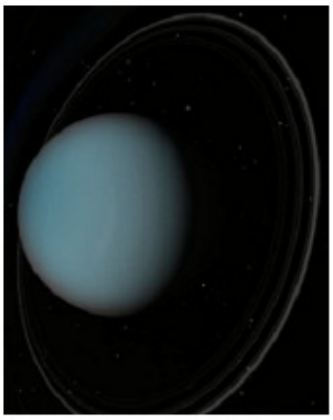
Select



Select



Select



Select

Question 3 of 7

Imagine that you are planning to build a settlement on a celestial body. Please complete the following questions (Question 3 to Question 7) based on **what you have learned in the VR tour**.

Can you name any characteristics about the geology, climate, or orbit characteristics of Triton? (please give at least 3 ideas)

Show word count

Question 4 of 7

Can you name any characteristics about the geology, climate, or orbit characteristics of Pluto? (please give at least 3 ideas, they should be from the learning content)

Show word count

Question 5 of 7

What would life be like on Triton? (please give at least 3 ideas with explanations, they should be from the learning content)

Show word count

Question 6 of 7

What would life be like on Pluto? (please give at least 3 ideas with explanations, they should be from the learning content)

Show word count

Question 7 of 7

Which celestial body do you think has greater potential for humans to live: Triton or Pluto? Please explain your answer by comparing and contrasting the two celestial bodies (please give at least 3 ideas, they should be from the learning content).

Show word count

APPENDIX D

EXPERIMENTAL INSTRUCTIONS

This document includes:

1. Go-Version Training Instruction
2. Go-Version Learning Instruction in the First Session
3. Go-Version Learning Instruction in the Second Session
4. Go-Version Learning Instruction in the Third Session
5. Rift-Version Training Instruction
6. Rift-Version Learning Instruction in the First Session
7. Rift-Version Learning Instruction in the Second Session
8. Rift-Version Learning Instruction in the Third Session

They were used at the beginning of each session to ensure that participants knew their training and learning tasks.

Go-Version Training Instruction

This game is a tour through the solar system, an immersive exploration of the celestial bodies. When you enter the VR environment, you will find that you are standing in a spaceship in the universe, and there is a dashboard in front of you.

The first celestial body you will visit is the earth. Please complete the training task by following these steps.

Step 1:

Click the “READY” button on the dashboard twice until the black empty place on the left side of the dashboard is filled by a “more information” button.

Step 2:

Click the “more information” button on the left side of the dashboard, and read the text information on dashboard by pages

Step 3:

Observe the surface characteristics of the earth.

Step 4. Click the “READY” button or the blue arrow button on the dashboard to fly to the next celestial body.

Note:

1. Please do not click other buttons on the dashboard in the virtual tour. They are not useful in this study.

2. *Autopilot* means that the spaceship is flying from one celestial body to the next. You must reach your destination to gain control again.

Go-Version Learning Instruction in the First Session

The celestial bodies you will experience in this session are:

Ceres, Jupiter, Callisto, Ganymede, Io, and Europa

When the spaceship stops in front of one (or several) celestial body(ies), take the same steps as you did in the training part.

1. Click the “more information” button and read text information by pages.
2. Observe the surface characteristics of the celestial body(ies).
3. Click the “READY” button or the blue arrow button on the dashboard to fly to the next celestial body.

When you reach Saturn, please inform the experimenter.

Go-Version Learning Instruction in the Second Session

The celestial bodies you will experience in this session are:

Saturn, Pan, Enceladus, Techys, Dione, Rhea, Titan, Iapetus

When the spaceship stops in front of one (or several) celestial body(ies), take the same steps as you did in the training part.

1. Click the “more information” button and read text information by pages.
2. Observe the surface characteristics of the celestial body(ies).
3. Click the “READY” button or the blue arrow button on the dashboard to fly to the next celestial body.

When you reach Chariklo, please inform the experimenter.

Go-Version Learning Instruction in the Third Session

The celestial bodies you will experience in this session are:

Chariklo, Uranus, Ariel, Umbriel, Titania, Oberon, Neptune, Triton, Pluto, and Charon

When the spaceship stops in front of one (or several) celestial body(ies), take the same steps as you did in the training part.

1. Click the “more information” button and read text information by pages.
2. Observe the surface characteristics of the celestial body(ies).
3. Click the “READY” button or the blue arrow button on the dashboard to fly to the next celestial body.

Pluto and Charon are in the last scenario. When you have learned Pluto and Charon, please inform the experimenter.

Rift-Version Training Instruction

This game is a tour through the solar system, an immersive exploration of the celestial bodies. When you enter the VR environment, you will find that you are standing in a spaceship in the universe, and there is a dashboard in front of you.

The first celestial body you will visit is the earth. Please complete the training task by following these steps.

Step 1:

Click the “READY” button on the dashboard twice until there are buttons filling in the five black empty places on the dashboard.

Step 2:

Click the “more information” button on the left side of the dashboard, and read the text information on dashboard by pages

Step 3:

Click the two buttons on the left side of the dashboard to rotate the earth clockwise or anticlockwise. Observe the surface characteristics of the earth.

Step 4:

Click the first button on the right side of the dashboard to change the map of the earth. Observe the surface characteristics of the earth.

Step 5:

Click the second button on the right side of the dashboard to zoom the earth in and out. Observe the surface characteristics of the earth.

Step 6. Hold the real trigger button labeled “HOLD for Close-Up” on the controller to rotate the earth freely. Observe the surface characteristics of the earth.

Step 7.

Click the third button on the right side of the dashboard (labeled “0g”) to switch the system to the “EVA” mode.

Step 8.

Fly toward the earth through controlling the direction with joysticks on the controllers. There are labels for the functions of other buttons on these two controllers, which help you control flying speed, such as thrust, boost, and brake.

When your virtual hands touch the surface of the earth, holding the rear trigger button (labeled “grab”) and grab the surface with your virtual hands.

If you want to fly back to the spaceship, click the button labeled “Go Back to Tour” on the left controller.

Step 9. When you are in the spaceship again, click the “READY” button or the blue arrow button on the dashboard to fly to the next celestial body.

Note:

1. Please do not click other buttons on the dashboard in the virtual tour. They are not useful in this study.

2. *Autopilot* means that the spaceship is flying from one celestial body to the next. You must reach your destination to gain control again.

Rift-Version Learning Instruction in the First Session

The celestial bodies you will experience in this session are:

Ceres, Jupiter, Callisto, Ganymede, Io, and Europa

When the spaceship stops in front of one (or several) celestial body(ies), take the same steps as you did in the training part.

1. Click the “more information” button and read text information by pages.
2. Click the four yellow-outline buttons on the two sides of the dashboard and observe the surface characteristics of the celestial bodies.
3. Hold the real trigger button labeled “HOLD for Close-Up”, rotate the celestial body and observe.
4. Click the “0g” button, fly towards the celestial body, grab its surface, and observe its characteristics.

When you reach Saturn, please inform the experimenter.

Rift-Version Learning Instruction in the Second Session

The celestial bodies you will experience in this session are:

Saturn, Pan, Enceladus, Techys, Dione, Rhea, Titan, Iapetus

When the spaceship stops in front of one (or several) celestial body(ies), take the same steps as you did in the training part.

1. Click the “more information” button and read text information by pages.
2. Click the four yellow-outline buttons on the two sides of the dashboard and observe the surface characteristics of the celestial bodies.
3. Hold the real trigger button labeled “HOLD for Close-Up”, rotate the celestial body and observe.
4. Click the “0g” button, fly towards the celestial body, grab its surface, and observe its characteristics.

When you reach Chariklo, please inform the experimenter.

Rift-Version Learning Instruction in the Third Session

The celestial bodies you will experience in this session are:

Chariklo, Uranus, Ariel, Umbriel, Titania, Oberon, Neptune, Triton, Pluto, and Charon

When the spaceship stops in front of one (or several) celestial body(ies), take the same steps as you did in the training part.

1. Click the “more information” button and read text information by pages.
2. Click the four yellow-outline buttons on the two sides of the dashboard and observe the surface characteristics of the celestial bodies.
3. Hold the real trigger button labeled “HOLD for Close-Up”, rotate the celestial body and observe.
4. Click the “0g” button, fly towards the celestial body, grab its surface, and observe its characteristics.

Pluto and Charon are in the last scenario. When you have learned Pluto and Charon, please inform the experimenter.

APPENDIX E

PREDICTOR MODEL

Predictors for Items Measuring Presence

Table D1 is the predictor matrix for presence. Its columns are the set of target variables that measured participants' presence across three sessions. For example, "Pr_1.2" means this column variable is participants' scores to item Pr_1 (Please rate your sense of being in the universe. I had a sense of being in the universe .....) in the second session. Each row represents one candidate predictor variable. A cell containing "1" means that the row variable was used to impute the column variable.

Table D1

Predictors for Items Measuring Presence

	Pr_1.1	Pr_1.2	Pr_1.3	Pr_2.1	Pr_2.2	Pr_2.3	Pr_3.1	Pr_3.2	Pr_3.3	Pr_4.1	Pr_4.2	Pr_4.3
Group	1	1	1	1	1	1	1	1	1	1	1	1
Gender	0	0	0	0	0	0	1	1	1	0	0	0
Age	1	1	1	1	1	1	1	1	1	1	1	1
Race	1	1	1	1	1	1	1	1	1	1	1	1
Major	1	1	1	1	1	1	1	1	1	1	1	1
PRE_MRT	0	0	0	0	1	0	0	0	1	1	0	0
PRE_AST	0	0	0	0	0	0	1	1	1	0	0	0
Cond_1	1	0	0	1	0	0	0	0	0	1	0	0
Cond_2	0	1	0	0	1	0	0	0	0	0	1	0
Cond_3	0	0	1	0	0	1	0	0	0	0	0	1
Pr_1.1	0	1	1	1	0	0	1	0	0	1	0	0
Pr_1.2	1	0	1	0	1	0	0	1	0	0	1	0
Pr_1.3	1	1	0	0	0	1	0	0	1	0	0	1
Pr_2.1	1	0	0	0	1	1	1	0	0	1	0	0
Pr_2.2	0	1	0	1	0	1	0	1	0	0	1	0
Pr_2.3	0	0	1	1	1	0	0	0	1	0	0	1
Pr_3.1	1	0	0	1	0	0	0	1	1	1	0	0
Pr_3.2	0	1	0	0	1	0	1	0	1	0	1	0
Pr_3.3	0	0	1	0	0	1	1	1	0	0	0	1
Pr_4.1	0	0	0	1	0	0	1	0	0	0	1	1
Pr_4.2	0	1	0	0	1	0	0	1	0	1	0	1
Pr_4.3	0	0	1	0	0	1	0	0	1	1	1	0
Em_2.1	0	0	0	1	0	0	0	0	0	0	0	0
Em_2.2	0	0	0	0	1	0	0	0	0	0	1	0
Em_2.3	0	0	0	0	0	0	0	0	0	0	0	0
Em_4.1	0	0	0	1	0	0	1	0	0	1	0	0
Em_4.2	0	0	0	0	0	0	0	0	0	0	0	0
Em_4.3	0	0	1	0	0	1	0	0	1	0	0	1
Em_5.1	0	0	0	1	0	0	0	0	0	1	0	0
Em_5.2	0	0	0	0	0	0	0	0	0	0	1	0
Em_5.3	0	0	0	0	0	1	0	0	0	0	0	1
Mo_1.1	1	0	0	1	0	0	1	0	0	1	0	0
Mo_1.2	0	1	0	0	1	0	0	1	0	0	1	0
Mo_1.3	0	0	1	0	0	1	0	0	1	0	0	0
Mo_3.1	1	0	0	1	0	0	0	0	0	0	0	0
Mo_3.2	0	1	0	0	1	0	0	0	0	0	1	0
Mo_3.3	0	0	1	0	0	1	0	0	1	0	0	0
Mo_4.1	1	0	0	1	0	0	0	0	0	1	0	0
Mo_4.2	0	1	0	0	1	0	0	1	0	0	1	0
Mo_4.3	0	0	1	0	0	1	0	0	1	0	0	1
En_1.1	1	0	0	0	0	0	1	0	0	0	0	0
En_1.2	0	1	0	0	0	0	0	0	0	0	0	0
En_1.3	0	0	0	0	0	0	0	0	0	0	0	0
En_2.1	1	0	0	0	0	0	1	0	0	0	0	0
En_2.2	0	1	0	0	1	0	0	1	0	0	0	0
En_2.3	0	0	1	0	0	0	0	0	0	0	0	1
En_3.1	0	0	0	0	0	0	1	0	0	1	0	0
En_3.2	0	0	0	0	0	0	0	0	0	0	0	0

En_3.3	0	0	0	0	0	0	0	0	0	0	0	1
En_4.1	1	0	0	0	0	0	0	0	0	0	0	0
En_4.2	0	0	0	0	0	0	0	0	0	0	0	0
En_4.3	0	0	0	0	0	1	0	0	0	0	0	0
En_5.1	1	0	0	1	0	0	0	0	0	0	0	0
En_5.2	0	1	0	0	1	0	0	1	0	0	0	0
En_5.3	0	0	1	0	0	1	0	0	0	0	0	0
Cl_1.1	1	0	0	0	0	0	1	0	0	0	0	0
Cl_1.2	0	0	0	0	0	0	0	0	0	0	1	0
Cl_1.3	0	0	0	0	0	0	0	0	0	0	0	0
Cl_2.1	1	1	1	1	1	1	1	1	1	1	1	1
Cl_2.2	0	0	0	0	0	0	0	0	0	0	0	0
Cl_2.3	0	0	1	0	0	0	0	0	0	0	0	0
Quiz_1	0	0	0	0	0	0	0	0	0	0	0	0
Quiz_2	0	1	0	0	0	0	0	1	0	0	1	0
Quiz_3	0	0	0	0	0	0	0	0	1	0	0	1
Post_MRT	0	0	0	0	0	0	0	0	0	1	0	0
Post_AST	0	0	0	0	0	0	0	1	0	0	0	1

Predictors for Items Measuring Sense of Embodiment

Table D2 is the predictor matrix for sense of embodiment. Its columns are the set of target variables that measured participants' sense of embodiment across three sessions. For example, "Em_1.2" means this column is the variable that collects participants' scores to item Em_1 (I felt as if the virtual body or body part was my own body or body part.) in the second session. Each row represents one candidate predictor variable. A cell containing "1" means that the row variable is used to impute the column variable.

Table D2

Predictors for Items Measuring Sense of Embodiment

	Em_2.1	Em_2.2	Em_2.3	Em_4.1	Em_4.2	Em_4.3	Em_5.1	Em_5.2	Em_5.3
Group	1	1	1	1	1	1	1	1	1
Gender	0	0	0	0	0	0	0	0	0
Age	1	1	1	1	1	1	1	1	1
Race	1	1	1	1	1	1	1	1	1
Major	1	1	1	1	1	1	1	1	1
PRE_MRT	1	1	0	0	1	0	0	1	0
PRE_AST	0	0	1	1	1	0	1	0	1
Cond_1	1	0	0	1	0	0	1	0	0
Cond_2	0	1	0	0	1	0	0	1	0
Cond_3	0	0	1	0	0	1	0	0	1
Pr_1.1	0	0	0	0	0	0	1	0	0
Pr_1.2	0	1	0	0	0	0	0	0	0
Pr_1.3	0	0	1	0	0	1	0	0	0
Pr_2.1	1	0	0	1	0	0	1	0	0
Pr_2.2	0	1	0	0	1	0	0	1	0
Pr_2.3	0	0	0	0	0	1	0	0	1
Pr_3.1	0	0	0	0	0	0	0	0	0
Pr_3.2	0	0	0	0	0	0	0	0	0
Pr_3.3	0	0	0	0	0	0	0	0	0
Pr_4.1	1	0	0	1	0	0	1	0	0
Pr_4.2	0	1	0	0	1	0	0	1	0
Pr_4.3	0	0	1	0	0	1	0	0	1
Em_2.1	0	1	1	1	0	0	1	0	0
Em_2.2	1	0	1	0	1	0	0	1	0
Em_2.3	1	1	0	0	0	1	0	0	1
Em_4.1	1	0	0	0	1	1	1	0	0
Em_4.2	0	1	0	1	0	1	0	1	0
Em_4.3	0	0	1	1	1	0	0	0	1
Em_5.1	1	0	0	1	0	0	0	1	1
Em_5.2	0	1	0	0	1	0	1	0	1
Em_5.3	0	0	1	0	0	1	1	1	0
Mo_1.1	1	0	0	1	0	0	1	0	0
Mo_1.2	0	0	0	0	1	0	0	0	0
Mo_1.3	0	0	0	0	0	0	0	0	0
Mo_3.1	0	0	0	1	0	0	1	0	0
Mo_3.2	0	1	0	0	0	0	0	0	0
Mo_3.3	0	0	1	0	0	1	0	0	0
Mo_4.1	0	0	0	1	0	0	1	0	0
Mo_4.2	0	1	0	0	1	0	0	0	0
Mo_4.3	0	0	1	0	0	1	0	0	0
En_1.1	1	0	0	0	0	0	0	0	0
En_1.2	0	1	0	0	1	0	0	0	0
En_1.3	0	0	0	0	0	0	0	0	1
En_2.1	1	0	0	1	0	0	0	0	0
En_2.2	0	0	0	0	0	0	0	1	0
En_2.3	0	0	0	0	0	1	0	0	0
En_3.1	1	0	0	0	0	0	0	0	0
En_3.2	0	0	0	0	0	0	0	0	0

En_3.3	0	0	1	0	0	1	0	0	1
En_4.1	1	0	0	0	0	0	0	0	0
En_4.2	0	0	0	0	0	0	0	1	0
En_4.3	0	0	0	0	0	0	0	0	0
En_5.1	0	0	0	1	0	0	0	0	0
En_5.2	0	1	0	0	1	0	0	0	0
En_5.3	0	0	1	0	0	1	0	0	1
Cl_1.1	0	0	0	0	0	0	0	0	0
Cl_1.2	0	0	0	0	0	0	0	0	0
Cl_1.3	0	0	0	0	0	0	0	0	0
Cl_2.1	1	1	1	1	1	1	1	1	1
Cl_2.2	0	0	0	0	0	0	0	1	0
Cl_2.3	0	0	0	0	0	0	0	0	0
Quiz_1	0	0	0	0	0	0	0	0	0
Quiz_2	0	0	0	0	0	0	0	1	0
Quiz_3	0	0	1	0	0	0	0	0	0
Post_MRT	0	0	0	0	0	0	0	0	1
Post_AST	0	0	0	0	0	0	1	1	1

Predictors for Items Measuring Motivation

Table D3 is the predictor matrix for motivation. Its columns are the set of target variables that measured participants' motivation across three sessions. For example, "Mo_1.2" means this column variable is participants' scores to item Mo_1 (I enjoyed doing this activity very much.) in the second session. Each row represents one candidate predictor variable. A cell containing "1" means that the row variable is used to impute the column variable.

Table D3

Predictors for Items Measuring Motivation

	Mo_1.1	Mo_1.2	Mo_1.3	Mo_3.1	Mo_3.2	Mo_3.3	Mo_4.1	Mo_4.2	Mo_4.3
Group	1	1	1	1	1	1	1	1	1
Gender	0	0	0	0	0	0	0	0	0
Age	1	1	1	1	1	1	1	1	1
Race	1	1	1	1	1	1	1	1	1
Major	1	1	1	1	1	1	1	1	1
PRE_MRT	0	1	0	0	0	0	0	0	0
PRE_AST	0	0	1	0	0	1	0	0	1
Cond_1	0	0	0	0	0	0	0	0	0
Cond_2	0	0	0	0	0	0	0	0	0
Cond_3	0	0	0	0	0	0	0	0	0
Pr_1.1	1	0	0	1	0	0	1	0	0
Pr_1.2	0	1	0	0	1	0	0	1	0
Pr_1.3	0	0	1	0	0	1	0	0	1
Pr_2.1	1	0	0	0	0	0	0	0	0
Pr_2.2	0	1	0	0	1	0	0	0	0
Pr_2.3	0	0	1	0	0	1	0	0	1
Pr_3.1	0	0	0	0	0	0	0	0	0
Pr_3.2	0	0	0	0	0	0	0	1	0
Pr_3.3	0	0	1	0	0	1	0	0	1
Pr_4.1	0	0	0	0	0	0	0	0	0
Pr_4.2	0	1	0	0	1	0	0	1	0
Pr_4.3	0	0	0	0	0	0	0	0	1
Em_2.1	0	0	0	0	0	0	0	0	0
Em_2.2	0	0	0	0	1	0	0	0	0
Em_2.3	0	0	0	0	0	0	0	0	0
Em_4.1	1	0	0	1	0	0	1	0	0
Em_4.2	0	0	0	0	0	0	0	0	0
Em_4.3	0	0	1	0	0	1	0	0	1
Em_5.1	1	0	0	1	0	0	1	0	0
Em_5.2	0	0	0	0	0	0	0	0	0
Em_5.3	0	0	0	0	0	0	0	0	0
Mo_1.1	0	1	1	1	0	0	1	0	0
Mo_1.2	1	0	1	0	1	0	0	1	0
Mo_1.3	1	1	0	0	0	1	0	0	1
Mo_3.1	1	0	0	0	1	1	1	0	0
Mo_3.2	0	1	0	1	0	1	0	1	0
Mo_3.3	0	0	1	1	1	0	0	0	1
Mo_4.1	1	0	0	1	0	0	0	1	1
Mo_4.2	0	1	0	0	1	0	1	0	1
Mo_4.3	0	0	1	0	0	1	1	1	0
En_1.1	1	0	0	1	0	0	1	0	0
En_1.2	0	1	0	0	1	0	0	1	0
En_1.3	0	0	0	0	0	0	0	0	0
En_2.1	1	0	0	1	0	0	1	0	0
En_2.2	0	1	0	0	1	0	0	1	0
En_2.3	0	0	0	0	0	1	0	0	0
En_3.1	0	0	0	1	0	0	0	0	0
En_3.2	0	0	0	0	0	0	0	0	0

En_3.3	0	0	1	0	0	1	0	0	1
En_4.1	1	0	0	1	0	0	1	0	0
En_4.2	0	0	0	0	1	0	0	1	0
En_4.3	0	0	0	0	0	0	0	0	0
En_5.1	1	0	0	1	0	0	1	0	0
En_5.2	0	1	0	0	0	0	0	1	0
En_5.3	0	0	1	0	0	1	0	0	1
Cl_1.1	0	0	0	0	0	0	0	0	0
Cl_1.2	0	0	0	0	1	0	0	0	0
Cl_1.3	0	0	0	0	0	0	0	0	0
Cl_2.1	1	1	1	1	1	1	1	1	1
Cl_2.2	0	0	0	0	0	0	0	0	0
Cl_2.3	0	0	0	0	0	0	0	0	0
Quiz_1	0	0	0	0	0	0	0	0	0
Quiz_2	0	0	0	0	0	0	0	0	0
Quiz_3	0	0	0	0	0	0	0	0	0
Post_MRT	0	1	0	0	0	0	1	1	0
Post_AST	0	0	1	0	0	0	0	0	0

Predictors for items measuring engagement

Table D4 and Table D5 are the predictor matrices for engagement. The columns are the set of target variables that measured participants' engagement across three sessions. For example, "En_1.2" means this column variable is participants' scores to item En_1 (I felt that I was competent enough to meet the demands of the learning task.) in the second session. Each row represents one candidate predictor variable. A cell containing "1" means that the row variable was used to impute the column variable.

Table D4

Predictors for Items Measuring Engagement (1)

	En_1.1	En_1.2	En_1.3	En_2.1	En_2.2	En_2.3	En_3.1	En_3.2	En_3.3
Group	1	1	1	1	1	1	1	1	1
Gender	0	0	0	0	0	0	0	0	0
Age	1	1	1	1	1	1	1	1	1
Race	1	1	1	1	1	1	1	1	1
Major	1	1	1	1	1	1	1	1	1
PRE_MRT	0	1	1	0	0	0	0	1	0
PRE_AST	0	0	0	0	0	0	0	0	0
Cond_1	0	0	0	0	0	0	0	0	0
Cond_2	0	0	0	0	0	0	0	0	0
Cond_3	0	0	0	0	0	0	0	0	0
Pr_1.1	1	0	0	1	0	0	0	0	0
Pr_1.2	0	0	0	0	1	0	0	0	0
Pr_1.3	0	0	0	0	0	1	0	0	0
Pr_2.1	0	0	0	0	0	0	0	0	0
Pr_2.2	0	0	0	0	0	0	0	0	0
Pr_2.3	0	0	0	0	0	0	0	0	0
Pr_3.1	1	0	0	0	0	0	0	0	0
Pr_3.2	0	0	0	0	0	0	0	0	0
Pr_3.3	0	0	0	0	0	0	0	0	0
Pr_4.1	0	0	0	0	0	0	1	0	0
Pr_4.2	0	0	0	0	0	0	0	1	0
Pr_4.3	0	0	0	0	0	1	0	0	1
Em_2.1	0	0	0	0	0	0	0	0	0
Em_2.2	0	0	0	0	0	0	0	0	0
Em_2.3	0	0	0	0	0	0	0	0	0
Em_4.1	0	0	0	0	0	0	0	0	0
Em_4.2	0	0	0	0	0	0	0	0	0
Em_4.3	0	0	1	0	0	1	0	0	1
Em_5.1	0	0	0	0	0	0	0	0	0
Em_5.2	0	0	0	0	0	0	0	0	0
Em_5.3	0	0	0	0	0	0	0	0	0
Mo_1.1	1	0	0	1	0	0	1	0	0
Mo_1.2	0	1	0	0	1	0	0	0	0
Mo_1.3	0	0	0	0	0	0	0	0	0
Mo_3.1	1	0	0	1	0	0	1	0	0
Mo_3.2	0	0	0	0	1	0	0	0	0
Mo_3.3	0	0	0	0	0	1	0	0	0
Mo_4.1	1	0	0	1	0	0	0	0	0
Mo_4.2	0	1	0	0	1	0	0	1	0
Mo_4.3	0	0	0	0	0	0	0	0	1
En_1.1	0	1	1	1	0	0	1	0	0
En_1.2	1	0	1	0	1	0	0	1	0
En_1.3	1	1	0	0	0	1	0	0	1
En_2.1	1	0	0	0	1	1	1	0	0
En_2.2	0	1	0	1	0	1	0	1	0
En_2.3	0	0	1	1	1	0	0	0	1
En_3.1	1	0	0	1	0	0	0	1	1
En_3.2	0	1	0	0	1	0	1	0	1

En_3.3	0	0	1	0	0	1	1	1	0
En_4.1	1	0	0	1	0	0	1	0	0
En_4.2	0	1	0	0	1	0	0	1	0
En_4.3	0	0	1	0	0	1	0	0	1
En_5.1	1	0	0	1	0	0	1	0	0
En_5.2	0	1	0	0	1	0	0	1	0
En_5.3	0	0	1	0	0	0	0	0	1
Cl_1.1	0	0	0	1	0	0	1	0	0
Cl_1.2	0	1	0	0	1	0	0	1	0
Cl_1.3	0	0	1	0	0	1	0	0	1
Cl_2.1	1	1	1	1	1	1	1	1	1
Cl_2.2	0	1	0	0	0	0	0	1	0
Cl_2.3	0	0	1	0	0	0	0	0	1
Quiz_1	1	0	0	1	0	0	1	0	0
Quiz_2	0	0	0	0	0	0	0	0	0
Quiz_3	0	0	1	0	0	1	0	0	1
Post_MRT	0	1	1	0	1	0	0	1	0
Post_AST	0	0	0	0	0	1	1	0	0

Table D5

Predictors for Items Measuring Engagement (2)

	En_4.1	En_4.2	En_4.3	En_5.1	En_5.2	En_5.3
Group	1	1	1	1	1	1
Gender	0	0	0	0	0	0
Age	1	1	1	1	1	1
Race	1	1	1	1	1	1
Major	1	1	1	1	1	1
PRE_MRT	0	1	1	0	1	1
PRE_AST	0	1	1	0	0	1
Cond_1	0	0	0	0	0	0
Cond_2	0	0	0	0	0	0
Cond_3	0	0	0	0	0	0
Pr_1.1	1	0	0	1	0	0
Pr_1.2	0	0	0	0	1	0
Pr_1.3	0	0	0	0	0	0
Pr_2.1	0	0	0	0	0	0
Pr_2.2	0	0	0	0	1	0
Pr_2.3	0	0	1	0	0	1
Pr_3.1	0	0	0	0	0	0
Pr_3.2	0	0	0	0	1	0
Pr_3.3	0	0	0	0	0	0
Pr_4.1	0	0	0	0	0	0
Pr_4.2	0	0	0	0	0	0
Pr_4.3	0	0	0	0	0	0
Em_2.1	1	0	0	0	0	0
Em_2.2	0	0	0	0	0	0
Em_2.3	0	0	0	0	0	0
Em_4.1	0	0	0	1	0	0
Em_4.2	0	0	0	0	0	0
Em_4.3	0	0	1	0	0	1
Em_5.1	0	0	0	0	0	0
Em_5.2	0	0	0	0	0	0
Em_5.3	0	0	0	0	0	0
Mo_1.1	1	0	0	1	0	0
Mo_1.2	0	0	0	0	1	0
Mo_1.3	0	0	0	0	0	1
Mo_3.1	1	0	0	1	0	0
Mo_3.2	0	1	0	0	0	0
Mo_3.3	0	0	1	0	0	1
Mo_4.1	1	0	0	1	0	0
Mo_4.2	0	1	0	0	1	0
Mo_4.3	0	0	0	0	0	1
En_1.1	1	0	0	1	0	0
En_1.2	0	1	0	0	0	0
En_1.3	0	0	1	0	0	1
En_2.1	1	0	0	1	0	0
En_2.2	0	1	0	0	1	0
En_2.3	0	0	1	0	0	0
En_3.1	1	0	0	1	0	0
En_3.2	0	1	0	0	1	0

En_3.3	0	0	1	0	0	1
En_4.1	0	1	1	1	0	0
En_4.2	1	0	1	0	1	0
En_4.3	1	1	0	0	0	0
En_5.1	1	0	0	0	1	1
En_5.2	0	1	0	1	0	1
En_5.3	0	0	1	1	1	0
Cl_1.1	1	0	0	1	0	0
Cl_1.2	0	1	0	0	0	0
Cl_1.3	0	0	1	0	0	1
Cl_2.1	1	1	1	1	1	1
Cl_2.2	0	0	0	0	0	0
Cl_2.3	0	0	0	0	0	0
Quiz_1	0	0	0	0	0	0
Quiz_2	0	0	0	0	0	0
Quiz_3	0	0	0	0	0	0
Post_MRT	0	1	0	0	1	0
Post_AST	0	0	0	0	0	0

Predictors for Items Measuring Cognitive Load, Quiz Scores, and Scores in the Post-Test.

Table D6 is the predictor matrix for cognitive load, quiz scores, and scores in the post-test. The columns are the set of target variables that measured participants' cognitive load, quiz score, and scores in the post-test respectively. For example, "Cl_1.2" means this column variable is participants' scores to item Cl_1 (The learning content was difficult.) in the second session. Each row represents one candidate predictor variable. A cell containing "1" means that the row variable was used to impute the column variable.

Table D6

Predictors for Items Measuring Cognitive Load, Quiz Scores, and Scores in the Post-Test

	Cl_1.1	Cl_1.2	Cl_1.3	Cl_2.1	Cl_2.2	Cl_2.3	Quiz_2	Quiz_3	Post_MRT	Post_AST
Group	1	1	1	1	1	1	1	1	1	1
Gender	0	0	0	0	0	0	0	0	1	0
Age	1	1	1	1	1	1	1	1	1	1
Race	1	1	1	1	1	1	1	1	1	1
Major	1	1	1	1	1	1	1	1	1	1
PRE_MRT	0	1	1	0	1	1	1	1	1	1
PRE_AST	0	0	0	1	0	0	1	1	0	1
Cond_1	0	0	0	0	0	0	0	0	0	0
Cond_2	0	0	0	0	0	0	0	0	0	0
Cond_3	0	0	0	0	0	0	0	0	0	0
Pr_1.1	1	0	0	0	0	0	0	0	0	0
Pr_1.2	0	0	0	0	0	0	1	0	0	0
Pr_1.3	0	0	0	0	0	1	0	0	0	0
Pr_2.1	0	0	0	0	0	0	0	0	0	0
Pr_2.2	0	0	0	0	0	0	1	0	0	0
Pr_2.3	0	0	0	0	0	0	0	0	0	0
Pr_3.1	1	0	0	1	0	0	0	0	0	0
Pr_3.2	0	0	0	0	0	0	1	0	0	0
Pr_3.3	0	0	0	0	0	0	0	0	0	0
Pr_4.1	0	0	0	0	0	0	0	0	0	0
Pr_4.2	0	1	0	0	1	0	1	0	0	0
Pr_4.3	0	0	1	0	0	1	0	1	0	0
Em_2.1	0	0	0	0	0	0	0	0	0	0
Em_2.2	0	0	0	0	1	0	0	0	0	0
Em_2.3	0	0	0	0	0	1	0	1	0	0
Em_4.1	0	0	0	1	0	0	0	0	0	0
Em_4.2	0	0	0	0	0	0	0	0	0	0
Em_4.3	0	0	0	0	0	0	0	0	0	0
Em_5.1	0	0	0	1	0	0	0	0	0	0
Em_5.2	0	0	0	0	1	0	0	0	0	0
Em_5.3	0	0	0	0	0	0	0	0	0	1
Mo_1.1	1	0	0	0	0	0	0	0	0	0
Mo_1.2	0	1	0	0	1	0	0	0	1	0
Mo_1.3	0	0	0	0	0	0	0	0	0	0
Mo_3.1	1	0	0	0	0	0	0	0	0	0
Mo_3.2	0	1	0	0	0	0	0	0	0	0
Mo_3.3	0	0	0	0	0	0	0	0	0	0
Mo_4.1	1	0	0	0	0	0	0	0	1	0
Mo_4.2	0	0	0	0	0	0	0	0	1	0
Mo_4.3	0	0	0	0	0	0	0	0	0	0
En_1.1	1	0	0	1	0	0	0	0	0	0
En_1.2	0	1	0	0	1	0	1	0	1	0
En_1.3	0	0	1	0	0	1	0	1	1	0
En_2.1	1	0	0	1	0	0	0	0	0	0
En_2.2	0	1	0	0	1	0	0	0	0	0
En_2.3	0	0	1	0	0	0	0	1	0	1
En_3.1	1	0	0	1	0	0	0	0	0	1
En_3.2	0	1	0	0	1	0	0	0	0	0

En_3.3	0	0	1	0	0	1	0	1	0	0
En_4.1	1	0	0	0	0	0	0	0	0	0
En_4.2	0	1	0	0	0	0	0	0	1	0
En_4.3	0	0	1	0	0	0	0	0	0	0
En_5.1	1	0	0	1	0	0	0	0	0	0
En_5.2	0	0	0	0	0	0	0	0	0	0
En_5.3	0	0	1	0	0	1	0	0	0	1
Cl_1.1	0	1	1	1	0	0	0	0	0	0
Cl_1.2	1	0	1	0	1	0	1	0	0	0
Cl_1.3	1	1	0	0	0	1	0	1	1	0
Cl_2.1	1	1	1	0	1	1	1	1	1	1
Cl_2.2	0	1	0	1	0	1	0	0	0	0
Cl_2.3	0	0	1	1	1	0	0	1	0	0
Quiz_1	0	0	0	0	0	0	1	1	0	1
Quiz_2	0	0	0	0	0	0	0	1	1	1
Quiz_3	0	0	1	0	0	1	1	0	0	1
Post_MRT	0	0	1	0	1	1	1	0	0	1
Post_AST	0	1	0	1	0	0	1	1	1	0