REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. REPORT DATE (DD-MM-YYYY)
10-07-2015

2. REPORT TYPE
Final Report

3. DATES COVERED (From - To)
11-Apr-2011 - 10-Apr-2015

4. TITLE AND SUBTITLE
Final Report: Evaluating the Effectiveness of Immersive Interfaces for Combat Training

5a. CONTRACT NUMBER
W911NF-11-1-0126

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
206022

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

6. AUTHORS
Alley Butler, Richard Fowler, Mark Winkel

7. PERFORMING ORGANIZATION NAMES AND ADDRESSES
University of Texas-Pan American
1201 W. University Drive

Edinburg, TX 78539 -2909

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSOR/MONITOR'S ACRONYM(S)
ARO

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
59095-CS-REP.2

12. DISTRIBUTION AVAILIBILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not contrued as an official Department of the Army position, policy or decision, unless so designated by other documentation.

14. ABSTRACT
This project focuses directly on immersive science. It answers basic questions about interfaces and their characteristics with respect to human responses with these interfaces. In this context, it studies immersion and the measurement of presence as it relates to the virtual world paradigm. Additionally, this proposal answers questions regarding effectiveness of virtual worlds versus traditional training, as posed by an Army laboratory and a Marine Corps program office. As part of this proposal, an Immersive Environments laboratory is established at The University of Texas-Pan American (UTPA). At the heart of this proposal are quantitative measures of human

15. SUBJECT TERMS
virtual environments, immersion, human subjects, combat effectiveness, computer interfaces, human factors

16. SECURITY CLASSIFICATION OF:
a. REPORT UU
b. ABSTRACT UU
c. THIS PAGE UU

17. LIMITATION OF ABSTRACT
UU

15. NUMBER OF PAGES

19a. NAME OF RESPONSIBLE PERSON
Alley Butler

19b. TELEPHONE NUMBER
956-665-2534

Report Title

Final Report: Evaluating the Effectiveness of Immersive Interfaces for Combat Training

ABSTRACT

This project focuses directly on immersive science. It answers basic questions about interfaces and their characteristics with respect to human responses with these interfaces. In this context, it studies immersion and the measurement of presence as it relates to the virtual world paradigm. Additionally, this proposal answers questions regarding effectiveness of virtual worlds versus traditional training, as posed by an Army laboratory and a Marine Corps program office. As part of this proposal, an Immersive Environments laboratory is established at The University of Texas-Pan American (UTPA). At the heart of this proposal are quantitative measures of human subject responses to differences in immersive interfaces and virtual world training regimens. Physiological measurements and statistical methods are employed to reach conclusions regarding presence in immersive environments and the effectiveness of virtual world paradigms as it relates to support for combat operations.

Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received Paper

TOTAL:

Number of Papers published in non peer-reviewed journals:

(c) Presentations

Number of Presentations: 0.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received	Paper

TOTAL:

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received	Paper

TOTAL:

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

Received	Paper

TOTAL:

Number of Manuscripts:

Books

Received Book

TOTAL:

Received Book Chapter

TOTAL:

Patents Submitted

Patents Awarded

Awards

Graduate Students

NAME	PERCENT_SUPPORTED	Discipline
Benjamin Peters, 1st MS Degree	1.00	
Benjamin Peters, 2nd MS Degree	1.00	
Moises Carrillo	1.00	
Ricardo Hernandez	1.00	
Rahul Varshney	0.50	
FTE Equivalent:	**4.50**	
Total Number:	**5**	

Names of Post Doctorates

NAME	PERCENT_SUPPORTED
FTE Equivalent:	
Total Number:	

Names of Faculty Supported

NAME	PERCENT_SUPPORTED	National Academy Member
Alley Butler	0.06	No
Richard Fowler	0.00	
Mark Winkel	0.04	
FTE Equivalent:	**0.10**	
Total Number:	**3**	

Names of Under Graduate students supported

NAME	PERCENT_SUPPORTED	Discipline
Aaron Hunsaker	1.00	
FTE Equivalent:	**1.00**	
Total Number:	**1**	

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period: 0.00

The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:...... 0.00

The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:...... 0.00

Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):...... 0.00

Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:...... 0.00

The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense 0.00

The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields:...... 0.00

Names of Personnel receiving masters degrees

NAME	
Moises Carrillo	
Benjamin Peters	
Benjamin Peters	
Total Number:	3

Names of personnel receiving PHDs

NAME	
Total Number:	

Names of other research staff

NAME	PERCENT_SUPPORTED
Jessica Sanchez	0.03
Ken Sailor	0.01
FTE Equivalent:	**0.04**
Total Number:	**2**

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

Please see attached report.

Technology Transfer

Final Report for

Evaluating the Effectiveness of Immersive Interfaces for Combat Training

Grant W911WF-11-1-0126
April 11, 2011 to April 10, 2015

Alley Butler, PhD, PE
Richard Fowler, PhD
Mark Winkel, PhD

University of Texas-Pan American

Abstract

This project focuses directly on immersive science. It answers basic questions about interfaces and their characteristics with respect to human responses with these interfaces. In this context, it studies immersion and the measurement of presence as it relates to the virtual world paradigm. Additionally, this proposal answers questions regarding effectiveness of virtual worlds versus traditional training, as posed by an Army laboratory and a Marine Corps program office. As part of this proposal, an Immersive Environments laboratory is established at The University of Texas-Pan American (UTPA). At the heart of this proposal are quantitative measures of human subject responses to differences in immersive interfaces and virtual world training regimens. Physiological measurements and statistical methods are employed to reach conclusions regarding presence in immersive environments and the effectiveness of virtual world paradigms as it relates to support for combat operations.

Table of Contents

Introduction

This project included three experiments conducted as quantitative, statistically based measures of immersion and the effectiveness of virtual environments for military and potential civilian use. The first experiment includes measurement of the sense of immersion by human subjects based on the size of the screen and the use of stereoscopic viewing when compared to simple monocular viewing. The second experiment was inspired by the STTC office of the Army's RDECOM in Orlando, FL. In this experiment, training using virtual worlds methods (Second Life software) was used and compared to classical case studies for the defusing of IED's (Improvised Explosive Devices). This study included bio-metric measurements of subjects, and statistically compared the effectiveness of each training method at the 95% and 90% level. The third experiment was requested by the Marine Corps Training Systems Office in Orlando, FL. This experiment was designed to provide statistical proof that the Alelo Software in use by the Marine Corps was (or was not) equivalent to classical, classroom training. Again, statistically valid conclusions are reached at the 95% level. In addition to these three experiments, the University of Texas-Pan American research team put together an Immersive Systems laboratory which included the commissioning of a virtual environment workbench, the completion of a "development laboratory," and the erection and commissioning of a CAVE system along with the acquisition and use of a system for acquitting biometric data from human subjects. Each of these pieced of work are discussed separately in the following report sections. This is followed by a brief conclusion.

Evaluating the Effects of Stereoscopy and Screen Size on Feelings of Presence in Virtual Environments

Summary

Though the efficacy of stereoscopy and large screen size is well documented for human factors tasks, the effects of stereoscopy and screen size on users' feeling of presence, or subjective feelings of "being in the virtual environment." are much less well known. The current work addresses this lack of knowledge through an experiment in which users interacted in virtual environments either with or without stereoscopic display using displays of varying sizes. It was found that users had higher feelings of presence with stereoscopic viewing than without, and display size was less important, with a smaller enhancement of presence provided by larger screen sizes.

Introduction

Among the earliest and best known experiments to demonstrate the efficacy of virtual environments in task performance was Ware and Franck's (1996) exploration of the effects of stereoscopy in path finding. In this series of experiments users followed paths in graphs, and performance, as measured by number of links traversed in paths, was found to be enhanced by stereoscopic display. Ni, Bowman, and Chen (2006) found enhanced task performance in information rich enviroments with large displays compared to desktop monitors. Minkov, Perry and Oron-Gilad, (2007), describe more effective target identification in a military setting with larger displays, though increased cognitive effort can overcome some advantages of larger displays.

However, as Bowman and McMahan (2007) point out, evoking the feeling of "being in the (virtual) environment, or presence, can be accomplished by the considered selection of "real-world" cues, rather than simply including all elements of veridical representation possible. Recently, for example, they (McMahan, Bowman, Zielinski, and Brady, 2012) examined the effects of display fidelity and interaction fidelity to examine their effects on strategy and performance. Similarly, Laha, Sensharma, Schiffbauer, and Bowman, (2012) investigated the effects of head tracking, field of regard, and stereoscopic rendering on visualization tasks with x-ray tomography visualization and found that benefits do not require all three components simultaneously.

The current study continues this delineation of the role of particular immersive elements on users by examining the role of display size and stereoscopy on users subjective feelings of presence in virtual environments.

Methodology

In the experiment 10 subjects performed exploration tasks in virtual environments using stereoscopic and monoscopic viewing with three different display sizes: 20", 4', and 8' widths. After each exploration task with a particular combination of stereoscopy and display size, the subject rated his or her subjective feels of "being in the environment," or presence, using a questionnaire adapted from Slater et al. (1994) and Widmer and Singer (1998). The questionnaire is included as an appendix.

Virtual environments

Two examples of virtual environments (VEs) used in the study are shown below. The VEs were a mix of architectural scenes developed using AutoCad and similar tools and "fantasy" scenes developed using the Unreal Development Kit (UDK), which is designed for video game development. The process of VE authoring typically entailed embellishment of toolkit provided templates, e.g., adding more streets to a city or rooms to a castle, or by applying material textures to polygons to create a more realistic VE. In this way the “richness” and complexity of VEs was varied, and in the context of this study allowed the creation of distinct regions, e.g., rooms, streets, buildings, the exploration of which was the subject task. The VEs were created or selected based on their utility in providing an environment that supports exploration (navigation) of distinct areas. Examples are provided in Figure 1.

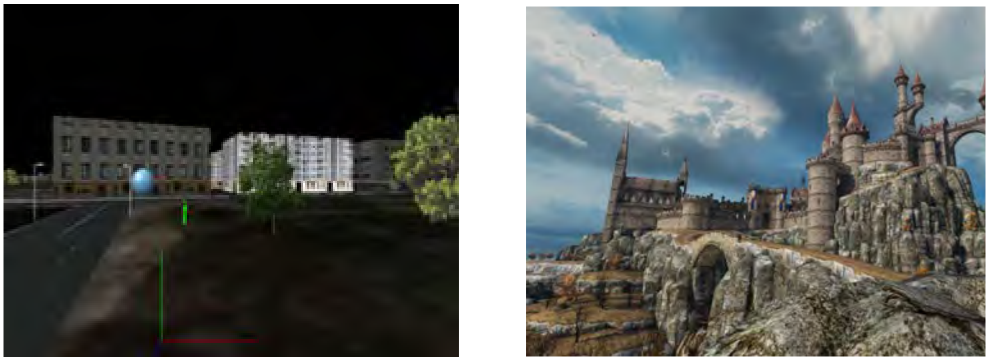


Figure 1. Two examples of the architectural and fantasy virtual environments used in the study.

The three display sizes, 20 inch, 4 feet, and 8 feet horizontal width, subtended viewing angles of 32^{o,} 68^{o}, and 108^{o}, respectively. The smallest size, 20 inch width, was provided by a standard 22 inch, as diagonally measured, computer monitor. The larger sizes were supplied by front projected image to a projection screen. Resolution, 1980 x 1020, was the same for all sizes. For each of the three display sizes, subjects viewed the displays stereoscopically (3D) or monoscopicly (2D). Thus, each subject viewed six different conditions of display size and stereoscopy. Subjects were instructed to move through the VE visiting different locations. Interaction, movement in the VE, was accomplished by moving the desktop mouse with one hand to control movement speed, while simultaneously depressing a key with the other hand to indicate direction. This is a common technique for video game interaction, and all subjects were well practiced with it. Following viewing in each condition, subjects completed the presence questionnaire. Subjects navigated in (explored) each VE for eight minutes. Total time was 90 minutes or less.

Details of experimental activities

Prospective subjects were seated in the laboratory and provided the informed consent form with the experimenter available to answer questions. After signing the informed consent form or opting not to participate (none declined), subjects completed a pre-experiment demographic and computer experience questionnaire, taking about two minutes. Introduction to the experiment and a practice task preceded data collection trials. Subjects were first introduced to the experimental setup (desk with monitor and screen displays, with keyboard and mouse for interaction). They were then instructed in the interaction technique, which entailed utilization of the keyboard with one hand to indicate direction of movement and the use of the mouse in the other to indicate velocity of movement. This interaction techniques is common in video games and was quite familiar to all subjects who participated in the study. Next, subjects were instructed to explore, or navigate through, a VE displayed on the monitor. Subjects were encouraged to ask questions about the task, and the experimenter provided further explanation, as requested. Finally, subjects performed a practice VE exploration task lasting about five minutes using the desktop monitor display with the experimenter answering subjects' questions as they might arise.

After the practice task, Ss then performed a series of six eight minute interactions in VEs during which they navigated through the environments. Subjects were instructed to navigate through as much of the VE as time permits. After each VE interaction the presence questionnaire was completed, taking about two minutes. Six different VEs were available and a different VE was used for each of the six display size / stereoscopy conditions. VEs were displayed in both stereoscopic and normal viewing modes, with each of the three display sizes. All combinations of stereoscopy (S-yes, S-no) and display size (small-20 inch, medium-4 feet, large-8 feet) were used with random order of pairings. For example, a subject might view a sequence of six stereoscopy / display size pairings such as: S-yes / medium, S-no / large, S-no / small, S-yes / small, S-no / medium, S-yes / large.

Results

Ten subjects, eight male, two female, performed the navigation and ratings tasks. All were students at the University of Texas - Pan American and seven were majoring in computer science. All used computers over 12 hours per week, and all had extensive experience with video games, most being at least weekly players. As all subjects performed the navigation task with each of the six display size (20 inch, 4 feet, 8 feet) x stereoscopy (monoscopic, stereoscopic) conditions, a within subjects analysis of variance (ANOVA) was performed. A subject's raw scores on the presence questionnaire were transformed to z-scores for the analysis. Table 1 provides the results of the ANOVA.

Table 1. Summary of within subjects ANOVA for subject's rating of presence for stereoscopy (monoscopic, stereoscopic) and display sizes (20 inch, 4 feet, and 8 feet).

ANOVA Summary

A = stereoscopy, B = display size, Subj = subjects

Source	SS	df	MS	F	P
Subjects	1.4923	9			
Within Subjects					
A	10.0123	1	10.0123	18.4797	0.001994
Subj x A		4.8762	9	0.5418	
B	2.4318	2	1.2159	1.818	0.196924
Subj x B		12.0378	18	0.6688	
A x B	1.1866	2	0.5933	0.8196	0.456392
Subj x A x B		13.0306	18	0.7239	
TOTAL	45.0676	59			

As shown in Table 2 reporting average presence scores for the six conditions, and in Figure 2 showing a graph of these average presence scores, there is a strong effect for stereoscopy ($p < 0.002$), with stereoscopic viewing leading to subjects reporting a higher subjective feeling of presence. Also, the data indicate a difference in subjects' feeling of presence for the larger 4 feet and 8 feet display sizes compared to the 20 inch display, though the result is not significant ($p < 0.20$).

Table 2. Subjects' presence ratings. Higher values indicate a stronger feeling of presence, or "being in the virtual environment." A strong effect of stereoscopy is evident. Results suggest that compared to a desktop monitor the larger displays enhanced presence, though with no difference in 4 feet and 8 feet displays.

		Display Size			
		Monitor	4'	8'	
Stereoscopy:	Monoscopic	-0.76	-0.36	-0.08	-0.40
	Stereoscopic	0.21	0.69	0.35	0.42
		-0.28	0.16	0.13	

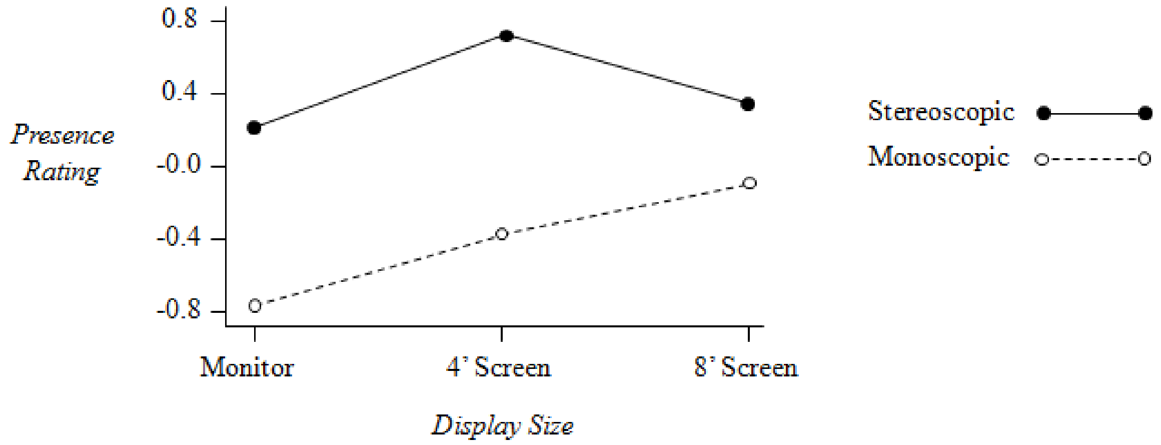


Figure 2. Subjects' feeling of presence are shown on the *x*-axis and display size on the *y*-axis.

Discussion and Conclusions

The result that stereoscopic viewing leads to an enhanced feeling of presence, i.e., that the subject felt "in the (virtual) environment" compared to monoscopic viewing confirms both popular belief and earlier studies. In the context of the presence questionnaire developed for this study, the item "How much did your experiences in the virtual environment seem consistent with your real-world experiences?" captures this notion succinctly. The particular exploration, or navigation, task subjects performed also naturally leads subjects to utilize depth cues, which are augmented by stereoscopy, and so in the VE conditions of this experiment are in accord with "real-world" experience. The result indicates that the experimental conditions and manipulations were able to change the user experience and to measure that change through the experiment's questionnaire.

Display size results, though only suggestive from these data, both confirm earlier findings and provide interesting information for further investigation. As noted earlier, in general larger display size leads to enhanced feeling of presence and task performance. This is evident in this study for the small monitor display, compared to the larger 4 feet and 8 feet displays. For both conditions of stereoscopy the larger displays lead to higher feelings of presence than the monitor. However, whereas with monoscopic display, feelings of presence increase monotonically with display size, for stereoscopic display feelings of presence do increase for the 4' display compared to the monitor display, but *decrease* with the 8 foot display compared to the 4 feet display. Further work in which variables such as information and scene density are manipulated as a function of display size, as well as the investigation of the role of full body movement versus head alone to provide orientation, are examples of studies that could help to further understand this outcome of the experiment.

References

Bowman, D., & McMahan, R. P. (2007). Virtual reality: how much immersion is enough?. *Computer, 40*(7), 36-43.

Laha, B., Sensharma, K., Schiffbauer, J. D., & Bowman, D. (2012). Effects of immersion on visual analysis of volume data. *Visualization and Computer Graphics, IEEE Transactions on*, *18*(4), 597-606.

McMahan, R. P., Bowman, D., Zielinski, D. J., & Brady, R. B. (2012). Evaluating display fidelity and interaction fidelity in a virtual reality game. *Visualization and Computer Graphics, IEEE Transactions on*, *18*(4), 626-633.

Minkov, Y., Perry, S., & Oron-Gilad, T. (2007, October). The effect of display size on performance of operational tasks with UAVs. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 51, No. 18, pp. 1091-1095). SAGE Publications.

Ni, T., Bowman, D. A., & Chen, J. (2006, June). Increased display size and resolution improve task performance in information-rich virtual environments. In *Proceedings of Graphics Interface 2006* (pp. 139-146). Canadian Information Processing Society.

Slater, M., Usoh, M., & Steed, A. (1994). Depth of presence in virtual environments. *Presence*, *3*(2), 130-144.

Ware, C., & Franck, G. (1996). Evaluating stereo and motion cues for visualizing information nets in three dimensions. *ACM Transactions on Graphics (TOG)*, *15*(2), 121-140.

Witmer, B. G., & Kline, P. B. (1998). Judging perceived and traversed distance in virtual environments. *Presence*, *7*(2), 144-167.

Witmer, B. G., & Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoperators and virtual environments*, *7*(3), 225-240.

Appendix: Presence Questionnaire

1. How much did your experiences in the virtual environment seem consistent with your real-world experiences?
2. How compelling was your sense of moving through space?
3. How much did the visual aspects of the environment involve you?
4. How aware were you of events occurring in the real world around you?
5. How aware were you of your display and control devices?
6. How responsive was the environment to actions that you initiated (or performed)?
7. How completely were you able to actively survey or search the environment using vision?
8. How compelling was your sense of moving around inside the virtual environment?
9. How closely were you able to examine objects?
10. How well could you examine objects from multiple viewpoints?
11. To what degree did you feel confused or disoriented at the beginning of the task?
12. How involved were you in the virtual environment experience?
13. How quickly did you adjust to the virtual environment experience?
14. How much did the visual display quality interfere or distract you from performing assigned tasks or required activities?
15. Were you involved in the experimental task to the extent that you lost track of time?

A Study Comparing the Pedagogical Effectiveness of Virtual Worlds and of Traditional Training Methods

Summary

The use of virtual worlds for training and education are reported in the literature, but it is recognized that there is a sparse amount of analysis regarding the effectiveness of training and education using virtual worlds. The experiment reported in this section provides data and statistical analysis to allow the reader to reach conclusions about the effectiveness of virtual worlds for a problem of Army interest, the proper defusing of IED's (improvised explosive devices). The data includes demographic information, training effectiveness assessments, timing, personality, anxiety propensity, and measures of effectiveness in handling a situation with a mock-up of an IED and hostage. It was found at the 90% statistical level that the virtual worlds training was superior in effectiveness. This experiment was inspired by and designed to provide information to the STTC Army RDECOM in Orlando, FL.

Introduction

1.1 Virtual Reality

Virtual reality is a fairly new form of technology finding its way into the world of business, government, and other organizations. While it has been a staple of the entertainment industry, virtual reality has struggled to find favor when used as a training tool. Due to cost, skepticism, or a resistance to change, organizations have been slow to embrace using virtual reality for training. With the creation of Linden Lab's Second Life in 2003, the perception of virtual reality training has improved. Expensive training tools have been replaced by digital replicas in a virtual island. Through the proper programming, these replicas work like their real-world counterparts, but they lack the safety concerns involved in having an unskilled operator using potentially dangerous machinery. However, there are concerns about whether this new technology is more effective than traditional training methods at training operators. These concerns are valid since making less effective changes would be costly to an organization. Therefore, proponents and skeptics alike need concrete evidence to determine if in fact virtual reality is more effective at training people than traditional training methods.

The factors playing a central role in the effectiveness of virtual reality training are presence and immersion. Presence refers to a person's feeling of being physically present in a virtual environment even when in reality they are not physically present. The level of presence depends on the level of realism in the environment. However, even if the environment simulatesreal life perfectly, it can be useless as a training tool unless the trainee is immersed in the environment. Immersion refers to how involved a person is with an environment whether the environment is real or synthetic. Asking if a person is immersed in an environment is the same as asking if the person believes that they are in that environment. Therefore, when designing a training exercise it is important to make sure the exercise holds the interest of those being trained. The more interested a person stays in the exercise, the more effective it should be. Over the years, different companies have attempted to use these concepts to introduce innovative training exercises. While not perfect, they have laid the groundwork for future projects attempting to determine the effectiveness of virtual training.

The types of virtual reality experiments conducted range from simple online instructional videos to fully developed 3D virtual rooms. These experiments were developed with the thought of improving the more traditional styles of training workers, athletes, or military personnel.

While traditional training methods might be adequate for certain types of work, such as welding or most athletic sports, they have the potential to become too costly, or even too dangerous for the trainees. Take the welding example for instance. The metal being welded in a training exercise is still useable metal. If the weld is prepared poorly, then the metal is wasted and cannot be reused. If in some way a virtual training tool is developed, then no resources would be wasted during training. The training is determined to be effective if the welder can provide as good of a weld as someone who was trained traditionally. This example shows the rationale used by companies when developing their own innovative training methods.

Innovation in training now is rooted in using the computer as a powerful tool to train workers in a variety of different ways. The simplest examples are programs where a trainee sits and watches an online tutorial of how to do something. The trainee is quizzed following the tutorial to see what they have learned. Slightly more immersive are third-person simulations exist. These are simulations where trainees take control of avatars in order to perform a specific task. One example is the use of a simulation to train people for emergency evacuations of buildings. In order to add some realism to the training, smoke and vibrations can be used to simulate the stages of a collapsing building.

The level of immersion increases in simulations where sensors are attached to the trainees. These sensors correspond to an avatar on the computer that follows the movements of the trainee. The trainee can move their hands in a matter similar to movement they would use if they were performing a particular task, say carpentry. In this example a block of wood is positioned on a table in a virtual world. By moving the hand avatar over the wood, the trainee can hold the wood in place while sawing it using their other hand. They do this by grabbing a virtual saw and moving their arm in a motion identical to how they saw in real life.

3D simulations are the most advanced when it comes to technology. These are built by designing 3D virtual environments that a trainee can step into and interact with a simulated environment. Through the use of projectors and mirrors, walls of 3D images surround the trainee creating the mirage of a different world designed to the creator’s choosing. However, even with all these advances, one question still persists. How is the effectiveness of these training methods determined?

1.2 Assessment Tools

In addition to creating these different training tools, researchers have used different methods to determine if their tools are effective or not. These methods lean towards using subjective questionnaires. The trainees are given questionnaires to fill out following the conclusion of their training. These questionnaires tend to use 5-point, 7-point, or 10-point scales for each question. The points are tallied and the results are reported. Often, the trainees report they liked the training and felt a sense of presence. A more sophisticated way of reporting results comes from using statistical methods. Statistics are taken on the sample itself for more detailed information on the people taking the training. Age, gender, history with virtual reality, history with the company, etc. are all reported. The results of the trainings are reported as statistical means and variances and through hypothesis testing, they are compared to a control sample. However, what are lacking are any physiological studies on the trainees. While there are studies about how the body reacts to stimuli when it is immersed in an environment, virtual training experiments tend not to include this type of study. In order to determine the effectiveness of virtual reality training, a quantitative method for determining presence and immersion needs to be used. Overall, the research done on virtual reality for training purposes is lacking because this type of investigation has been omitted.

1.3 Purpose of Experiment

The purpose of this experiment is to determine whether virtual reality training is more effective for training human subjects than traditional training methods. As previously stated, concrete evidence is needed to affirm or refute this hypothesis. Evidence is gathered by performing an experiment using human subjects. The subjects are split into two groups. Both groups are trained in defusing an explosive device. One group receives virtual reality training and the other receives only the traditional style training. The subjects are tested in defusing a mockup of a bomb in a live simulation. They are timed and these times determine the success or failure of the virtual training procedure. In addition, the subjects' physiological responses are monitored during the final day of the experiment in order to determine if there is a difference in the responses for the two instruction types. In other words, the feeling of presence and immersion of the subject in the virtual world is being tested scientifically and not just by a qualitative questionnaire. Through these methods the hypothesis that virtual training is more effective than a traditional training method is evaluated.

1.4 Outline of Report

Before diving into the experiment though, an extensive review on the literature covering the topics of virtual reality training, sense of presence during training, and training effectiveness evaluation is provided in section 2. This literature review covers previous research undertaken in order to provide a good picture of what has been accomplished in the field of virtual reality training. It also provides information on how particular training methods are evaluated. These evaluation tools are useful for creating a plan on how to evaluate the experiment highlighted in this report. As will be seen, some of these methods rely primarily on qualitative analysis. However, there are a few experimenters which use quantifiable data in their evaluation methods.

In section 3, the experiment upon which this report is based is explained in detail. This includes a summary of how the subjects are separated into two different classes – one which trains the subjects in the defusing of a bomb using virtual reality training and another which trains the subjects using traditional training methods. The testing simulation the subjects undertake is explained in detail and the tools by which the data is gathered are explained. Included in this explanation is a description of the arrangement of physiological sensors. Section 4 focuses on the analysis of the data. It gives a full analysis on what the results of the experiment are. This provides an answer to the question of whether virtual reality is more effective than traditional methods. Section 5 serves as a conclusion to the report of this experiment. It summarizes the results from the experiment and explains the impact of these findings.

Literature Review

2.1 Introduction

The use of virtual reality as a tool for training purposes has been studied since the latter half of the twentieth century. Unfortunately, the conclusions linked to these studies tend to say that the research done on virtual reality is inadequate. Dickey (2003) claims, "more research needs to be done to fully explore the potential [of virtual reality.]" When it comes to the topic of virtual reality, Gaimster (2008) says there has been "Little research in this specific area …." Goel (2009) supports this assertion by saying, "There is little research that addresses what features of virtual worlds support … applications." While much of the research is in need of further testing, there have been major advancements in the use of virtual reality for training purposes. These advancements are backed with statistical proof of virtual reality's effectiveness.

This section is devoted to literature on virtual reality technology. While some of the projects presented may not be completed, the section still provides an idea of what has been done so far. It also persuades readers to see the benefits of using virtual reality. But before moving to the experiment, a review on the literature about virtual reality needs to be explored. The review includes an overview of virtual reality and critical thinking and learning. Included in the overview are projects which support the use of virtual reality and its effect on learning. Following the overview, the ideas of presence and immersion are discussed with articles related to these ideas. Since presence and immersion tie into physiological response, a section is dedicated to that topic. Afterwards, there is a section dedicated to the types of experiments run and the corresponding results of those experiments. Finally, projects still in progress are discussed and the section is completed with a conclusion regarding the literature.

When virtual reality or virtual worlds are mentioned, the thought that comes to mind is video games. While video games have been created using virtual worlds to represent towns, forests, oceans, etc., this study focuses on the use of virtual reality for training purposes, not gaming. While video games could be used to create a training program, this examination dives deeper than just a 3D video game. This is about determining if virtual reality can simulate presence for a person who is not actually in the environment in which they see themselves. Then, once knowing the person is immersed in a virtual reality, whether this reality can be used to train them for a task better than traditional training methods needs to be determined. Before getting into that, the question, “What is Virtual Reality?” must be asked. Liu and Hao (2004) describe virtual reality as a powerful technology for creating an interactive virtual environment for the purpose of education and training. According to Cheng et al. (2010), virtual reality can improve learning performance by offering hands-on experience. As a matter of fact, Kelly and Cheek (2008) looked to revolutionize the way people learn using virtual worlds. Their goal was to create a market for virtual training in the future. Their idea to complete this goal was to “collaboratively build and test a meta-layer compatible with a subset of leading virtual world platforms that provides the robust administrative tools necessary for adoption in educational and work settings.” In other words, they wanted to use virtual worlds as a teaching tool for students and workers.

This use of virtual reality is fairly recent. In fact, virtual reality is a fairly recent concept. Holm and Priglinger (2008) trace the history of virtual reality as a technology to Ivan Sutherland who wrote *The Ultimate Display* in 1965. This resulted in the first virtual reality system: *The Sword of Damocles*. The first head mounted display was then developed in 1970. Since then, a multitude of simulators, games, and training devices have been created. One of the most popular versions of virtual reality is the use of virtual worlds. “Virtual worlds are computer-based simulated environments designed to allow users to inhabit and interact via avatars, the human agent’s in-world representative” (Monahan et al. 2009). One of the most prominent virtual worlds is Second Life. Second Life is a 3D virtual world created by Linden Labs. This virtual world consists of “islands,” servers that customers or private organizations can customize to how they see fit (Heiphetz and Woodill 2010). While Second Life has been used for gaming, it also has been used for marketing, classroom instruction, and training simulations. When it comes to education over 1500 universities are using Second Life (Ondrejka 2007). As far as training goes, companies believe using virtual reality for training is an efficient way to teach their employees.

2.2 Critical Thinking and Learning

The perceived strength of virtual reality is that it fosters an environment for learning and critical thinking. Scriven and Paul (1992) define critical thinking as the “intellectually disciplined process of actively and skillfully conceptualizing, applying, analyzing, synthesizing, and/or evaluating information gathered from, or generated by, observation, experience, reflection, reasoning, or communication, as a guide to belief and action.” Scriven and Paul go on to say that critical thinking is based on two components: a set of skills to process and generate information and beliefs, and the habit of using those skills to guide behavior. While technical skills are necessary to perform tasks, a worker with good technical skills and strong critical thinking skills can stand out against a worker with only great technical skills. This idea can also pertain to combat situations. According to Hammond (2004), “Even in combat, how well you think is more important to how well you fight than how physically fit you are. A wrong decision, an unasked question, a forgotten task, an incomplete analysis, or a poor synthesis can kill you.”

In order to prevent people from harming themselves in potentially dangerous activities due to a lack of critical thinking skills, virtual reality projects have been designed to foster learning and critical thinking. Interactive tutorials are a common usage for virtual reality. These tutorials start by walking the user through a set of instructions and then allow the user to practice using whatever tools they need to master. Three tutorials include: the Oil-field Safety Operation Training Interactive Virtual Environment, the Virtual Assembly System on Automobile Engines, and a tutorial to teach hand hygiene to hospital employees.

Liu and Hao (2004) designed a virtual environment called the Oil-field Safety Operation Training Interactive Virtual Environment (OSOTIVE). In this environment, the trainee goes through a series of levels, or modes, which increase in difficulty. The first mode is the Close Demonstration Mode (CDM). In this mode, the trainee is given a situation and has the system demonstrate the correct steps to perform in that situation. The second mode is the Guided Operation Mode (GOM). In this mode, the trainee controls an avatar, their representative in the virtual world, to complete the steps of a particular job. However, the program rejects any wrong moves so the user knows when they make a mistake. Finally the third and final mode is the Operation Mode (OOM). In this mode, the trainee is allowed to explore freely without interruption. If they make a mistake, the system reacts accurately to the situation. The mode encourages the trainee to learn from their mistakes.

Just like OSOTIVE, the Virtual Assembly System on Automobile Engines (AEVAS) uses a level system. Designed by Cheng et al. (2010), AEVAS consists of four rooms: Knowledge Room, Assembly Room, Expert Room and Checking Room. The Knowledge Room is the first room the trainee sees after logging in. They can view an assortment of material on the engine parts. The purpose is to gain knowledge of the different parts to help in the later rooms. The Assembly Room follows and consists of four inner-rooms: Crankshaft Assembly Room, Head Assembly Room, Tensioner Assembly Room and the Whole Engine Assembly Room. In each room, there is an exhibition of an assembly for each section. With this knowledge, the trainee is ready to try to assemble the engine in the next room. The Expert Room gives the trainee the chance to test what they learned. The trainee can assemble parts in the virtual environment and gain feedback on whether they did well or not. The final room serves as a bonus room. The Checking Room removes a part of the engine and tests the trainee to see if they know where it goes and the steps to get it there. Wen et al. (2009) created a similar environment fit with an oil drilling rig. The trainee can see the rig from a 3D perspective and learn how to operate it.

Similarly to the previous three groups, Bertrand et al. (2010) created a virtual training tutorial system. In this case, they designed a tutorial to train hospital workers in proper hand hygiene procedures. In the virtual training, they introduce five times when employees should wash their hands. Those are:

- Before touching a patient.
- Before clean/aseptic procedures.
- After body fluid exposure/risk.
- After touching a patient.
- After touching the patient's surroundings

There are three levels, called "phases" in the simulation. The first is the Tutorial Phase. Here the trainee is given an overview of the five moments for hand hygiene by a virtual doctor named Dr. Evan. The doctor prepares the trainee by telling them that they are to be tested in the five moments of hand hygiene. The doctor demonstrates the situations and uses voice and expression to do so. The Interactive Training Phase shows a virtual nurse named Simon who interacts with patients. The trainee must answer whether or not Simon followed the correct hand hygiene procedures. Finally in the Feedback Phase, the trainee is scored on how well they answered the questions. It should be noted this is not a very immersive system as the trainee does not interact through an avatar.

A similar scoring method is used in a driving simulator (Liang 2007). The simulator is created as a project for a Computer-Aided Design class. The simulator allows the trainee to take a simulated driving test. For every mistake the trainee makes, their score is reduced by a certain amount. If the trainee drops below 70 points the simulation ends and is reset. Consequently, the trainee passes the test if their score is over 70 after they complete the exam. Liang is not unique in his idea to use a scoring system in a simulator. The idea of using feedback to aid in a simulation is also used by Marcos de Moraes and dos Santos Machado (2009) in a bone marrow harvest simulator.

In their book, *Training and Collaboration with Virtual Worlds*, Heiphetz and Woodill (2010) support the use of Second Life by referring to two particular success stories. The first story is that of Michelin. Michelin had been dealing with increased competition and therefore created the foundation of a global delivery model for information systems. In order for this to work, they needed to train 200 workers across three continents. First, Michelin tried using traditional training methods, but they failed. As a result, Michelin looked to use virtual technology to train their employees. They settled on Second Life due to Second Life's offering of a virtual classroom and an arena for the participants to train further. To start off the training, the workers took a virtual class where they had pre-made avatars. After the class, the trainees used their knowledge to create proposals for potential business targets without exceeding budget constraints. The training was considered a success due to a low cost of about $100,000 and taking much less time than the traditional training methods. In the other example, the Kansas University Medical Center (KUMC) developed a virtual training simulator for training nurses in an induction sequence. The training focused on the sequence of events of preparing a patient for surgery. While the training was supposed to teach the students the sequence of events of the induction process, an unexpected benefit also occurred. Since the hospital room was modeled accurately, the students did not just learn the processes of induction, they learned about the layout of the hospital. The students reported they felt the training was very effective, and they felt the training would be useful for future students.

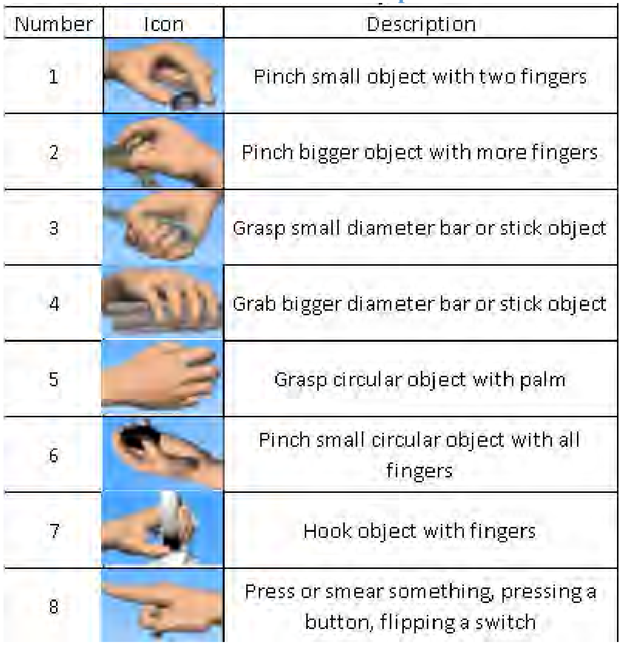
2.3 Presence, Immersion, and Engagement

Those simulations are beneficial for helping workers learn how to use tools or just to follow instructions in general. They combine uses of levels which increase in difficulty with a feedback scoring system to aid in the trainees' learning. But, trainees need more than just a simulation to really learn using virtual reality. If a user of these simulations does not feel immersed in the environment and only sees it as a simple game they may not learn from it. However, if the trainee is able to block out their surroundings and hone in on the training they are going through, their chances of learning from it increases dramatically. This section focuses on this idea of feeling immersed in the environment. It discusses the ideas of presence and immersion and how these ideas can aid in learning.

Witmer and Singer (1998) define presence in virtual worlds as "the sense of being there (in a virtual environment), even when one is physically situated in another place." When training someone, it is important that those being trained feel a sense of presence. Unfortunately, the sense of presence appears to be a very subjective feel which is based on the opinions of the trainees. Sanchez-Vives and Slater (2005) reported that this view of presence as being subjective has led to the widespread use of self-report of user experience. The danger with this type of reporting is that a tester has no way of knowing whether or not the trainee is being truthful. This leads to skepticism of the effectiveness of a test and of virtual reality training in general. Sanchez-Vives and Slater (2005) help with this problem by offering an idea of when presence occurs. According to them, "Presence occurs when there is a successful substitution of real sensory data by computer generated sensory data, and…the person responds to the virtual stimuli as if they were real." The ways to test for this is discussed in the next section. For now, a project dealing with presence is discussed.

One type of virtual reality project related to the idea of presence is a virtual reality environment used for oil-field safety training. Liu and Hao (2004) designed the world. One particularly interesting feature is their use of a disembodied hand as an avatar. The hand is seen as if it belongs to the trainee, as the trainee sees the hand and the world from a first-person point of view. The trainee chooses from several options of hand shapes, the best hand shape to grab the tool they needs. Table 2.1 shows the different hand shapes and a description of each one.

Table 2.1: Hand Shapes



Number	Icon	Description
1		Pinch small object with two fingers
2		Pinch bigger object with more fingers
3		Grasp small diameter bar or stick object
4		Grab bigger diameter bar or stick object
5		Grasp circular object with palm
6		Pinch small circular object with all fingers
7		Hook object with fingers
8		Press or smear something, pressing a button, flipping a switch

This attention to detail is important for simulating the methods of using small tools that workers use in real life. While the necessity of this amount of detail depends on what type of simulation is being done, the thought is the trainee must feel more excited, and therefore immersed, by the detailed hand grasps presented in Table 2.1.

Immersion plays a key role in how much presence a trainee feels while they are training. A couple common tools used to immerse trainees are head-mounted displays (HMD) and immersive rooms. As previously stated, the former were introduced in the 1970's. However their use has not gone away. Holm and Priglinger (2008) have designed a simulator which involves an HMD for refinery workers. The training consists of two computers, one which projects the virtual environment to the HMD and another which is controlled by the tester. The tester can trigger occurrences to the trainee forcing them to react. The tester can then assess the performance of the trainee.

The immersive room is a more recent idea being used. This consists of a room surrounded by monitors or screens which project images giving the trainee the feel that they are in a room that they are not actually in. The trainee is surrounded by projections which are supposed to give the feeling of presence. The Virtual Environment Radiotherapy Training (VERT) is an award winning innovation. VERT consists of an auditorium with a large screen that projects the images to the trainee. The trainee can walk on the stage and perform tasks which are assigned as part of the training.

A huge benefit of an immersive environment like the one proposed by VERT, is the ability to train as long as you want without endangering someone's life. VERT is meant to be a training tool for radiotherapy. When dealing with medicine, mistakes can be deadly. Therefore, the ability to train without worrying about mistakes being critical is very useful for selling the idea of virtual reality-based training. As Liang (2007) says, “Immersive VR is able to provide a rich, interactive and engaging training context that in reality would be too dangerous, too expensive or simply impossible to access.” Nowhere is this more evident than in an evacuation setting. Orr et al. (2008) created a simulation for mine evacuation. The simulation included “smoke that significantly obscures the trainees' VR vision in some areas” and moments when the ground would collapse under the avatars' feet. Obviously, this type of training would be unrealistic in real life but in a virtual environment it is very plausible.

While the previous simulations focus on visual and audio feedback to give the feeling of presence, Ruffaldi et al. (2009) expands on those senses by adding an element of touch. The training simulation is meant for competitive rowers. The rowers train by using oars in a simulated environment. While the trainee is rowing, large fans are used to simulate the force feedback that would be felt if rowing in the water. This way, the trainee can practice in an environment that is as similar as possible to rowing in actual water. The effectiveness of this training and the other ones like it is dependent on how engaged the trainee is in the training.

Engagement is a feeling one gets while interacting with someone or something. If presence is the “sense of being there,” and immersion is interacting in a world while losing sense of where one is, then engagement occurs when one feels so involved in a simulation that they feel a sense of control of the simulation and also lose track of time (Cooper 2010). In Karen Cooper's article “Go with the Flow,” Cooper references two authors to define engagement. Astin (1984) defines engagement as the amount of physical and psychological energy that the student devotes to the academic experience. Hornik (2008) says student engagement may be associated with increased time on task, and the development of deep learning, resulting in better classroom performance. While the relationship between engagement and immersion with presence and learning cannot be overstated, it is important to know how much of these factors does it take to produce an adequate amount of presence. Bowman and McMahan (2007) advocate "investigating multiple components of immersion simultaneously with multiple levels per component while still maintaining a high degree of experimental control." According to that logic, designing a highly interactive training method yields the best learning results. However, there is a need for tests which give credibility to the effectiveness of a training tool and to the notion that the trainees feel a sense of presence. The next section mentions several tests performed to give credibility to the effectiveness of their respective projects.

2.4 Testing

In the previous sections, several projects related to virtual reality have been discussed. However, these projects tend to be experimental in that they are innovations which are believed to be helpful in aiding in learning. Unfortunately, they lack data to support their assumptions. This section includes projects completed using objective measures. They fall in one of three categories: comparisons of two types of tools, questionnaires, and statistical test results.

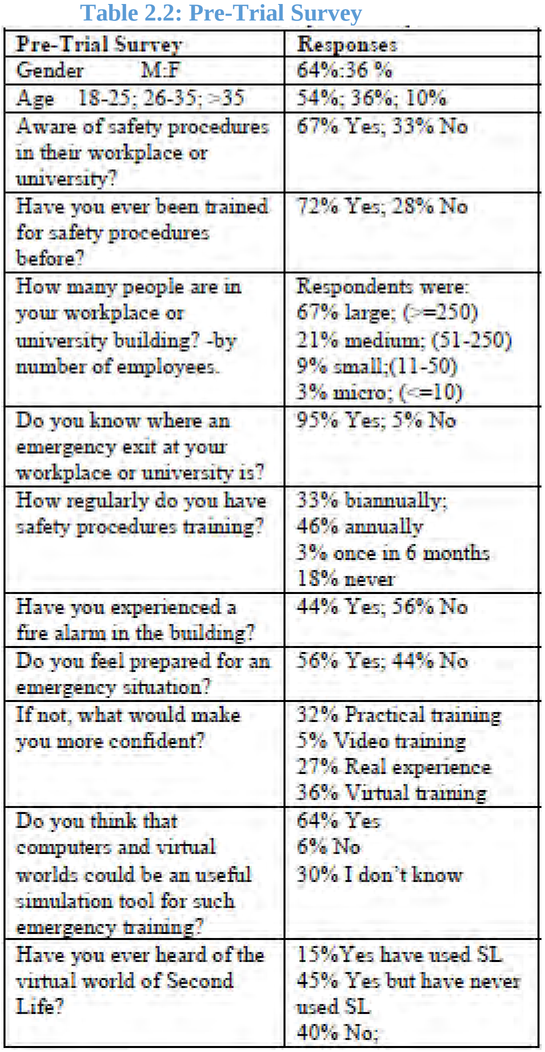
Comparisons are an important aspect of determining the effectiveness of a new innovation at least in a relative sense. In order to convince a market that it should use a different technology than it has been using for a long time, the market needs to see that there is improvement from the old way. The best way to do this is to use Hypothesis testing which can be used to determine whether there is a significant improvement by a new technology over an old technology. An example of this is by Gruchalla (2004) who compared the effectiveness of the task of oil well path tracking using a desktop computer with that of a CAVE. Despite his attempts to make the two mediums the same in respects to the program and screen resolution, the task was performed better in the CAVE technology.

Sometimes though, while technology seems like it should make things better, the results do not turn out to be what was originally expected. Datey (2001) also used a desktop in a comparison. In this case, he was comparing it to an HMD. The test was for information visualization tasks with spatial components but Datey came to the conclusion that there was no statistical difference in the two methods. In a similar case, Pausch et al. (1997) predicted that an HMD with head tracking would be better for locating items than a stationary HMD with hand controls. However, there was no statistical advantage either way. These tests are important for companies potentially investing in developing innovative software or products. The chance of losing money selling a product not any more effective than the current technology could be very high.

While these methods compare different tools to determine which one is the effective option, other methods test to determine if their virtual reality tools can be used for training. One of the most common types of tests is evacuation simulation. As stated before, virtual simulations can provide situations which would be impossible to simulate in real life. For instance, there are health risks involved in putting people in an evacuation setting with real fire. At the same time, a trainee might not feel a sense of urgency if their training is just walking around in a safe building. Orr et al. (2008) created a simulation for evacuating an underground mine fire as a response to this issue. Thirty-two people in groups of four went through the training simulation. Half of the groups were given a sample route to examine first as a way to introduce them to the virtual environment. They were then given a tougher scenario to go through. The other group was given the tougher route first. Those that were given the easiest route first performed 37% faster than the other group. With this, it was concluded that the virtual training aided the trainees in completing the route. While there were issues about whether or not the trainees were thinking for themselves (since they worked in groups and one could easily just follow the leader), a majority of the trainees reported that they felt the training was effective.

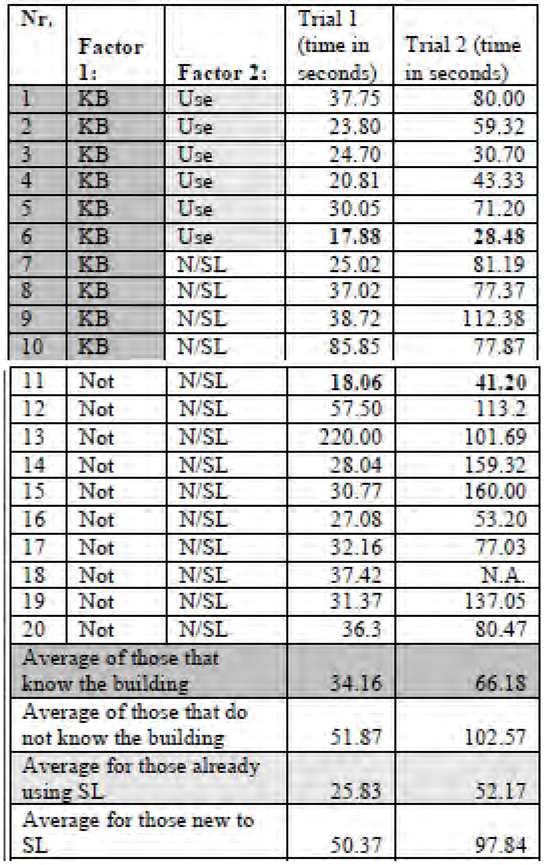
In a separate evacuation training program, Molka-Danielson and Chabada (2010) created a replica of the first floor of a university college building in Second Life to use for an evacuation simulation. Their hopes were that this simulation "could contribute to an evacuation plan for the college and to more effective evacuation training exercises by raising interest." The simulation included putting an individual into a burning building and timing their ability to escape the building through a predetermined exit. Tables 2.2 and 2.3 show a pre-simulation survey taken by the participants and the results of the simulation respectively.

Table 2.2: Pre-Trial Survey



Pre-Trial Survey	Responses
Gender M:F	64%:36 %
Age 18-25; 26-35; >35	54%; 36%; 10%
Aware of safety procedures in their workplace or university?	67% Yes; 33% No
Have you ever been trained for safety procedures before?	72% Yes; 28% No
How many people are in your workplace or university building? -by number of employees.	Respondents were: 67% large; (>=250) 21% medium; (51-250) 9% small;(11-50) 3% micro; (<=10)
Do you know where an emergency exit at your workplace or university is?	95% Yes; 5% No
How regularly do you have safety procedures training?	33% biannually; 46% annually 3% once in 6 months 18% never
Have you experienced a fire alarm in the building?	44% Yes; 56% No
Do you feel prepared for an emergency situation?	56% Yes; 44% No
If not, what would make you more confident?	32% Practical training 5% Video training 27% Real experience 36% Virtual training
Do you think that computers and virtual worlds could be an useful simulation tool for such emergency training?	64% Yes 6% No 30% I don't know
Have you ever heard of the virtual world of Second Life?	15%Yes have used SL 45% Yes but have never used SL 40% No;

Table 2.3: Trial Results

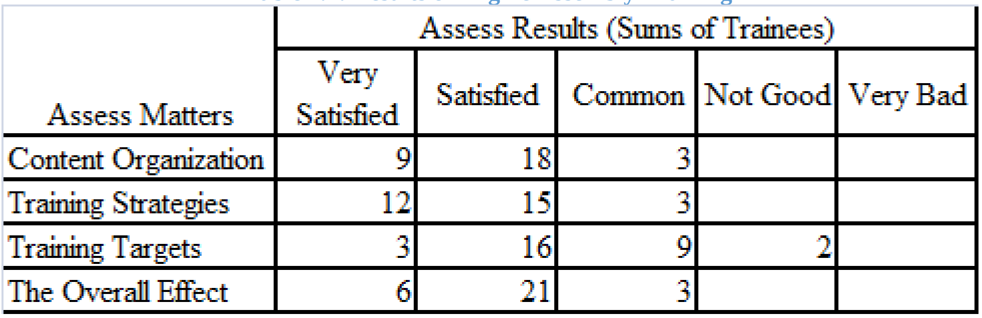


Nr.	Factor 1:	Factor 2:	Trial 1 (time in seconds)	Trial 2 (time in seconds)
1	KB	Use	37.75	80.00
2	KB	Use	23.80	59.32
3	KB	Use	24.70	30.70
4	KB	Use	20.81	43.33
5	KB	Use	30.05	71.20
6	KB	Use	**17.88**	**28.48**
7	KB	N/SL	25.02	81.19
8	KB	N/SL	37.02	77.37
9	KB	N/SL	38.72	112.38
10	KB	N/SL	85.85	77.87
11	Not	N/SL	**18.06**	**41.20**
12	Not	N/SL	57.50	113.2
13	Not	N/SL	220.00	101.69
14	Not	N/SL	28.04	159.32
15	Not	N/SL	30.77	160.00
16	Not	N/SL	27.08	53.20
17	Not	N/SL	32.16	77.03
18	Not	N/SL	37.42	N.A.
19	Not	N/SL	31.37	137.05
20	Not	N/SL	36.3	80.47
Average of those that know the building			34.16	66.18
Average of those that do not know the building			51.87	102.57
Average for those already using SL			25.83	52.17
Average for those new to SL			50.37	97.84

Table 2.2 provides a view of the results of the pre-trial survey. It provides an example of how to report the sample data and provides a model for a questionnaire prior to using a simulation. Prior to the administration of the training, the trainees were split on virtual training and more common practical training, but a majority felt computers would be helpful. Table 2.3 shows the effect of virtual training. ‘KB’ means the user knew the building whereas ‘Not’ means they were unfamiliar with it. ‘Use’ means the user was familiar with Second Life and ‘N/SL’ means the user was not familiar with Second Life. Trial 1 and 2 represent different routes with the second one being more difficult. The times recorded for those who were trained with Second Life showed they were able to complete both escape routes in virtually half the time as those who went without it. This type of comparison is useful for determining if virtual training can be useful. From this simulation, it is clear that virtual reality had a significantly positive effect.

While the previous methods use statistical comparisons to confirm the effectiveness of a certain technology, others use a more subjective approach. Instead of comparing, they use questionnaires. These questionnaires are given after a user practices with a technology. The user answers the questions which usually are along the lines of, “To what level did you feel a sense of presence?” The tester than analyzes all the users according to their responses and draws conclusions on them. One of the most common questionnaires is the University College London (UCL) questionnaire (Slater et al., 1994; Usoh et al., 1999). This questionnaire contains seven questions measuring presence, three measuring behavioral presence, and three measuring ease of locomotion. Cheng et al. (2010) use a questionnaire to analyze the effectiveness of their virtual car engine assembly training. They used thirty trainees consisting of twenty employees from three different companies and ten undergraduate engineering students. The results of the questionnaire are shown in Table 2.4:

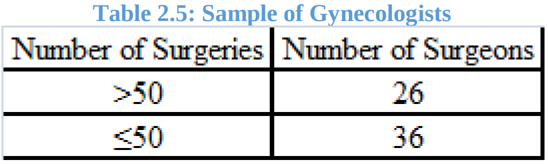
Table 2.4: Results of Engine Assembly Training



	Assess Results (Sums of Trainees)				
Assess Matters	Very Satisfied	Satisfied	Common	Not Good	Very Bad
Content Organization	9	18	3		
Training Strategies	12	15	3		
Training Targets	3	16	9	2	
The Overall Effect	6	21	3		

Table 2.4 shows how a strong majority of the sample felt the virtual training was a very good method. In a similar test, Bajka et al. (2008) used a questionnaire to evaluate their virtual simulation of a hysteroscopy. They took sixty-two gynecologists and put them all through twenty minutes of hands on virtual training. The sample of surgeons can be summarized in Table 2.5.

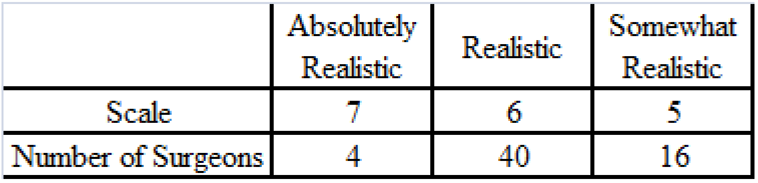
Table 2.5: Sample of Gynecologists



Number of Surgeries	Number of Surgeons
>50	26
≤50	36

These participants were asked to rate on a 7-point Likert scale, how effective they felt the simulation was. Table 2.6 shows the results of the questionnaire but only includes the first three levels of the scale, as no one rated the simulation worse than a "5."

Table 2.6: Likert Scale



	Absolutely Realistic	Realistic	Somewhat Realistic
Scale	7	6	5
Number of Surgeons	4	40	16

Those that were less experienced rated the training at an average of 6.48 while the more experienced surgeons rated the training at an average of 6.08. For this study, 95.2% believed that the training was adequate, and 85.5% suggested the training to be given to all inexperienced surgeons.

2.5 Physiology

While the previous authors claim their tests give credibility to the effectiveness of virtual reality, they mostly fail to cover the idea of physiological factors. The test results are from subjective questionnaires asking questions like, "Did you feel a sense of presence?" Even if the trainee answers to the highest level possible depending on the scale, there is no real proof if they truly felt immersed in the virtual environment. A true test on physiological response is still missing.

While testing has been scarce, the ability to test for physiological responses is well within the realm of possibility. Slater et al. (2010) say that when a trainee is going through a virtual training environment, they should have physiological responses and should show behavior that supports the idea that they are immersed in their environment. The experimenter just needs to know what to test. Stress tends to be the most common sensation tested. Durrani and Geiger (2008) claim that the introduction of stress into an experiment causes an increase in engagement and subsequently, in training effectiveness. If that is true, physiological factors correlating to stress can be tested. Slater et al. (2009) states three factors which correlate to stress are skin conductance, heart-rate, and heart-rate variability. Meachen et al. (2002) used these factors, sans heart-rate variability, to study stress using a visual cliff. In something that can be used in the future, Durrani and Geiger (2008) propose a statistical approach to study knowledge transfer and skill development by comparing virtual training and traditional training based on physical cue fidelity and neurophysiologic response.

2.6 Future Projects

Durrani and Geiger's approach is simply a proposal. There are a few more projects underway which relate to virtual trainings and presence. For instance, Monahan et al. (2009) want to develop training environments for emergency preparedness training exercises. According to them, there are two alternatives to virtual training: live exercises and tabletop exercises. Live exercises saw an increase in use for emergency training following the September 11th attacks on the World Trade Center and the Pentagon. Unfortunately they face the drawbacks of being costly and needing a large number of volunteers to help train a small number of people. Tabletop exercises suffer from lack of visual stimulation. Virtual worlds do not suffer from these drawbacks as the cost of transportation and loss of work days can be solved by performing trainings in a virtual world in which a person can connect to wherever they are located. Additionally, virtual worlds offer a visual stimulus not present in tabletop exercises. Therefore, Monahan et al. believe the use of virtual worlds for emergency training is more cost efficient and more stimulating than the possible alternatives.

Another future endevor is being conducted by Mott and Rajaei (2010). Their goal is to design a system which uses a standard webcam and advanced computer vision techniques to allow the learners use of their hands when interacting with the virtual objects. This approach is a response to the use of virtual gloves which are used for the same effect. However, they feel this is a less expensive alternative. Cai (2008) wants to create a 3D environment that simulates the real business world while at the same time provide interfaces to 3D application users for them to interact with the virtual world. These future projects should add to the available literature on virtual reality. Even without proper testing procedures, the literature can still provide useful information.

2.7 Conclusion Regarding Literature

Because virtual reality is a fairly recent innovation, there has not been a large amount of research done on the topic, as it relates to learning effectiveness. While there have been a fair share of simulations using virtual reality, they have lacked an adequate amount of testing to give credibility to their effectiveness. Creators of devices, trainings, or simulators using virtual reality tend to confirm the effectiveness of their creations with assumptions or subjective questionnaires. Objective evidence of the effectiveness of virtual reality is still lacking. A mixture of test data and physiological data can go a long way in giving credibility to virtual reality as a testing method. Bronak at al. (2006) state, "We believe that virtual worlds support deep learning and can help learners make meaning in ways similar to outside…environments. Our experience suggests that virtual worlds offer participants a sense of presence, immediacy, movement, artifacts, and communication unavailable within traditional Internet-based learning environments." Now the only thing to do is attain data that can give credibility to this assertion.

Methodology

3.1 Experiment Overview

In order to verify if virtual worlds can be as effective as more traditional methods, a comparative study is created in which two instruction types, Second Life and a traditional set of paper-and-pen exercises called case studies, are used. The subjects in the study are evenly separated into the two groups and then tested on the final day of the experiment. The results from this final assessment are used to determine if one of the instruction types is more effective than the other. The procedure for this experiment is outlined in Figure 3.1.

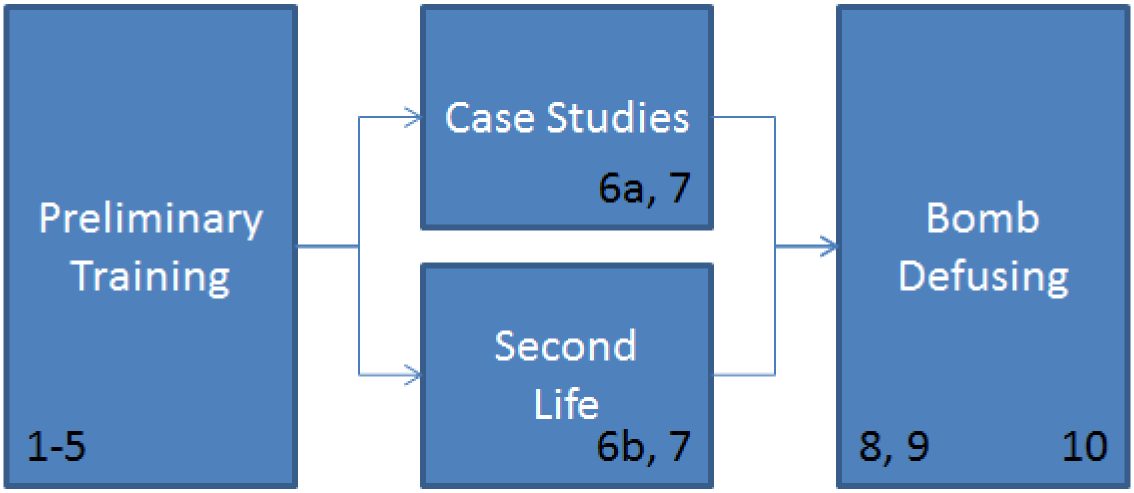
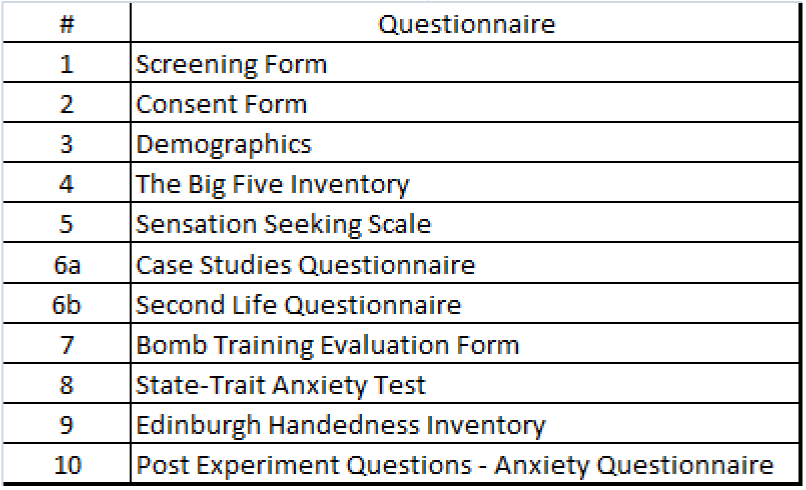


Figure 3.1: Block Diagram of Experiment

On Day 1, the subjects go through a preliminary training where they learn about improvised explosive devices (IED's). The subjects learn how IED's work and the damage they can cause. In addition they learn the procedures for defusing a bomb. In Day 2 the subjects are separated into two groups: a paper and pen training group using classical case studies and a virtual environment training group using the Second Life software. In this part of the training, the subjects apply what they learned in Day 1 to simulated situations and are provided with feedback in order to gain experience. The difference is in the medium in which they experience the simulations. On the final day, the subjects, one at a time, apply what they have learned in a live scenario with a victim played by an actor or actress. They are given a time limit and are graded on how well they perform the necessary procedures.

In addition to the training, each subject fills out a series of questionnaires throughout the course of the experiment. The numbers on the bottom of the blocks in Figure 3.1 refer to the questionnaires completed during the experiment. The placement of the numbers (left and right) indicates when in the day the subjects take the questionnaires (before and after the training respectively). The questionnaires are listed below in Table 3.1.

Table 3.1: List of Questionnaires



#	Questionnaire
1	Screening Form
2	Consent Form
3	Demographics
4	The Big Five Inventory
5	Sensation Seeking Scale
6a	Case Studies Questionnaire
6b	Second Life Questionnaire
7	Bomb Training Evaluation Form
8	State-Trait Anxiety Test
9	Edinburgh Handedness Inventory
10	Post Experiment Questions - Anxiety Questionnaire

The details of the experiment are organized by the day in which the activities occur. First is the preliminary training in Day 1.

3.2 Day 1: Preliminary Training

At the start of the experiment the subjects are introduced to the experiment’s conductors and are told about how the payment for their time is administered. The subjects are also warned not to try to utilize this training in a real situation. The material presented to them is for research purposes only, and it lacks enough depth to fully qualify the subjects as bomb experts. After the initial introduction, each subject is required to complete the Screening Form. This form is used to screen out any subjects that would have a high chance of being harmed by the content of the experiment or subjects that do not meet the age or school registration requirements. After the screening is completed, the subjects are given the Consent Form in which they grant permission to be used as human subjects. The nature of the experiment is described and the subjects are told they may opt out of the experiment whenever they wish. After signing the Consent Form the subjects are given note cards with their Subject Numbers and their log-in information. The numbers are provided randomly and range from 1 to 65. With their note card, each subject logs into the experiment’s web page and begins their preliminary training by completing additional questionnaires.

The three questionnaires completed by the subjects prior to beginning their training are the Demographics Questionnaire, The Big Five Inventory, and the Sensation Seeking Scale. The Demographics Questionnaire asks questions about their history with Second Life along with some general questions about age and college grade level. The Big Five Inventory is a personality test that rates subjects based on five factors: openness, conscientiousness, extraversion, agreeableness, and neuroticism. The Sensation Seeking Scale is another personality test. It is used to assess the subjects’ tendency to engage in spontaneous and potentially dangerous behavior. After taking the questionnaires, the subjects can move on to the actual training.

3.2.1 Preliminary IED Training

To begin the training, the subjects read through a file that contains seven links. The first link contains some basics about explosives. The second gives information about how bombs cause damage. The third provides the subject with an option to open a file containing information on render safe procedures as either a Word document or a PDF file. The fourth link opens a PDF document where the render safe procedures are detailed. After reading the PDF document, the subjects open a Soft Chalk file in the fifth link which has more information regarding the effects of IED explosions and the process that initiates the explosive. Using this knowledge, the subject learns how to defuse an IED by learning the components of the bomb and the order to defuse it without accidently causing it to explode. The subject is also given several matching exercises to practice learning the components of IED's and the procedures for defusing explosives. Additionally, the subjects are given note cards with the render safe procedures typed out in order to help them remember the procedures. After they have completed the training, the subjects take an assessment test in the sixth link to assess how well they learned the material.

3.3 Day 2: Case Studies or Second Life Training

On the second day, the subjects are split into two groups: a paper and pen training group (case studies) and a virtual environment training group (Second Life software). The subjects in the paper and pen training group are the ones that received an odd number for their subject number. The even-numbered subjects take the virtual environment training. The two types of training are administered separately. Therefore they are discussed in individual sections aptly labeled Case Studies and Second Life respectively.

3.3.1 Case Studies

Before starting the case studies, the subjects are given a chart which helps them evaluate the blast radius of an IED and the corresponding distance that is safe for evacuation. The subjects are presented with three case studies. In each case, the subject is provided with information about the situation. The subject completes a series of matching exercises and multiple choice questions. After answering the questions, the subject assesses how the explosive looks and proceeds to draw the explosive and label its components. The subject completes the case by completing an ordering exercise in which they label each wire of the explosive in the order it is to be cut. After each case study is completed, the subjects are provided with the answers to the questions and then the next case begins. The three cases are described below:

3.3.1.1 Case Study 1: Unvacated Business Building

Early Thursday morning, a security officer of BioPharm Corp. finds a bomb on the first floor stairway while doing a routine security walk through of the 30 story building. The building is located in the center of downtown Dallas, TX. Upon discovering the explosive, the officer contacted 911 for help. This is where the subject comes into the case study. When they arrive, the subject is met by the security officer who provides them further details.

Details

Bomb Location: Hidden Behind first floor Stairway

Civilians: The building cannot be totally evacuated. There are a number of labs that require personnel to remain in order to protect the community.

Possible Suspects: Police reinforcements are patrolling the surrounding area and reviewing security tapes, but have not found any suspects.

Bomb Visual: The case study is accompanied with pictures of the bomb shown in Figure 3.2.

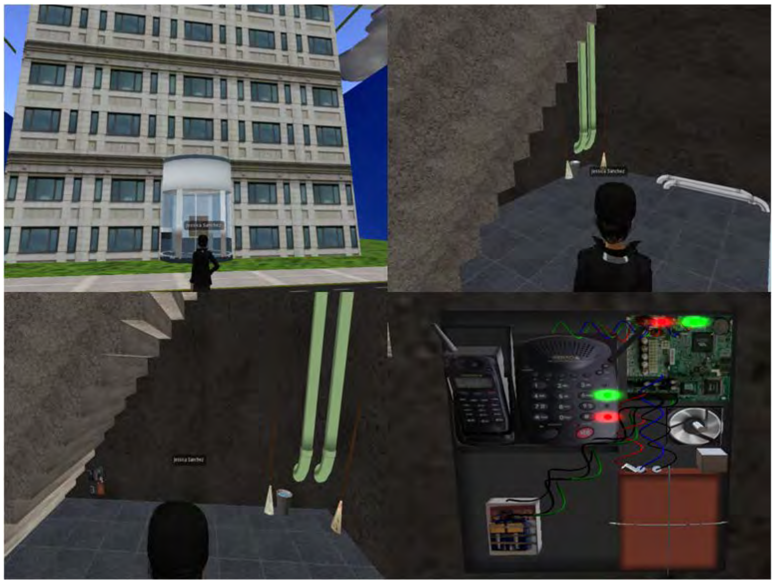


Figure 3.2: Bomb in Case Study 1

3.3.1.2 Case Study 2: Vacated Bus on Active Highway

A 60 passenger Greyhound bus which left from Houston, Texas en route to the McAllen Bus Station has stopped on the side of Highway 281 near by Monte Cristo. One of the passengers in the bus, having moved to a recently vacated seat, discovered a box with a bomb inside. When you arrive you are greeted by a police man who notifies you that there are current efforts on the way to block the highway.

Details

Bomb Location: Found on top of 3rd row seat

Causality Risks: Bus emptied, however cars are still occupying the road by the busses current location.

Possible Suspects: Passenger previously occupying the seat where the bomb is located exited on the previous stop. All efforts are being made to locate suspect.

Bomb Visual: The case study is accompanied with pictures of the bomb shown in Figure 3.3.

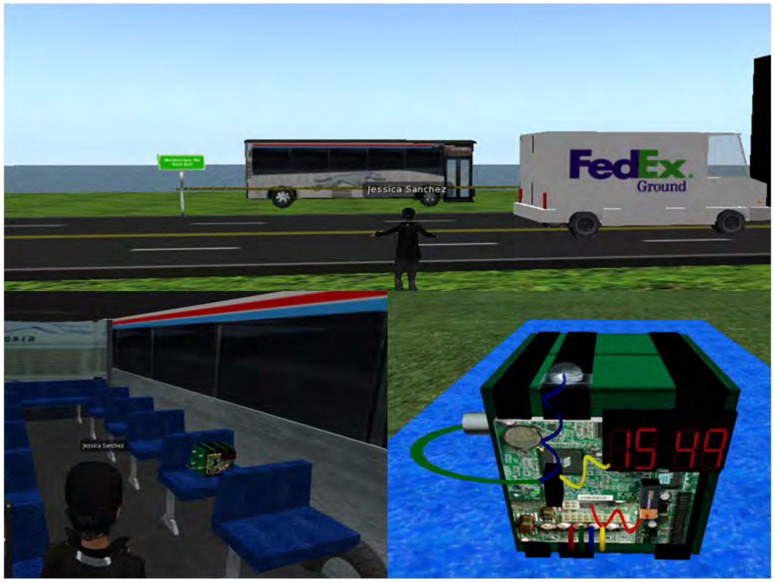


Figure 3.3: Bomb in Case Study 2

3.3.1.3 Case Study 3: School Janitor's Room

In the afternoon of April 5th, 2009, the Janitor of Edinburg Junior High went to the Janitor's room to retrieve a mop in order to clean up a spill that took place in the cafeteria. When reaching for the mop, the Janitor noticed a foreign object located on the floor of the room. Upon closer inspection he realized it was a bomb and quickly ran to the Administration office to notify the principal. The principal then notified authorities, which luckily had located you there for your son's Career Day. You are quietly asked to step out of the room and then notified of the situation by the principal.

Details:

Bomb Location: On 1st floor Janitor Room, placed on the floor.

Causality Risks: The school has barely begun the evacuating process of the Jr. High School. You are told that on fire drills it takes approximately 10 minutes for school to be evacuated.

Additional Information: Janitor had noticed that the timer seemed to stay on 15 minutes and 49 seconds. But unsure if that is a decoy. There is also a woman tied up near the bomb.

Suspects: No known suspects.

Bomb Visual: The case study is accompanied with pictures of the bomb shown in Figure 3.4.

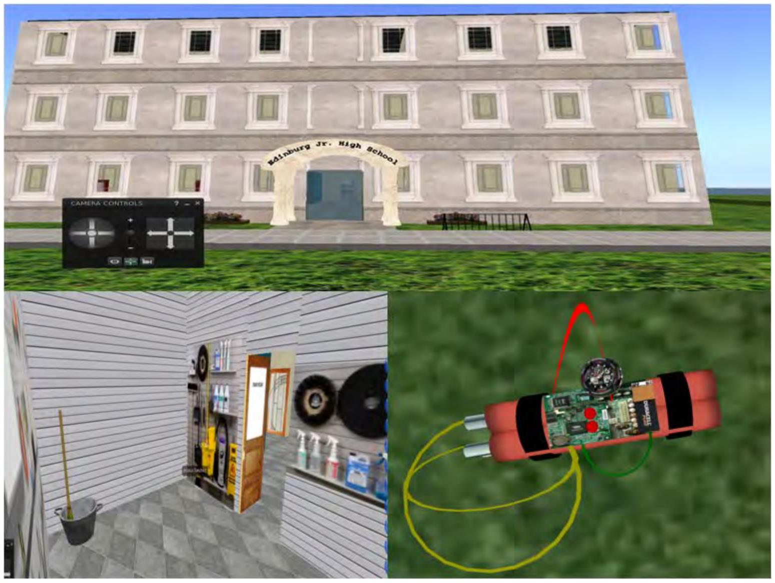


Figure 3.4: Bomb in Case Study 3

3.3.1.4 Questionnaires

Following the completion of the three case studies, the subjects complete two questionnaires. The first is the Case Study Questionnaire which assesses the effectiveness of the Case Studies in properly training the subjects. The second questionnaire is the Bomb Training Evaluation Form located in the seventh link of the file from the first day of training. This form assesses the effectiveness of the entire training according to the subjects. After they complete their questionnaires, the subjects are now ready to take their final test.

3.3.2 Second Life

Whereas the odd-numbered subjects participated in the case studies, the even-numbered subjects participate in the Second Life training. The Second Life training consists of the same scenarios used in the case studies and in the same order. However, the beginning of the training begins with a tutorial on how to use some basic functions of Second Life. Using the log-in information on their cards, the subjects enter the Second Life virtual world. The subjects start at the outside of a training facility where they first learn how to walk. A screen shot of this tutorial is shown in Figure 3.5.

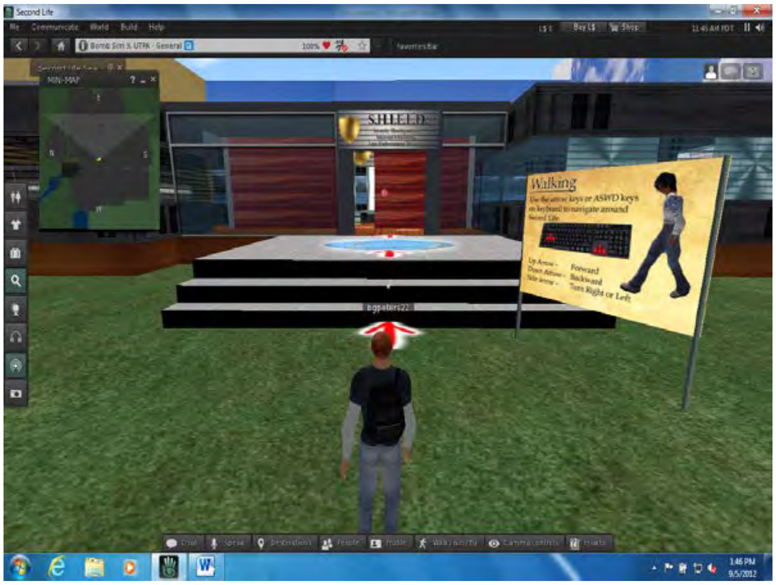

Second Life
Me Communicate World Build Help
SHIELD
Walking
Use the arrow keys or ASWD keys on keyboard to navigate around Second Life
Up Arrow - Forward
Down Arrow - Backward
Side arrow - Turn Right or Left

Figure 3.5: Second Life Walking Tutorial

After walking through the door, the subjects learn how to touch objects and zoom in and out to get a better view of them. Figure 3.6 and Figure 3.7 show this part of the tutorial.

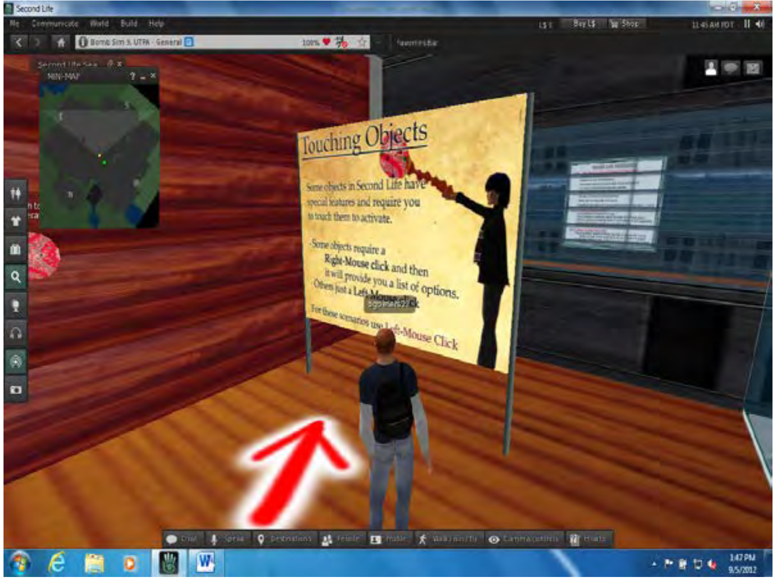
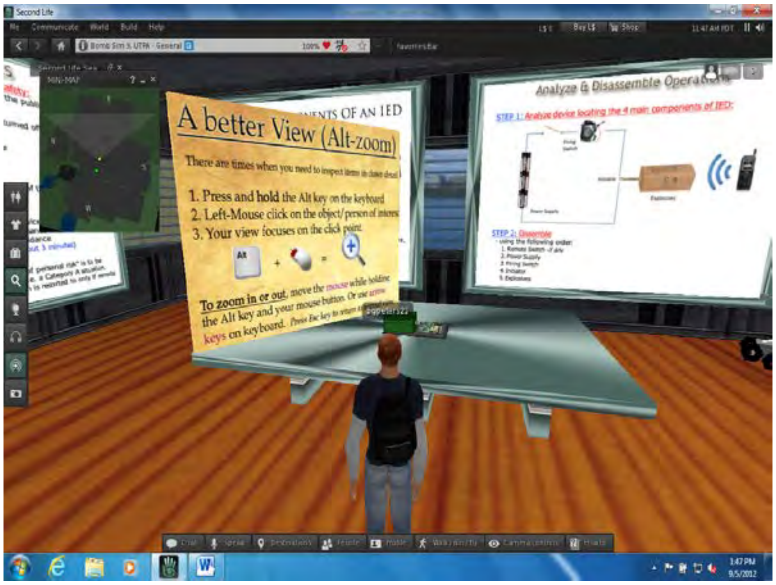

Second Life
Me Communicate World Build Help
Touching Objects
Some objects in Second Life have special features and require you to touch them to activate.
-Some objects require a Right-Mouse click and then it will provide you a list of options.
-Others just a Left-Mouse click
For these scenarios use Left-Mouse Click

Figure 3.6: Second Life Touching Objects Tutorial


Second Life
Me Communicate World Build Help
Bomb Sim 3, UTPA - General
100%
Favorites Bar
L$ Buy L$ Shop
11:47 AM PDT
MINI-MAP
Firing Switch
Power Supply
Initiator
Explosives
Chat Speak Destinations People Profile Walk / run / fly Camera controls Howto
1:47 PM
9/5/2012

A better View (Alt-zoom)

There are times when you need to inspect items in close detail

1. Press and hold the Alt key on the keyboard
2. Left-Mouse click on the object/person of interest
3. Your view focuses on the click point

Alt + =

To zoom in or out, move the mouse while holding the Alt key and your mouse button. Or use arrow keys on keyboard. Press Esc key to return

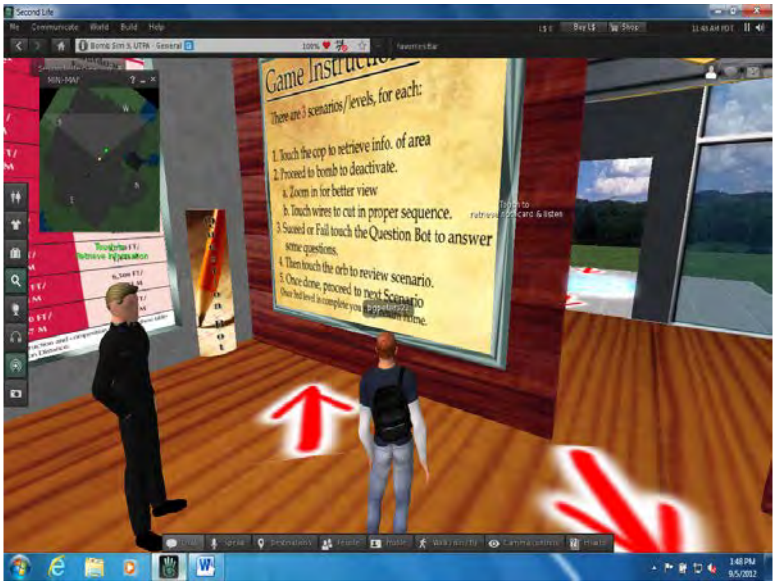
Analyze & Disassemble Operations

STEP 1: Analyze device locating the 4 main components of IED:

STEP 2: Dissemble
- using the following order:
1. Remote Switch - if any
2. Power Supply
3. Firing Switch
4. Initiator
5. Explosives

Figure 3.7: Second Life Zoom Tutorial

Using the zoom feature, the subjects can look at an IED and practice defusing it by clicking on the wires. After practicing with the IED, the subjects look at sign with instructions for progressing through the scenarios. A screenshot of this instructional sign is shown in Figure 3.8.


Second Life
Me Communicate World Build Help
Bomb Sim 3, UTPA - General
100%
Favorites Bar
L$ Buy L$ Shop
11:48 AM PDT
MINI-MAP
Touch to retrieve notecard & listen
Chat Speak Destinations People Profile Walk / run / fly Camera controls Howto
1:48 PM
9/5/2012

Game Instructions

There are 3 scenarios/levels, for each:

1. Touch the cop to retrieve info. of area
2. Proceed to bomb to deactivate.
a. Zoom in for better view
b. Touch wires to cut in proper sequence.
3. Succeed or Fail touch the Question Bot to answer some questions.
4. Then touch the orb to review scenario.
5. Once done, proceed to next Scenario
Once 3rd level is complete you

Figure 3.8: Second Life Instructions

The instructions on the sign tell the subjects to first click on the police officer to gather information on the scenario. To practice this, a police officer is placed near the sign. If the subjects click on him, they are shown a window containing different types of information that the subjects can gather from the officer. This is shown in Figure 3.9. Next the subject answers the same multiple choice questions provided to the Case Study group. While a programmed ‘Question Bot’ was to be used for the questions, multiple problems led to the questions being typed on a Word document and displayed to the subject. After answering the multiple choice questions, the subject defuses the IED. After defusing the IED, the subject reviews the scenario before moving on to the next scenario. The review includes the proper evacuation distance, the threat level, the steps to ensure public safety, and the procedure to defuse the IED.

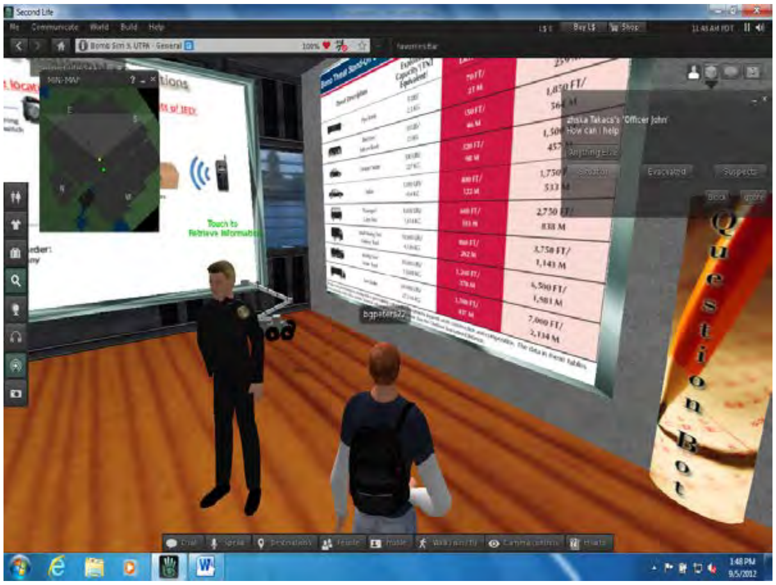


Figure 3.9: Second Life Police Officer Tutorial

After completing the tutorial, the subjects continue to the scenarios which are identical to those in the Case Studies.

3.3.2.1 Questionnaires

Following the completion of the Second Life training, each of the subjects complete two questionnaires. The first is the Second Life Questionnaire which assesses the effectiveness of Second Life in properly training the subjects. The second questionnaire is the Bomb Training Evaluation Form. This form assesses the effectiveness of the entire training according to the subjects. After the subjects complete their questionnaires, they are ready to for the final day of the experiment.

3.4 Day 3: Final Scenario

3.4.1 Preliminary Questionnaires

The final day of the training begins with the test subject taking the State-Trait Anxiety Inventory. In this inventory the subjects are given several statements. The subject rates themselves on a scale of 1 to 4 with 1 being that the statement does not describe them at all and 4 being that the statement describes them “very much so.” After taking the State-Trait Anxiety Inventory, the subject is verbally given the Edinburgh Handedness Inventory. In this questionnaire, the subject is asked whether they use their right or left hand for a series of everyday activities such as brushing their teeth or writing. The purpose of the questionnaire is to determine the subject’s dominant hand so not to interfere with attachment of electro-dermal sensors. After completing the questionnaires, the sensors for the physiological measurement detailed in the next section are placed on the subject.

3.4.2 Physiological Arrangement

The physiological equipment being used in this experiment is provided by BIOPAC Systems, Inc. The equipment consists of biotelemetry modules designed to receive signals from battery charged transmitters. The transmitters that are connected to the physiological transducers relay the signals wires that are either clipped or attached with an adhesive to the subject. The physiological measurement being taken is the subject's heart rate using an Electrocardiogram (ECG).

3.4.2.1 ECG

Figure 3.10 displays electrode placement for the ECG. First the skin is prepared by using an alcohol swab to clean off any excess dirt or oil from the skin. Next three pads are attached to the subject's chest: one under each pectoral muscle, and another underneath the right clavicle. A red, white, and black clip lead is used to set up the ECG. The red wire is clipped under the left pectoral muscle and the black wire is clipped under the right pectoral muscle. Finally, the white wire is clipped under the right collarbone. The three-wire lead is attached to an ECG transmitter and the transmitter is taped to the subject's shirt.

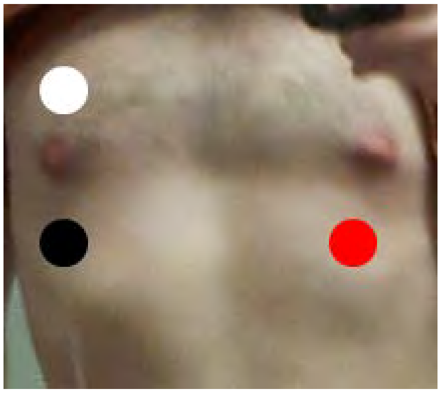


Figure 3.10: ECG Arrangement

3.4.3 Final Test

To begin the test, the subject sits still for ten minutes in order to get a preliminary baseline of their physiological measurements and to acclimate to the situation. After the ten minutes are completed, one of the researchers enters the room and presents the scenario to the subject. The subject reads the scenario and is given a notepad, a pen, and a flashlight as their resources. When they are ready, they are taken into the scenario room with a victim who has a mock bomb attached to them. The subject has 15 minutes to defuse the bomb. While the subject is defusing the bomb, they are graded on how they perform the actions they were trained to perform. After the scenario is completed, the subject has the sensors removed and is taken outside to take one final questionnaire. This questionnaire assesses the anxiety the subject says they felt during the scenario.

The completion of the final questionnaire marks the conclusion of the experiment. Following the conclusion of the experiment, all the data was collected and organized for analysis. In the next section, the data from this experiment is analyzed in order to draw inferences about whether one instruction type was more effective than the other at teaching the subjects how to defuse an explosive device.

Data Analysis

4.1 Outline of Data Analysis

As explained in the previous section, this experiment is performed over a three day span. Therefore, the data gathered from this experiment is presented in the order in which it is collected. An overview of the subjects' demographic information is provided first as are the results of the assessment the subjects complete following the conclusion of their first day of training.

On the second day of their training, the subjects are broken up into two groups and receive a specialized instruction based on their grouping. The even-numbered subjects received the Second Life instruction while the odd-numbered subjects received the case studies instruction. Following the completion of their training, each subject completes a questionnaire assessing their opinion of the effectiveness of their particular training. The results of the questionnaires are provided along with the result of the Bomb Training Evaluation Form, a questionnaire assessing the effectiveness of the training overall.

On the final day, each subject completes their final assessment. Prior to completing the assessment, the subjects each take the State-Trait Anxiety Questionnaire, a questionnaire designed to assess the difference between a subjects' normal anxiety and the anxiety the current situation is eliciting. The purpose of this experiment is to determine if the instruction type has an effect on the subjects' ability to learn and perform a task. The result of testing whether the instruction type affected the subjects' performance in terms of completing render safe procedures and correctly defusing the bomb is provided. In addition, a test is conducted to assess whether subjects that took one instruction type were calmer than the subjects that took the other. This is done by using physiological measurements recorded during the final day of the experiment. For this experiment, the measurements analyzed are the subjects' heart rates. Finally an overview of the subjects' perception of the final assessment is provided.

4.2 Day 1

4.2.1 Demographics

This experiment was limited to current male students of the University of Texas-Pan American. In addition, the students had to be eligible to work in the United States. Potential subjects that noted a history of epilepsy, seizures, post-traumatic stress disorder, or panic attacks were disqualified from the experiment. Sixty-five male participants participated in this experiment. On the first day of their training, the subjects were given a survey asking for demographic information as well as their familiarity with Second Life. The questions asked are as follows:

- Question 1: What is your age?
- Question 2: What is your classification?
- Question 3: Please indicate your current level of familiarity with computer games and gaming.
- Question 4: Please indicate your current level of awareness about Second Life, on a scale of 1 to 6.
- Question 5: Approximately how many months ago did you first use Second Life? (Please enter only whole numbers. For example, if you started using it 3 years ago, enter 36; if you have never used it before, enter 0):___months ago.
- Question 6: Please indicate how frequently you use Second Life for personal leisure or communication, on a scale of 1 to 6:

The responses for Question 1 are summarized in Figure 4.1.

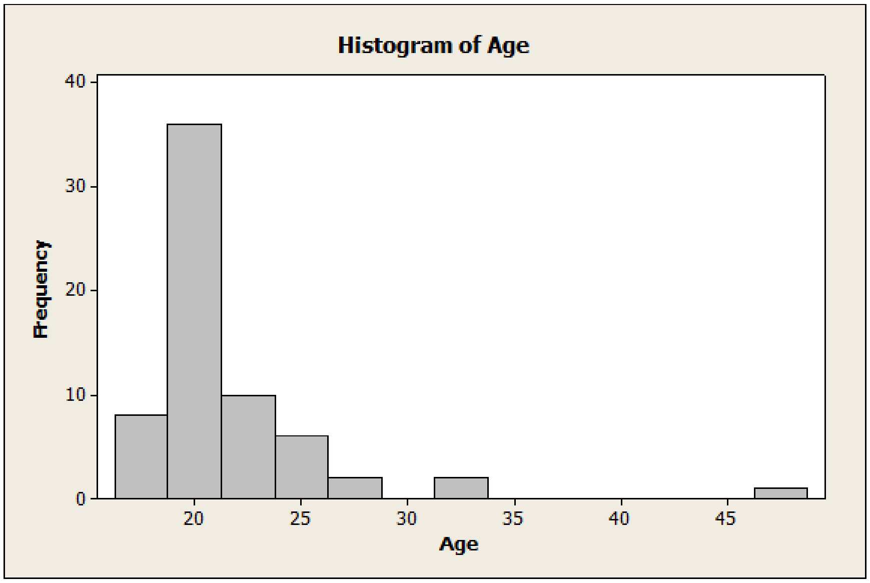


Figure 4.1: Subjects' Ages

Figure 4.1 shows that a strong majority of the subjects were between the ages of 18-20 which is expected as those are typically the years that people attend college. The classification of each of the subjects was examined next. Figure 4.2 summarizes the subjects' responses in regards to their college classification.

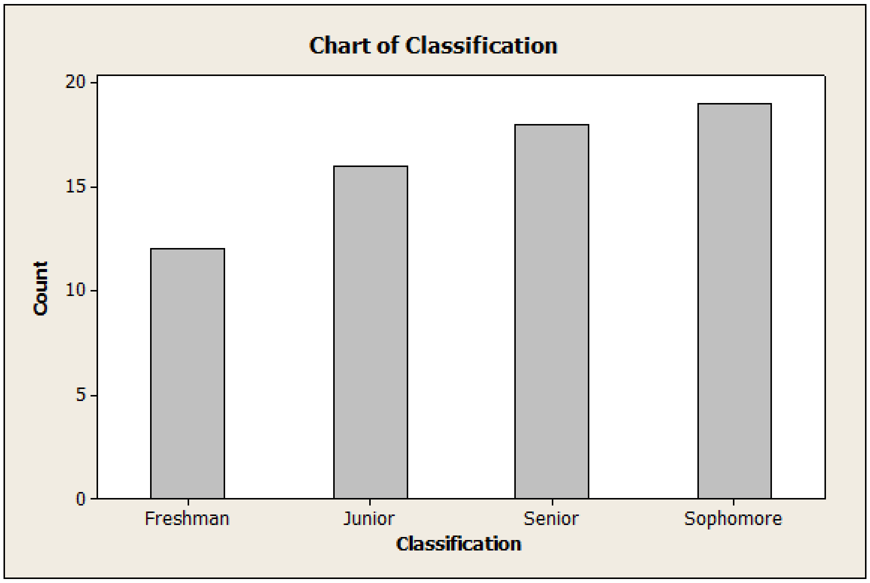


Figure 4.2: Subjects' Classifications

The distribution of classifications is almost uniform. Seniors and sophomores make up a majority of the subjects while there were an almost equal number of freshmen and juniors. Figure 4.3 displays how much familiarity with computer games the subjects have. They were asked to respond based on the following scale:

- 1: Never played computer games
- 2: Little experience with computer games
- 3: Some experience with computer games
- 4: Fair amount of experience with computer games
- 5: Moderate experience with computer games
- 6: Significant experience with computer games

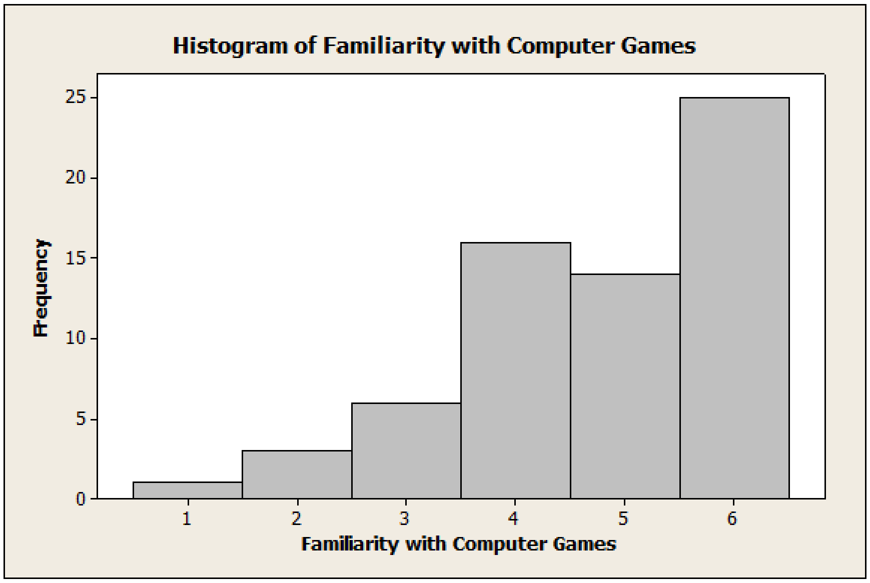


Figure 4.3: Subjects' Familiarity with Computer Games

According to Figure 4.3, most of the subjects rated at a 3 or above with only three subjects mentioning that they had little to no computer game experience. Their awareness with Second Life was quite different. The subjects were asked to rank their awareness on 1-6 scale with 1 meaning that they never heard of Second Life and 6 meaning they are significantly familiar with Second Life. Figure 4.4 shows that over half of the subjects did not know about Second Life and only one subject reported they were significantly familiar.

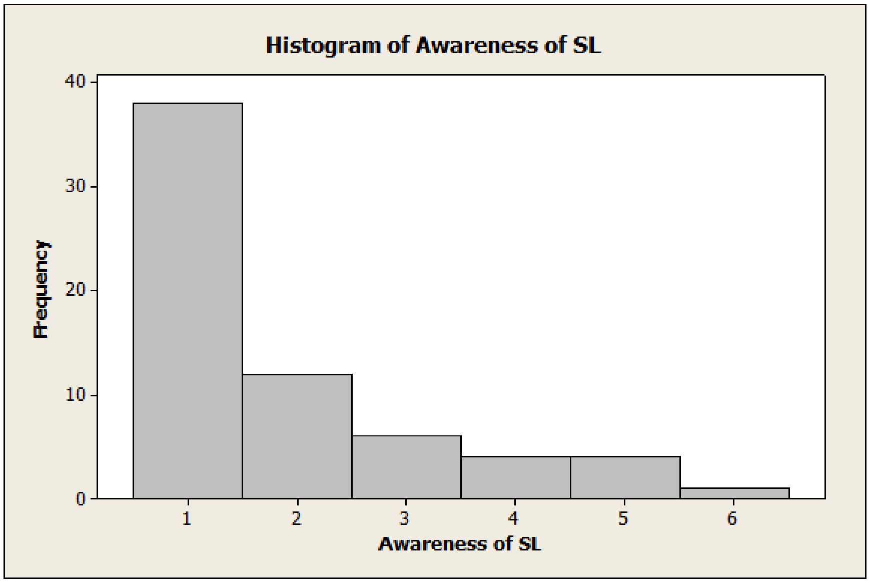


Figure 4.4: Subjects' Awareness of Second Life

Due to the majority of the group lacking any experience with Second Life, a majority of the group reported that they had no experience using Second Life as almost all of them recorded a response of 0. Figure 4.5 shows this result.

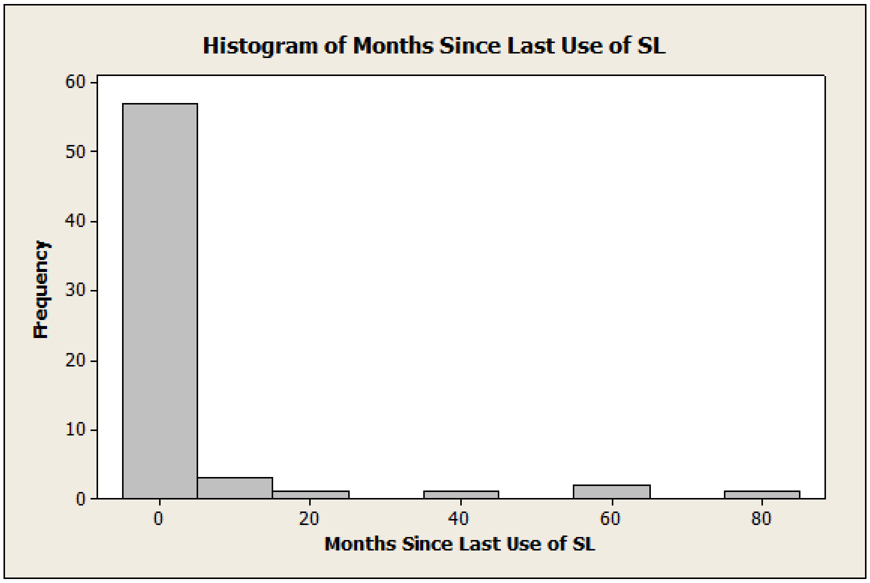


Figure 4.5: Months since Subjects' Last Usage of Second Life

Unsurprisingly, as is shown in Figure 4.6, all but one of the subjects recorded a 1 for their frequency of use of Second Life for leisure. Since this was also supposed to be on a 1-6 scale with 1 meaning ‘I do not use it at all’ and 6 meaning ‘I use Second Life very frequently’ the results indicate that almost no one in the group uses Second Life.

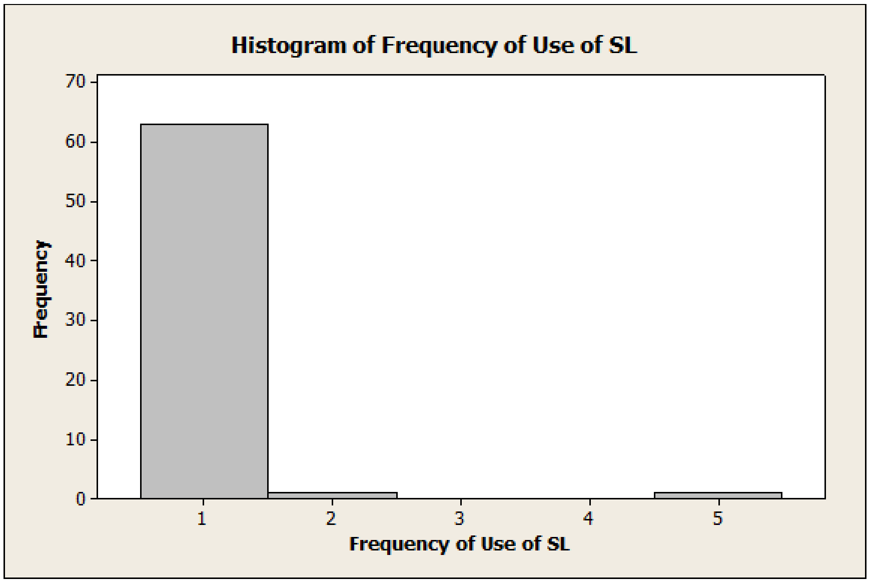


Figure 4.6: Subjects' Frequency of Use of Second Life

The results of the demographics survey reveal that a strong majority of the subjects would be introduced to something new if they were to receive the Second Life training.

4.2.2 Render Safe Assessment Summary

Prior to concluding their first day of training, the subjects complete the Render Safe Assessment. This assessment is meant to test the subjects' knowledge of what they read during the first day of training. A summary of the results for the subjects is shown in Figure 4.7.

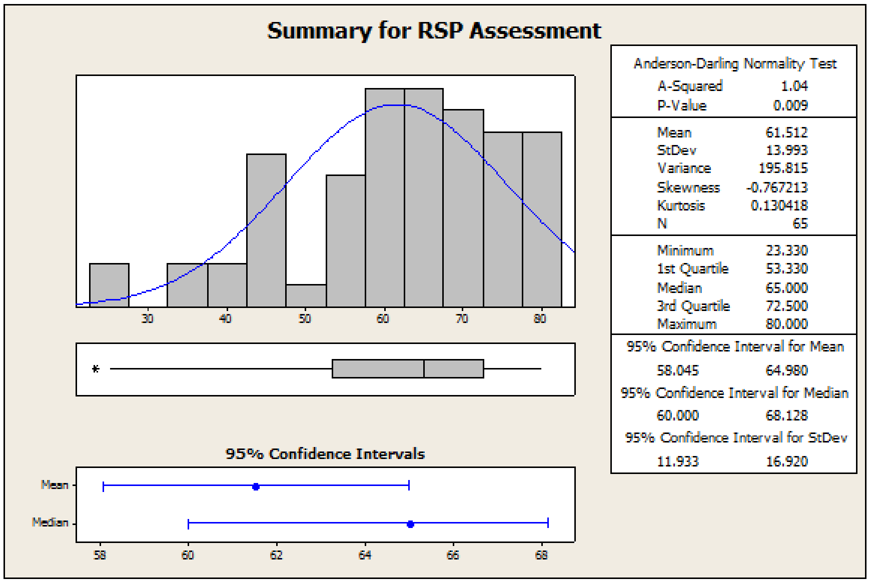


Figure 4.7: Summary for Render Safe Assessment

The results show that on average the subjects obtained a score of about 61.5 out of a possible 80 points. The lowest score was a 23.33 while the highest was the maximum score of 80. The confidence interval for the mean shows that with 95% confidence, the mean falls between 58.04 and 64.98.

4.3 Day 2

As previously mentioned, the subjects are divided into their respective groups on the second day of their training. Day 2 consists of each group going through a specialized training with odd-numbered subjects receiving the case studies instruction and the even-numbered subjects receiving the Second Life instruction. Following their training, each subject completes a survey for the instruction type under which they are trained and then completes a survey assessing their opinion of the training overall.

4.3.1 Case Studies Questionnaire

The Case Studies Questionnaire consists of six questions with responses recorded on a 6-point Likert scale where 1 means 'Strongly Disagree' and 6 means 'Strongly Agree.' The questions provided are as follows:

- Question 1: I experienced a high level of interaction in the case studies scenarios.
- Question 2: The case studies did a very poor job of using a story to explain tasks.
- Question 3: I believed that I was a character in the scenarios.
- Question 4: The case studies were very unrealistic.
- Question 5: The writing was very descriptive.
- Question 6: The case studies training was not at all engaging.

The questions alternate between positively and negatively connoted questions to avoid subjects just choosing the same answer for every question. Figures 4.8 through 4.13 summarize the subjects' responses to each of the questions. Note that instances of an asterisk denote that a subject failed to answer one of the questions.

Figure 4.8 displays the results of the subjects' responses to the first question.

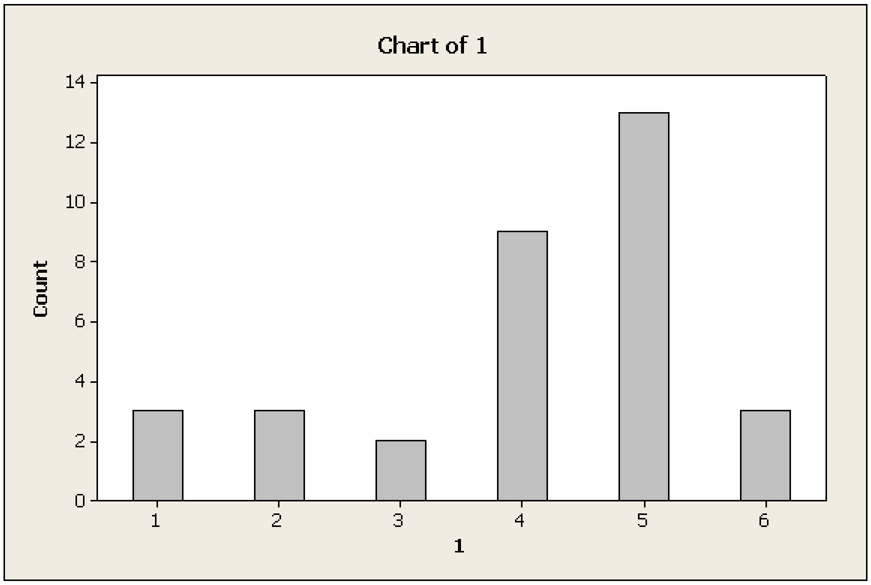


Figure 4.8: Summary of Case Studies Question 1

Figure 4.8 shows that a majority of the subjects found the case studies to be interactive with a rating of 5 being the most common response. Since the positively connoted question was mostly answered with high marks it can be expected that a negatively connoted question would be answered with mostly low marks. Figure 4.9 displays exactly this trend.

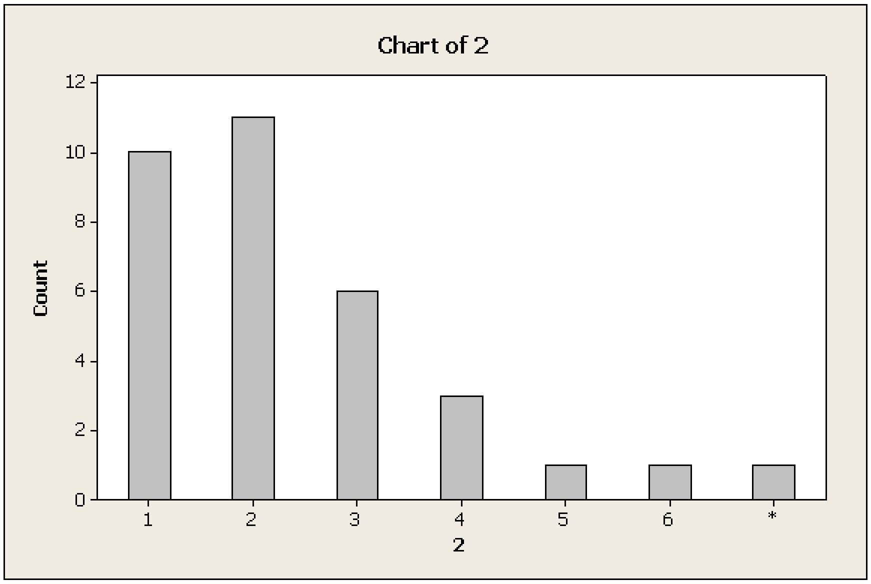


Figure 4.9: Summary of Case Studies Question 2

According to Figure 4.9, most of the subjects disagreed strongly about the assertion that the case studies did a poor job of using a story to explain the tasks they were to complete. Figure 4.10 displays the results of the third question which asked if the subjects felt they were a character in the scenario. The results are positive, but the highest rated response is a neutral ‘Somewhat Agree.’

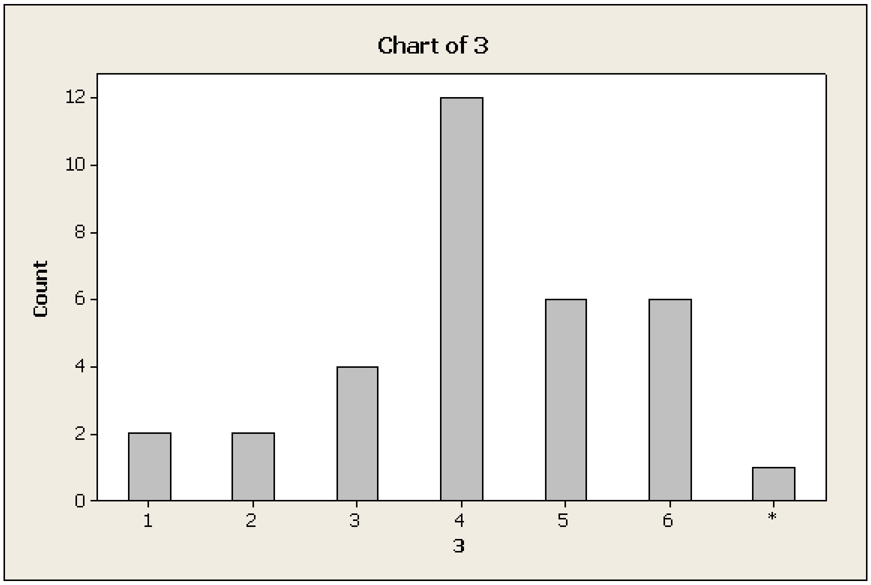


Figure 4.10: Summary of Case Studies Question 3

Figure 4.11 shows that the subjects disagreed with the assertion that the case studies were unrealistic showing that they felt the scenarios could take place in a real-life environment.

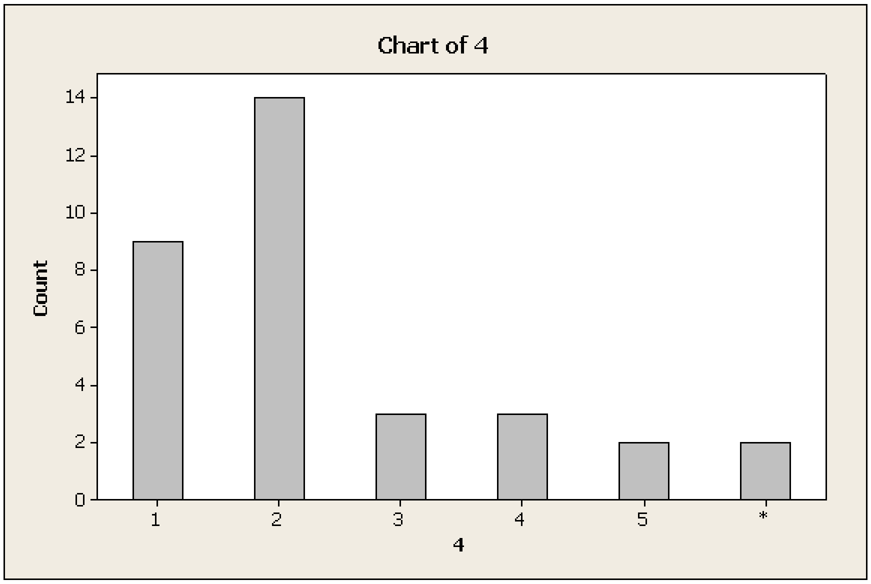


Figure 4.11: Summary of Case Studies Question 4

Figure 4.12 shows that the subjects' thoughts on the descriptiveness of how the cases were written vary widely. While more subjects agreed with the assertion than disagreed, there was still a strong contingency that disagreed.

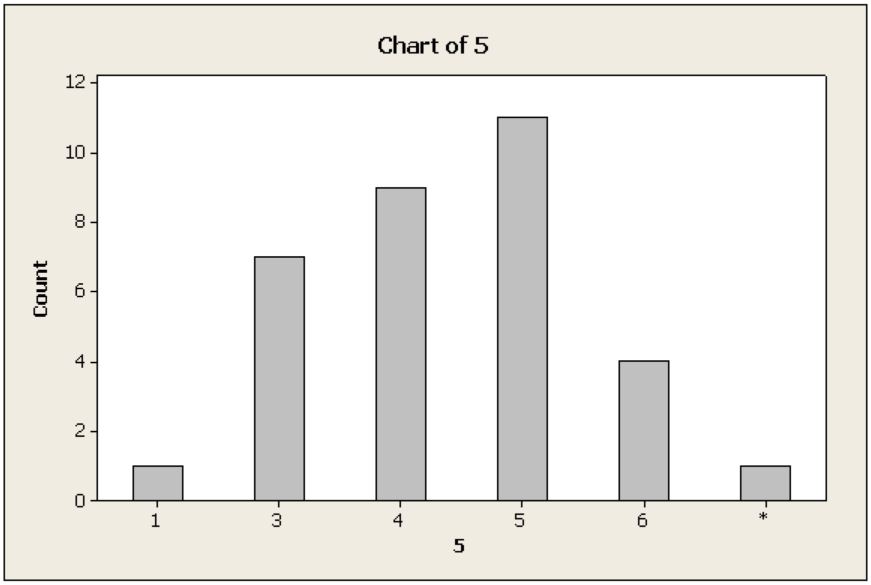


Figure 4.12: Summary of Case Studies Question 5

The results of the final question are displayed in Figure 4.13.

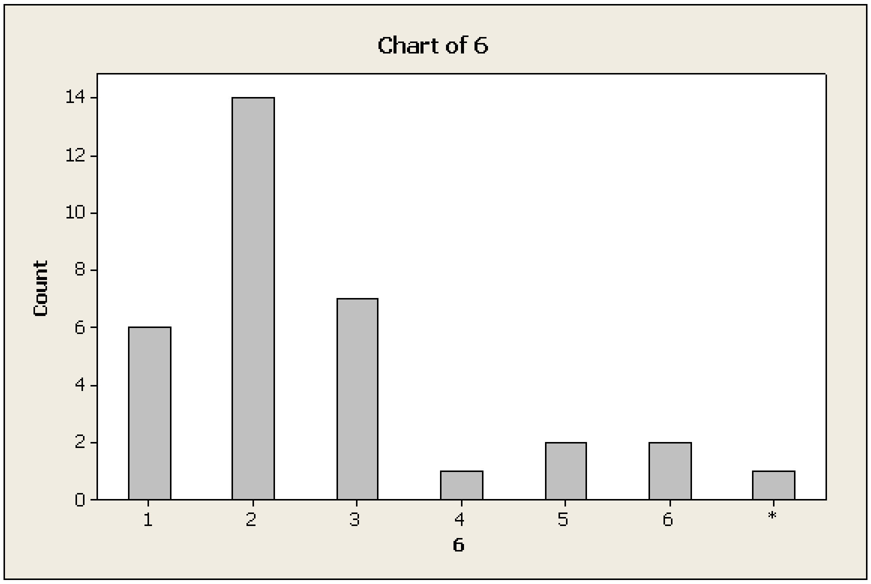


Figure 4.13: Summary of Case Studies Question 6

The results indicate that the subjects found the case studies engaging although there were a couple of subjects that strongly agreed with the assertion that the case studies were not at all engaging. The results of the questionnaire imply that there was a general good feeling towards the case studies. In the next section, a similar questionnaire assessing the subjects' opinion about the Second Life training is provided.

4.3.2 Second Life Questionnaire

While one group took the case studies training, the other group was taking the Second Life training. After completing their training, the subjects completed a questionnaire similar to that of the case studies group. The subjects responded to the questions using the same 6-point Likert scale as used in the case studies training. The questions asked in the questionnaire are listed below:

- Question 1: I experienced a high level of interaction in the Second Life scenarios.
- Question 2: The Second Life did a very poor job of using a story to explain tasks.
- Question 3: I believed that I was a character in the scenarios.
- Question 4: The Second Life scenarios were very unrealistic.
- Question 5: The sound effects were very realistic.
- Question 6: The Second Life training was not at all engaging.

Just as with the case studies questionnaire, the questions in the Second Life questionnaire alternate between positively and negatively connoted questions to avoid subjects just choosing the same answer for every question. Figures 4.14 through 4.19 summarize the subjects' responses to each of the questions. Again, note that instances of an asterisk denote that a subject failed to answer one of the questions.

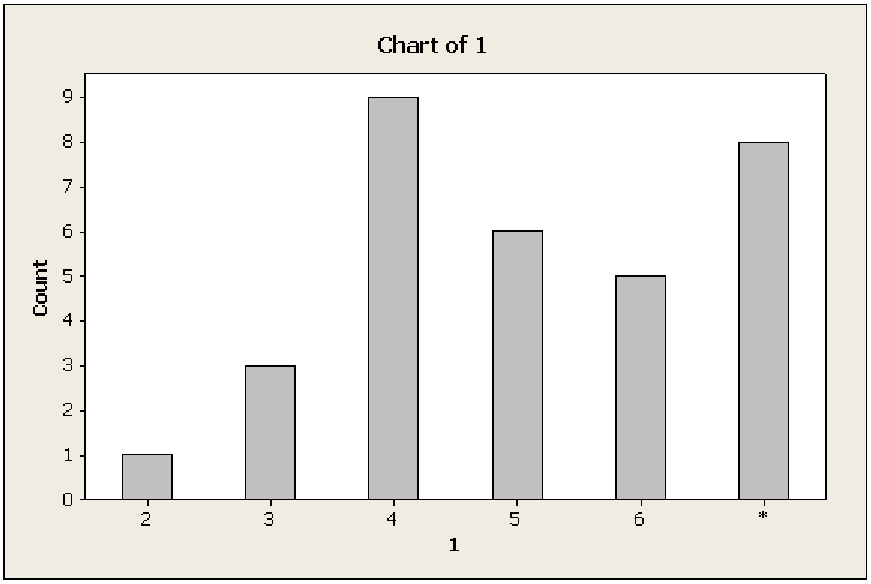


Figure 4.14: Summary of Second Life Question 1

According the Figure 4.14, a majority of the subjects reported that they felt a high level of interaction while they were training in Second Life. A majority of the subjects also reported that they felt the training did a good job of using stories to explain the tasks as indicated with the subjects disagreeing with the negative wording of the question. This is shown in Figure 4.15.

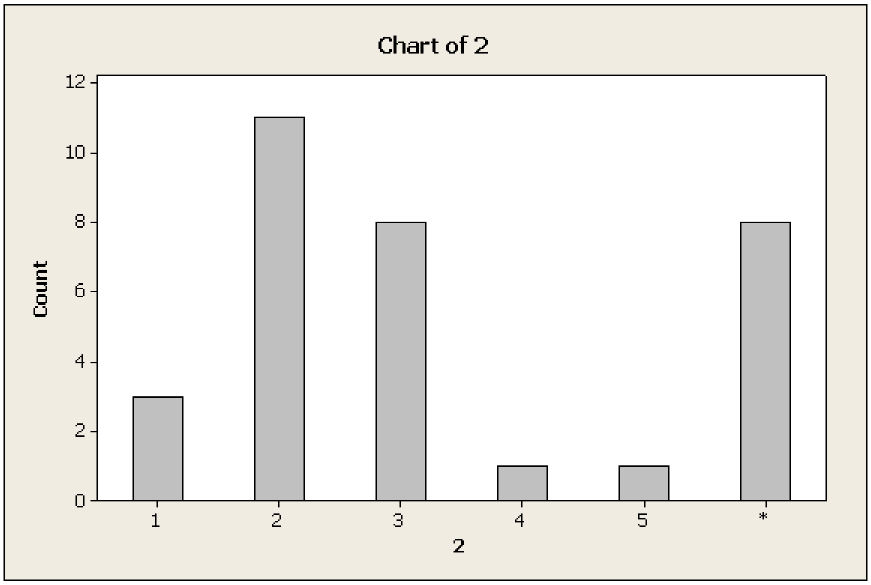


Figure 4.15: Summary of Second Life Question 2

The next question received mixed responses. About the same amount of people reported that they felt they were the character in the scenarios as the people who said they did not feel this way. The results for Question 3 are shown in Figure 4.16.

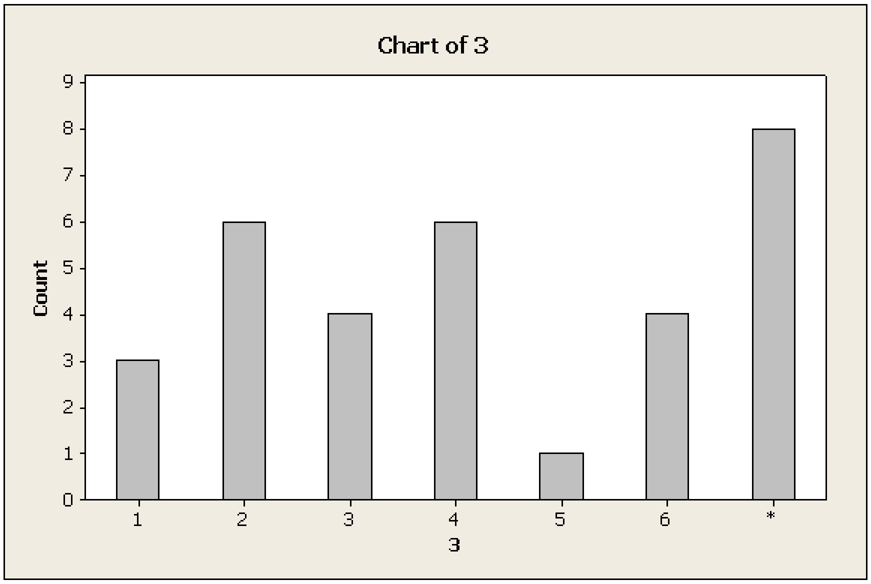


Figure 4.16: Summary of Second Life Question 3

In terms of level of realism, the subjects were also mixed in their responses. As shown in Figure 4.17, most of the responses fall under the two most neutral categories 'Agree' and 'Disagree.' Therefore it can be concluded that the perceived level of realism was not very high for most of the subjects.

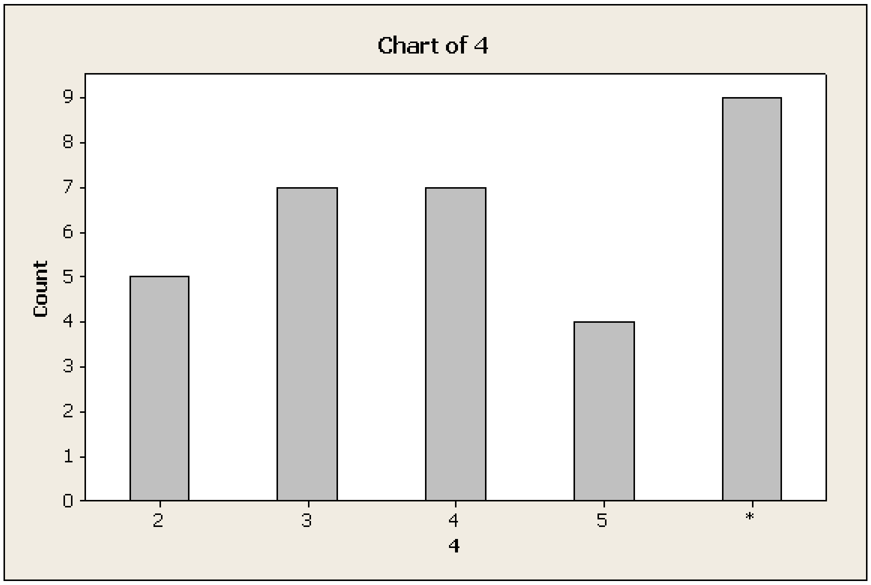


Figure 4.17: Summary of Second Life Question 4

In terms of the sound effects, the reviews were also mixed. This is shown in Figure 4.18. It is to no surprise that the realism of the sound effects would be perceived similarly as the realism of the entire training as a whole.

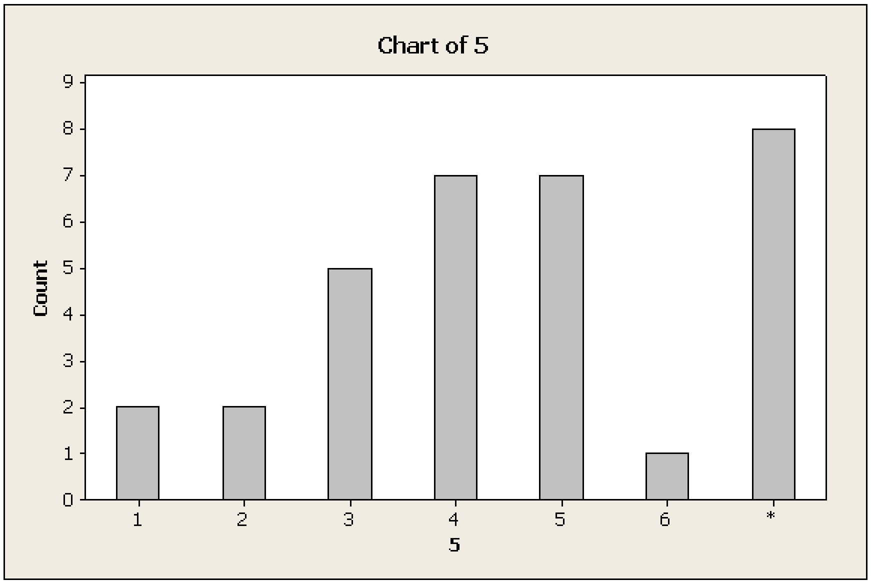


Figure 4.18: Summary of Second Life Question 5

The results for the final question are shown in Figure 4.19.

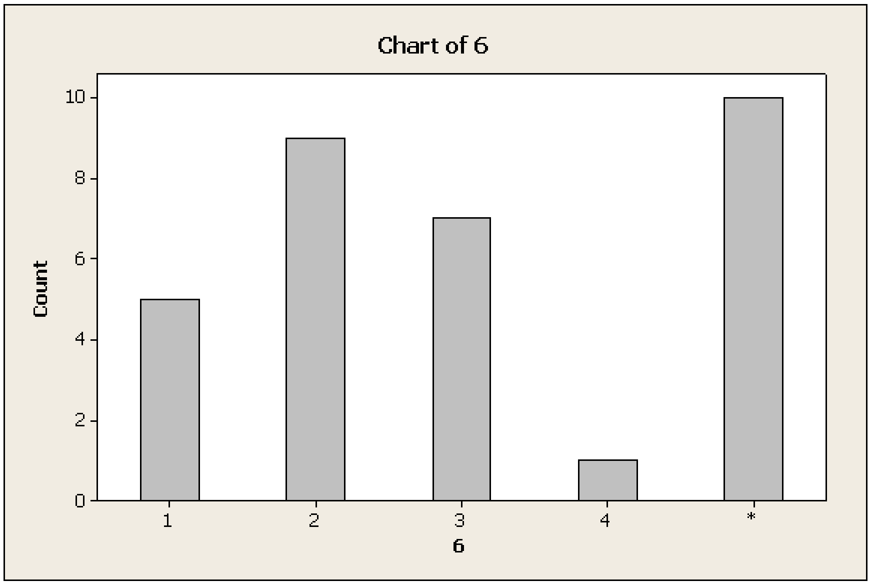


Figure 4.19: Summary of Second Life Question 6

The results indicate that the subjects disagreed with the assertion that the Second Life training was not engaging. Overall, the surveys indicate generally positive perceptions of the Second Life training although some questions tended to be answered rather neutrally, especially when it came to the realism of the training. This indicates some improvements can be made to the program which could improve the effectiveness of the Second Life training. The next section contains the results of the final survey taken during the second day of training.

4.3.3 Post Training Student Survey

The final survey administered on the second day of training is the Post Training Student Survey. This survey assesses the perceived quality of the study as a whole. In total there were 9 questions in the survey, but one question was repeated. Therefore, the responses to question 4 have been removed. The questions were answered using a 6-point Likert scale ranging from a ‘1’ meaning ‘Strongly Disagree’ to ‘6’ meaning ‘Strongly Agree.’ The questions of the survey are as follows:

1. The training met my expectations.
2. I will be able to apply the knowledge learned.
3. The content was organized and easy to follow.
5. The materials distributed were pertinent and useful.
6. The trainer was knowledgeable.
7. The quality of instruction was good.
8. The trainer met the training objectives.
9. Class participation and interaction were encouraged.

The results of the surveys are shown in Figures 4.20 through 4.27. The results indicate that the subjects had a generally positive opinion of the training as all the surveys have as their majority 4’s and 5’s for responses. Figure 4.20 through 4.22 display the results for Questions 1 through 3 respectively. The results indicate that the subjects felt that there expectations of the training were met. They also showed confidence in their ability to apply what they learned to their final day of the experiment. Figure 4.22 indicates that the subjects felt the information was organized and easy to follow.

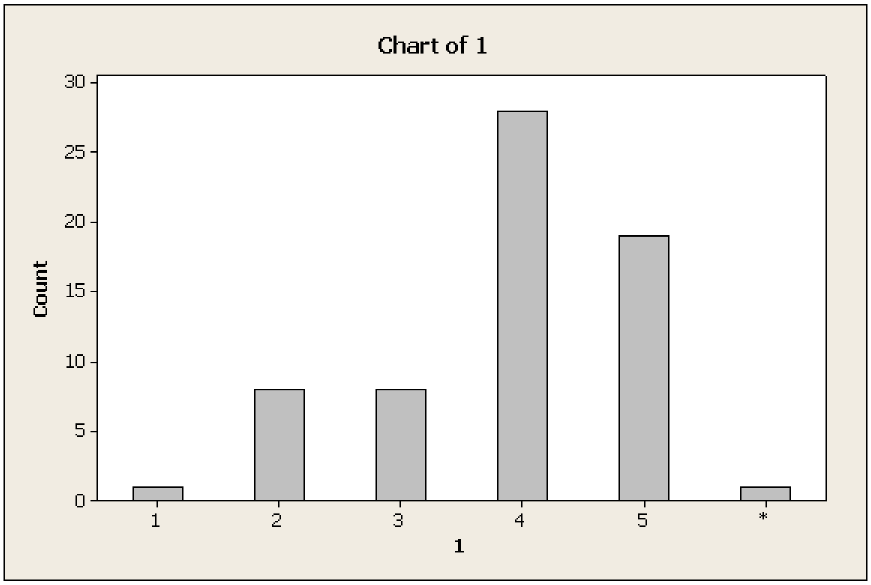


Figure 4.20: Summary of Post Training Student Survey Question 1

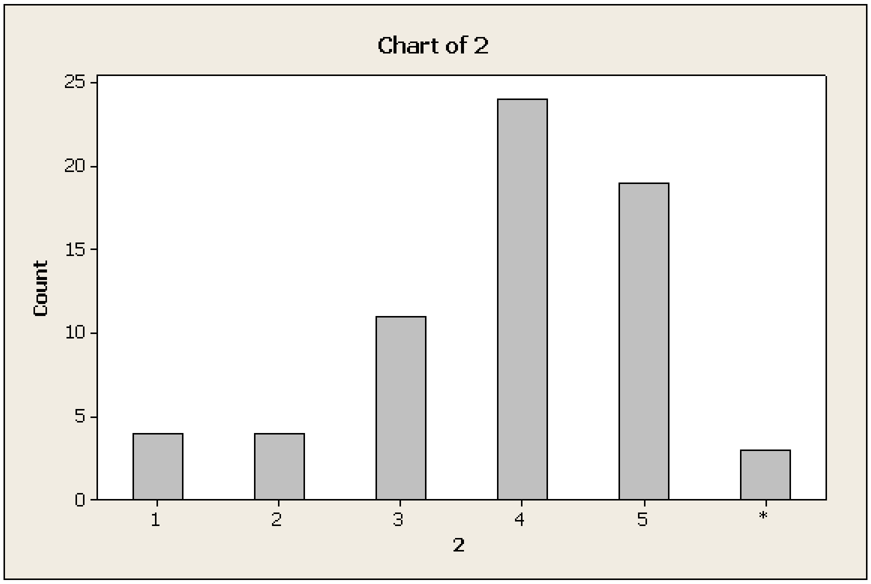


Figure 4.21: Summary of Post Training Student Survey Question 2

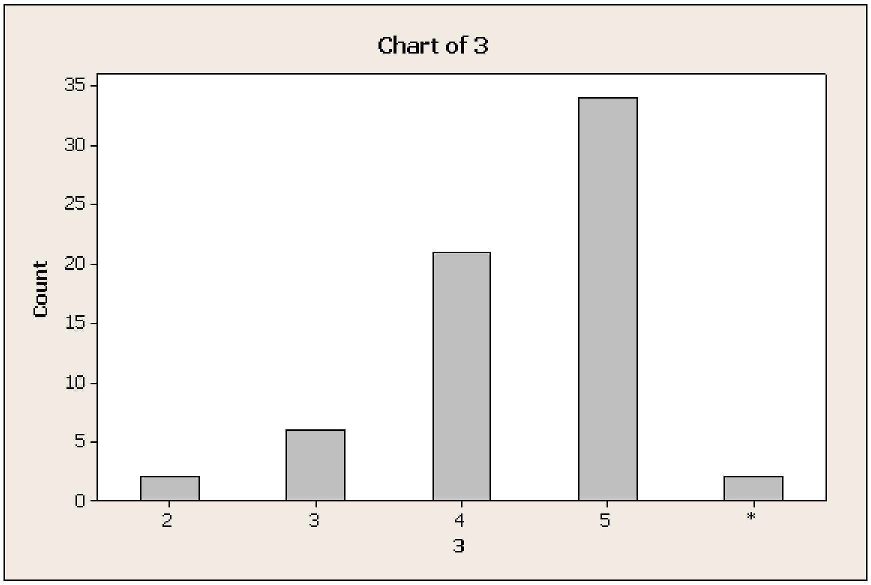


Figure 4.22: Summary of Post Training Student Survey Question 3

As for Figures 4.23 through 25, they indicate that the subjects felt the information provided to them was useful and pertinent to what they were studying. They had mostly high reviews for the instructor proctoring their second day of training and they were generally pleased with the quality of the instruction.

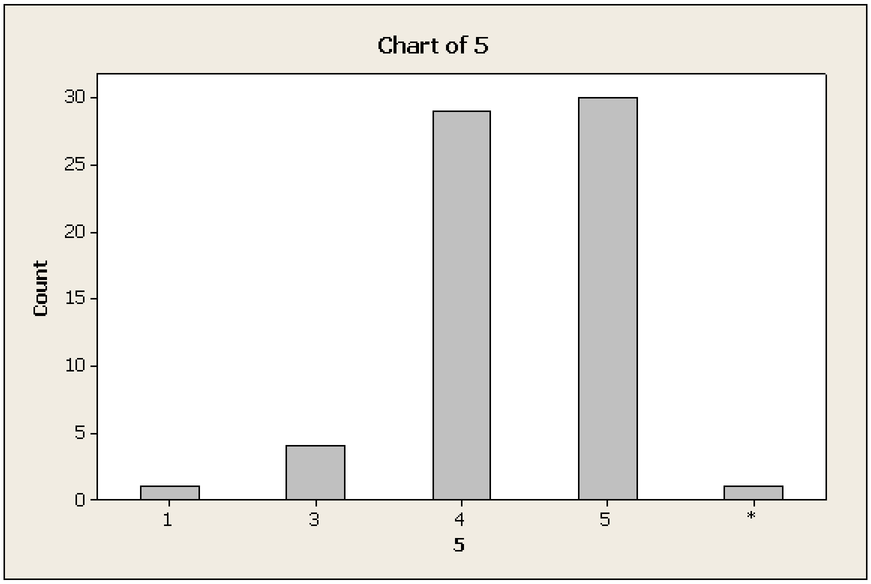


Figure 4.23: Summary of Post Training Student Survey Question 5

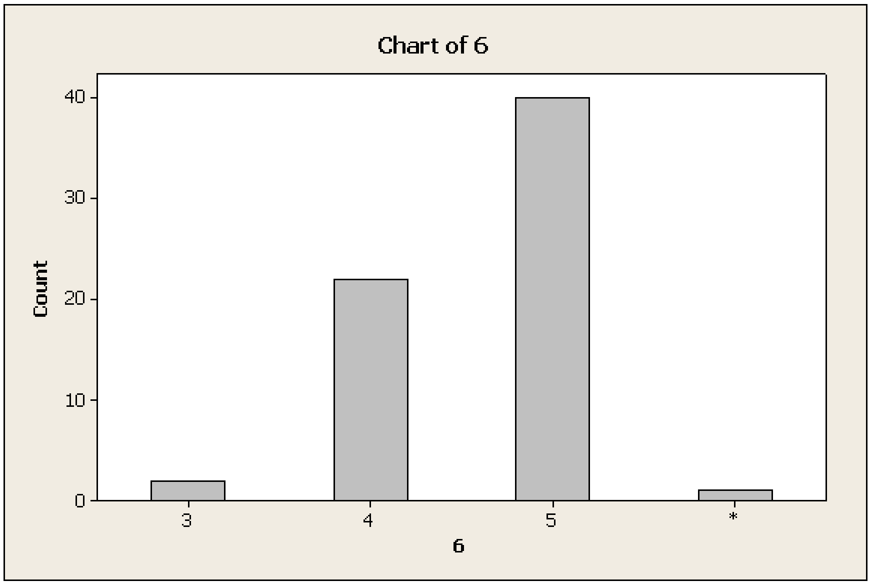


Figure 4.24: Summary of Post Training Student Survey Question 6

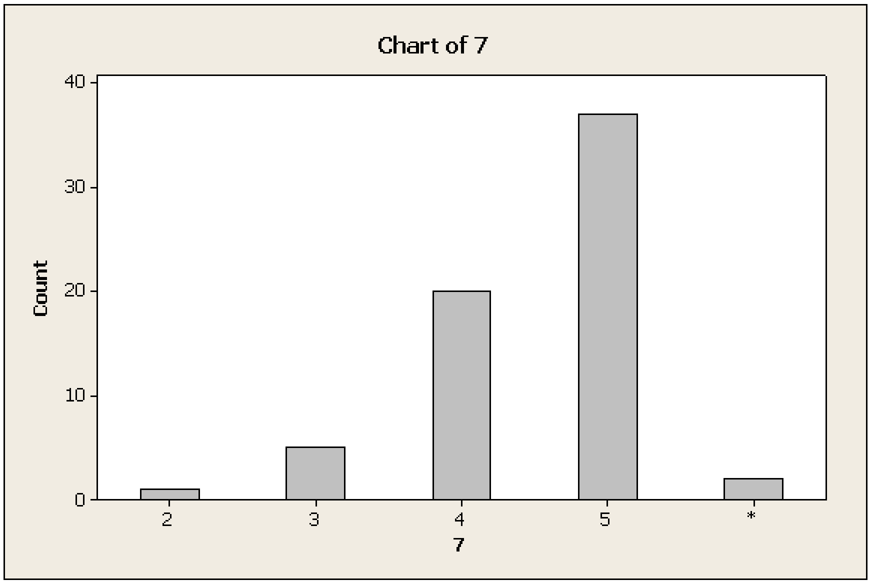


Figure 4.25: Summary of Post Training Student Survey Question 7

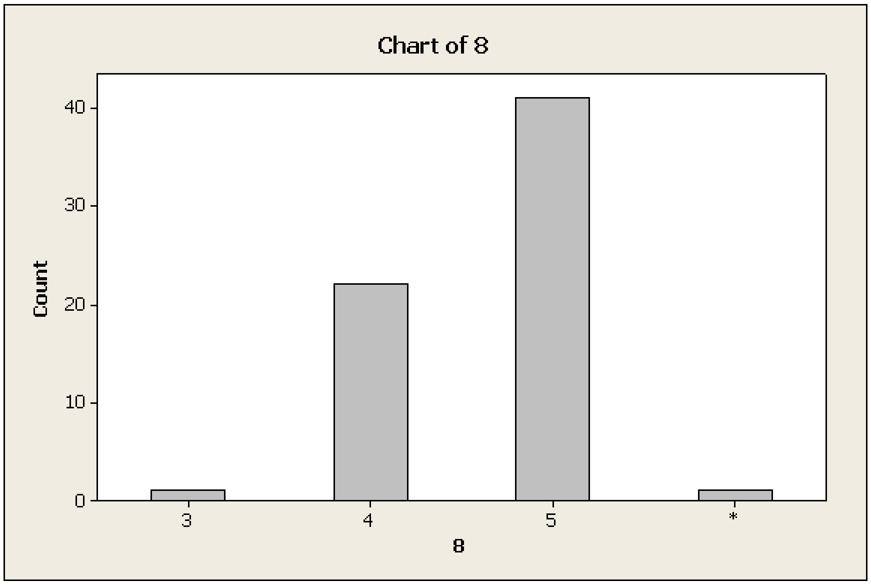


Figure 4.26: Summary of Post Training Student Survey Question 8

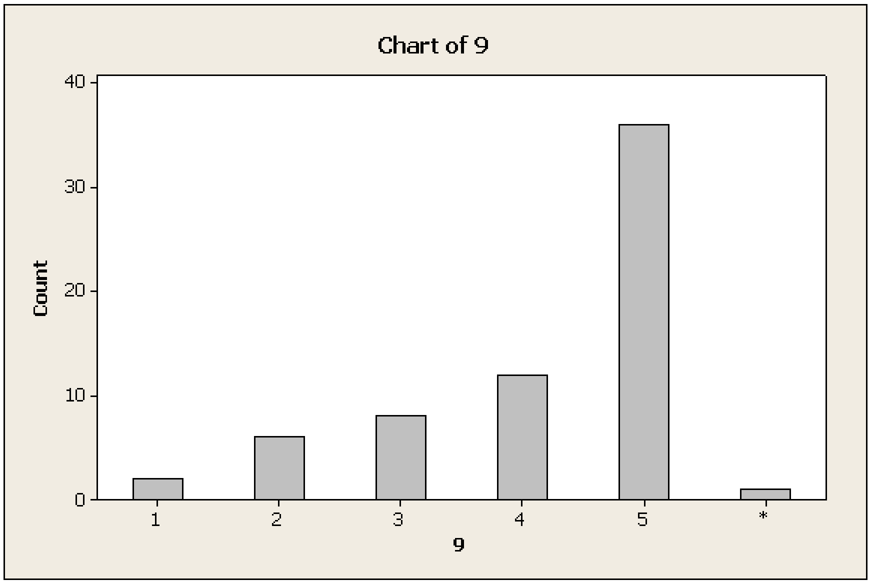


Figure 4.27: Summary of Post Training Student Survey Question 9

The results for the final two questions are shown in Figure 4.26 and 4.27 respectively. Figure 4.26 shows that the subjects felt the instructor adhered to the learning objectives provided during the first day of training. The results shown on Figure 4.27 indicate that the subjects felt they were allowed to participate in the instruction. Since this was an individual study, it can be assumed that they interpreted this question as they felt they were allowed to ask questions and get feedback. The feedback provided though was only to help with problems understanding a word or a sentence and not with answering questions. The results of the three surveys indicate that the subjects had a generally positive reaction to their training. In the next section, the results of how they utilized their training are analyzed.

4.4 Day 3

On the final day of training, each subject follows the same procedure. Their first step is to take the State-Trait Anxiety Questionnaire. Following completion of this questionnaire, the subject is taken to a room to have the physiological sensors attached. After the physiological equipment is attached to the subject, the subject attempts the final scenario which consists of a victim with a mockup of an explosive device attached to their back. The subject's job is to defuse the device within 15 minutes while performing the render safe procedures taught in the first day of training and reinforced during the second day. The purpose of this experiment is to determine whether the instruction type had an effect on the performance of the subjects when it came to defusing the explosive and displaying knowledge of the render safe procedures. However, before answering this question, there is interest in determining if the test taken at the conclusion of Day 1 is predictive of how the subjects perform during Day 3.

4.4.1 Final Scenario Scoring

The final scenario is graded in three ways. One way is while the subjects are performing their final scenario, a grader marks whether the steps of the render safe procedures are followed correctly. If a subject performs a step, they are given 10 points. In total, there are 19 render safe procedures which corresponds to a maximum score of 190 (the minimum score is 0). A summary of the scores is displayed in Figure 4.28.

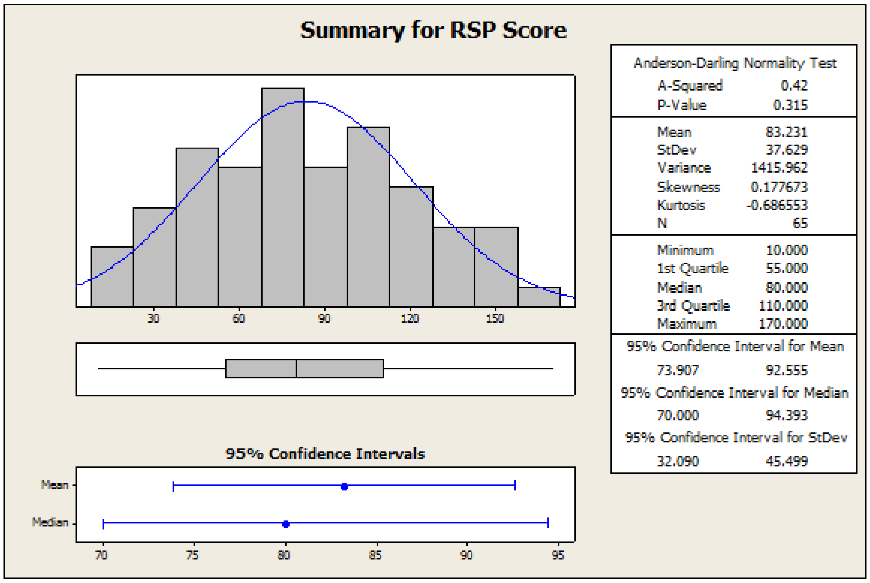


Figure 4.28: Summary for Render Safe Procedure Score

The summary shows that on average the subjects did not perform very well when it came to completing the render safe procedures. The average was only a score of 83.23 which is less than half of the total available points. The mean score is expected to fall with 95% confidence within the values of 73.91 and 92.55. The upper level of this confidence interval is still less than half of the maximum score. This indicates that the training did not do a good job of teaching the render safe procedures to the subjects.

The other two scoring methods are linked together. The first simply assesses whether the subjects properly defused the bomb or not (‘Yes’: 1, ‘No’: 0). The second, called the Defusing Score, is a 5 point scale based on what components of the explosive the subject properly disconnected. The explosive used in the final scenario consists of 4 primary components.

- Radio transmitter (walkie-talkie)
- 1st battery
- 2nd battery
- Initiators (4 blasting caps)

When being scored, the subject starts with a 1 and gains a point for every component they properly disconnect. The key is to disconnect the components in the proper order which is the order the components are listed in (the batteries can be disconnected in any order). For example, if a battery is disconnected prior to the radio transmitter being disconnected the subject receives only a 1 and does not receive credit for any other actions performed. However, if the subject disconnects the radio transmitter but then pulls the blasting caps they still get the point for disconnecting the radio transmitter. They are not given any points for disconnecting a battery after they disconnect the initiators. A score of 4 or 5 indicates that the subject prevented an explosion. Scores from 1 to 3 indicate that the explosive detonated, and the subject failed the assessment. Two bar charts in Figures 4.29 and 4.30 show a graphical representation of the raw data for these two scoring methods.

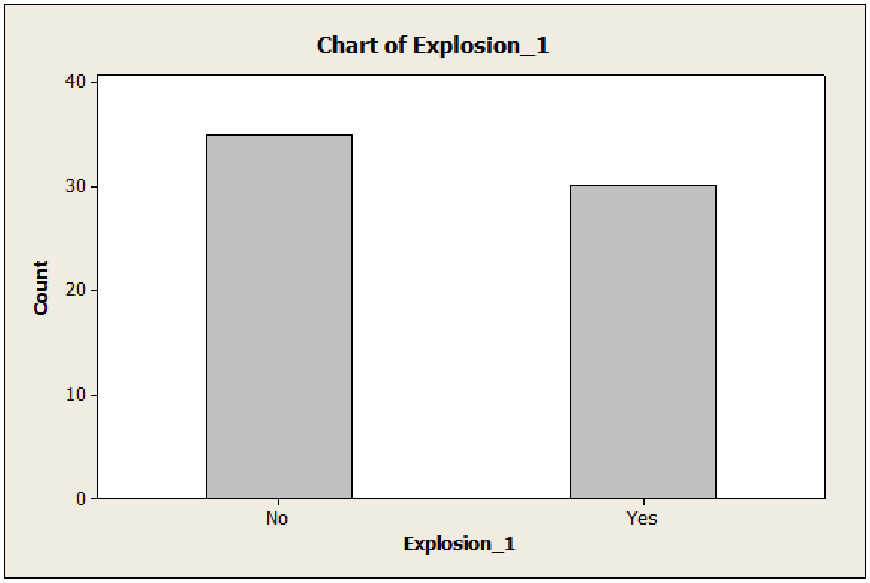


Figure 4.29: Bar Chart of Whether the Explosive Detonated

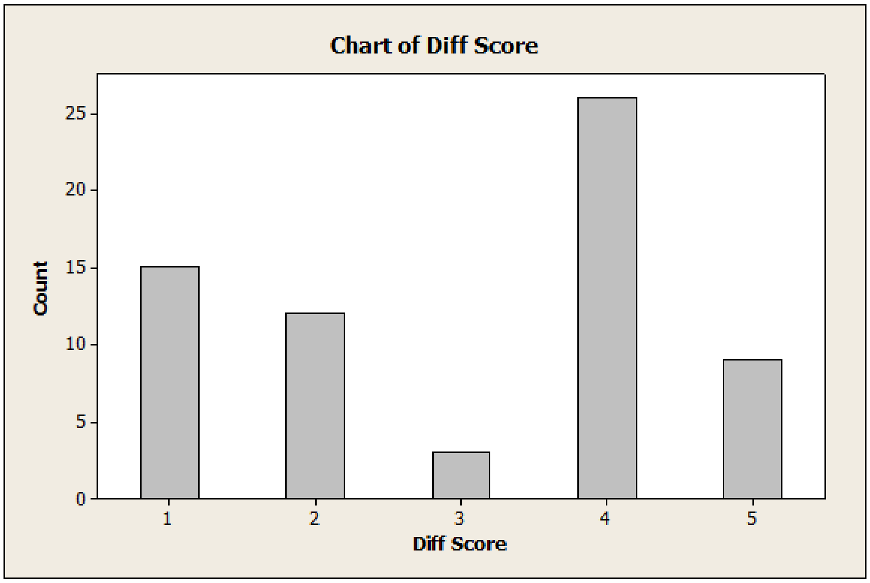


Figure 4.30: Bar Chart of Defusing Scores

In Figure 4.29, ‘Yes’ and ‘No’ refer to whether there was an explosion during the experiment. As can be seen the numbers are about equal with the ‘No’s’ having a slight edge. Figure 4.30 indicates that of the subjects that passed, few actually completed the final step of pulling the blasting caps. Figure 4.30 also indicates that there were quite a few subjects who failed to properly perform any operations. The data in these charts only shows how the subjects performed as a whole. The purpose is to determine if the instruction type the subjects received affected their performance on the final day. This is to be determined in the following sections.

4.4.2 Final Scenario vs. Render Safe Assessment

Before determining the effect of the instruction type on the subjects’ final day of performance, a reference to Day 1 is necessary. In section 4.2.2, the results of the Render Safe Assessment, completed by each subject during Day 1, are summarized. There is interest in determining if success on this assessment is predictive of success during the final scenario. Three tests of hypothesis, one for each scoring technique, are used to determine if this is the case.

- Render Safe Procedure Score vs. Render Safe Assessment
- Explosion (Yes or No) vs. Render Safe Assessment
- Defusing Scores vs. Render Safe Assessment

4.4.2.1 Render Safe Procedure Score vs. Render Safe Assessment

In order to determine whether performance on the Render Safe Assessment completed during Day 1 is predictive of success in completing the render safe procedures tested for during the final day of the experiment, a linear regression model is fitted. Figure 4.31 displays a scatter plot of the data with the Render Safe Procedure Score plotted against the subjects' Render Safe Assessment score.

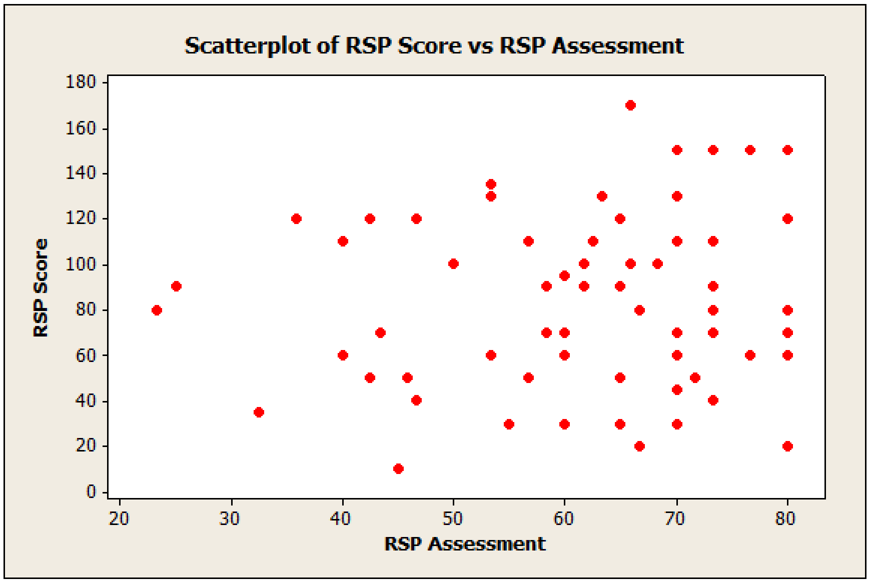


Figure 4.31: Scatter Plot of Render Safe Procedure Score vs. Render Safe Procedure Assessment

The scatter plot indicates that there is not a linear relationship between the two scores. In order to verify, the linear regression model is displayed below:

Regression Analysis: RSP Score versus RSP Assessment

```
The regression equation is
RSP Score = 66.9 + 0.266 RSP Assessment


Predictor          Coef  SE Coef     T      P
Constant          66.86    21.26  3.14  0.003
RSP Assessment   0.2662   0.3371  0.79  0.433


S = 37.7405   R-Sq = 1.0%   R-Sq(adj) = 0.0%
```

The regression analysis verifies that the Render Safe Assessment is not predictive of performance in regards to displaying knowledge of render safe procedures during the final day of training as the p-value is well above 0.05 for the predictor 'RSP Assessment'.

4.4.2.2 Explosion (Yes or No) vs. Render Safe Assessment

In order to test whether performance on the Render Safe Assessment is predictive of the subject properly defusing the bomb a different type of regression is necessary. This is due to the fact that the data for the response variable ‘Explosion’ is binary which causes the assumptions of linear regression to fail. The model used in this case is called Binary Logistic Regression which is not limited by the assumption of normality and independence of the residuals. Instead, the only assumptions are that the dependent variable is binary and that the independent variable is what is called “linear in the logit.” Montgomery et al. (2012) include a discussion on logistic regression on pages 421-430 in their work *Introduction to Linear Regression Analysis, 5th Edition*.

The binary logistic regression method fits the data to the function

$$\hat{y} = \frac{1}{1 + e^{-\boldsymbol{x}\hat{\boldsymbol{\beta}}}}$$

where $\hat{\boldsymbol{\beta}}$ is the column vector of estimated coefficients for the linear part of the equation and $\boldsymbol{x}$ is the row vector of the predictor variables. In order to test the linearity in the logit assumption, the linear portion of the function (the part contained in the exponent) is plotted against the independent variable. If a linear relationship exists than the assumption is validated. Note that this only applies to continuous independent variables. Discrete independent variables are not confined by this assumption. The results of the binary logistic regression model are shown below:

Binary Logistic Regression: Explosion_1 versus RSP Assessment

```
Link Function: Logit


Response Information

Variable     Value  Count
Explosion_1  Yes       30  (Event)
             No        35
             Total     65


Logistic Regression Table

                                                    Odds     95% CI
Predictor             Coef    SE Coef      Z      P  Ratio  Lower  Upper
Constant         -0.344563    1.13293  -0.30  0.761
RSP Assessment  0.0030943  0.0179533   0.17  0.863   1.00   0.97   1.04


Log-Likelihood = -44.847
Test that all slopes are zero: G = 0.030, DF = 1, P-Value = 0.863
```

The results show that there is not a relationship between whether the explosive detonated and the results of the Render Safe Assessment. A similar test is used to determine whether the subjects’ Render Safe Assessment score is predictive of the score the subject receives from disconnecting particular components of the explosive device.

4.4.2.3 Defusing Score vs. Render Safe Assessment

The primary difference between the model used for this test and the test in the previous model is that the Defusing Score is not binary. There are five possible scores a subject can receive. Since a higher score is preferred over a lower score, the response variable is considered to be ordinal. Therefore, an ordinal logistic model is used. The results are shown below:

Ordinal Logistic Regression: Diff Score versus RSP Assessment

```
Link Function: Logit


Response Information

Variable    Value  Count
Diff Score  1         15
            2         12
            3          3
            4         26
            5          9
            Total     65


Logistic Regression Table

                                                     Odds     95% CI
Predictor            Coef    SE Coef      Z      P  Ratio  Lower  Upper
Const(1)         -1.75957    1.05170  -1.67  0.094
Const(2)        -0.891404    1.03437  -0.86  0.389
Const(3)        -0.703641    1.03217  -0.68  0.495
Const(4)          1.27814    1.04806   1.22  0.223
RSP Assessment  0.0090765  0.0162441   0.56  0.576   1.01   0.98   1.04


Log-Likelihood = -92.954
Test that all slopes are zero: G = 0.320, DF = 1, P-Value = 0.571
```

The results once again indicate that the Render Safe Assessment is not predictive of performance. In this case it is not predictive of performance on the Defusing Score. The results of all the tests run thus far indicate that the Render Safe Assessment did not have a significant relationship with the final assessment conducted on the final day of the study. This indicates that any effect on the results of the subjects' performance on the final day would be from the instruction type and potentially the gender of the victim in the final assessment.

4.4.3 Final Scores vs. Instruction Type and Victim's Gender

The primary goal of the experiment is to assess whether the instruction type had an effect on the final scores during the final day of the experiment. Since a subject could be assigned to either a female or a male victim, Victim's Gender is included as a potential predictor variable for the model. The experiment consisted of 65 male subjects which were separated into a group of 32 for Second Life and 33 for case studies. Each group was administered a particular instruction type (either Second Life or the case studies). Of the 65 subjects, 33 were assigned to a female victim while 32 were assigned to a male victim. The number of male and female victims for the two instruction types was approximately equal.

4.4.3.1 Render Safe Procedures Score vs. Instruction Type and Victim's Gender

The first of the scoring types is again the Render Safe Procedures Score. A matrix scatter plot in Figure 4.32 displays the relationship between the two predictor variables (Instruction Type and Victim's Gender) and the Render Safe Procedures Score.

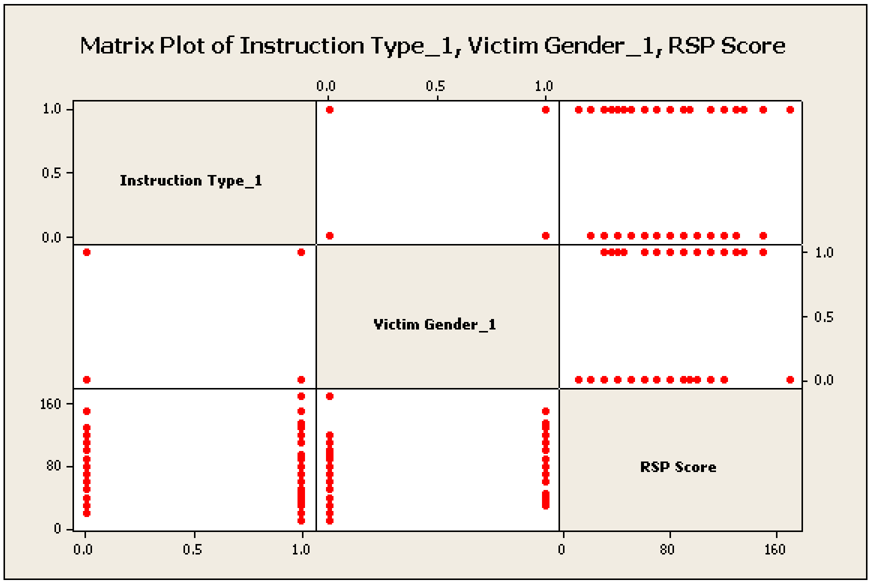


Figure 4.32: Matrix Plot of Render Safe Procedure Score vs. Predictor Variables

As far as the Instruction Type is concerned, the means appear to be located at around the 80 for both instruction types (0: case studies, 1: Second Life). The only difference is in the dispersion of scores. For victim gender, the mean appears slightly higher for male subjects (1) than for female subjects (0). A linear regression model is used to assess if any of the predictor variables significantly affect the Render Safe Procedures Score.

Regression Analysis: RSP Score versus Instruction Type, Victim Gender

```
The regression equation is
RSP Score = 67.4 + 5.34 Instruction Type + 26.8 Victim Gender


Predictor             Coef  SE Coef     T      P
Constant            67.398    7.697  8.76  0.000
Instruction Type_1   5.345    8.848  0.60  0.548
Victim Gender_1     26.815    8.848  3.03  0.004


S = 35.6261   R-Sq = 13.2%   R-Sq(adj) = 10.4%
```

The results indicate that the only significant factor is the Victim’s Gender. As such, the Instruction Type can be eliminated from the model to produce the model below:

Regression Analysis: RSP Score versus Victim Gender

```
The regression equation is
RSP Score = 70.2 + 26.6 Victim Gender


Predictor        Coef  SE Coef      T      P
Constant       70.152    6.170  11.37  0.000
Victim Gender  26.567    8.794   3.02  0.004


S = 35.4461   R-Sq = 12.7%   R-Sq(adj) = 11.3%
```

The low R^2 value implies that there are other variables accounting for the variance in the scores as the Victim's Gender only accounts for 12.7%. However, the model still indicates that the Victim's Gender had an effect on the Render Safe Procedures Score. Additionally, the coefficient is positive indicating that the subject did better at performing the Render Safe Procedures with a male victim than a female victim.

4.4.3.2 Model Adequacy

With a model established, it is important to test the assumptions of the regression model. The assumptions of linear regression are that the residuals are normally and independently distributed with a mean of zero and constant variance. The normality and zero mean assumption can be determined by using a normal probability plot of the residuals as shown in Figure 4.33. It should be noted that all residuals used in the adequacy check are standardized residuals

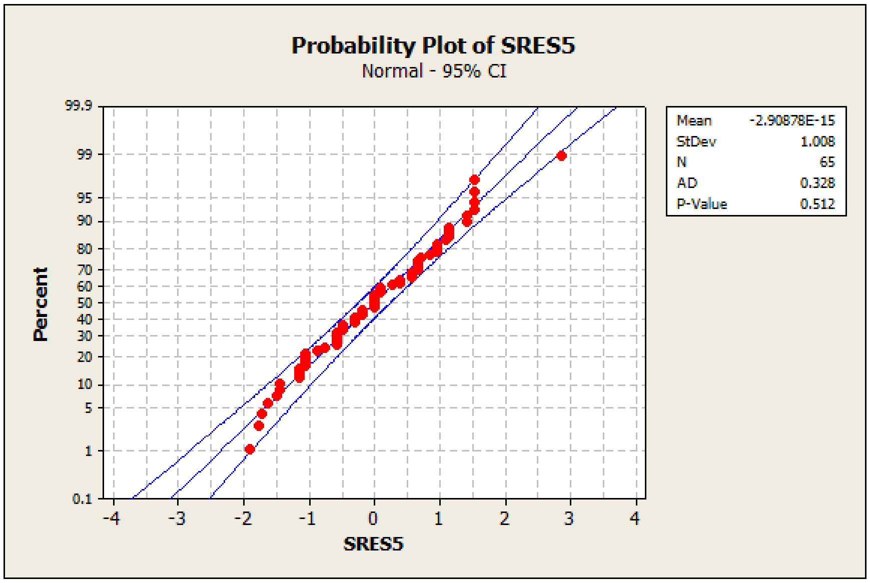


Figure 4.33: Normal Probability Plot of Residuals

Figure 4.33 indicates that the normality assumption is valid at that 95% confidence level. As can be seen all of the data points fit within the confidence bands and the p-value is well over 0.05 leading to the conclusion that the hypothesis that the data fits the normal distribution fails to be rejected. Figure 4.33 also shows that mean is very close to zero which gives validity to the zero mean assumption. As for the constant variance assumption, two types of plots are used. One of the plots shown in Figure 4.34 is a plot of the residuals vs. the fitted values of the model. The other plot is shown in Figure 4.35 and displays the relationship between the residuals and the predictor variable, Victim's Gender.

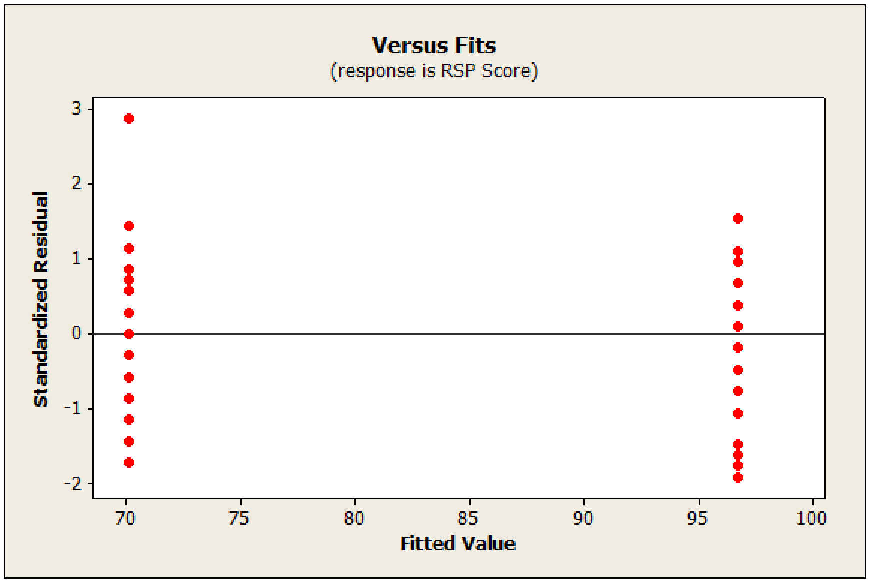


Figure 4.34: Residuals vs. Fitted Values

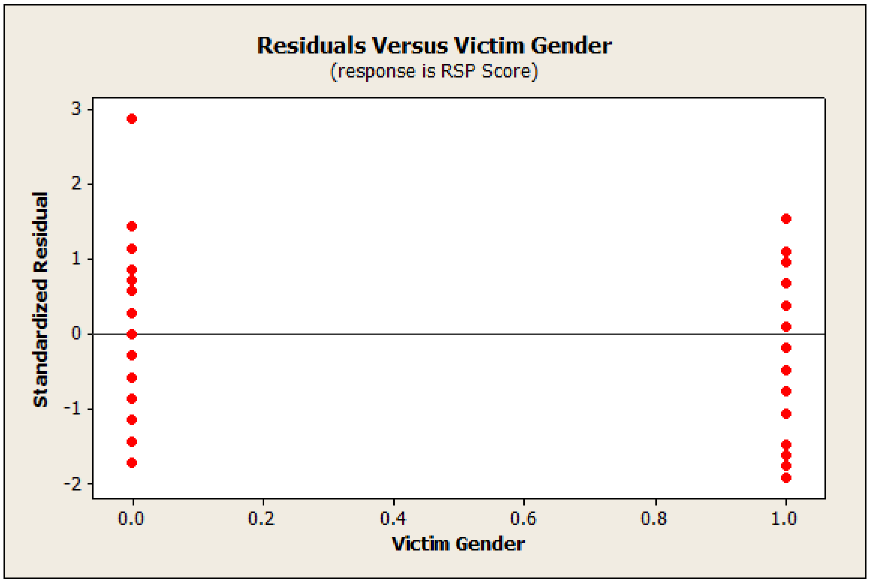


Figure 4.35: Residuals vs. Victim's Gender

The two plots do not show any indication of a pattern between the residuals and neither the fitted values nor the Victim’s Gender. The final diagnostic checks to see if the independence assumption is adequate. This check consists of checking for patterns in the plot of the residuals vs. the time sequence the data was collected. In this experiment, while subjects were assigned numbers they were not necessarily tested in order of those numbers. Instead the testing schedule was dictated by their convenience. Figure 4.36 shows a plot of the residuals vs. the order the data was recorded.

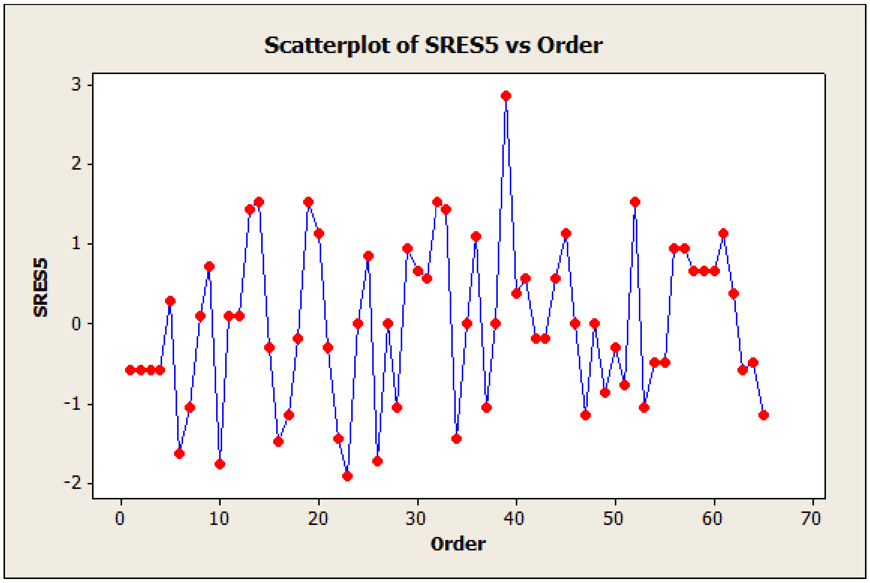


Figure 4.36: Residuals vs. Order

The graph in Figure 4.36 does not appear to have any pattern. The data points appear to be random and they do not simply oscillate about the line $y = 0$. Therefore, there does not appear to be a problem with the independence assumption. Thus, the assumptions appear to be validated and the conclusions drawn from the linear model can be properly inferred. In the next two models, the response variables are no longer continuous. Therefore, logistic regression is used once again.

4.4.3.3 Explosion (Yes or No) vs. Instruction Type and Victim's Gender

In section 4.4.2, binary logistic regression was introduced as a method for modeling data with binary response variables. This method is employed here to determine if the two predictor variables have a significant effect on whether the subject defused the bomb. A bar chart displaying how subjects in particular groups performed is shown in Figure 4.37.

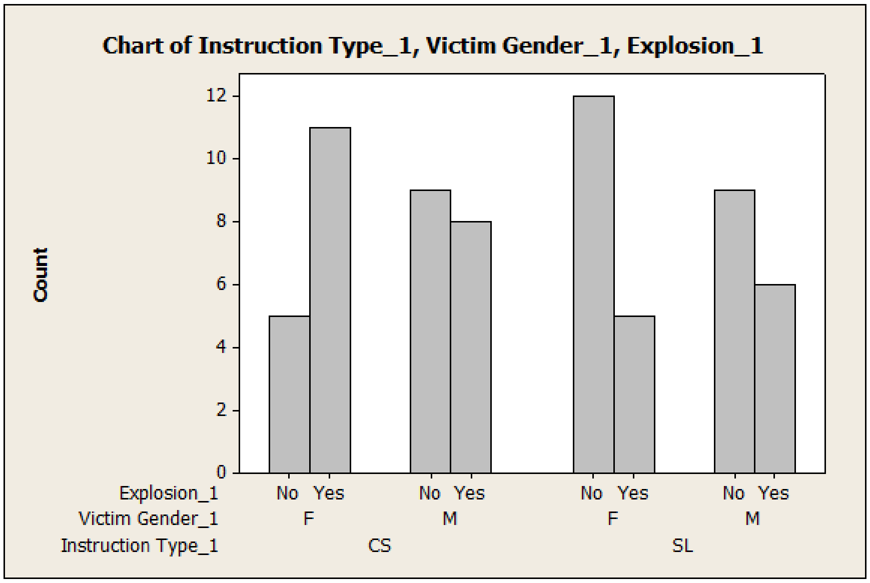


Figure 4.37: Bar Chart of Explosions by Victim's Gender and Instruction Type

The chart shows that subjects taking the Second Life instruction type tended not to have explosions more so than those taking the case studies. As for Victim's Gender, subjects taking the case studies with a male victim tended to outperform those that took case studies and had a female victim. However, the trend is reversed for the subjects using Second Life. Overall the subjects with male victims outperformed the subjects with female victims by only one successful defusing. Also of interest is: if the length of time the subject took to complete the final assessment had an effect of whether they successfully defused the explosive.

In order to verify the inferences drawn from simply looking at the graphs, a binary logistic regression analysis is performed.

Binary Logistic Regression: Explosion versus Instruction, Victim Gender

```
Link Function: Logit


Response Information

Variable   Value  Count
Explosion  Yes       30  (Event)
           No        35
           Total     65


Logistic Regression Table

                                                         Odds     95% CI
Predictor              Coef   SE Coef       Z       P  Ratio  Lower  Upper
Constant           0.434564  0.444192    0.98   0.328
Instruction Type
 SL               -0.967136  0.514820   -1.88   0.060   0.38   0.14   1.04
Victim Gender
 M                -0.248477  0.514166   -0.48   0.629   0.78   0.28   2.14


Log-Likelihood = -42.968
Test that all slopes are zero: G = 3.788, DF = 2, P-Value = 0.150
```

The results indicate that none of the predictor variables are significant at the 95% significance level. However, the Instruction Type is close to being considered significant as the p-value is 0.06. If the significance level is set at 90%, the Instruction Type is considered significant at the 90% significance level. The results of the test indicate that the subjects training in Second Life performed better than those taking the case studies. Of note is the odds ratio which represents how likely the "Event" (an explosion) occurs for one option of a predictor variable relative to the other. In the case of the Instruction Type, the odds of a subject taking Second Life failing to defuse the bomb is 0.38 times that of the odds of a subject taking the case studies failing to defuse the bomb. This shows that the odds of an explosion are less for the Second Life instruction type than for the case studies instruction type. However, the upper part of the confidence interval for the odds ratio is over 1 which indicates that there is a chance that the odds of an explosion for subjects taking Second Life can be higher than the odds for those taking the case studies.

This model analyzes how well the subjects performed from on a pass/fail basis. But just as in section 4.4.2, an ordinal logistic model is appropriate for determining how significantly

Instruction Type and Victim's Gender affect the subjects' performance in terms of their Defusing Score.

4.4.3.4 Defusing Score vs. Instruction Type and Victim's Gender

In this model, the effect the three predictor variables have on performance in regards to the subjects' Defusing score is analyzed. Figure 4.38 displays a bar graph with a count of subjects receiving a particular score. The bar graph is separated by the Instruction Type and Victim's Gender.

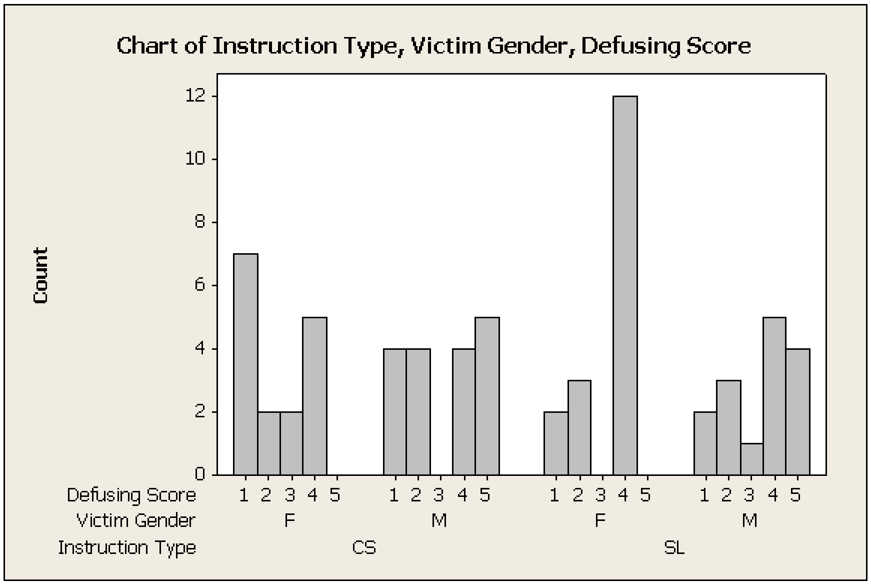


Figure 4.38: Bar Chart of Defusing Scores by Instruction Type and Victim's Gender

Figure 4.38 indicates that the most instances of the subject failing to defuse any component of the explosive correctly fall within the case where a subject received the case studies instruction type and had a female victim. The case with the highest amount of successes at defusing is the one where a subject receives the Second Life training and is assigned a female victim. However, all the maximum scores were attained when the victim was male. Consistent with the previous section, the subjects receiving Second Life training had a higher amount of 4's and 5's which constitute successes. Conversely, the higher amount of 1's, 2's, and 3's occur when the subject receives the case studies. Ordinal Logistic Regression is used to analyze this data just as it is used to analyze whether the Render Safe Assessment score is predictive of success in the final part of the experiment. The results of the analysis are shown below:

Ordinal Logistic Regression: Defusing Sco versus Instruction , Victim Gender

```
Link Function: Logit


Response Information

Variable        Value  Count
Defusing Score  1         15
                2         12
                3          3
                4         26
                5          9
                Total     65


Logistic Regression Table

                                                          Odds     95% CI
Predictor              Coef   SE Coef      Z      P  Ratio  Lower  Upper
Const(1)          -0.507047  0.416843  -1.22  0.224
Const(2)           0.392586  0.410015   0.96  0.338
Const(3)           0.586142  0.413494   1.42  0.156
Const(4)            2.66984  0.537472   4.97  0.000
Instruction Type
 SL               -0.794209  0.461914  -1.72  0.086   0.45   0.18   1.12
Victim Gender
 M                -0.761026  0.461214  -1.65  0.099   0.47   0.19   1.15


Log-Likelihood = -90.400
Test that all slopes are zero: G = 5.429, DF = 2, P-Value = 0.066
```

Once again, the test shows there are no significant predictor variables at the 95% significance level. However, both variables are significant at the 90% significance level. Instruction Type has a p-value of 0.078 and an odds ratio of 0.44. This indicates that lower scores most likely occur with subjects receiving the case studies instruction type. However, the confidence interval encapsulates the value of 1 indicating that there is a chance that the chance of lower scores can be higher for subjects receiving Second Life. As for the Victim's Gender, the p-value is at the very edge of being considered significant at the 90% significance level. The odds ratio is 0.47 which indicates lower scores are more likely to occur with female victims. The next section provides some reasoning for these results and their implications. Before looking into those implications, the effect of the predictor variables on the time to complete the final scenario is analyzed in the next section. The final two sections contain the results of the physiological data and the final survey taken by the subjects respectively.

4.4.4 Time (min) vs. Instruction Type and Victim's Gender

In addition to the scoring metrics, it is of interest to determine if the length of time it takes for the subject to complete the experiment is affected by the instruction type or the victim's gender. A relationship between time and either of the two factors can provide insight as to whether one of the two predictors affects the subjects' performance during the final scenario. The time to complete the experiment is measured from the time the subject has entered the room with the victim until the time that the subject says they have completed defusing the explosive. The time is recorded (in minutes) by the victim who has a clock near them in order to record the time lapse. Figure 4.39 shows a matrix plot displaying the graphical relationship between time and the two predictors.

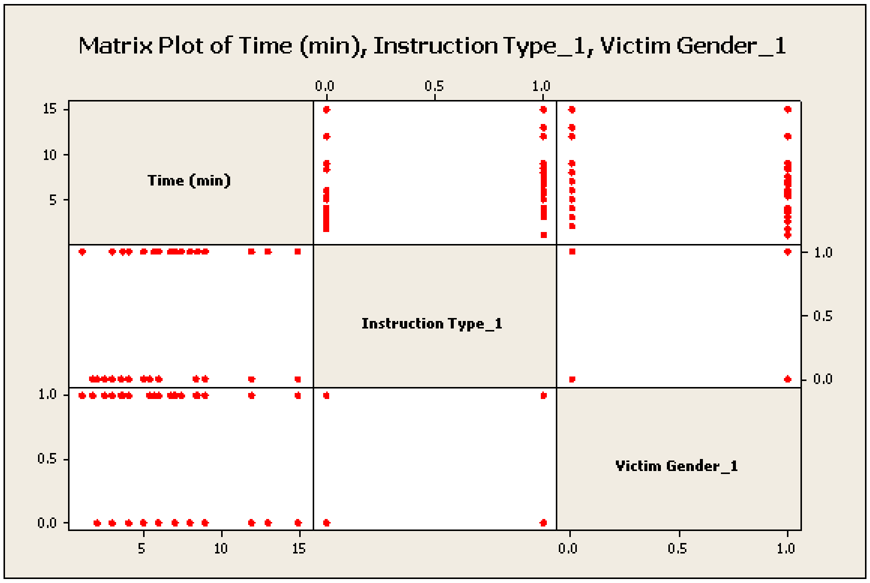


Figure 4.39: Matrix Plot of Time (min), Instruction Type, and Victim's Gender

There does not appear to be a relationship between Instruction Type and Time (min). However, there does appear to be a relationship Victim's Gender and Time (min). A linear regression model is used to determine if there is a relationship. The results are shown below:

Regression Analysis: Time (min) versus Instruction Type, Victim Gender_1

```
The regression equation is
Time (min) = 7.15 + 0.834 Instruction Type_1 - 1.56 Victim Gender_1


Predictor              Coef  SE Coef      T      P
Constant             7.1463   0.7553   9.46  0.000
Instruction Type_1   0.8336   0.8683   0.96  0.341
Victim Gender_1     -1.5605   0.8683  -1.80  0.077


S = 3.49615   R-Sq = 6.5%   R-Sq(adj) = 3.5%
```

The results indicate that there is no relationship between Time (min) and Instruction Type. However, the effect of Victim's Gender on Time (min) is significant at the 90% significance level. Since subjects with male victims are coded with a '1' and the coefficient of Victim's Gender is negative, the subjects with female victims took a significantly longer time to complete the final scenario than the subjects with male victims.

4.4.5 ECG Data

In section 3, the arrangement for the ECG is explained. The purpose of recording this measurement is based on the assumption that a better-prepared subject would elicit lesser signs of anxiety when faced with the final scenario. In the case of this study, elevations in heart rate are being associated with higher levels of anxiety. In a study by Elwess and Vogt (2005), they found that students performing commonly stressful college activities such as oral presentations and written exams were found to have elevated heart rates. Therefore, elevated heart rate is used as an indicator of anxiety.

The data used to analyze the ECG is obtained by averaging the heart rate responses in beats-per-minute (BPM) during the ten minute baseline period and then averaging the heart rate responses during the actual scenario. In this section, three tests are performed. The first test determines whether there is a difference in the baseline BPM for the two instruction types. A difference would indicate that subjects from one instruction type would be more anxious than the subjects from the other type. The next thing to be tested is whether there is a difference in the average BPM from the baseline to the actual scenario. Finally it is of interest to determine if the difference in BPM is affected by the two predictor variables: Instruction Type and Victim's Gender.

4.4.5.1 Baseline Average Heart Rate vs. Instruction Type

The first test is meant to determine whether there is a difference between the two instruction types in average heart rate prior to the start of the experiment. The results are shown below:

Regression Analysis: Baseline versus Instruction Type_1

```
The regression equation is
Baseline = 121 + 3.87 Instruction Type_1


Predictor              Coef  SE Coef      T      P
Constant            121.374    3.549  34.20  0.000
Instruction Type_1    3.873    5.058   0.77  0.447


S = 20.3886   R-Sq = 0.9%   R-Sq(adj) = 0.0%
```

The results indicate that there is no difference in the baseline heart rate between the subjects of each instruction type. Therefore, it can be concluded that neither instruction type reduced the subjects' anxiety prior to the start of the experiment.

4.4.5.2 Difference in Baseline Heart Rate and Scenario Heart Rate

In order to test if there was a change in average heart rate, a Paired t-test is used. The result of the Paired t-test is shown below:

Paired T-Test and CI: Avg Experiment, Baseline

```
Paired T for Avg Experiment - Baseline

                 N    Mean  StDev  SE Mean
Avg Experiment  65  118.71  12.74     1.58
Baseline        65  123.28  20.32     2.52
Difference      65   -4.57  16.22     2.01


95% upper bound for mean difference: -1.21
T-Test of mean difference = 0 (vs < 0): T-Value = -2.27  P-Value = 0.013
```

The p-value of the test is less than 0.05. Therefore, the conclusion is to reject the hypothesis that the difference between the average heart rate during the baseline and the average heart rate during the experiment is the same. Using a one-tailed test, the conclusion is that the average heart rate during the experiment was less than it was in the baseline.

4.4.5.3 Change in Average Heart Rate vs. Instruction Type and Victim's Gender

The final test determines if there was a difference in the change of average heart rate between the two different instruction types and the two victim genders. In order to determine if relationships exist between either of the predictor variables, a linear regression model is used. The result of the model is shown below:

Regression Analysis: Change versus Instruction Type_1, Victim Gender_1

```
The regression equation is
Change = - 2.48 - 2.88 Instruction Type_1 - 1.36 Victim Gender_1


Predictor              Coef  SE Coef      T      P
Constant             -2.479    3.544  -0.70  0.487
Instruction Type_1   -2.881    4.075  -0.71  0.482
Victim Gender_1      -1.361    4.075  -0.33  0.739


S = 16.4059   R-Sq = 0.9%   R-Sq(adj) = 0.0%
```

The model indicates that there is no significant relationship between the change in average heart rate and the two predictors. Therefore it can be concluded that the instruction type taken during the second day is not related to the change in heart rate during the final day. Likewise, the victim's gender does not have an effect on the subjects' change in heart rate. While there was not a difference between the two predictor variables and the change in heart rate, there may be a difference in how the subjects perceived their level of stress during the experiment and the two instruction types. The next section provides an analysis of the final questionnaire.

4.4.6 Post Experiment Questions – Anxiety Questionnaire

Following the conclusion of the experiment, each subject receives one final questionnaire. This questionnaire is used to analyze how the subjects perceived the scenario from the final day of the experiment. The questions asked are listed below:

1. I felt extremely anxious when I was defusing the bomb.
2. The bomb-defusing situation seemed very artificial to me.
3. My attention was extremely focused when defusing the bomb.
4. I felt anxious only because I was a subject in an experiment.
5. I thought the explosive device would actually explode.
6. The bomb defusing training helped me very much to complete the task.
7. I wasn't at all nervous when working on the explosive device.
8. The electrodes connected to me caused me to feel very nervous.
9. After a few minutes, I stopped paying attention to the electrodes.
10. I felt the scenario training helped me to be very effective in defusing the bomb.

Subjects responded to the questions using a 6-point Likert scale with a '1' meaning 'Very Strongly Disagree' and a '6' meaning 'Very Strongly Agree.' The results of the survey are shown in Figures 4.40 through 4.49.

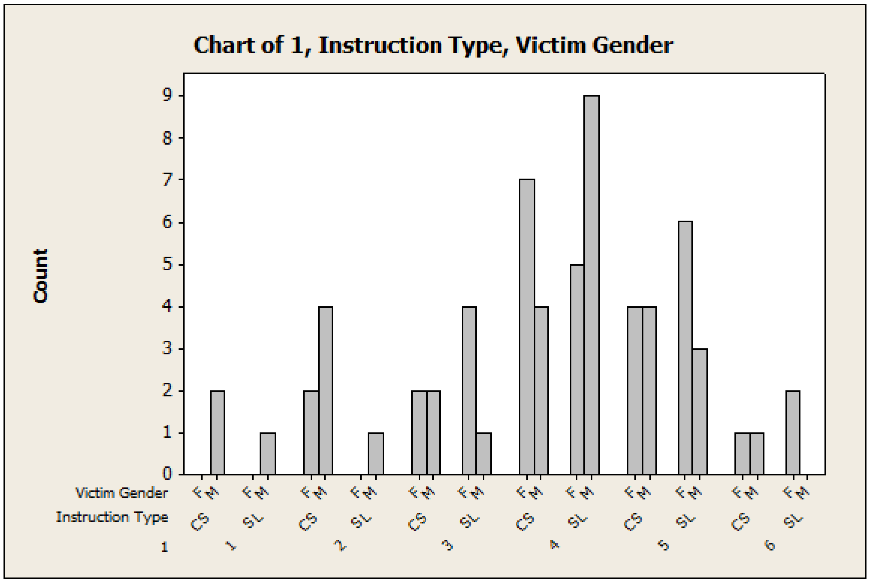


Figure 4.40: Results to Anxiety Questionnaire Question 1

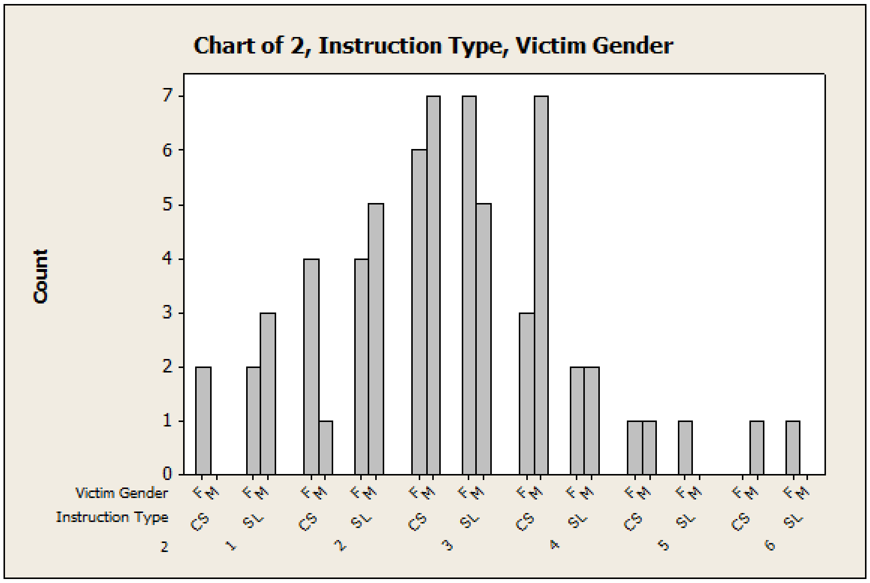


Figure 4.41: Results to Anxiety Questionnaire Question 2

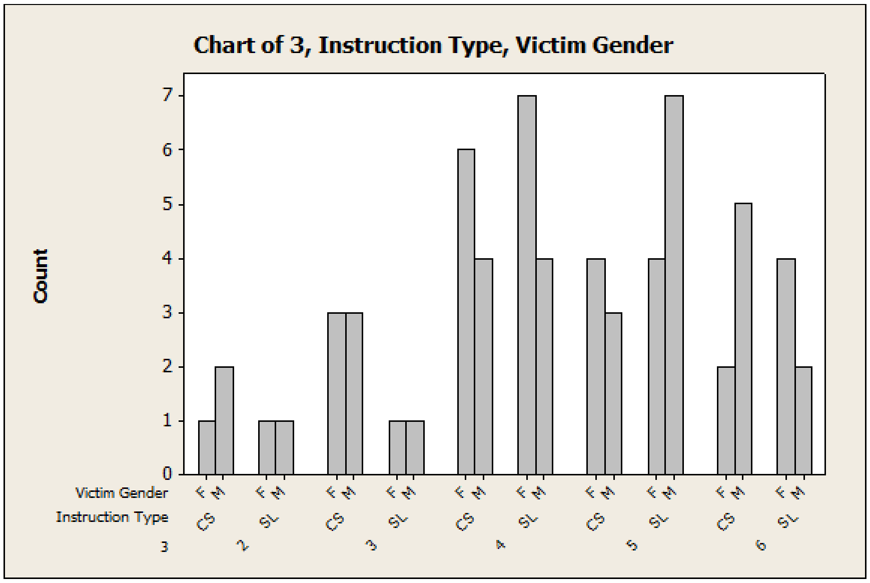


Figure 4.42: Results to Anxiety Questionnaire Question 3

The results of Questions 1 to 3 are displayed in Figures 4.40 to 4.42 respectively. The results indicate that many subjects felt anxious when going through the final scenario. Most subjects noted that the final scenario did not seem artificial to them. Additionally, the subjects indicated that they were focused during the scenario despite the victim's constant pleads for help and questioning of the subjects' abilities.

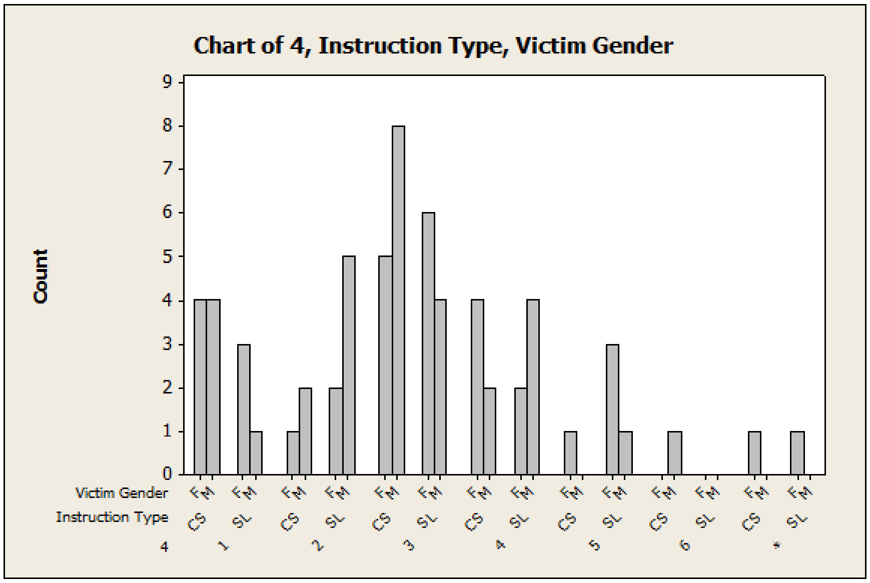


Figure 4.43: Results to Anxiety Questionnaire Question 4

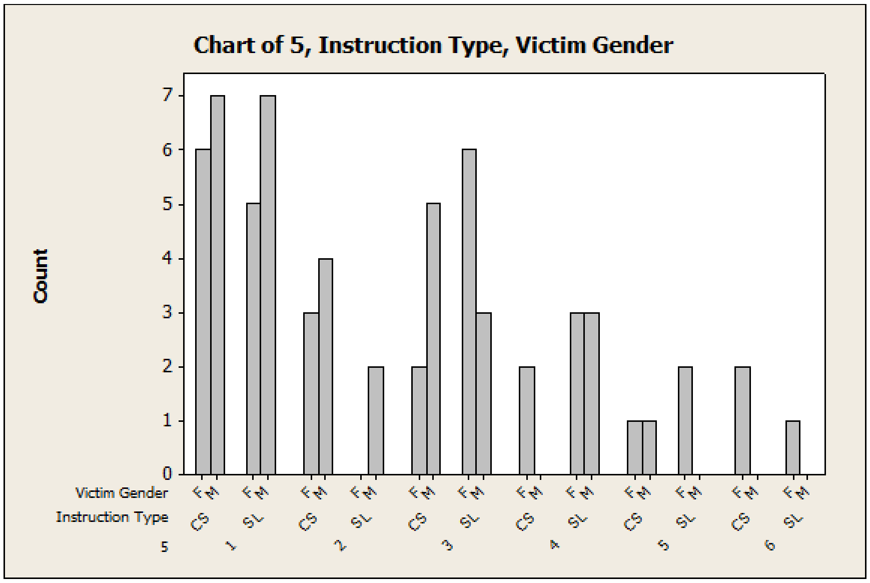


Figure 4.44: Results to Anxiety Questionnaire Question 5

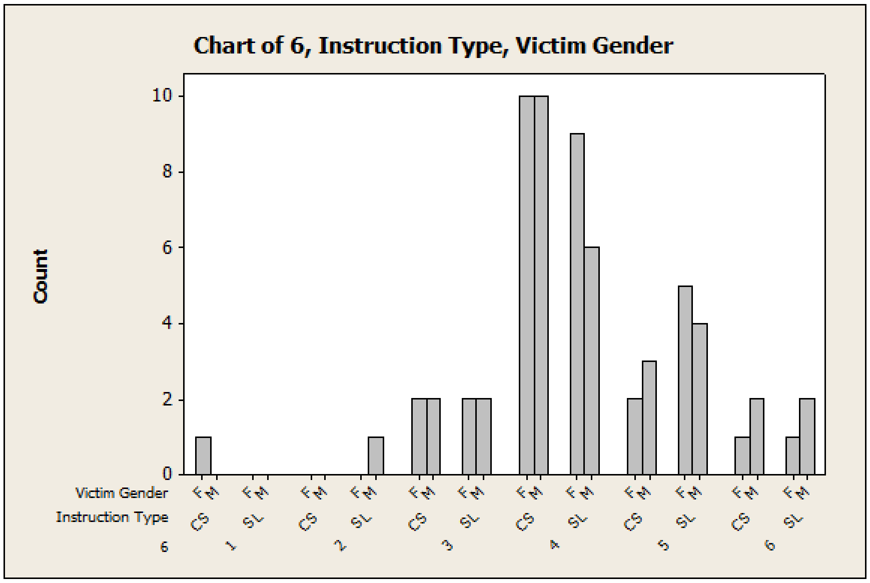


Figure 4.45: Results to Anxiety Questionnaire Question 6

Figure 4.43 indicates that the subjects felt that their anxiety was not just a result of being part of an experiment. However, as Figure 4.44 shows, a majority of the subjects did not feel that the mock explosive would actually detonate. Figure 4.45 shows that a majority of the subjects felt that their training sufficiently prepared them for the scenario.

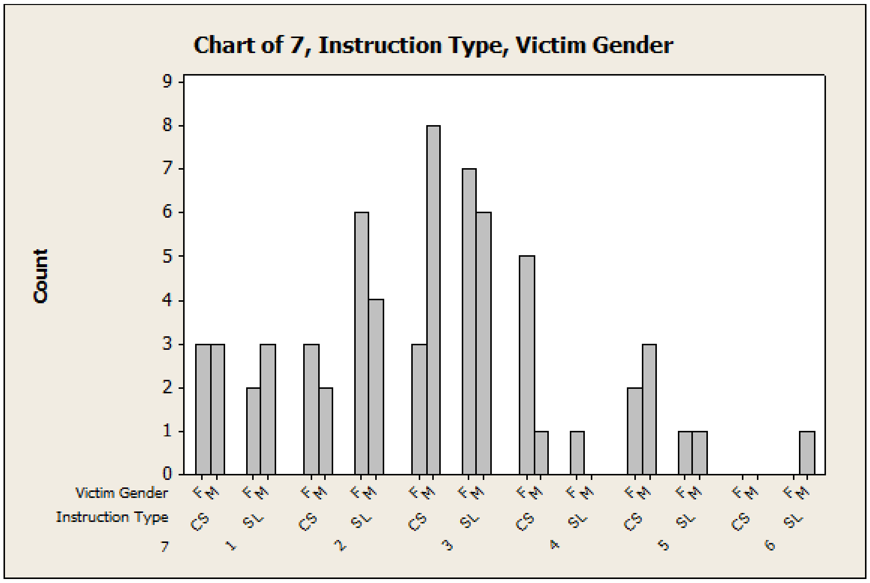


Figure 4.46: Results to Anxiety Questionnaire Question 7

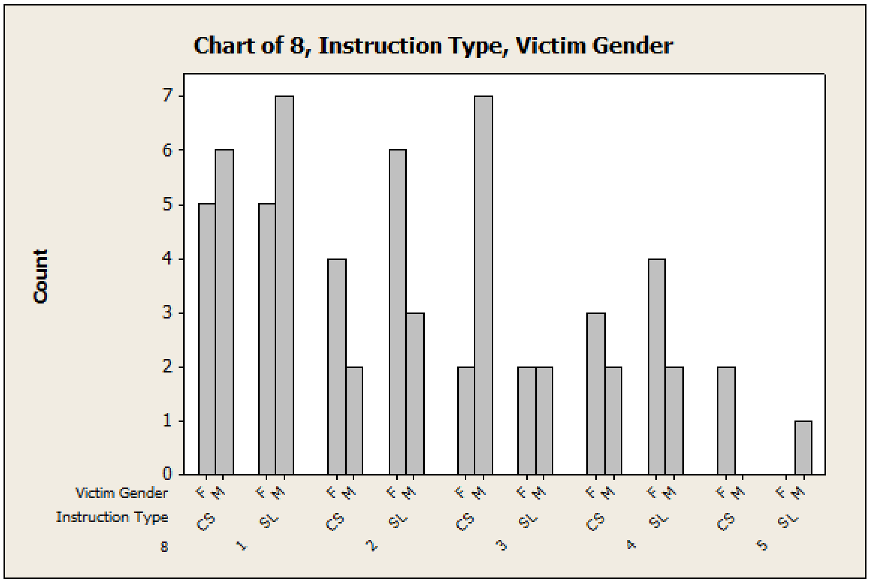


Figure 4.47: Results to Anxiety Questionnaire Question 8

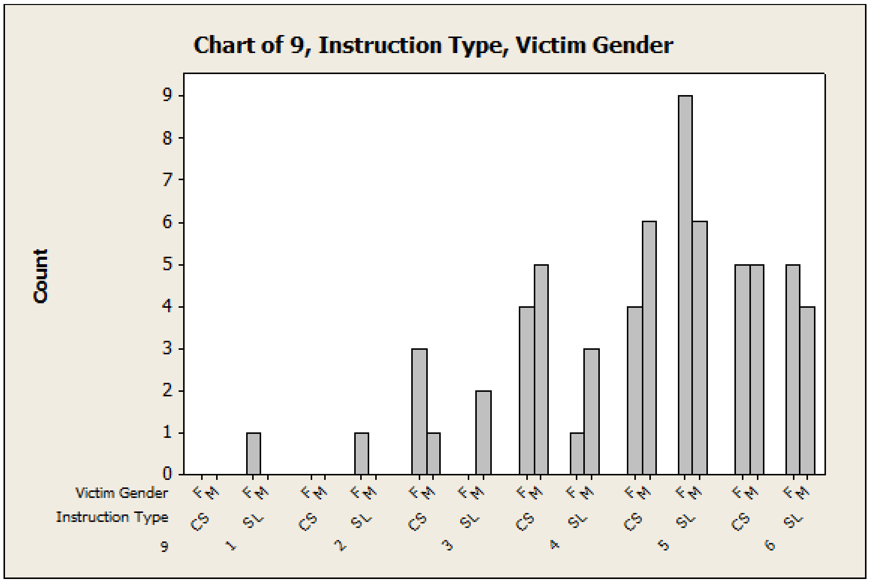


Figure 4.48: Results to Anxiety Questionnaire Question 9

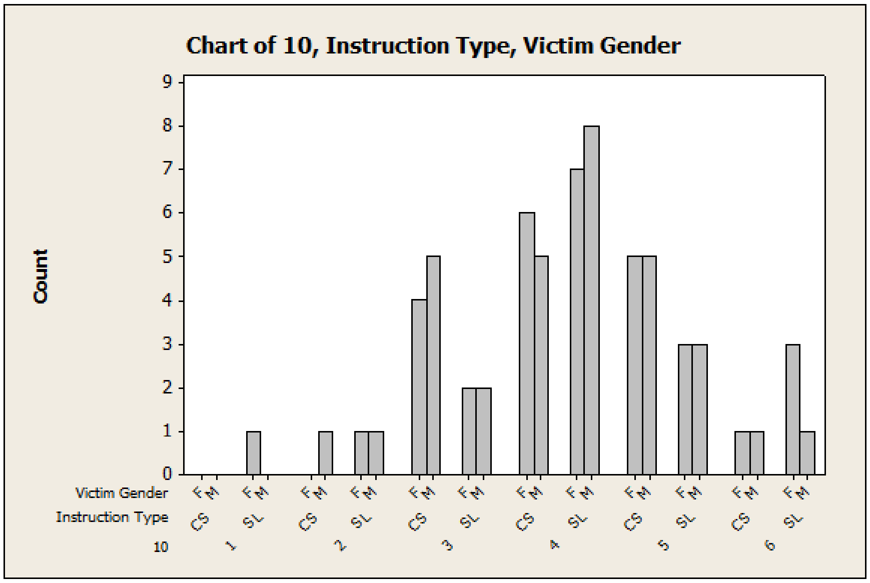
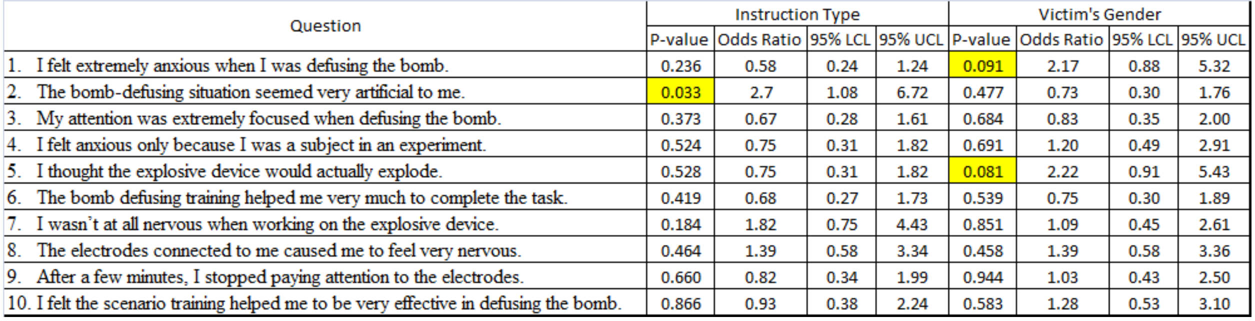


Figure 4.49: Results to Anxiety Questionnaire Question 10

As shown in Figure 4.46, a majority of the subjects felt nervous during the scenario. However, as Figure 4.47 shows, the nervousness was usually not a result of the sensors. Figure 4.48 shows a majority of the subjects stopped focusing on the sensors when they started the scenario. The final question's responses are shown in Figure 4.49. In this figure, most subjects believed that their training the past two days prepared them to defuse the explosive device.

Just as with the Day 2 surveys, the surveys for the final day are generally positive and reflect that the subjects experienced a level of anxiety as a result of the experiment and not just from being wired to physiological sensors. It should be noted that the reported scores are broken up by Instruction Type and Victim's Gender. The purpose is to not only show how subjects tended to report on the aggregate, but also how they reported within the different groups created in the experiment. Table 4.1 contains the results of using an ordinal logistic regression model to analyze each question.

Table 4.7: Results of Anxiety Questionnaire



Question	Instruction Type				Victim's Gender			
	P-value	Odds Ratio	95% LCL	95% UCL	P-value	Odds Ratio	95% LCL	95% UCL
1. I felt extremely anxious when I was defusing the bomb.	0.236	0.58	0.24	1.24	0.091	2.17	0.88	5.32
2. The bomb-defusing situation seemed very artificial to me.	0.033	2.7	1.08	6.72	0.477	0.73	0.30	1.76
3. My attention was extremely focused when defusing the bomb.	0.373	0.67	0.28	1.61	0.684	0.83	0.35	2.00
4. I felt anxious only because I was a subject in an experiment.	0.524	0.75	0.31	1.82	0.691	1.20	0.49	2.91
5. I thought the explosive device would actually explode.	0.528	0.75	0.31	1.82	0.081	2.22	0.91	5.43
6. The bomb defusing training helped me very much to complete the task.	0.419	0.68	0.27	1.73	0.539	0.75	0.30	1.89
7. I wasn't at all nervous when working on the explosive device.	0.184	1.82	0.75	4.43	0.851	1.09	0.45	2.61
8. The electrodes connected to me caused me to feel very nervous.	0.464	1.39	0.58	3.34	0.458	1.39	0.58	3.36
9. After a few minutes, I stopped paying attention to the electrodes.	0.660	0.82	0.34	1.99	0.944	1.03	0.43	2.50
10. I felt the scenario training helped me to be very effective in defusing the bomb.	0.866	0.93	0.38	2.24	0.583	1.28	0.53	3.10

The response variable is the range of scores from 1-6. The predictor variables are the Instruction Type and Victim's Gender. For all questions, the odds ratio for the Instruction Type refers to the odds, relative to the case studies, of a subject receiving the Second Life training would disagree with the given statement. Likewise, the odds ratio for the Victim's Gender refers to the odds, relative to the subjects with female victims, of a subject with a male victim would disagree with the given statement. The highlighted cells in Table 4.1 indicate the significant results at the 90% significance level. They indicate that subjects with female subjects tended to report a high sense of anxiety more often than subjects with male victims. The same is the case when it comes to reporting whether they thought the bomb would explode. The only question with a significant response for the Instruction Type was Question 2. Table 4.1 indicates that subjects who received the Second Life were almost three times as likely to not agree that the situation seemed artificial as the subjects receiving the case studies.

This section contains analysis for all three days of the experiment. More specifically, it contains the results of the tests used to determine the whether one instruction type was more effective than the other as well as the results of the heart rate measurements. In the next and final section, some explanations of the results reported in this section are provided as well as possible implications of the results and some recommended best practices.

Study Conclusions

5.1 Summary of Results

The purpose of this experiment was to compare two instruction types in regards to their effectiveness at teaching some basic skills in properly defusing improvised explosive devices. In addition, there is interest in determining if the gender of the victim had any effect on the subjects' ability to defuse the explosive. Overall there were eleven tests of hypothesis performed.

1. RSP Score vs. RSP Assessment
2. Binary Explosion vs. RSP Assessment
3. Defusing Score vs. RSP Assessment
4. RSP Score vs. Instruction Type and Victim's Gender
5. Binary Explosion vs. Instruction Type and Victim's Gender
6. Defusing Score vs. Instruction Type and Victim's Gender
7. Time vs. Instruction Type and Victim's Gender
8. Baseline Heart Rate vs. Instruction Type
9. Paired t-test between Baseline Heart Rate and Scenario Heart Rate
10. Change in Heart Rate vs. Instruction Type and Victim's Gender
11. Questions from Anxiety Questionnaire vs. Instruction Type and Victim's Gender

The eleven tests can be broken up into four groups. The first group consists of testing for a linear relationship between each of the particular assessment measures and the RSP Assessment completed after the first day of training. The second group consists of testing for a linear relationship between each of the particular assessment measures and the three predictor variables (Instruction Type, Victim's Gender, and Time). The third group consists of analyzing the heart rate data. The final group consists of the results from the Anxiety Questionnaire.

5.1.1 Group 1: RSP Assessment

The results indicate that the RSP Assessment completed following the first day of training is not predictive of the subjects' performance during the final day of the experiment for any of the assessment measures. It can be concluded that just because a subject performs well on a test does not mean they can be as successful in the field. Conversely, the opposite should hold as well. Just because a subject performs poorly on a test does not mean they would perform poorly in the field. This indicates that success would be dependent on some other variable.

5.1.2 Group 2: Assessment Measures

In this experiment, the primary variable being tested is the instruction type the subjects receive on the second day of training. The results of the experiment show that there is no significant difference between the two instruction types when it comes to performance in any of the assessment measures at the 95% significance level. However, Instruction Type is significant at the 90% significance level. Furthermore, the subjects trained in Second Life outperformed those trained using case studies 21 successes to 14. This gives credibility to the idea that Second Life is an effective tool for learning.

One interesting result was that the Victim's Gender was found to significantly affect the subjects' RSP Score. In particular, subjects with a male victim outperformed subjects with a female victim. This could be an indication that the subjects' memory of the render safe procedures could have been compromised, if they were adversely affected by the sight of a female victim. The implication of this hypothesis is that subjects with male victims are not adversely affected and are able to remember their tasks better. While Victim's Gender was not found to be significant when it came to the subjects' ability to simply defuse the bomb, there was a significant relationship at the 90% significance level between Victim's Gender and the Defusing Score. This is due to the fact that there are two possible scores for successfully defusing the explosive. The highest score of '5' was only attained by subjects with male victims. Therefore, subjects with female victims were more likely to receive lower scores.

Additionally, subjects with female victims tended to take longer to complete the scenario. This phenomenon coupled with the fact that subjects with female victims performed worse at remembering Render Safe Procedures and at obtaining a high Defusing Score indicates that female victims adversely affected the subjects' performance. There was no significant difference in the time taken to complete the final scenario between the two instruction types.

5.1.3 Group 3: ECG

The results of the ECG indicate that there was no difference in baseline heart rate between the two instruction types. In addition the change in heart rate from baseline to the scenario was not found to be affected by Instruction Type or Victim's Gender. While the predictor variables were not found to significantly affect the heart rate, the heart rate did change from baseline to scenario. However, it unexpectedly dropped. The reasoning for this is that the subject may have calmed down once the scenario began. While the subject did not experience a raise in heart rate, the average baseline heart rate was really high. An average adult has a resting heart rate between 60-100 BPM (Mann 2011). The subjects had an average heart rate of over 120 BPM during the base line. This indicates that the subjects were experiencing a high level of anxiety before starting the experiment. Therefore the anxiety appears to be a result of the subjects being in an experiment. The final grouping summarizes the results of the Anxiety Questionnaire.

5.1.4 Group 4: Anxiety Questionnaire

The results of the questionnaire appear to reflect the hypothesis that the subjects were in fact feeling a sense of anxiety during the scenario. However, the subjects' general feeling that the anxiety was not simply a result of being in an experiment appears to be contradicted by the results of the heart rate measurements. While there was no difference in the change in heart rate between either the two instruction types or victim genders, there were differences between the two predictor variables when it came to how the subjects responded on the questionnaire. Subjects with female victims tended to report anxiety during the scenario more often than subjects with male victims. They also reported that they thought the bomb would actually explode more than the subjects with male victims. These higher levels of reported anxiety and fear of explosion could be indicative of why the subjects with female victims performed worse than those with male victims. As far as the instruction type is concerned, subjects receiving Second Life tended to disagree with the assertion that the scenario seemed artificial more often than the subjects receiving the case studies.

5.2 Implications

When it came to Instruction Type, there was a significant difference found between the effectiveness of the two instruction types at the 90% significance level. This finding suggests that a virtual simulation can teach a task more, or at least as effectively as traditional paper-and-pen training. Therefore, if an organization is trying to decide whether to invest in virtual reality for pedagogical purposes, they have evidence to support the notion that their students can learn what they need to learn with the same or even better proficiency as they could have learned with a traditional paper-and-pen study. There is a potential drawback to using virtual reality and that is in the initial cost of its implementation. If an organization wants to design a virtual world in order to specifically teach something unique to their organization, the cost of creating the world can be expensive at first. The cost of contracting developers to create the software is more than printing out a handout with instructions. However, if the end result is that those receiving training (employees, students, etc.) learn better with the virtual world then the payoff can be seen in fewer mistakes. This leads to higher quality work which in turn makes the organization more profitable.

5.3 Best Practices

When conducting this experiment, there were some important lessons learned. The primary lesson is to have a properly updated computer lab available through the duration of the experiment. There were multiple instances when certain functions of the training did not work properly due to certain computer programs not being downloaded. While downloads were eventually able to be made, it delayed the subjects' instruction. Having an updated computer lab is important in order to maximize the benefit of a virtual world. If the computers are not properly updated and maintained, certain problems such as rendering issues and freezing can occur which occurred during the experiment.

Data collection for this experiment was conducted using the Blackboard Learning System. Each subject was assigned a separate profile corresponding to their subject number. The data was available to be downloaded by the project leaders who were given administrator access.

However, an issue arose due to the original data collection settings. All of the questionnaires and surveys were set to "Survey" mode in the data collection settings. As a result, when trying to collect the data, the data was randomized making it impossible to know to which subject each row of data belonged. In addition, there were dummy profiles set up for practice purposes. The scores from those results would get mixed in with the actual data making it impossible to determine which data was real and which data was practice data. As a result, the data was gathered by going into each profile one by one until all data was properly gathered. In order to avoid this problem, the data collection setting should be set to "Test." This allows the data to be associated with their subject number in order to properly analyze the data.

The final recommendations deal with the final day's experiment. When taking physiological data, it is important to keep outside factors such as lighting and ambient sound as constant as possible. This means turning on all the lights in the experiment room as well as limiting access to the room to those directly involved in the experiment. A method used in the experiment that aided in data collection for the final day was the implementation of a video camera. The subjects were recorded during the final day of the experiment allowing for review of the film for any missed data. The tape was not used as the primary method of data collection, but it is a useful tool in case some data is lost or accidently not recorded. These best practices serve as recommendations for improving the experiment.

REFERENCES

Astin, A. (1984) “Student Involvement: A Developmental Theory for Higher Education,” *Journal of College Student Personnel*, Volume 25, pp. 297-308.

Bajka, M., Tuchschmid, S., Streich, M., Fink, D., Szekely, G., and Harders, M. (2008) “Evaluation of a new virtual-reality training simulator for hysteroscopy,” Springer Science and Business Media.

Bertrand, J., Babu, S., Polgreen, P., and Segre, A. (2010) “Virtual Agents based Simulation for Training Healthcare Workers in Hand Hygiene Procedures,” *IVA'10 Proceedings of the 10th International Conference on Intelligent Virtual Agents.*

Bowman, D. & R. McMahan (2007) “Virtual Reality: How Much Immersion is Enough?” *IEEE Computer,* Volume 40, Issue 7, pp. 36-43.

Bronak, S. R. Riedl, & J. Trashner (2006) “Learning in the Zone: A Social Constructivist Framework,” *Interactive Learning Environments,* Volume 14, Issue 3, pp. 219-232.

Cai, H. (2008) "Service Design for 3D Virtual World Learning Applications," 2008 IEEE International Conference on Web Services, pp. 795-796.

Cheng, X., Gu, R., Chen, M., and Weng, Y. (2010) "A Virtual Assembly System on Automobile Engine for Assembly Skills Training," American Society for Engineering Education.

Cooper, K. (2010) "Go With the Flow: Engagement Factors for Learning in Second Life," SCS pp. 1-9.

Datey, A. (2001) "Experiments in the Use of Immersion for Information Visualization," Masters Thesis, Virginia Tech, <http://scholar.lib.vt.edu/theses/available/etd-05092002-151043/>

Dickey, M. (2003) "Teaching in 3D: Pedagogical Affordances & Constraints of 3D Virtual Worlds for Synchronous Learning," *Distance Education*, Volume 24, Issue 1, pp. 105-121

Durrani, S., Geiger, C., Jones, D., and Hale, K. (2008) "An Approach for Assessing Training Effectiveness in Virtual Reality Environments," Proceedings of the 2008 Industrial Engineering Research Conference, pp. 452-456.

Gaimster, J. (2008) "Reflections on Interactions in Virtual Worlds & Their Implications for Learning Art & Design," *Art, Design, & Communication in Higher Ed*, Volume 6, Issue 3, pp. 187-199.

Goel, L. (2009) "Situated Learning in Virtual Worlds," PhD Dissertation, University of Houston, Houston, TX.

Gruchalla, K. (2004) "Immersive Well-Path Editing: Investigating the Added Value of Immersion," *Proc of IEEE Virtual Reality*, pp. 157-164.

Hammond, G. (2004) "Fit to Think: Conceptual, Critical & Creative Thinking, Retrieved from <http://www.au.af.mil/au/awc/awcgate/awc-thkg.htm>

Heiphetz, A. and Woodill, G. (2010) *Training and Collaboration with Virtual Worlds: How to Create Cost-Saving, Efficient, and Engaging Programs*, pp. 52-56, 105-122, 165-172.

Holm, R., and Proglinger, M. (2008) "Virtual Reality Training as a Method for Interactive and Experience Based Learning," Intelligent Energy 2008, pp. 25-27.

Hornik, S. (2008) "Seventeenth Annual Research Workshop on: Artificial Intelligence and Emerging Technologies in Accounting, Auditing and Tax," *Really Engaging Accounting: Second Life as a Learning Platform*.

Kelly, H. and Cheek, D. (2008) "Designing an Online Virtual World for Learning and Training," Fifth IEEE International Conference on Wireless, Mobile, and Ubiquitous Technology in Education, 2008, pp. 208-209.

Liang J. (2007) “Generation of a Virtual Reality-Based Automotive Driving Training System for CAD Education,” *Computer Applications in Engineering Education*, Vol.17, Issue 2, pp. 148–166.

Liu, X. and Hao, A. (2004) “An Interactive Virtual Environment Inhabited Virtual Agents for Oil-field Safety Operation Training,” Web Technology, pp. 51-60.

Marcos de Moraes, R. and dos Santos Machado, L. (2009) “Gaussian Naive Bayes for Online Training Assessment in Virtual Reality-Based Simulators,” *Mathware & Soft Computing,* Volume 16, pp. 123-132.

Molka-Danielson, J. and Cabada, M. (2010) “Application of the 3D Multi User Virtual Environment of Second Life to Emergency Evacuation Simulation,” 43rd Hawaii International Conference on System Sciences (HICSS), 2010, pp. 1-9.

Monahan, C., Ullberg, L., and Harvey, K. (2009) “Virtual Emergency Preparedness Planning Using Second Life,” SOLI '09. IEEE/INFORMS International Conference on Service Operations, Logistics and Informatics, 2009, pp. 22-24.

Montgomery, D., Peck, E., & Vining, G. (2012). Generalized Linear Models. In *Introduction to Linear Regression Analysis* (5th ed., pp. 421-430). Hoboken, New Jersey: John Wiley & Sons, Inc.

Mott, M. and Rajaei, H. (2010) “Hand Detection and Tracking for Virtual Training Environments,” SCS, pp. 1-5.

Ondrejka, C. (2007) “Education Unleashed: Participatory Culture, Education, and Innovation in Second Life,” *The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning*, pp. 229-251.

Orr, T., Mallet, L. and Margolis, K. (2008) “Enhanced Fire Escape Training for Mine Workers Using Virtual Reality Simulation.”

Pausch, R., Proffitt, D., & Williams, G. (1997) “Quantifying Immersion in Virtual Reality.” SIGGRAPH'97.

Ruffaldi, E., Filippeschi, A., and Avizzano, C. (2011) “Feedback, Affordances, and Accelerators for Training Sports in Virtual Environment,” *Presence; Teleoperators and Virtual Environments*, Volume 20, Issue 2, pp. 33-46.

Sanchez-Vives, M. & Slater, M. (2005) “From Presence to Consciousness through Virtual Reality,” *Nature Neuroscience*, Volume 6, Issue 4, pp. 8-16.

Scriven, M., & Paul, R. (1992) “Critical Thinking Defined,” Handout given at Critical Thinking Conference, Atlanta, GA.

Slater, M., M. Usoh and A. Steed (1994) “Depth of Presence in Virtual Environments,” *Presence: Teleoperators and Virtual Environments*, Volume 3, Issue 2, pp. 130-144.

Slater, M., Spanlang, B., & Corominas, D. (2010) “Simulating Virtual Environments within Virtual Environments as the Basis for a Psychophysics of Presence” *ACM Transactions on Graphics*, Volume 29, Issue 4, Article 92.

Slater M., Khanna, P., Mortensen, J., & Yu, I. (2009) “Visual Realism Enhances Realistic Response in an Immersive Virtual Environment,” *Computer Graphics & Applications* Volume 29, Issue 3, pp. 76–84.

Usoh, M., K. Arthur, M. Whitton, R. Bastos, A. Steed, M. Slater & F. Brooks (1999). “Walking > Walking-in-place > Flying in Virtual Environments,” *Proc. of SIGGRAPH 99*.

VERT: Virtual Environment Radiotherapy Training. <http://www.virtalis.com/files/articles/flier_whatisvert.pdf>

Wen, G., Xu, L., Chen, H., Shang, X. (2009) “Horizontal Directional Drill Rig Operating Training System based on Virtual Reality Technology,” *ICPPT 2009: Advances and Experiences with Pipelines and Trenchless Technology for Water, Sewer, Gas and Oil Applications*, Volume 10, pp. 1093-1103.

Witmer B. & M. Singer. (1998) “Measuring Presence in Virtual Environments: A Presence Questionnaire,” *Presence: Teleoperators &Virtual Environments*, Volume 7, Issue 3, pp. 225-240.

A Quantitative Competitive Study of Virtual World Language Learning Compared to Classical Classroom Training

Summary

This study reports on a quantitative, competitive analysis regarding virtual world learning and classical (teacher centric) classroom training. The Computer Assisted Language Learning (CALL) system that is used in this study is provided by Alelo (www.alelo.com) as their Tactical Language Training (TLT) system. This package focuses on French spoken in Sub-Saharan Africa and the culture of the people of Sub-Saharan Africa. As a comparison, a classroom setting is structured with a teacher to provide training over the same subject matter. Sixty four college students over the age of 18 were used in two experimental groups: virtual world training (n = 31) and classical classroom training (n = 33) were statistically compared at the 95% level based on pre-test and post-test improvement. Both test groups improved, but at the 95% level neither method of instruction could be proved to be superior. Therefore, the CALL methods are assumed to be equal to the classical classroom instruction methods in effectiveness.

1.0 Introduction

Technology is changing rapidly, and Computer Assisted Language Learning (CALL) is represented by multiple forms of new technology to support language learning. The pedagogical progress includes the rise of computer language learning systems such as the commercially successful Rosetta Stone language learning system. Another successful CALL package is produced by Alelo (www.alelo.com) as their Tactical Language Training (TLT) system. The development of the Alelo TLT system was supported by the US DARPA (Defense Advanced Research Projects Agency) as a collaborative project with the USC Information Sciences Institute. In 2002, soldiers returning from Afghanistan reported on the additional challenges they faced when entering potentially hostile villages, unprepared for dealing with the dialects and etiquette specific to the region. The result was a series of language and culture learning packages that supported military operations in Afghanistan, Iraq and other theaters to "ensure that no American soldier or Marine would again ride blind into a foreign culture." This paper represents a competitive study to quantitatively measuring the effectiveness of Alelo TLT when compared with classical classroom training covering identical subject matter.

The Alelo TLT System for French language followed the previous computer-based systems produced by Alelo as a blend of human-computer interaction technology and task-based language techniques using learner-centered methodologies. The development of hybrid and online courses in the last years at the post-secondary level in US shows the appeal of technology and its numerous advantages. As of 2012 there were 5.6 million college students taking courses online or hybrid (Russell, 2012). One of these advantages is that technology can simulate real-world environments and engage the learners linguistically and culturally mainly by using the language to complete a specific task. For example the Virtual Cultural Awareness Trainer (Johnson, 2010) analyzed how much language training should be in the pre-deployment training and concluded that even basic language skills can significantly improve a mission outcome. It has been proved that learners using SLEs (synthetic learning environments) have a higher self-efficacy, retention and verbal performance than those using traditional classroom methods (Sitzmann and Ely, 2011).

In the case of the French training language package the learners acquire enough language skills and culture immersion to carry out tasks abroad such as obtaining directions, going through the customs, getting a cab or befriending locals. The SLEs attempt to recreate the real life world in order to capture cognitive and affective skills to complete a task in real life situations. The artificial intelligence part of the program processes oral communication through speech recognition; it controls the non-players action and assists in the the instructional process. Further the immersion using the SLEs has been proven to raise the knowledge acquisition and retention of language skills (Ricci, Salas, and Cannon-Bowers, 1996).

Created with these specific targets in mind, the Alelo TLT (2005) is based on a practical and theoretical technology-based individual training started at University of Southern California in 2003. Initially the project was funded by the Defense Advanced Research Projects Agency, the U.S. Special Operations Command followed by the U.S. Marine Corps, Army, and the Australian Defense Forces.

Before developing the French program, Alelo and USC developed successfully the Tactical Levantine Arabic, Tactical Iraqi, and the Tactical Pashto so they were able to predict the technical issues and to improve with the new language program. These systems were systematically under external formative evaluation in order to improve different features. The Tactical French was designed for personnel going in Sahel Africa with a particular focus on Chad. This training project was planned in combination with a following development of a 3D virtual world and game.

The structure of the French Alelo TLT contains a couple of modules satisfying basic language acquisition requirements such as the introduction of grammar structures within a task-based curriculum; the insertion of game-based and interactive multimedia practice as a serious game; the cultural knowledge and the intercultural competence skills. The modules can be generally relevant or mission-specific with a couple of initial selective choice to place the learner on the right module. The program uses standard characters (the avatar, the locals) meant to individualize the learning experience.

The first developments of previous TLT’s showed a series of technical challenges such as the impossibility to anticipate all the learner’s intents and the dialog moves. For example the game-based technology puts in place a series of both task-based and mission-based scenarios and with the tremendous help of voice-recognition it builds a large interactive dimension to the entire learning process to insure the language use in context and culture. While the task-based scenario is focused on acquiring a certain skill the mission-game scenario focuses on acquiring skills and knowledge collectively. This is meant to simulate real life conditions and to replace the more traditional peer exchange or tutor-student exchange. Promoting interactive engagement enhances active learning (Marsella, 2003) and the learner’s self-confidence while identifying as a character in the scenario on three educational levels: the cognitive content, the motivational and social dimensions. In order to enhance the communicative skills the social input during simulations is central by providing personal interaction. A conversational politeness hypothesis has been formulated stating that learners engage with computers tutors as social partners and use the same social norms as in the human conversational etiquette (Reeves and Nass, 1996; Mayer, 2005;

Wang, 2008). Formulating dialogues and question with politeness strategies works because of the human desire to be approved of by others. Using politeness strategies is a matter of reinforcing the learner's confidence and sense of control.

The Alelo TLT's through linguistic and social simulations have been proven to enhance the learner's confidence and motivation (Emonts, 2012). Its software includes reputed learning technologies such as speech recognition, artificial intelligence, dialogue modeling and game-based learning. Before starting a scenario the learner receives a summary of the task proposed and the objectives to complete. These objectives follow the steps building the scenario and improving the learner's language and culture-oriented skills. The final score underlines the linguistic and the cultural errors, tailoring the activities to each learner's skills. The cognitive psychology theory states that every learner is constructing his own knowledge structures and ways of thinking based on their previous baggage of knowledge. This theory of learning, combined with the ideas of experiential learning, encourages the involvement of the learner in the process and his critical input.

The learner engages through clicking the microphone icon and speaking into the headset microphone. An additional tool is the help menu to offer a choice of different topics and dialogues. This gives a realistic belief in one's own efficacy and power to chose the right strategy for a given scenario.

The sequence generator helps with practicing oral pronunciation. The scenarios teach about cultural knowledge or combine the language structures learning with cultural tips. All of these scenarios are using role-playing to ensure the interactive immersion in a responsive and adaptive environment.

The use of digital pictures and game-like videos have been used for their engaging output while a reader would try use the visuals and make sense of the text or the question presented. Visualization has been identified as a powerful tool to enhance language comprehension and critical thinking (Bell, 1991). Keeping the learners engaged is what makes the serious games a great tool combining the challenge to win with the instant visual feedback. In addition it has been noticed that learners tend to engage more in active learning when the program uses a conversational style than a formal style. The players' interaction is based on a computer-human interaction and makes for an excellent animated pedagogical tool of tutorial dialogue in a virtual world simulating real world. The bigger the number of dialogue simulations the more efficient is the architecture of the tutorial. For the French, Alelo TLT more than 50 dialogues structures were created (Johnson, 2009).

To really enhance the cultural immersion a series of anthropological methods have been used to raise data on subtle decisions and linguistic interpretations in cross-cultural encounters. This data was subsequently under review by cultural consultants as bilingual natives with life experience in the United States and the country of the target language. The use of computer-based programs with a social dimension is one of the most important advantages in second language education as shown in the Media Equation hypothesis (Wang, 2008). This continues to be a viable alternative to the in class peer exchange or oral practice with a virtual tutor in addition to other benefits like increasing the learner’s participation and time on task, reducing anxiety, developing oral communication, and tailoring the task on individual pace and needs (Sanders, 2005).

The main tool to enhance social interaction is the speech recognition as an automated words and phrases interpreter. This advanced technology also interprets the cultural tips such as body language and contexts. With regards to The French Alelo TLT, where the speech is irregular, the speech recognizer is designed to decode the speech into a context and disregard speech errors. According to evidence-based cognitive and learning theories such as the Vygotsky’s 1978 learning model the role of the environment and social interaction is essential in acquiring new information. Using this new technology also allowed registering the vocal reproduction as necessary in order to improve speech production and proper pronunciation. In addition to these features, the locomotion, the gaze, and the gestures simulation of standard characters are focusing the learner’s attention and enhance the lifelike scenarios. This is how the computer engages and keeps the learner motivated. In order to do this a high quality and dramatic sense of communication must be achieved.

Within the task-based practice there is the skill builder that teaches vocabulary and expressions, the exercises, and the testing through quizzes to enhance learner modeling. The mission-based scenarios use games directing the learner through a series of commands in the target language and the mission games demanding linguistic production. It has been noticed by the developers that the task-based practice gives faster, good results in improving immediate language skills. This is due to the self-monitoring option as an active involvement in the learning process. To check the progress the learner is tested through different tools: performance on quizzes, games, dialogue simulations.

Both the task-based and the mission-based scenarios activate linguistic contents and grammatical structures while informing the learner on the micro sociological concepts and customs to enhance cross-cultural communication. The task-oriented practice is based on the repeated training as a well-known tool to strength the memory and structures retention (Johnson, 2012). A commonsense theory of microsociology based on concepts such as commitment, shared plans, good will, and friendship was one of the components on the Alelo TLT French application as it has been previously for the other languages application (Hobbs, 2011). In these scenarios the learner is prompted to speak with a non-player character using a headset microphone. The speech recognition compares this string representation with the microsociology and its communication protocol in order to react at the learner linguistic production.

As the Tactical Iraqi the French program’s architecture includes two sections or modules. The first one, the skill builder module, includes authentic materials, audio materials followed by multiple-choice evaluations. In the second module presents a free writing part based on a variety of sources such as articles, tables or graphics and free oral utterances as a part of a simulated conversation or a cultural presentation. The skill builder provides content in vocabulary, grammar structures and phonetics as well as cultural tips. To pass the exercises and the tests the learners have to listen, speak, and write. The skill builder includes vocabulary pages and exercise pages about utterance formation, multiple-choice and match-item exercises.

The task-game is helpful in improving pronunciation and listening skills related to colors, numbers, directions. For example the Tactical French offers a first unit lesson preparing the learner on how to greet a person or how to react and talk in different situations: arrival at an airport, passport control, clearing customs, or getting transportation.

The mission games immerse the learner in direct interactions with non-player local characters within simulation of real-life scenarios. The artificial intelligence and the dialogue simulations with non-players are proven to raise the learner's self-confidence. This interpersonal exchange with virtual characters had good results in many fields such as the medical one through the Carmen's Bright IDEAS in August of 2004 (Johnson, 2004). The Tactical French would immerse the learner in different scenarios about: leaving the airport, getting a hotel room or presenting himself and his destination.

Of course, the Alelo TLT is only one of many CALL software packages using new technologies. In order to improve these CALL software packages and to understand their role in the world of language learning, the evaluation of effectiveness must be undertaken. This is the role of this paper.

In the course of the investigation that is the focus of this paper, the authors observed many papers announcing the use of new technology for language learning in a new way. Examples include task based language teaching in a blended learning context (Thomas, 2013), learning French using a digital kitchen (Seedhouse et al. 2013), video gaming (Benson & Chik, 2011), digital story telling (Kimura, 2012), learning using mobile devices (Viswanathan, 2012), and blended learning with an e-book (Hwa et al., 2012). In each case a new technology is offered, and in some cases the new technology is comparatively evaluated.

Many of the comparative evaluations represent the use of questionnaires to evaluate effectiveness (Hamel, 2013), qualitative studies (Egbert & Huff, 2011), transcripts, observations, video clips, interviews, and focus groups (Thomas, 2013), regression analysis to identify related factors (Chen, 2012), and other non-direct or non-quantitative measures. In contrast, Hwa et al. (2012) use pre-test and post-test on a competitive basis with two samples of ten subjects to prove effectives. This paper also uses a pre-test and post-test regimen on a competitive basis with two approximately equal samples totaling 64 subjects. This quantitative comparative study reaches conclusions at the 95% confidence level using rigorous statistical methods. These methods are believed by the authors to provide a rigorous scientific basis to the work reported in this paper.

As a means to form context, the next section of this paper provides a brief literature review. This literature review is followed by a section on experimental methodology, and then followed by a section on data analysis. Finally, the last section provides a set of conclusions.

2.0 The Literature

A review of the literature regarding the use of computers to aid in language learning was conducted. While there are those that are excited to use the new technology, others are more cautious. In addition, there are people who reject the notion that computers can be used effectively for language training. The negative opinions tend to originate from traditionalism and a fear of the role of educators being lessened due to the use of computers. The role of the educator in learning is a topic of debate between followers of psychologists Jean Piaget and Lev Vygotsky. As a result, this literature review focuses on the philosophical debate between Piaget and Vygotsky and how it relates to experiments conducted in the field of using virtual reality for language education. In section 2.1, the debate is explained in full. Section 2.2 analyzes the role of tutor in virtual reality language training. Section 2.3 discusses the advantages and disadvantages of virtual reality as a training tool. Section 2.4 discusses methods used to evaluate the results of the experiment discussed in this paper, and lastly, section 2.5 concludes the literature review and transitions to the experiment reported on in this paper. Before moving on, the philosophies of Piaget and Vygotsky are discussed in order to demonstrate how they relate to computers being used for language learning.

2.1 Piaget vs. Vygotsky

In his paper highlighting the debate between Piaget and Vygotsky , Levy (1998) explains that both psychologists share one thing in common: they are constructivists. Constructivism is based on the theory that people learn by gathering information from their surroundings and piecing the information, or constructing it, in their minds in order to better understand it. It should be noted that both Piaget and Vygotsky were both studying how children learn their first language. However, their followers have taken their work and applied it to how people learn in general. While Piaget and Vygotsky agree on the idea of constructivism, the similarities between the two psychologists end there. In order to understand how these two opposing philosophies relate to the field of using computers for language training, their respective philosophies need to be explained.

Jean Piaget was a Swiss psychologist and, according to Levy, is considered the founder of constructivism. As the founder of constructivism, Piaget (1980) asserts that knowledge is not accumulated from simply observing the environment, but from “a structuring activity on the part of the subject.” Levy states that Piaget sees this subject as a “lone, inventive scientist trying to make sense of the world.” As can be seen, Piaget believes that the learner is individualistic in their learning process and not dependent on a teacher or tutor.

Vygotsky, a Russian psychologist whose theories serves and the foundation of sociocultural theory, believes that the key to a person learning is social interaction. While Piaget sees learning as a byproduct of adaptation to the outside world, Vygotsky sees learning as a result of communication with other people and sees language as a tool to be used by people to communicate with their environment. Whereas Piaget sees the relationship between a learner and their environment as parasitic, the learner gathers information from the environment in order to understand it, Vygotsky sees the same relationship as symbiotic. Through language, the learner and the environment can communicate and benefit each other.

While the use of the words “parasitic” and “symbiotic” are used to describe the relationship between the learner and the environment might indicate that this paper is championing Vygotsky’s point of view, it is not. As Levy states, “the question of whether learning should consist of mainly of social or individualistic activity is an ideological issue.” However, this ideological issue presents some interesting possibilities when it comes to conducting research in the area of using computers for language training. In the next section, the two psychologists’ theories on learning are used to shape their views on the role of the tutor in language learning. Research conducted championing both sides of the debate are then presented to demonstrate how different researchers have tested these theories.

2.2 The Role of the Tutor

Just as Piaget and Vygotsky disagree when it comes to how they see the relationship between the learner and the learner's environment, so too do they disagree when it comes the role of the tutor in learning. By referencing Jones and Mercer (1993), Levy states that from Piaget's view the tutor's role in the learning process is to provide the learning environment in which the learner can develop. In the contest of using computers to teach language skills, the tutor's responsibility would be to provide the tool with which the learner can use to develop their skills. Whether the tutor develops the software or simply passes on the software that was developed somewhere, the role is still the same: give the software to the student and let them work. The tutor may continue to provide resources for the learner, but the responsibility after that is minimal.

On the contrary, Vygotsky believes that the tutor's role "is central to the learning process" (Levy, 1998). Levy describes how Vygotsky developed the theory of the zone of proximal development which refers to how learners develop from completing assignments which are, at first, outside their skill level. In order to accomplish the tasks, the learner must build their skills up. This is where Vygotsky says the role of the teacher exists. Vygotsky's belief is that the student cannot bridge the gap between what they know and what they need to know in order to complete their assignment alone. Therefore, they need the help of a more knowledgeable tutor to provide them with the information to complete the assignment (Vygotsky, 1978).

After explaining each side of the argument, Levy (1998) critiques each side. As stated, Levy believes that neither philosophy is correct. The matter of "right" depends on the individual who is trying to learn something. He also states that if Piaget's theory is to be used to develop a computer-based tool for language learning, the tool's success depends on how well the tool manages the student's learning and the feedback it provides. In critiquing Vygotsky's point of view, Levy states that according to Anderson et al. (1996), many studies comparing the two philosophies have resulted in no significant differences being discovered.

While findings from comparative studies are said to show no favoritism to one philosophy over the other, this has not stopped people from continuing to test the theories. Goodfellow and Lamy (1998) developed the Lexica On-Line project which created a learning environment for students learning French. The objectives of the project were to:

- Test whether the students would be able to use the lexical tools without fact-to-face supervision
- Create self-sustaining interaction among the students on-line, with minimal intervention from tutors, and
- Introduce the students to the Francophone Web in a controlled way, ultimately guiding them toward the completion of a constructive task.

These objectives fall in line with the theories presented by Piaget. The result of the project was that the students found the project to be engaging. The problem was that the workload was too much considering it was an extracurricular project and the students had a full curricular workload. Another project testing Piagetian theory were conducted by Ibanez et al. (2011).

They looked to “deploy an engaging learning experience to foster communication skills within a 3D multi-user virtual world with minimum teacher’s help.” The virtual world was comprised of a 3D model of Madrid, Spain. In the world, the students would go through simulations of real conversations in Spanish with native Spaniards who were a part of the 3D model. It is important to note though that with the tutor being virtually removed from the experiment, the students began collaborating with each other in order climb the learning curve. While this project was intended to determine how the student would perform without a tutor, the students used each other as tutors at the beginning. Thus, they gave credibility to Vygotsky’s theory that learning is a collaborative effort.

Vygotsky’s theory on the role of the tutor was also supported by a paper by Fox (1998). In this paper, Fox focused on “teaching and learning issues related to technology mediated distance language acquisition, with particular emphasis on the role of the teacher.” Fox asserts that with a distance learning class, communication and support from the tutor is very important. For the study, Fox put student through a fifteen week course in which all the lectures were provided online. While there was support from a tutor, there was no fact-to-face contact with the tutor. Pre-course surveys indicated that the students were excited by the possibility for self-study. However, at the end of the course, half of the students who finished the course were asking for “greater regimentation and imposition of targets by the tutors.” Fox claims that technology, while having benefits, is not alone suitable for successful language learning. According to Fox, the role of the tutor is “to perform needs analyses, to set objectives, analyze language, develop and select and prepare materials, carry out assessment procedures, aid with learning strategies, carry out administration and management, and act as a librarian of potential resources.” Murray (1998) adds to the role of the teacher in his paper about integrating a software package to teach French to undergraduate students. According to Murray, tutors should gain as much knowledge as they can in order to choose, or create, software and integrate it into a language curriculum.

Schwienhorst (1998) further supports Vygotsky in his paper about integrating tandem language learning with a multi-user, object oriented (MOO) domain. Tandem language learning involves pairs of students, each with a different native language, communicating with each other to learn each other’s language. In the case of Schwienhorst’s work, students from England worked with students from Germany to learn German and English respectively. Schwienhorst integrated this type of strategy with the MOO domain called Diversity University. The goal was to combine the tandem learning strategy with the immersive effects of the virtual world. In a way, Schwienhorst is integrating Piaget and Vygotsky’s theories. He provides students with an immersive world to conduct their conversation, but also requiring the student to work together to learn from each other. Schwienhorst mentions immersion as a benefit of using virtual worlds for learning purposes. In the next section, the advantages and disadvantages of virtual worlds (and technology in general) for language learning are discussed.

2.3 Advantages and Disadvantages of Virtual Language Learning

Though the use of virtual environments for language learning has not been perfected, the reported research infers that the advantages of using virtual environments outweigh the disadvantages. This is the stance taken by Felix (1998). In his paper about a course teaching students Vietnamese, Felix states that the issues of using CD-ROM’s and the internet deal on the more technical side, and can be mitigated. For example, students might not have fast enough internet access or adequate training using the technology. However, these issues can be dealt with in the design of the experiment. Using a computer lab with computers fast enough to support the technology along with providing the necessary training in the use of the software goes a long way in eliminating these difficulties. While technical issues may exist, the benefits of virtual language learning cannot be ignored. In an excerpt from Fox's paper, he states that the benefits of virtual language learning are as follows:

- **Student Diversity**
    - Offering resources on CD-ROM or the World Wide Web (www) can allow for student differences – ability, interest, learning strategies, time spent on learning, attention span, and prior knowledge – to be dealt with more systematically and more easily than in a classroom
- **Pedagogy**
    - Firstly, they can provide large amounts of linked material on language, literature and culture in the form of tutorials, games, lectures, and contextualized exercises using video, audio, and text – all in one flexible resource that students can work with alone or in pairs or take home (if they have the appropriate hardware)
    - Secondly, the WWW in particular provides opportunities for truly interactive language teaching at the highest level. Students can be involved in co-operative exercises in which they are engaged in a task or quest in true-to-life situations in which they have some sort of influence over the outcome.
- **Delivery**
    - Direct and instant links to the tutor
    - Bringing groups of student together
    - Extending learning communities
    - Potential for co-operative work among students that is task or project oriented
    - Possibility of a wide variety of feedback and assessment formats

Von Der Emde et al. (2001) also champions the use of virtual language learning. In a course similar to Schwienhorst's course, Von Der Emde et al. develops a MOO to aid American and German student to learn each other's language using tandem learning strategies. From their observations, Von Der Emde et al. describe five benefits from using virtual language learning:

- Authentic communication and content occurred,
- Autonomous learning and peer teaching in a student-centered classroom occurred,
- Individualized learning occurred,
- The importance of experimentation and play was demonstrated, and
- Students developed as researchers.

While Felix and Von Der Emde et al. are mostly positive about learning language in virtual environments, Zhang (2013) takes a more cautious look at the subject. In his paper, Zhang conducts an empirical study about the factors that can impede the learning of English as a second language in a virtual environment. The factors from Zhang's paper are listed below:

- Distracting factors in opening simulations,
- Time zone lags,
- Absence of nonverbal cues,
- Lack of sufficient formal language input for linguistic competence,
- Irregular availability of language partners,
- Lack of familiar conversation of discussion topics,
- Lack of equal practice opportunities for students of varying proficiency levels,
- Uncontrollable group size, and
- Instructor's increasing workload.

It is important to note that many of these factors can be solved by designing a virtual world and password protesting it so only a select group of people may enter. However, this causes a couple of new issues. For example, the creation takes time and skills that teachers may not have. In addition, the students would need to be trained in the world for it to be effective. Finally, the computers need to be fast enough to support the virtual world or the immersive effected can be compromised. Zhang is not dismissing the use of Second Life or any other world training tool. Instead, he states the factors that could limit the effectiveness of the medium. He says, "The success of enabling effective spoken English learning requires timely elimination of technical problems and a systemic optimization of learning environments from both pedagogical and administrative perspectives."

2.4 Evaluation Methods

As Felix and Von Der Emde et al. demonstrate, there are significant benefits from using virtual environments for language learning. Unfortunately, there have not been many studies that have been able to quantitatively demonstrate the effectiveness of virtual language learning. Instead, most rely on observations of their experiments and questionnaires to gauge the effectiveness of their virtual environment learning experiments. This method is used by many of the papers previously mentioned along with a paper from Little and Ushioda (1998). In their paper, Little and Ushioda discuss their tandem learning project using Irish and German students. Like Schwienhorst, they conduct an experiment where students from an English speaking country, in this case Ireland, try to learn German by communicating bilingually with German students trying to learn English. To assess the success of this effort, Little and Ushioda use a questionnaire. While the results are considered "positive and encouraging," there is no quantitative method used to prove that the method used was really effective. Subjects might find a method appealing, but that does not mean it is effective at teaching the new language.

Another method of evaluating a project's success was to track the number of times the program was accessed or the number of "hits" received. This is the method used by Nesi (1998) to evaluate their project which involved an online center developed to aid non-English speakers. The issue they faced was not being able to track how long people used the site making the number of "hits" not all that important. Additionally, they could not determine on what the student who accessed the site focused their time. As a result, they could not determine which parts of the project were successful.

While some researchers have already concluded their projects, others were still in the preliminary stages when they published. Sims (2007) was a good example of this situation.

Sims reports on the status of a virtual language simulator created to aid American soldiers to verbally and non-verbally with the Iraqi population. The simulation is designed to react to the soldiers' actions by either causing the situation to stay calm or escalating into violence if the wrong things are said or done. Currently, there has not been an assessment of the potential success of this project, only a report of its development. In another case, a paper reporting on a similar project has reported results of their experiment. Dunard (2008) conducted a study using the Tactical Iraqi Language and Culture Training System (TILTS). Dunard states that the system is designed to teach United States Marines "a usable grasp of Iraqi culture, gestures, and situational language." However, the results of the experiment are still limited to questionnaires. An interesting observation was that the Marines using the system mentioned that the program could be more successful if complemented by a teacher. This again seems to give credibility to the Vygotskyian point of view as it relates to the importance of the teacher or tutor.

Extending past the use of questionnaires and moving toward the use of pre-experiment and post-experiment exams was Berns' effort (2013). Berns tested a video game used to teach basic German to students. The results of the experiment showed that the game increased the student's vocabulary and listening comprehension. The students were given four exams at the beginning and end of the experiment. The results showed that the students improved dramatically in three of four exams. The other exam proved to be too easy initially and therefore improvement was minimal. However, Berns provides an excellent model that has not been used much in research on virtual environments for language learning. In order to give credibility to virtual environments for language learning, students' opinions and empirical results need to be used.

2.5 Conclusions from the Literature Review

In this literature review, two opposing philosophies related to the field of education are presented. These philosophies, the Piagetian and Vygotskian, relate to virtual environment language learning in that they give the basis of how the experiments can be conducted. The literature in the field of virtual environments for language learning shows that researchers are looking to determine which method is better. While there is no clear cut answer as people have different learning styles, the literature indicates that virtual tools are acceptable, if they are incorporated with a teacher. Whether this is a result of people being accustomed to the traditional teacher-student roles or it is an axiom of learning not yet known. What is known is that virtual environments for language learning has proven to be successful from a student's perspective. However, the field of study lacks empirical, quantitative evidence to back up the claims that virtual environments for language learning is a viable and efficient option. In the next section, the methodology for conducting this experiment is provided in detail. The methodology provides the framework for determining how successful virtual environments for language learning can be on a quantitative basis.

3.0 Methodology

In the previous section, two differing perspectives from Piaget and Vygotsky are presented. The purpose of presenting these two different philosophies is to establish the foundation for what this experiment. Piaget believed that people learn by being immersed in an environment whereas Vygotsky believed that people learn by having the material taught to them by a tutor. In a modern sense, Piaget's view on individual learning is being represented by a computer program named Alelo. This computer program uses voice recognition software and simulated scenarios in order to provide feedback to a user attempting to learn a new language. The Vygotskian view is represented by a traditional classroom environment where a teacher teaches the class and students take notes.

Following completion of experiment, each method is analyzed in order to determine if it effectively improved the subjects' second language proficiency. Additionally, the two methods are compared to each other to determine whether one method performed better than the other. This section presents a detailed description of the methodology used to perform this experiment whereas the next section presents a data analysis where an answer to the question of whether or not the Alelo software improved the subjects' second-language proficiency when compared to the Teacher/Tutor is determined.

3.1 Alelo Software

The purpose of this experiment is to determine whether the software by Alelo 1) improves the second-language proficiency of its users and 2) is more effective at improving second-language proficiency than a traditional classroom. It should be noted that the US Marine Corps uses Alelo in order to train marines preparing to deploy to other countries. Their goal is for marines to have the ability to communicate effectively and respectfully with the citizens of the countries in which they are conducting missions. Therefore, it is important to know if the virtual environments training provided by the Alelo software effectively teaches a second language to people when compared to more classical methods. Since marines and soldiers may be expected to be deploy to Sub-Saharan Africa, the language taught in this study is Sub-Saharan French. In addition to language training, Alelo provides Sub-Saharan African cultural training as well. In order to determine if the software improves the subjects' second-language proficiency, the subjects each take a pre-test and post-test. These tests are developed by a college French teacher who is the second author. The next section explains the teacher's role in detail.

3.2 Role of the Teacher

In order to compare the effectiveness of virtual environment training provided by Alelo to that of a traditional classroom, a teacher is needed to properly teach the material to the class. Using the Alelo software, the teacher created 110 questions. From these questions, the teacher created two 20-question exams with the same format and the same level of difficulty. In addition to creating the exams, the teacher teaches the class for the group of subjects learning Sub-Saharan French in a traditional classroom environment. The class covers the same material that the Alelo virtual environments software covers. The class also has the same time restrictions as the students using the Alelo software have – ten hours of instruction over three days. After the experiment is complete, the teacher grades each subject's pre-test and post-test and reports the grades for analysis.

3.3 Experimental Methodology

In order to conduct this experiment, two groups were created: the Alelo Group and the Classroom Group. The groups consisted of similar sample sizes. However, due to a licensing issue, only four computers were used for the Alelo software at a time. Therefore, the subjects in the Alelo Group were trained as approximately 4 subjects at a time while the subjects in the Classroom group were trained with more stuents, usually about 8 students. In addition, the Alelo

Group was trained first as one part of this blocked study. After the tests from that group were collected, the Classroom Group was trained. The experiment was completed following the conclusion of the Classroom Group. After the experiment was completed, the exams were graded and the results were analyzed.

The following subsections provide a detailed explanation of the experimental process. They are provided in the order their respective groups were taught with an explanation of the Alelo Group's experimental process followed by that of the Classroom Group.

3.3.1 Alelo Group The Alelo Group's experiment lasts for three days for each subject. The three days are not necessarily consecutive. However, the three days are completed within a full week. The first day of the Alelo Group's experiment lasts four hours. The first hour is used to sign a consent form and to explain the nature of the experiment. The subjects are told they are being taught Sub-Saharan French language and culture using the computer software, Alelo. In addition, they are given a subject number and a set of headphones with a built in microphone.

As mentioned, each subject takes a pre-test. Before taking the test, the subject views a tutorial on how to answer the different types of questions on the test. The test has three types of questions: multiple choice, matching, and voice response. A multiple choice question provides a question with different answer choices, but only one correct choice. Matching problems require the subjects to match a list of words or phrases to their respective translations. Finally, the voice response questions provide a phrase or a list of phrases to speak into a microphone. The responses are recorded using Microsoft Voice Recorder and saved using the format *Subject M Question N*, where *M* is an integer representing the subject's designated subject number and *N* is the question number. The test they receive is based on the subject number they are given. For example, the odd-numbered subjects take Test 1 while the even-numbered subjects take Test 2. The subjects are then instructed to use the remaining three hours to work with the Alelo software.

The second day consists of four more hours of using the Alelo software. There is no testing on this day. On the final day, each subject uses the Alelo software for three more hours for a total of ten hours for the entire experiment. After completing the ten hours, each subject takes their post test. The post test is simply the other version of the test that the subject did not take at the beginning. For example, odd-numbered subjects take Test 1 at the beginning and Test 2 as their final test and even-numbered subjects take Test 2 as the beginning and Test 1 upon completion of the training. Upon completing their post-test, the subjects are dismissed.

3.3.2 The Classroom Group The experimental process for the Classroom Group is similar to that of the Alelo Group. For example, on the first and last day, the subjects take their test with the version of the test depending on whether they are odd-numbered or even-numbered. Again they are provided with headphones with built-in microphones. The other ten hours of the experiment are used for the class. During the class, each subject is provided a spiral notebook to take notes in class. On the first day, the subjects are in class for three hours following their pre-test. On the second day, the subjects have class for four hours. On the final day, the subjects have class for another three hours to complete the class. They then conclude with a post-test where they do not have access to their notes. Again, the odd numbered subjects take Test 1 as the pre-test and Test

2 as the post test, while the even numbered subjects take Test 2 as the pre-test and Test 1 as the post test. The subjects are also not allowed to take their notebooks home for review until the completion of the entire experiment.

3.3.3 Test Grading In order to eliminate bias in the grading process, the tests were coded so the teacher would not know whether the subject was from the Alelo Group or the Classroom Group. The teacher grades all the exams, both the pre-tests and post-tests without knowing to which group any exam belonged. The results are then delivered for analysis. The code is reversed and the appropriate statistical analysis is conducted. The next section presents the results of this experiment and provides an answer as to whether the virtual environments training offered by Alelo software is an effective tool at improving second-language proficiency when compared to traditional classroom training.

4.0 Data Analysis

Before the subjects received their training, they were administered a pre-test. Upon concluding their training, the subjects were administered a post-test. Once all the subjects had completed their training, the exams were coded and given to the Classroom training teacher for grading. Both the pre-test and the post-test consist of 20 questions (4 voice recording questions and 16 multiple choice or matching questions). Each question is worth 5 points and the subject must answer all components of the question correctly in order to receive credit. Since no partial credit is awarded, the subjects' scores are all multiples of 5 from 0-100. In order to determine if this training improved the subjects' French language proficiency, the scores from the tests were analyzed. The analysis of these scores consists of three parts: testing if either method effectively improved the subjects' French language proficiency, testing if one class type performed better than the other, and testing if the improvement in scores is correlated to any experimental factors (i.e. gender, subject number, and days to complete testing). Before analyzing the test scores, the experiment's demographic data is provided. The experimental factors are related to the demographic data.

4.1 Demographics and Procedure Related Features

Demographic and procedure related factors could have had an influence on exam results for this study. As a result, demographic and procedure related factors are examined to determine if some form of bias existed because of these issues.

4.1.1 Recruitment and Participation Criteria The subjects participating in this experiment were recruited primarily through the following means: advertising on campus Facebook pages, advertising on a personal Facebook page, advertising in class rooms, personal requests, and asking subjects to recruit their friends. In order to participate, potential subjects had to fulfill the following criteria: current University of Texas-Pan American student, 18 years of age or older, valid social security number (so a stipend of $10 per hour could be paid), and not having a current knowledge of French exhibited by a High School or College French course or other proficiency in French.

4.1.2 Gender Originally, 32 subjects participated in and completed the Alelo training and 33 subjects participated in and completed the classroom training. One subject was removed from the Alelo training group because their data existed as an extremely large outlier, due to suspected misconduct on the pre-test. At the conclusion of the experiment, 31 subjects (17 male, 14 female) completed the Alelo training and 33 subjects (15 male, 18 female) participated in the classroom training.

4.1.3 Subject Number As mentioned in section 3, each subject was assigned a subject number. The type of subject number (odd or even) dictated which test was administered as a pretest to the subject. Odd-numbered subjects received Test 1 as their pre-test and Test 2 as their post-test. Even-numbered subjects received Test 2 as their pre-test and Test 1 as their post-test. The mixing of the two tests was done in order to add a level of randomness to the experiment and to reduce the chance of subjects copying the answers from a subject sitting next to them. There was no statistical testing conducted on the equality of the exams. However, they were designed to be similar in format, in material tested, and in the order in which the material is presented. Differences between the two tests are for example: "Say *Bonjour*." for Test 1 and "Say *Bonsoir*." for Test 2. Therefore, the tests were assumed to be equal in content, structure, and difficulty. Since numbers were assigned prior to subjects being removed from the experiment for non-attendance, there is not an equal amount of odd- and even-numbered subjects. The Alelo training group consisted of 16 odd-numbered subjects and 15 even-numbered subjects. The Classroom training group consisted of 17 odd-numbered subjects and 16 even-numbered subjects.

4.1.4 Days to Completion The final factor to be considered is the days to complete the training for each subject. While the training is only for three days, the three days were at times stretched out due to scheduling conflicts. For this experiment, the days for a subject to complete their training ranged from 3 to 8 days. In the following sections, the relationship between these factors and the improvement from pre-test to post-test is determined.

4.2 Results

Using the data and statistical testing, one can reach mathematically valid conclusions. This is a process of extracting information based on the scientifically gathered data.

4.2.1 Test for Difference between Pre-test and Post-test for Alelo Training In order to determine whether the Alelo training significantly improved the subjects' French language proficiency, a paired-t test is used. The null hypothesis is that there is no difference between the pre-test and the post-test while the alternative hypothesis is that there is a difference. The hypothesis test uses a significance level of 0.05. Therefore, the null hypothesis is rejected if the p-value is less than 0.05. The p-value for this test is 0.00. Therefore, the null hypothesis is rejected and the conclusion is the Alelo training significantly improved the subjects' French proficiency.

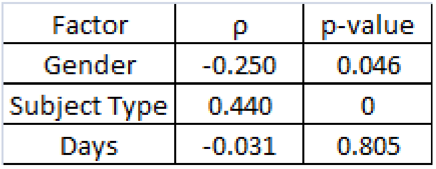
4.2.2 Test for Difference between Pre-test and Post-test for Classroom Training A paired-t test is used to determine whether the Classroom training significantly improved the subjects' French language proficiency. The null hypothesis is that there is no difference between the pre-test and the post-test while the alternative hypothesis is that there is a difference. The hypothesis test uses a significance level of 0.05. The p-value for this test is 0.00. Therefore, the null hypothesis is rejected and the conclusion is the Classroom training significantly improved the subjects' French proficiency.

4.2.3 Comparison of Means to Test for Difference between Alelo and Classroom Training The previous tests gave credibility to the hypothesis that both training types significantly improved the subjects' proficiency in French. However, whether one training regimen performed better than the other training regimen needs to be determined. Before testing if the average difference in test scores for the Alelo training is greater than that of the Classroom training, a test to determine if the variances between the two samples are equal needs to be performed. The null hypothesis is that the variances are equal, while the alternative hypothesis is that the variances are not equal. The significance level is 0.05. Since the p-value is 0.94, the conclusion is to fail to reject the null hypothesis that the variances are equal.

Now that the equal variance assumption (that the variances cannot be proved unequal) is validated, a hypothesis test to determine whether the test difference between the Alelo training is greater than the test difference for the Classroom training is performed. The null hypothesis is that there is no difference between the two training regimens while the alternative hypothesis is that there is a difference. Using a significance level of 0.05, a Student-t test assuming equal variances is used. The p-value for this test is 0.526. Therefore the conclusion is to fail to reject the null hypothesis that there is no difference between the two training types.

4.2.4 Test for Correlation between Difference in Tests and Experimental Factors The purpose of this section is to test whether there is a correlation between the test score difference and the experimental factors, discussed earlier. Table 1 shows the Pearson's correlation coefficient (ρ) and the corresponding p-value between the factors and the test score difference.

Table 1 – Correlation Coefficient between Factors and Difference in Test Scores



Factor	ρ	p-value
Gender	-0.250	0.046
Subject Type	0.440	0
Days	-0.031	0.805

The assumption is that these factors do not have an effect on the results. However, this is not the case. Using a significance level of 0.05, both Gender and Subject Type show a significant correlation with test score difference. Since odd-numbered subjects were coded with the number 1 and even-numbered subjects were coded with the number 0, the positive correlation indicates that the odd-numbered subjects improved more than the even-numbered subjects. By performing a correlation test on the factors and the results on the pre-test, it is discovered that there is a significant (p-value = 0.000) correlation coefficient of -0.599 indicating that odd-numbered subjects performed worse on their pre-test than even-numbered subjects. However there is no difference between the two subject types on the post-test (ρ = 0.000, p-value = 0.999). Since the pre-test (and post-test) version depends on the subject's number type (odd or even), it is difficult to assess whether one test was more difficult than the other. While on one hand, the subjects performed worse on Test 1 than on Test 2 when the tests were administered as pre-tests. On the other hand, the two tests performed equally when administered as post-tests.

A similar, although not as extreme, phenomenon occurs with the gender as the correlation coefficient indicates that males improved slightly more than females. However there is not a significant correlation between gender and the pre-test scores (ρ = 0.203, p-value = 0.108) or between gender and the post-test scores (ρ = -0.126, p-value = 0.320). While not significant, the correlation coefficients indicate that female subjects performed better on the pre-test than male subjects but worse on the post-test. There is no significant correlation between the number of days to complete the training and the difference in test scores. With the experimental factors explained, two extra factors exclusive to each training type need to be analyzed.

4.2.5 Test for Correlation between Difference in Tests and Alelo Quizzes A feature exclusive to the Alelo training is the quizzes at the end of the chapters that were meant to test how the subjects were progressing. Since the quizzes could be taken as many times as the subject chose, a parallel version of that type of free access quiz is difficult to conduct in a traditional classroom. For the Classroom training, the material on the quizzes was provided throughout the training but the subjects were never given grades through the duration of the training (except for the pre-test and post-test). Since the Alelo software stored the quiz scores, the scores were correlated with the change in test scores. In order to do this, the average quiz score for each subject was used as a data point and the entire set was correlated with the test score difference. The result shows that there was no significant correlation between the quiz score average and the test score difference (ρ = 0.206, p-value = 0.267).

4.2.6 Tukey's Test to Test for Difference in Means between Classroom Groups Just as the quizzes were difficult to replicate for the Classroom training, so was the scheduling method used for the Alelo training an added factor. The subjects who were trained using the Alelo training attended the training in scattered instances, rather than at specific classroom times. Although there are those that argue that the asynchronous delivery times allowed by the software based training was an advantage, the authors have chosen to consider this issue as simply another factor to be examined. There was no group dynamic to potentially affect the subjects since their schedules did not always align with the other subjects. For example, a subject being trained in Day 1 could have attended the second day of training with a completely different group of subjects. Since the Alelo training was meant to be a solo training, this was not an issue with scheduling. However, since the Classroom training requires a teacher, it was not feasible to use this scheduling method. As a result, subjects were told to show up for specific days ahead of time with little to no flexibility in which days they could attend. Therefore, four distinct groups were created. In order to determine whether there was a difference between the test score difference for the different groups, Tukey's Test is used. The results of the test are shown in Table 2.

Table 2 – Results with Turkey’s Test

```
Factor        Type   Levels  Values
Grouping_1_1  fixed       4  1, 2, 3, 4


Analysis of Variance for Change_1_1, using Adjusted SS for Tests

Source        DF   Seq SS   Adj SS   Adj MS     F      P
Grouping_1_1   3  0.15147  0.15147  0.05049  1.25  0.309
Error         29  1.16868  1.16868  0.04030
Total         32  1.32015


S = 0.200747   R-Sq = 11.47%   R-Sq(adj) = 2.32%


Grouping Information Using Tukey Method and 95.0% Confidence

Grouping_1_1  N    Mean  Grouping
3             8  0.4125  A
2             8  0.3438  A
1             8  0.3438  A
4             9  0.2278  A

Means that do not share a letter are significantly different.
```

As can be seen, the p-value is 0.309 indicating there is no significant difference between the four groups. In addition, Minitab, which was the software package used for analysis, categorizes the groups into distinct categories if they exist. However, the results show that no distinction needs to be made as all of the groups are listed under Category A.

4.3 Summary of Results

The result of this experiment is that there is no significant difference between the effect the Alelo training had on the subjects’ performance and the effect the Classroom training had on the subjects’ performance. However, both training types effectively improved the subjects’ proficiency in French. The scores on the Alelo quizzes are not correlated with improvement. This does not indicate that the quizzes do not work though. It only indicates that scoring poorly or scoring well on the quiz does not mean that a person is not improving or is improving. In addition factors such as group scheduling and days to complete training did not have an effect on the subjects’ improvement. However, gender and subject number type did have an effect. In the next section, conclusions are provided by explaining the consequences of this research.

5.0 Conclusions

In the literature review, two teaching philosophies were mentioned: the Piagetian view and the Vygotskian view. Piaget believed in self-teaching with the role of the tutor confined to providing an environment conducive to learning. Vygotsky believed the tutor’s role is to be more hands on and to teach the learner. These philosophies are represented by the Alelo software based training and Classroom training, respectively. While the philosophical argument can continue forever, the scientific results indicate that both approaches are equally valid. The question of whether one method is the correct method is most likely a matter of preference to the learner. The virtual environments learning offered by Alelo language learning software offers schedule and circumstance flexibility, not available in classical classroom learning. However, for the population in general, the same success rate can be found using either training type of training. From a practical standpoint, this means that if the right amount of effort is used in developing an immersive self-teaching tool, the tool is just as effective as placing the learner in a traditional classroom. The key factor is the amount of effort used in development. A poorly constructed tool can easily fail at its goal just as a poor teacher can fail to reach their students. Therefore, effort from the teachers' side is important to facilitate learning regardless of their amount of interaction with the learner.

The experiment described in this paper was limited by time and resources to 64 subjects and 10 hours of instruction. This was a constraint placed on this study by the circumstances with which it had to be constructed. With more students and with more time, better answers might be obtained. However, for the limited populations tested and the limited time of instruction (both limited by budget for subject pay), the authors offer conclusions that help to provide solid scientific evidence of the effectiveness of both training regimens, given the resource and time constraints.

Despite the issues with gender and subject number type correlating with test score difference, this experiment successfully demonstrated that Alelo software, a virtual training tool, designed for self-teaching, effectively improves a subjects' French proficiency to the same degree as a traditional classroom. Given the remoteness of a classroom to the deployed solider or marine, the Alelo software represents an outstanding tool for military operations. Therefore if the military wants to teach soldiers a new language, Alelo software is proven to improve language proficiency in a manner equivalent to classical classroom training.

REFERENCES

Anderson, J., Reder, L. M., and Simon H. A. (1996) "Situated Learning and Education," *Educational Researcher*, 25(4), 5-11.

Barrett, K.A. & Johnson, W.L. (2010). Developing serious games for learning language-in-culture. In Richard Van Eck (Ed), Interdisciplinary Models and Tools for Serious Games: Emerging Concepts and Future Directions. Hershey, PA: IGI Global. Provides an overview of the Alelo instructional design approach.

Benson, P. and Chink A. (2011), "Towards a More Naturalistic CALL: Video Gaming and Language Learning," *International Journal of Computer-Assisted Language Learning and Teaching*, 1(3) 1-13.

Berns, A., Gonzalez-Pardo, A. and Camacho. D. (2013) ""Game-like Language Learning in 3-D Virtual Envonments," *Computers and Education*, 60(1), 210-220.

Cahoon, M., Surface, E., Towler, A., & Dierdorff, E. (2012). A Comparison of Training Outcomes from a Simulated Learning Environment and a Traditional Classroom. *International Journal of Global Management Studies Professional (IJGMSP)* 3(2).

Chen, P-H, (2012) “E-Learner Characteristics and E-Learner Satisfaction: A Study of Taiwanese EFL University Students,” *International Journal of Computer-Assisted Language Learning and Teaching*,” 2(2) 1-15.

Dunard, M. E. (2008) “Tactical Iraqi Language and Culture Training (TILTS).

Egbert, J. and Huff, L. (2011) “’You’re a Winner’: An Exploratory Study of the Influence of Exposure on Teachers’ Awareness of Media,” *International Journal of Computer-Assisted Language Learning and Teaching*,” 1(4) 33-48.

Emonts, M., Row, R., Johnson, W.L., Thomson, E., Joyce, H., Gorman, G., Carpenter, R. (2012). Integration of Social Simulations into a Task-based Blended Training Curriculum. *Proceedings of the 2012 Land Warfare Conference*, Melbourne, Australia.

Felix, U. (1998) “Virtual Language Learning: Potential and Practice,” *ReCall*, 10(1), 53-58.

Fox, M. (1998) “Breaking Down the Distance Barriers: Perceptions and Practice in Technology Mediated Distance Language Acquisition,” *ReCall*, 10(1), 59-67.

Goodfellow, R., and Lamy, M. N. (1988) “Learning to Learn a Language – At Home and on the Web,” *ReCall*, 10(1) 68-78.

Hamel, M. (2013) “Questionnaires to Inform a Usability Test Conducted on a CALL Dictionary Prototype,” *International Journal of Computer-Assisted Language Learning and Teaching*,” 3(3) 56-76.

Hobbs, J. R., Sagae, A., Wertheim, S. (2012). Toward a Commonsense Theory of Microsociology: Interpersonal Relationships. *Proceedings of FOIS 2012*. July 24-27, Graz, Austria.

Hwa, S. Weei, P., and Len, L. (2012) “The Effects of Blended Learning Approach through and Interactive Multimedia E-Book on Student’s Achievement in Learning Chinese as a Second Language at the Tertiary Level,” *International Journal of Computer-Assisted Language Learning and Teaching*,” 2(1) 35-50.

Ibanez, M. B., Garcia, J. J., Galan, S., Moroto, D. Morillo, D. and Kloos, C. D. (2011) “Design and Implementation of a 3D Multi-user Virtual World for Language Learning,” *Educational Technology and Society*, 14(4), 2-10.

Johnson, W.L. (2014). Using Virtual Role-Play to Prepare for Cross-Cultural Communication. *Proceedings of the 5th International Conference on Applied Human Factors and Ergonomics* AHFE 2014, Kraków, Poland 19-23 July 2014, Edited by T. Ahram, W. Karwowski and T. Marek

Johnson, W.L. & Zaker, S.B. (2012). The Power of Social Simulation for Chinese Language Teaching. *Proceedings of the 7th Conference on Technology and Chinese Language*

Teaching. Honolulu. Describes Alelo's social simulation approach, as applied to Chinese language teaching.

Johnson, W.L., Friedland, L., Watson, A., & Surface, E. (2012). The art and science of developing intercultural competence. In Paula J. Durlach & Alan M. Lesgold (Eds.): *Adaptive Technologies for Training and Education*. New York: Cambridge University Press.

Johnson, W.L. (2012). Error detection for teaching communicative competence. *Proceedings of the International Symposium on Automatic Detection of Errors in Pronunciation Training*. Stockholm, Sweden.

Johnson, W.L. & Sagae, A. (2012). Personalized refresher training based on a model of skill acquisition and decay. *Proceedings of the 2nd International Conference on Applied Digital Human Modeling*. San Francisco, 2012.

Johnson, W.L., Sagae, A., Friedland, L. (2012). VRP 2.0: Cross-Cultural Training with a Hybrid Modeling Architecture. *Proceedings of the 2nd International Conference on Cross-Cultural Decision Making*. San Francisco.

Johnson, W.L., Friedland, L., Schrider, P., Valente, A., & Sheridan, S. (2011). The Virtual Cultural Awareness Trainer (VCAT): Joint Knowledge Online's (JKO's) Solution to the Individual Operational Culture and Language Training Gap. Proceedings of ITEC 2011. London: Clarion Events. Describes the instructional design and implementation of Alelo's VCAT courses.

Johnson, W.L. & Wang, N. (2011). Politeness in interactive educational software. In C. Hayes & C. Miller (Eds.), Human-Computer Etiquette. London: Taylor & Francis. Examines the effect of polite tutorial tactics in educational software, in multiple learning domains.

Johnson, W.L. (2010). Serious use of a serious game for language learning. Int. Journal of Artificial Intelligence in Education, 20(2). An in-depth description of the development of Alelo's Tactical Language courses, and their evaluation in the field.

Johnson, W.L. (2010). Using immersive simulations to develop intercultural competence. In T. Ishida (Ed.), Culture and Computing, LNCS 6259, 1-15. Berlin: Springer-Verlag. Describes Alelo's approach to teaching intercultural competence, as realized in its products.

Johnson, L.W., Ashish, N., Bodnar, S., Sagae, A. (2010). Expecting the unexpected: Warehousing and analyzing data from ITS field use. In V. Aleven, J. Kay, and J. Mostow (Eds.), International Conference on Intelligent Tutoring Systems, Part 2, (pp 352 – 354). Berlin: Springer-Verlag. Describes a study in user interaction data was collected for a Tactical Language course and used to evaluate training system effectiveness.

Johnson, W.L., & Valente, A. (2009). Tactical language and culture training systems: Using AI to teach foreign languages and cultures. AI Magazine, 30(2), 72-83. An overview of the artificial intelligence techniques underlying Alelo courses.

Johnson, W.L. & Wu, S.M. (2008). Assessing aptitude for learning with a serious game for foreign language and culture. In B. Woolf, E. Aimeur, R. Nkambou, & S. Lajoie (Eds.), International Conference on Intelligent Tutoring Systems (pp. 520-529). Berlin: Springer-Verlag. Reports a findings from a field study in which training performance with Tactical Language courses were used to predict future learning success.

Johnson, W.L. Valente, A., Heuts, R. (2008), Multi-platform delivery of game-based learning content. Proceedings of SALT. Describes multi-platform content authoring approach, that enables us to author once and deliver on multiple platforms and devices.

Johnson, W. L., & Valente, A. (2008). Collaborative authoring of serious games for language and culture. SimTecT 2008. Describes the authoring tools used to develop Alelo courses.

Johnson, W.L., Mayer, R.E., André, E., & Rehm, M. (2005). Cross-cultural evaluation of politeness in tactics for pedagogical agents. Proceedings of the conference on Artificial Intelligence in Education. Amsterdam: IOS Press. A research study that compared the effect of computer-generated tutorial feedback in two different cultures (US and Germany).

Johnson, W. L., & Beal, C. (2005). Iterative Evaluation of a Large-Scale, Intelligent Game for Language Learning. Proceedings of the Twelfth International Conference on Artificial Intelligence in Education (pp. 290-297). Amsterdam: IOS Press. Describes how an iterative series of formative evaluations guided the development of the first Tactical Iraqi courses.

Johnson, W. L., Marsella, S., Mote, N., Viljhálmsson, H., Narayanan, S., & Choi, S. (2004). Tactical Language Training System: Supporting the rapid acquisition of foreign language and cultural skills. Proceedings of InSTILICALL—NLP and Speech Technologies in Advanced Language Learning Systems. Describes advanced methods for language error detection, that were prototyped in the Tactical Language courses.

Johnson, W., LaBore, C., & Chiu, Y. C. (2004). A pedagogical agent for psychosocial intervention on a handheld computer. AAAI Fall Symposium on Dialogue Systems for Health Communication (p. 22–24). An experimental mobile learning solution employing pedagogical agents.

Johnson, W.L., Rickel, J.W., & Lester, J.C. (2000). Animated pedagogical agents: Face-to-face interaction in interactive learning environments. International Journal of Artificial Intelligence in Education, 11, 47-78. A seminal overview of pedagogical agent technology.

Jones, A. and Mercer, N. (1993) "Theories of Learning and Information Technology," in Scrimshaw P. (ed.), *Language, Classrooms, and Computers*, London: Routledge, 11-26.

Kimura, M. (2012) "Digital Storytelling and Oral Fluency in an Enginsh Reading Class at a Japanese University," *International Journal of Computer-Assisted Language Learning and Teaching*, 2(1) 1-12.

Kumar, R., Sagae, A., & Johnson, W.L (2009, July). Evaluating an authoring tool for Mini-Dialogs. Poster session presented at The 14th International Conference on Artificial Intelligence, Brighton, U.K. Describes a template-based approach to authoring dialogs in Tactical Language courses.

Levy, M. (1998) "Two Conceptions of Learning and Their Implications for Call at the Tertiary Level," *ReCall*, 10(1), 86-94.

Little, D., and Ushioda, E. (1998) "Designing, Implementing and Evaluating a Project in Tandem Language Learning Via E-mail," *ReCall*, 10(1), 95-101.

Marsella, S., Johnson, W.L., & LaBore, C. (2003) Interactive pedagogical drama for health interventions. Proceedings of the Eleventh International Conference on Artificial Intelligence in Education (pp. 341-348). Amsterdam: IOS Press. Describes the application of social simulation to health communication directed at caregivers of pediatric cancer patients.

Mayer, R., Johnson, W., Shaw, E., & Sandhu, S. (2006). Constructing computer-based tutors that are socially sensitive: Politeness in educational software. International Journal of Human-Computer Studies, 64(1), 36-42. Reports on an evaluation study that demonstrated that politeness tactics in pedagogical agents can affect learning outcomes.

Murray, L. (1998) "Wintergrate? Reactions to Tele-Textes Author 2, a CALL Multimedia Package," *ReCall*, 10(1), 102-108.

Nesi, H. (1998) "Using the Internet to Teach English for Academic Purposes," *ReCall*, 10(1), 1209-117.

Sagae, A., Johnson, W. L., & Bodnar, S. (2010). Validation of a Dialog System for Language Learners. Paper presented at the The 11th annual SIGdial Meeting on Discourse and Dialogue, Tokyo, Japan. Presents the results of a study in which the performance of the Tactical French spoken dialog system compared against human raters.

Sagae, A., Johnson, W. L., & Valente, A. (2011). Conversational Agents in Language and Culture Training. In D. Perez-Marin & I. Pascual-Nieto (Eds.), Conversational Agents and Natural Language Interaction: Techniques and Effective Practices. Madrid: IGI Global. Describes the natural language dialog techniques employed in Alelo courses.

Sagae, A., Hobbs, J. Ho, E. (2012) Efficient Cross-Cultural Models for Communicative Agents. *Proceedings of the 2nd International Conference on Cross-Cultural Decision Making*. San Francisco, July 2012.

Seedhouse, P., Preston, A., Oliver, P., Jackson, D., Heslop, P., Plotz, T., Balaam, M., and Ali, S., (2013), "The French Digital Kitchen: Implementing Task-Based Language Teaching Beyond the Classroom," *International Journal of Computer-Assisted Language Learning and Teaching*, 3(1) 50-72.

Sims, E. M. (2007) "Reusable, Life Like Virtual Humans for Mentoring," *Computers and Education*, 49(1), 75-92.

Thomas, M. (2013) "TBLT in Business English Communication: An Approach for Evaluating Adobe Connect and Second Life in a Blended Language Learning Format," *International Journal of Computer-Assisted Language Learning and Teaching*," 3(1) 73-89.

Viswanathan, R. (2012) "Augmenting the Use of Mobile Devices in Language Classrooms," *International Journal of Computer-Assisted Language Learning and Teaching*, 2(2) 45-60.

Von Der Emde, S., Schneider, J. and Kotter M. (2001) "Technically Speaking: Transforming Language Learning Through Virtual Environments (MOOS)," *The Modern Language Journal*, 85(2), 210-225.

Wang, N., Johnson, W.L., Mayer, R.E., Rizzo, P., Shaw, E., & Collins, H. (2008). The politeness effect: Pedagogical agents and learning outcomes. International Journal of Human Computer Studies, 66(2), 98-112.

Wertheim, S., Agar, M. (to appear) (2012) Culture that Works. Proceedings of the 2nd International Conference on Cross-Cultural Decision Making. San Francisco, July 2012. Motivation, methodology, and results for sociocultural research on the CultureCom project. Interviews with SMEs in two languages and lit review produced a working definition of "culture" appropriate for integration with the computational system

Zaker, S.B. (2012). Designing instruction for affect: The need for feelings of belonging in distance education. Proceedings of Teachers College Educational Technology Conference. May 2012, New York.

The Immersive Environments Laboratory at the University of Texas-Pan American

Summary

The Immersive Environments Laboratory at the University of Texas-Pan American (UTPA) consists of three distinct installations that work together to further the goals of research in visualization and the study of immersive environments. First, is a workbench which is based on a Mechdyne M-1 workbench provided by the Naval Undersea Warfare Center, Newport, RI to UTPA. Second, is a "development laboratory" which is found in the Engineering Building at a location that permits rapid development of hardware and software for use in the various virtual environments, and third, is the CAVE facility for full scale demonstration in a four surface (three walls – back projected, one floor – front projected) fully stereoscopic system. Additionally, biometric equipment has been purchased and utilized to gather data for experiments.

The Workbench Laboratory

The workbench laboratory consists of the projector from the Mechdyne M-1 system, an 8.5 inch Electrohome analog projector reconfigured to provide a 72 inch stereoscopic image on a vertical screen. This reconfigured system uses a wooden frame and a mylar mirror to provide folded optics shown in Figure 1. The system employs a Hewlett Packard XM 9400 workstation with an NVIDIA FX 5500 graphics card. Tracking is accomplished with an Ascension Flock of Birds system, and glasses are Crystal Eye glasses. Software is VR Juggler available from Iowa State University. A photograph of the workbench in action is provided as Figure 2.

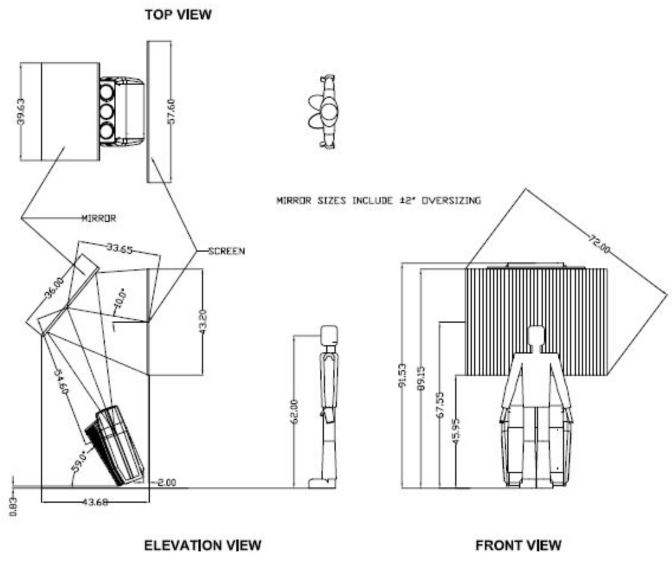


Figure 1 – Drawing of Reconfigured Workbench to Produce a 72 Inch Screen

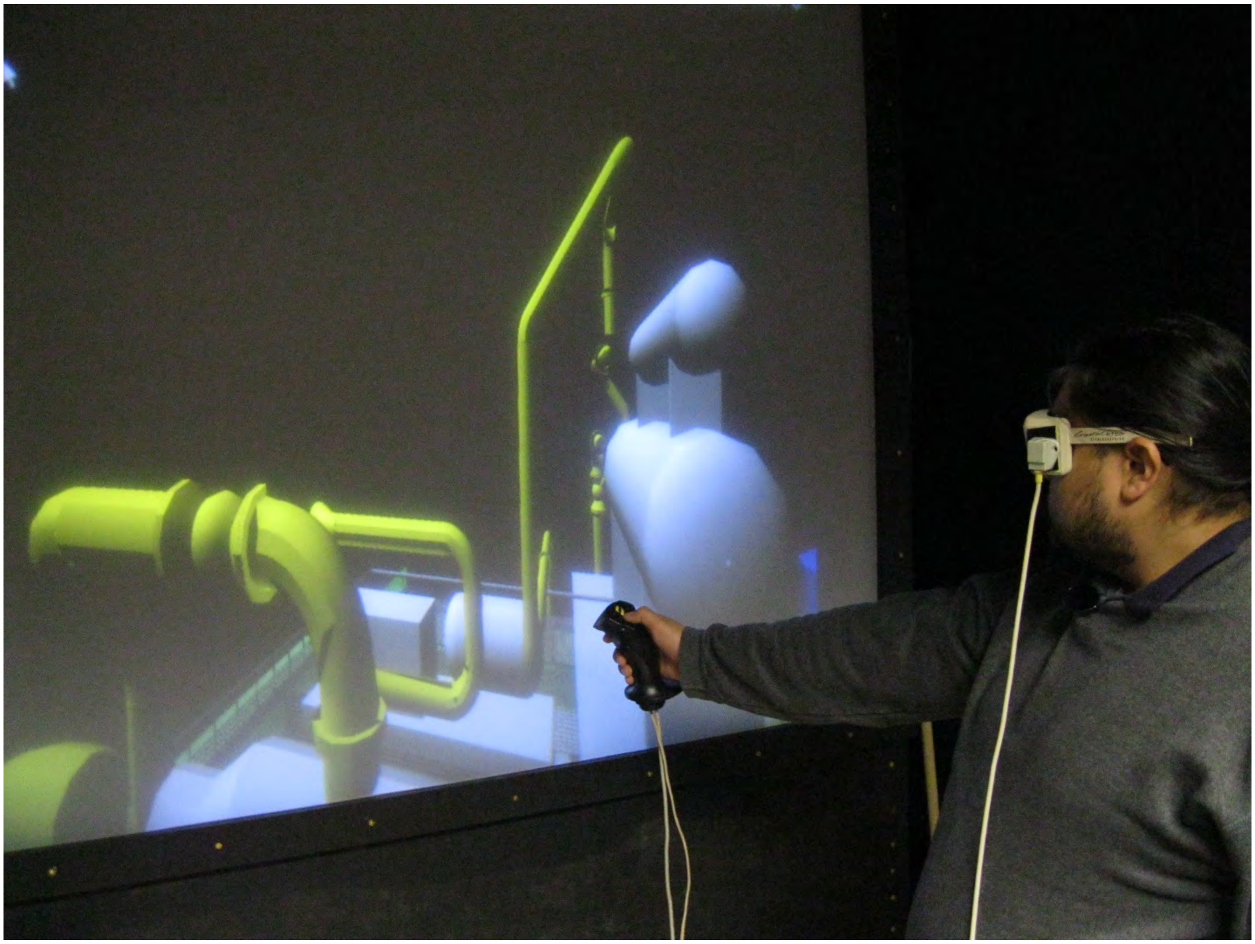


Figure 2 – Workbench with Virtual Reality Display

The Development Laboratory

The development laboratory has software and hardware identical to the CAVE system. This duplication permits development in the development laboratory followed by direct and efficient translation to the full scale CAVE system. This developmental laboratory is limited by space to three, reconfigurable 8 foot tall screens which are front projected rather than back projected. Further, due to the reduced space only four Bonita tracking cameras are used. Projector stands are movable, but all hardware and software is otherwise identical to that installed in the CAVE system.

The CAVE System

The CAVE System employs three 12 foot wide by 10.5 foot back projected screens configured as shown in the Diagram shown in Figure 3 below. The projectors are DPI Projectors from M Viion, and the screens are from Stewart Filmscreen. The mirrors measuring 8 feet by 6 feet are simply ordinary glass mirrors on frames designed by UTPA students and faculty. The frame for the CAVE system was designed and fabricated by UTPA students and faculty, and constructed by UTPA students and faculty. Computer systems include a five node network of HP Z800 with Nvidia 5,000 G graphics cards. The fourth surface is the floor, and it is projected directly using a

Depth Q projector. Software is the VR Juggler software from Iowa State University. Tracking is accomplished using 7 Bonita cameras configured by Vicon using their tracking software.

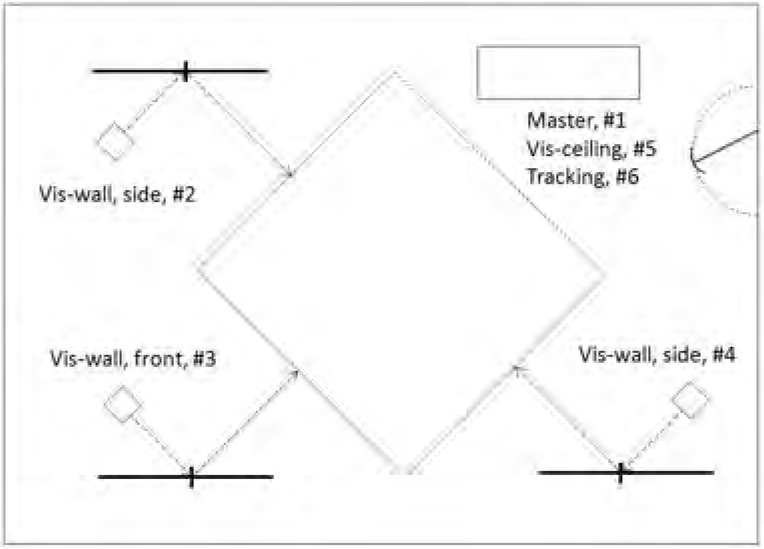


Figure 3 – Diagram of the CAVE Folded Optics System

The BioMetric Equipment

An entire system was purchased from BioPac Systems. This system included transmitters and receivers with software to allow bio-metric sensors to be placed on a human subject. This system uses telemetry to avoid interference with wires. Among the sensors that can be used with this system are: GSR (galvanic skin response), respiration, EEG (electroencephogram), and EKG (electrocardiogram). Included in this system is software to interpret the sensor information. Figures X and Y show the BioPac telemetry system and software in operation and a test subject with sensors on his body, respectively.

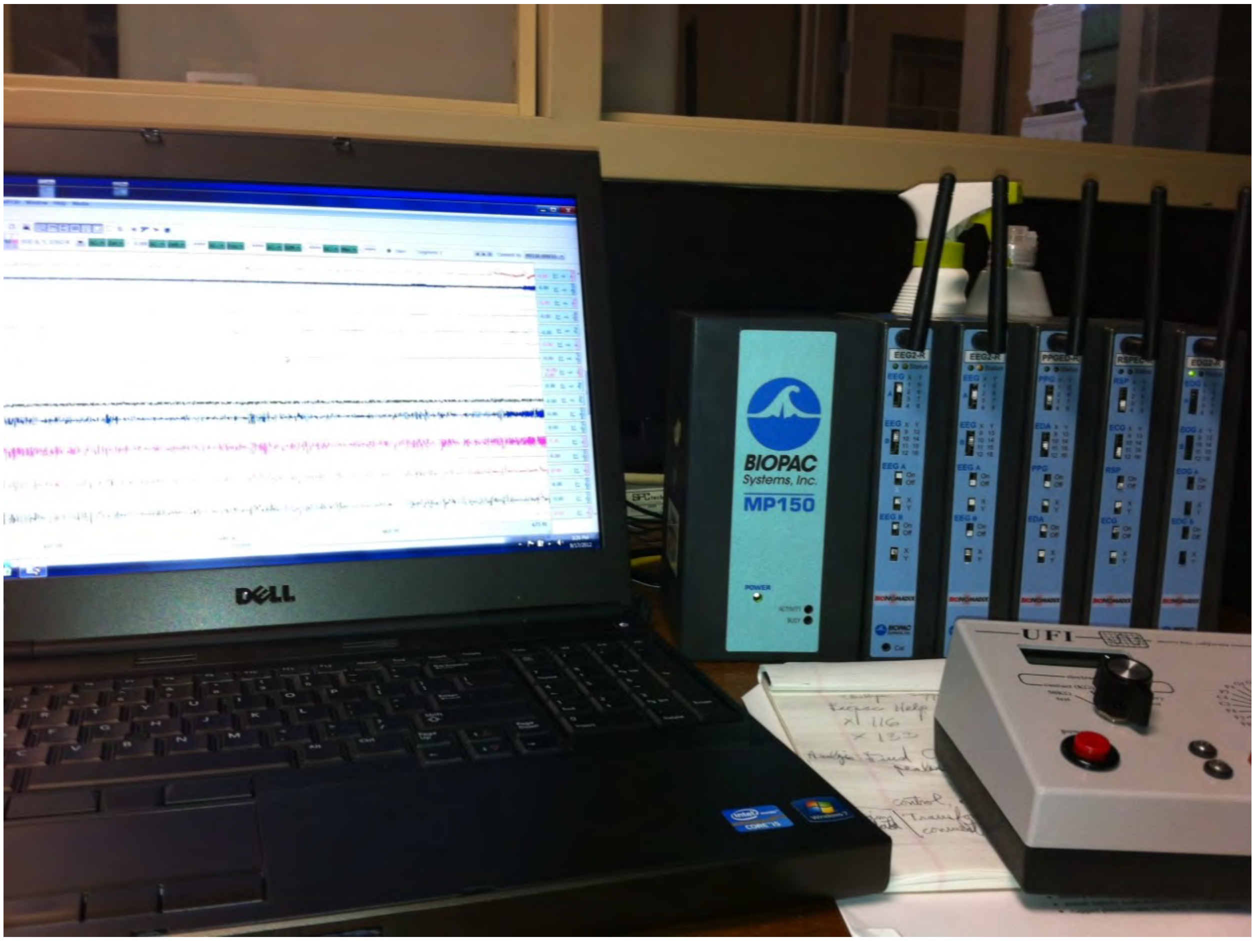


Figure 4 - The Bio Pac System with Data Receivers and Software

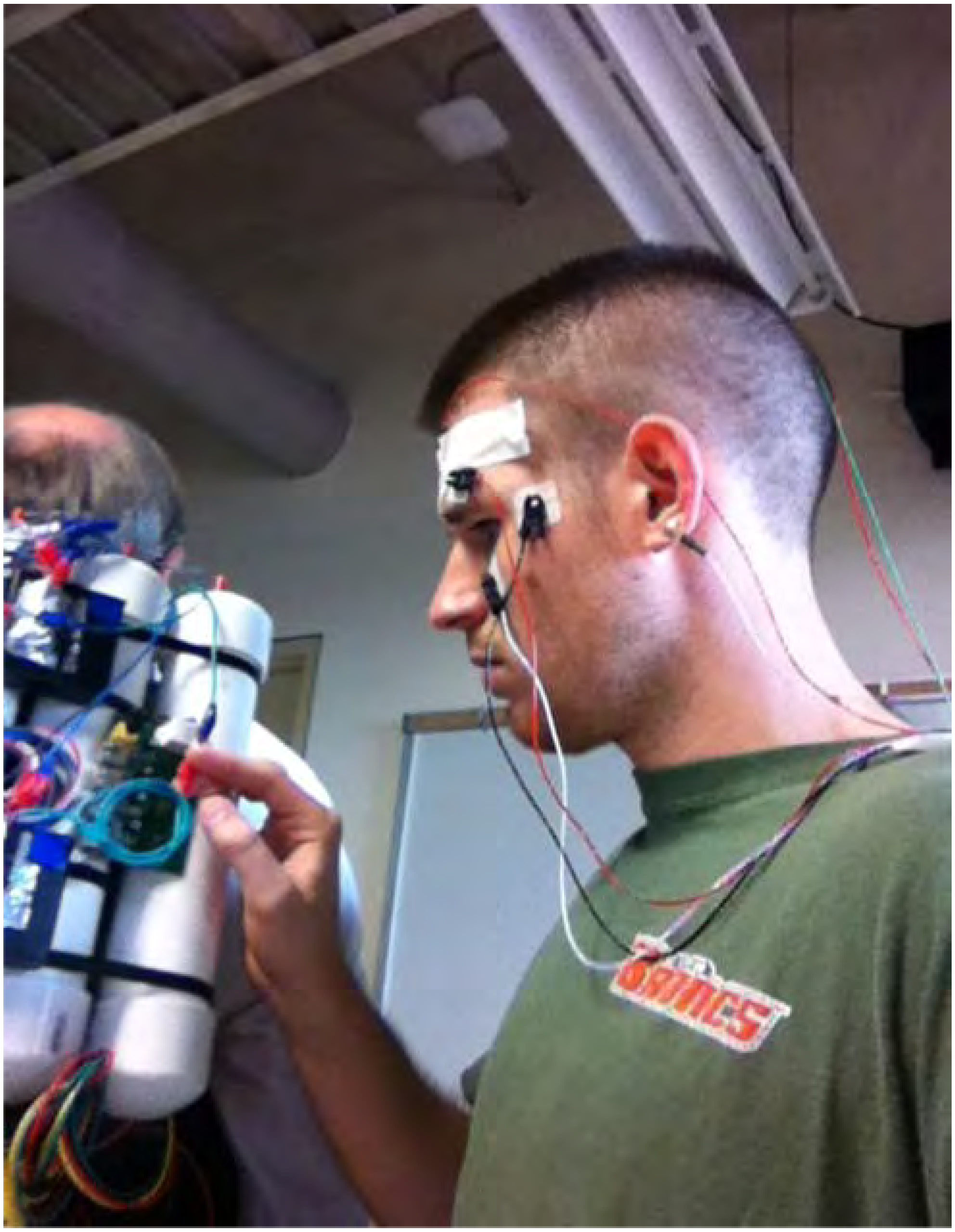


Figure 5 - A Test Subject Wired for the Acquisition of BioMetric Data

Conclusion

The literature is full of researchers reporting that a new system had been developed. This is particularly true of the literature regarding immersive environment or virtual environments. Unfortunately, many of these systems are taken on faith that they represent an improvement simply because they are new and/or represent new technology. The University of Texas-Pan American research team has taken a different approach. We have provided proof using quantitative, statistical methods that the new techniques are better or at least equal to the existing technology. Further, we have built a well-equipped laboratory that makes the completion of studies of effectiveness, possible in the future.