Usability Testing of a Virtual Reality Tutorial

Stephanie G. Fussell, Jessyca L. Derby, Joshua K. Smith, William J. Shelstad,
Jacob D. Benedict, Barbara S. Chaparro, Robert Thomas, and Andrew R. Dattel,
Embry-Riddle Aeronautical University, Daytona Beach, FL

Immersive simulation technology has transformed the training and learning environment. Virtual reality (VR) and augmented reality (AR) devices have been adopted by medical professionals, military forces, and marketing firms. Aviation training facilities are also integrating VR and AR technology into a variety of training. To ensure students begin training on equal footing, an engaging, guided tutorial for the virtual environment (VE) was created. A usability study was conducted to evaluate the tutorial's learnability, effectiveness, and satisfaction for two user groups varying in VR experience. Results show users found the tutorial enjoyable with high usability and playability. Novice users reported the tutorial as more mentally effortful than expert users and were less comfortable with self-maneuvering. Users successfully completed most tasks on the first attempt after completing the tutorial. Those who noted difficulty in completing tasks in a post-assessment reported user error and corrected themselves without instruction. The tutorial demonstrated learnability, effectiveness, and satisfaction ensuring that users will be able to enter the VE with more confidence after engaging with the tutorial.

INTRODUCTION

Aviation education and training has substantially evolved since flight training devices (FTDs) were first introduced for instrument instruction in the 1930s, the most well-known being the Link trainer (Allerton, 2009). This evolution has been gradual with the introduction of new technology and innovative design, ranging from personal computer aviation training devices (PCATDs) to high fidelity aviation training devices (ATDs) and FTDs to exact replicas of aircraft (full flight simulators [FFSs]). Technology has continued to progress beyond physical simulation devices to the virtual environment (VE). Augmented reality (AR) and virtual reality (VR) are used to navigate and explore VEs and develop skills outside of the live task environment.

Hale, Stanney, and Badcock (2015) note that the earliest forms of aviation simulators focused on near-identical operational values. Since then, a shift to learning construction has focused on psychological fidelity, physical fidelity, and high efficacy to generate an active learning environment (Hale et al., 2015). Training and educational institutions are embracing virtual and computing technologies as they have transformed training methods through the incorporation of interactive simulations. These technologies both enhance the learning process and diversify learning opportunities to promote knowledge retention and high skill transfer (Gray, 2017). In addition, the VE may be designed to stimulate cognitive and psychomotor behaviors similar to those stimulated in a live task environment. When effectively designed, "the virtual environment system capitalizes on a fundamental and distinctively human sensory, perceptual, information processing, or cognitive capability" (Stanney, Mourant, & Kennedy, 1998, p. 330).

However, VEs must have a foundation in theoretical learning constructs to enhance skill acquisition and cognition. The usability of VEs and the learning of the task within them must be considered. Both qualitative and quantitative measures may be used to assess the usability of a system. Nielsen (1996, p.12) identifies five attributes by which to measure the usability of a system:

- Learnability: How easily can users accomplish basic tasks and begin to work?
- Efficiency: How quickly can users perform tasks once they have learned the design?
- Memorability: How easily can users reestablish proficiency after returning to the design after a period of time?
- Errors: How many errors do users make and how easy is recovery from errors?
- Satisfaction: How pleasant is the design to use?

Usability testing allows educators, trainers, and developers to understand how a user interacts with the system and identify usability issues before implementation on a large scale (Sanders, 2010). Although Dumas and Redish noted in 1999 that many developers shifted to qualitative data to demonstrate usability, 21st century educators may require more substantial data to support the implementation of VR as a training device to support educational programs.

Assessments of immersive simulation in an aviation environment often utilizes quantitative data collected from the simulator to demonstrate the time and costs associated with using immersive technology for training. Transfer of efficiency and effectiveness, in terms of time to learn and complete a task using immersive simulation, and subsequently perform a task using in the live task environment, has been studied (Carretta & Dunlap, 1998; Macchiarella, Arban, & Doherty, 2006). Works by Fletcher have demonstrated that trainers can receive high return on investment by prioritizing resources to develop effective, high-quality, technology-based instruction (Cohn & Fletcher, 2010; Fletcher & Chatham, 2010; Foster & Fletcher, 2013). Vincenzi, Macchiarella, Opalenik, Gangadharan, and Majoros (2003) found that aviation maintenance students had significantly better immediate recall when AR was used as opposed to print and video instruction. Taylor et al. (1999) and Talleur et al.

(2003), among others, found that PCATDs generally increased performance levels for instrument related flight tasks and currency (positive transfer of training).

However, these studies focused on the transfer of training from an immersive simulation environment and did not address usability concerns. In creating a new and innovative learning tool, usability of the system must be analyzed and iteratively improved before it is incorporated into a training program. Training on complex tasks in a VE should not be adopted if the learning constructs do not equate to the live task environment.

PURPOSE

Training in an immersive virtual environment offers increased understanding as learners build knowledge at their own pace. The aviation industry and aviation educational institutions have historically embraced immersive simulation technology to efficiently and effectively train student in a low risk, high fidelity environment. VEs are quickly being adopted as more developers and hardware companies offer new products.

The College of Aviation at Embry-Riddle Aeronautical University (ERAU) has developed a state-of-the-art VR laboratory to train students on complex tasks related to flight, commercial space operations, aircraft maintenance, and other aviation topics. As the lab has evolved to support various programs, a tutorial was created to teach users how to navigate within and interact with a VE. The tutorial instructs the user on teleportation, manipulating objects, adding items to inventory, and other actions facilitated by using a controller. The purpose of this research is to test the usability of the tutorial using three of Nielsen's (1996) five attributes, specifically learnability, effectiveness, and user satisfaction across two user populations varying in VR experience.

METHOD

Participants

Participants included 23 college students (19 males and 4 females) with ages ranging between 18 and 26 (M = 21, SD = 3.28). The majority identified as having an aviation-related major (n = 13), 9 of which were pilots. Participants were asked to self-identify the level of gaming expertise as novice (n = 4), intermediate (n = 4), advanced (n = 8), or expert (n = 7). Based on experience with VR, eligible participants were categorized as novices (little to no experience, n = 15), or expert (a lot of experience or a great deal of experience, n = 8) and qualified to participate in the study. None of the participants reported visual impairments.

Experimental Design

A mixed-methods, between group (novice, expert) design was employed for this research. The attributes being considered for the study were learnability, effectiveness, and satisfaction. Learnability was measured as the difficulty ratings participants assigned to tasks. Effectiveness was measured by the amount of errors committed (e.g., omission of task, wrong control input and ability to correct, committing error but believing task was completed); these were recorded as researcher observations. Satisfaction was measured through a variety of quantitative assessments using Likert response items as well as verbal feedback for qualitative assessment.

Materials

The VR tutorial was designed in Unity by ERAU developers and was housed on Steam. The system uses an HTC Vive Pro head mounted display (HMD) and two hand controllers. The tutorial guides the user through a series of exercises to learn how to use hand controls to teleport, interact with objects, and add and remove objects from inventory. Participants also gaze at an object for a set amount of time. The tutorial prompts the user to complete a given task and repeats instructions until the task is successfully completed.

Figure 1. Screenshot of an interaction point in the VE.

In the tutorial, users were instructed on how to perform a variety of tasks within the VE. Teleport instruction began with a far-off, fixed target, which users had to learn to navigate to. Subsequent teleport tasks varied in range and to fixed targets or free navigation. A virtual room within the tutorial allowed users to push buttons, flip switches and levers, and select inventory. Users also gazed at fixed objects for varying amounts of time. Participants were asked to subjectively report their comfort and ease of navigation in the VE, as well as whether or not they thought the tasks in the tutorial were difficult to perform on a Likert scale from 1 (strongly disagree) to 7 (strongly agree).

A shortened version of the GUESS was modified for using the VR tutorial to measure usability, immersion, playability, enjoyment, personal gratification, and visual aesthetics (Phan, Keebler, & Chaparro, 2016). The GUESS, a "psychometrically validated and comprehensive gaming scale," has nine subscales to evaluate video games (Phan et al., 2016). The questions of four subscales (enjoyment, usability/playability, personal gratification, and visual aesthetics) were modified to reflect experience in a VE as opposed to a game. The remaining five subscales were deemed irrelevant for assessing the VR tutorial.

The System Usability Scale (SUS) was also modified to evaluate participant's perceived usability of specified input methods (Brooke, 1996). Participants rated statements associated with usability attributes (e.g., functionality, ease of use, confidence, complexity, integration) using a Likert scale

(1 = strongly disagree, 5 = strongly agree) based on the "tutorial" rather than a "system." For analysis, the final SUS score was calculated from 0-100. This score can be mapped to an adjective rating scale (from 0-25 = worst imaginable to 100 = best imaginable) (Bangor, Kortum, & Miller, 2009).

The NASA-TLX (Sharek, 2009) evaluates participants' perceived workload based on six subscales: physical demand, mental demand, temporal demand, performance, effort, and frustration. Pairwise comparisons were included in this study.

A post-tutorial exercise within the VE evaluated the participant's ability to complete the five tasks learned within the tutorial. Participants were instructed to navigate in the VE using teleporting and by walking around, gaze at targets, interact with toggle switches and a rotating device, and add and remove an item from their inventory. These tasks were given to participants in a random order. All participants had to complete each task successfully before moving to the next task. Tasks were rated from 1 (not at all difficult) to 5 (very difficult).

The Simulator Sickness Questionnaire (SSQ) (Kennedy et al., 1993) has participants rate 16 physical symptoms historically associated with prolonged activity in a simulator. Each symptom is ranked in order of effect on the user: no affect (none), slight affect, moderate affect, and severe affect. Each question is scaled and weighted for a nausea-related, oculomotor-related, and disorientation-related subscore, as well as a total score for severity of cyber sickness.

Procedure

Participants were placed into one of two groups (novice or expert) based on VR experience. Before participating in the study, participants signed an informed consent form, were provided instructions on how to wear the VR head mounted display, and reminded of the potential physical risks associated with being in a VE. Participants then completed the guided VR tutorial. Upon conclusion of the tutorial, participants completed the GUESS, SUS, SSQ, NASA-TLX, and self-report surveys. Participants then completed five tasks in a VE post-assessment. The order of tasks was randomized to control for order effects. Quantitative and qualitative data was collected by the researcher in the form of ranked difficulty and feedback for each task. Researchers took notes throughout the participant's time in the VE, noting if a task was attempted multiple times, recording qualitative feedback, etc. Upon completing the VE post-assessment, the participant completed a second SSQ and was debriefed.

RESULTS

VR Tutorial Task Performance

There were no significant differences in the number of attempts novices or experts took to complete any of the tasks in the tutorial $p > .05$. However, some participants did need to have the instructions repeated multiple times. When it came to learning how to teleport for the first time, instructions needed to be repeated more times for novices. A total of 7 out of 15 novices needed these instructions repeated ($n = 1$ had instructions repeated once, $n = 3$ repeated twice, and $n = 3$ repeated three times) while no experts had to have these instructions repeated. During all of the tasks in the tutorial, novices needed instructions repeated between 0-8 times ($n = 8$ had no repeated instructions, $n = 1$ repeated twice, $n = 3$ repeated three times, $n = 1$ repeated five times, $n = 2$ repeated 8 times). Only one expert needed any instructions repeated; this only occurred once during the inventory section.

There was also a significant difference in the total amount of times the instructions had to be repeated throughout all of the tasks between novices ($M = 2.13$, $SD = 2.88$) and experts ($M = 0.13$, $SD = 0.35$), $t(22) = 2.67$, $p = .018$, $d = 0.98$.

Self-Report and Satisfaction

No significant difference was found in how comfortable participants felt using VR technology in both the novice ($M = 6.4$, $SD = 0.91$) and expert ($M = 6.38$, $SD = 1.41$) groups, $p > .05$. No significant difference was also found in the reported difficulty in performing tasks in the tutorial between both the novice ($M = 1.33$, $SD = 0.62$) and expert ($M = 1.38$, $SD = 0.77$) groups, $p > .05$. However, a significant difference was found in how confident participants felt maneuvering in the VE between novices ($M = 5.93$, $SD = 1.22$) and experts ($M = 6.75$, $SD = 0.46$), $t(22) = -2.296$, $p = .033$, $d = 0.89$.

A significant difference was found for the enjoyment subscale of the GUESS between novices ($M = 6.62$, $SD = 0.53$) and experts ($M = 5.94$, $SD = 0.7$), $t(22) = 2.626$, $p = .016$, $d = 1.09$. No significant differences were found in the other subscales (usability/playability, personal gratification, or visual aesthetics) or total GUESS score $p > .05$.

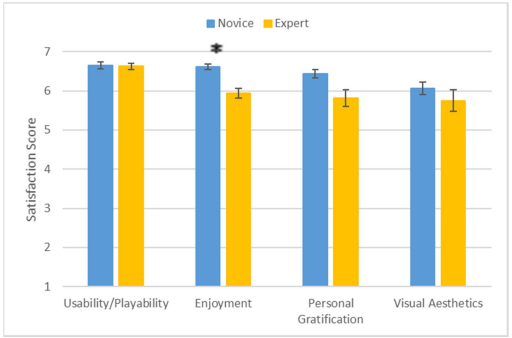

Figure 2. GUESS subscale comparison.

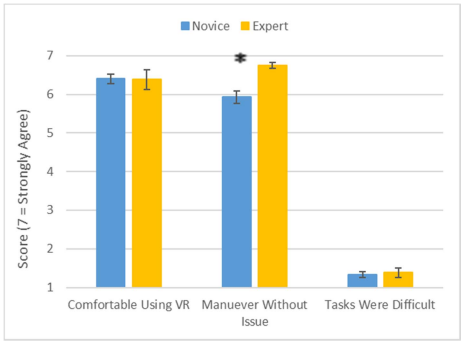

Figure 3. Self-report comparison.

Perceived usability. There was no significant difference between the SUS scores of novices (M = 87.67, SD = 8.84) and experts (M = 90.31, SD = 9.4), $p > .05$. Both of these scores correspond to a "Best Imaginable" adjective rating (Bangor et al., 2009).

Perceived workload. A significant difference was found within the mental demand subscale of the NASA TLX between novices (M = 33.53, SD = 27.99) and experts (M = 13.38, SD = 10.99), $t(22) = 2.456$, $p = .023$, $d = 0.95$.

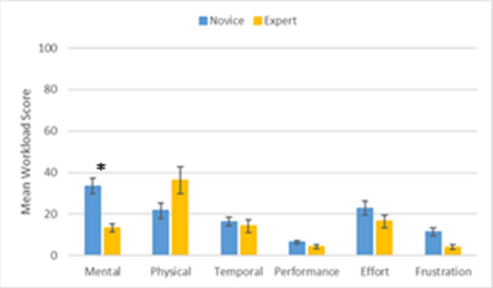

Figure 4. NASA-TLX Score Comparison.

VE post-assessment task performance. No significant differences were found in difficulty scores between both the novice and expert groups, $p > .05$. On average, participants rated all tasks (teleporting, gaze, add item to inventory, place item from inventory, flip toggle switch, and spin and rotate an object) between 1 and 2. Many participants reported that a task was easy because they had learned it in the tutorial and/or recalled how to complete the task from memory. No significant differences were found in difficulty scores between both the novice and expert groups, $p > .05$.

Researchers observed participants as they completed tasks within the VR tutorial and post-tutorial assessment. Most users did not encounter issues in the tutorial, but for the few who did reported difficulties with teleporting or pressing the correct button on the controller. In the virtual assessment, many users who ranked a task as higher than 1 reported that they attempted to perform the task incorrectly. For example, several participants pressed the wrong buttons and therefore could not complete the task without multiple attempts. Those who ranked the task as a 1 noted the task was easy because they had "learned how in the tutorial" or they had "done it many times." Observations revealed that although some participants struggled to perform a task the first time, subsequent iterations were performed more easily with low error rate. Table 4 highlights select gestures and tasks which novices found difficult to execute.

Task or Gesture	N	Researcher Observations
Choosing correct button for task	4	- Several participants pushed wrong buttons before choosing correct button(s).
Pressing all of the buttons in tutorial	6	- Participants had issue finding all of the buttons (mostly the bottom and ring finger buttons).
1st teleport in tutorial	3	- N = 2 were due to system error. - Participants had issue snapping teleport arc to target resulting in multiple attempts.
	12	- 3, 4, and 5+ attempts were observed.
Adding/ choosing inventory	7	- Participants reported that this was an easy interaction because item immediately added.
Moving in VE	11	- Participants would not move, only reach, resulting in multiple attempts to complete task.

Table 3. Select Observations of Task/Gesture Execution.

Participants were given the option to provide qualitative feedback on the tutorial. Seven participants gave positive feedback, saying the tutorial was user friendly, fun, and enjoyable. One participant said, "Not only the instruction was clear but the interactions given by the tutorials made it understandable." Additional comments included that the tutorial was "very user friendly and fun to use for a first timer enjoyed it and wanted to do more" and "It was easy and fun as well as well programmed." Two more participants gave feedback and stated that the HMD heated up over time, caused them to feel dizzy at first, and that the HMD's cable caused frustration as they physically moved in the defined space.

Simulator sickness. All participants rated each subscale low. This was true for novices' total SSQ score after the tutorial (M = 14.21, SD = 14.68) and after the virtual assessment (M = 14.03, SD = 17.51) as well as experts' total SSQ score after the tutorial (M = 16.36, SD = 14.68) and after the virtual assessment (M = 14.03, SD = 17.51). The maximum score that can have been achieved is 235.62. No significant differences were found in any of the scores between groups or time, $p > .05$.

DISCUSSION

The results of the usability test for a VR tutorial are overall positive. Users' satisfaction of the tutorial was high. Both novice and expert participants felt overall comfortable using VR, reported a low amount of simulator sickness and workload, and rated the tutorial highly on usability, personal gratification, and visual aesthetics. Novices reported that they were less comfortable maneuvering in the VE and rated the tutorial more mentally demanding and enjoyable than experts. This is likely due to a novelty effect from using VR for the first time.

Many participants had little to no issues in the tutorial and VE post-assessment. This demonstrated that the tutorial was effective and learnability was high as the majority of

participants rated tasks very low in difficulty and did not have to repeat any tasks. Many participants reported that a task was easy because they had learned it in the tutorial and/or recalled how to complete the task from memory.

However, when participants did encounter problems, they involved either moving in the VE or pressing buttons on the controller. Some participants had to attempt tasks multiple times because they tried to only teleport in the VE instead of walking around, could not snap the teleport arc to the target, could not find the correct button to complete an action, or in the case of two individuals, encountered a system error where some of the buttons did not appear in the VE. Observational evidence suggests that performance increases and error rate decreases as the user moves through the tutorial and completes tasks multiple times. Changing the first teleport task to a closer target may improve user understanding of how to teleport; subsequent teleports with farther targets and voice prompt to physically move in the space may further improve user navigation. Engaging with the VR tutorial allowed participants to build confidence in the VE and enabled efficient task completion in the VE post-assessment. All participants were able to complete the tutorial, survey assessments, and post-assessment in the VE; no participant experienced physical discomfort severe enough to end participation in the study.

Having demonstrated the learnability, efficiency, and effectiveness of the VR tutorial for novice and expert users, the tutorial may be employed to ensure all users engage in the VE with confidence. The tutorial may also be used by those interested in learning more about the VE. Future research within the lab is being planned by professors of commercial space operations, air traffic management, and aviation maintenance classes. Those wishing to use the VR laboratory for research purposes or to integrate VR into curriculum will have the option to incorporate the tutorial by fall of 2019. The methodology of the experiment may be modified for future usability studies with new VR programs.

A limitation of this study was the inability to measure efficiency, measured by time to complete task and time to complete tutorial. Efficiency could be measured using video and/or audio recording in a future study.

ACKNOWLEDGEMENTS

The research team wishes to acknowledge several groups of people in the creation of this project. First, the Deans of the College of Aviation, who have ever supported and promoted the establishment of a cross-reality laboratory within the College. The Research in User Experience Lab of the College of Arts and Sciences must be thanked for their guidance and support of this project every step of the way. The VR/AR/MR Committee is commended, as they have endeavored to create a lab with viable research projects that promote innovative, diverse learning opportunities for the students at ERAU. Finally, this project could not have been done without Special VFR Productions, the student lab assistants, and the developers who created the VR programs, assisted in data collection, and provided technical support every step of the way.

REFERENCES

Allerton, D. (2009). *Principles of flight simulation*. West Sussex, UK: John Wiley & Sons, Ltd.

Bangor, A., Kortum, P., T., & Miller., J., T. (2009). Determining what individual SUS scores mean: adding an adjective rating scale. *Journal of Usability Studies*, *4*(3), 114-123.

Brooke, J. (1996). SUS - A quick and dirty usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L .McClelland (Eds), *Usability Evaluation in Industry* (pp 189–94). London: Taylor & Francis.

Carretta, T. R., & Dunlap, R. D. (1998). Transfer of training effectiveness in flight simulation: 1986 to 1997. Air Force Research Lab, Brooks AFB, TX. Human Effectiveness Directorate.

Cohn, J., & Fletcher, J. D. (2010). What is a pound of training worth? Frameworks and practical examples for assessing return on investment in training. In *Proceedings of the 31st InterService/Industry Training, Simulation and Education Annual Conference* (I/ITSEC 2010 Paper Number 10013). Arlington, VA: National Training and Simulation Association.

Dumas, J. S. & Redish, J. C. (1999). *A practical guide to usability testing*. Exeter, UK: Intellect Books.

Fletcher, J. D., & Chatham, R. E. (2010). Measuring return on investment in military training and human performance. In J. Cohn & P. O'Connor (Eds.), *Human performance enhancements in high-risk environments* (pp. 106 –128). Santa Barbara, CA: Praeger/ABC-CLIO.

Foster, R. E., & Fletcher, J. D. (203). Toward training transformation. *Military Psychology, (25)*3, 308-317. DOI: 10.1037/h0094971

Gray, R. (2017) Transfer of training from virtual to real baseball batting. *Frontiers in Psychology*, 8, 2183.

Hale, K. S., Stanney, K. M., & Badcock, D. R. (2015). *Handbook of virtual environments: Design, implementation, and applications* (Second ed.). Boca Raton, Florida: CRC Press.

Kennedy, R., Lane, N., Berbaum, K., & Lilienthal, M. (1993). Simulator sickness questionnaire: An enhanced method of quantifying simulator sickness. *The International Journal of Aviation Psychology, 3*, 203-220.

Macchiarella, N. D., Arban, P. K., & Doherty, S. M. (2006). Transfer of training from flight training devices to flight for ab-initio pilots. *International Journal of Applied Aviation Studies, 6*(2). Retrieved from http://commons.erau.edu/publication/149

Nielsen, J. (1996). Usability metrics: Tracking interface improvements. *IEEE Software, 13*(6), 12. doi:10.1109/MS.1996.10031

Phan, M. H., Keebler, J. R., Chaparro, B. S. (2016). The development and validation of the game user experience satisfaction scale (GUESS). *Human factors, 58*(8), 1217-1247

Sanders, J. (2010). The importance of usability testing to allow e-Learning to reach its potential for medical education. *Education for Primary Care, 21*, 6-8. doi: 0.1080/14739879.2010.11493869

Sharek, D. (2009). NASA-TLX Online Tool (Version 0.06) [Internet Application]. Research Triangle, NC. Retrieved from http://www.nasatlx.com

Stanney, K. M., Mourant, R., & Kennedy, R. S. (1998). Human factors issues in virtual environments: A review of the literature. *Presence: Teleoperators and Virtual Environments, 7*(4), 327-351.

Taylor, H. L., Lintern, G., Hulin, C. L., Talleur, D. A., Emanuel Jr, T. W., & Phillips, S. I. (1999). Transfer of training effectiveness of a personal computer aviation training device. *The International Journal of Aviation Psychology*, *9*(4), 319-335. doi:10.1207/s15327108ijap0904_1

Talleur, D. A, Taylor, H. L, Emanuel Jr. T. W., Rantanen, E., & Bradshaw, G. L. (2003). Personal computer aviation training devices: Their effectiveness for maintaining instrument currency. *The International Journal of Aviation Psychology, 13*(4), 387-399, DOI: 10.1207/S15327108IJAP1304_04

Vincenzi, D. A., Macchiarella, N., Opalenik, C., Gangadharan, S. N., & Majoros, A. E. (2003). The effectiveness of cognitive elaboration using augmented reality as training and learning paradigm. *Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting*. Denver: Human Factors & Ergonomics Society.