Spring 2024

The Use of Virtual Reality Training Environments for Procedural Training in Fourth-Generation Airliners

Mark Edgar McCullins
Embry-Riddle Aeronautical University, mccullim@my.erau.edu

Scholarly Commons Citation

McCullins, Mark Edgar, "The Use of Virtual Reality Training Environments for Procedural Training in Fourth-Generation Airliners" (2024). *Doctoral Dissertations and Master's Theses*. 825.
https://commons.erau.edu/edt/825

The Use of Virtual Reality Training Environments for Procedural Training in Fourth-Generation Airliners

Mark Edgar McCullins

Dissertation Submitted to the College of Aviation in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Aviation

Embry-Riddle Aeronautical University
Daytona Beach, Florida
May 2024

The Use of Virtual Reality Training Environments for Procedural Training in Fourth-Generation Airliners

By

Mark Edgar McCullins

This dissertation was prepared under the direction of the candidate's Dissertation Committee Chair, Steven Hampton, Ed. , and has been approved by the members of the dissertation committee. It was submitted to the College of Aviation and was accepted in partial fulfillment of the requirements for the Degree of Doctor of Philosophy in Aviation.

Steven Hampton Digitally signed by Steven Hampton
Date: 2024.05.14 11:27:52 -04'00'

Steven Hampton, Ed.D.
Committee Chair

Kristine Kiernan Digitally signed by Kristine Kiernan
Date: 2024.05.14 14:12:31 -05'00'

Kristy Kiernan, Ph.D.
Committee Member

Steven Hampton Digitally signed by Steven Hampton
Date: 2024.05.14 11:28:20 -04'00'

Steven Hampton, Ed.D.
Associate Dean, School of Graduate Studies, College of Aviation

Jennifer E. Thropp Digitally signed by Jennifer E. Thropp
Date: 2024.05.15 13:24:09 -04'00'

Jennifer Thropp, Ph.D.
Committee Member

Alan J. Stolzer Digitally signed by Alan J. Stolzer
Date: 2024.05.16 08:59:41 -04'00'

Alan J. Stolzer, Ph.D.
Dean, College of Aviation

Committee Member

Stephanie G. Fussell Digitally signed by Stephanie G. Fussell
Date: 2024.05.15 15:37:50 -04'00'

Stephanie Fussell, Ph.D.
Committee Member (External)

May 3, 2024

Signature Page Date

Abstract

Researcher: Mark Edgar McCullins

Title: The Use of Virtual Reality Training Environments for Procedural Training in Fourth-Generation Airliners

Institution: Embry-Riddle Aeronautical University

Degree: Doctor of Philosophy in Aviation

Year: 2024

This study examined the effectiveness of using virtual reality training environments for procedural training in fourth-generation airliners. Its goal was to assess whether the training outcomes from a recurrent training course for FAA certificated Airframe and Powerplant (A&P) technicians, which used a Full Flight Simulator (FFS) to deliver and assess training, differed from the same training delivered using a Virtual Reality (VR) device.

The study used an experimental design with three groups and two within-group measures of training effectiveness. The control group followed the current training program and was assessed in the FFS, while the second group was trained using a VR device and was subsequently assessed in the FFS. A third group was formed as a sub-group of the second group, and it contained subjects who had prior VR experience. Training effectiveness was assessed using a modified Global Evaluative Assessment of Robotic Skills (GEARS) tool that measured cognitive and psychomotor aspects of learning along with the time to successful task completion.

The population sampled for the study were all FAA certificated A&P technicians who were engine-run qualified; a total sample of 100 was used. Ages ranged from 22 to

72 years old, with a mean age among all groups of 40.37 (*SD* = 11.50). Four out of 100 participants were female. A&P experience ranged from 1 to 42 years, with a mean experience among all groups of 14.79 years (*SD* = 10.34). Engine-run experience ranged from 0 to 35 years, with a mean experience among all groups of 9.01 years (*SD* = 8.02).

The hypothesis tested was that there is no difference in performance between the three groups. A MANCOVA analysis was performed using the GEARS scores and Time to Completion as variables. There was no significant difference in training effectiveness (GEARS Total Score and Time) based on Training Group (Control, VR, and VR with Experience), $F(4,190) = 1.307$, $p = .269$; Wilk's lambda = .946, partial eta squared = .027, and the null hypothesis was retained. Similar results were returned using individually the cognitive and psychomotor elements of the GEARS assessment. Engine-run Experience was significant in influencing both GEARS Psychomotor Score, $F(1, 95) = 5.732$, $p = .019$ and in influencing Time to task completion, $F(1, 95) = 9.346$, $p = .003$. Engine run experience was a significant covariate in this study, while overall A&P experience was not.

The VR system, as evaluated, was found to provide task performance that is equivalent to that of the traditional training method that used the FFS. Recommendations for future research and ongoing application of the specific experimental methodology are provided.

Keywords: aviation training, flight simulation, maintenance training, virtual reality, VR.

Dedication

To Kathryn, Kai, and Kalindi, with love.

Acknowledgements

“He who would learn to fly one day must first learn to stand and walk and run and climb and dance; one cannot fly into flying.”

- Friedrich Nietzsche

It is not possible to complete an endeavor such as this without the help, support, and love of a cast of hundreds. While I have space and time to mention a few, I would like to acknowledge the hundreds of colleagues and friends who have taken the time to share their experiences in learning over the years, and their insights into the new virtual worlds which we are beginning to explore.

Dr. Hampton, thank you for accepting me as a student and for being my dissertation chair. As you put it, “I always get the weird ones!” I have appreciated your input and advice over the years, as well as the prompting to, “Get on with it!” when needed. Thanks also to my committee for their time and insights: Doctors Kiernan, Thropp and Fussell. I am humbled that you took the time to work with me, and your input has made this a better study.

This study could not have been done without the support of the team at the Airbus Miami Training Center. In particular, I owe a huge debt of gratitude to Victor Liriano and Erik Marrero who ran the VR training program, and to Jimmy Torres, Dennis Zubizarreta, and Jose Barro who helped to collect the data that was used in this dissertation. Thank you for your time and efforts; this dissertation also belongs to you.

Finally, to my parents who instilled in me a love of learning and education, my wife Kathryn, and my children Kai and Kalindi. Your love and support every day encourage me to learn, grow, and become a better person.

Table of Contents

Page

Contents

List of Tables

List of Figures

Chapter I: Introduction

When asked about investing and wealth, Richard Branson said, “If you want to be a millionaire, start with a billion dollars and launch a new airline” (Brainy Quote, 2019)! Other aviation entrepreneurs have made similar comments, and they attest to the massive resources consumed by an aviation operation. Beyond fuel and capital acquisitions, the single biggest expense for an airline is personnel (International Civil Aviation Organization [ICAO], 2013). Trained and qualified personnel are required to fly and maintain the aircraft, and these personnel must be provided with both initial and recurrent training. These training costs are significant, require dedicated facilities and simulators, and make up a significant portion of an air operator’s fixed operating costs (Cao et al., 2024).

In addition to the material that has to be taught in any training event, airlines must also consider how the training will be delivered. Pilots cannot learn the skill of flying from a book, and no amount of lecturing will make a technician competent to repair and reinstall an aircraft engine. In this sense, aviation has much in common with fields such as automotive mechanics, medicine, dentistry, and marine operations: it requires continuous and competent application of theoretical concepts in a physical environment. No one would want to get a root canal from a dentist who had only read about the procedure, nor would one wish to fly with a pilot who had spent hundreds of hours in the classroom but has never physically landed an aircraft. Pilots and mechanics are required to not only demonstrate the required theoretical knowledge to do their jobs but are also required to demonstrate competency in the physical performance of all job-related tasks (Franks et al., 2014). It is axiomatic that an aircraft that is not flying is not earning money, so over the years airlines and flight schools have developed a suite of tools to give their personnel the maximum amount of hands-on experience and practical training without using a physical aircraft (McCullins, 2019).

There are a wide range of Flight Training Devices (FTDs), part-task trainers, workbenches, and other low-fidelity training devices that have served for years as effective and low-cost options for training aircrews and maintainers (Macchiarella et al., 2008). Despite this, certain tasks still require either the use of an aircraft, or the nearest possible thing, which is the Full Flight Simulator (FFS) (Airbus, 2016, 2019).

The global simulator population today is estimated to be approximately1500, serving a global airline fleet of approximately 26,000 transport aircraft, a number that is forecast to grow to 35,000 aircraft within 10 years (Fafard, 2020). These simulators must serve not only the pilots of the aircraft but also the technicians who must qualify to start engines, run engines, and taxi the aircraft. Simulator time is precious, and there is never enough of it to go around. The return to service of the Boeing 737 Max following almost two years of grounding is a case in point; airlines have been unable to put enough crews through the available simulators to bring the aircraft back into service rapidly (Lampert et al., 2021). The costs of using aircraft to train, the scarcity of flight simulators and associated costs of building more, and the intense competition for simulator time has led operators and aviation training providers to ask the question: Is there not a better, more effective, and cheaper way that systems training could be delivered to both aircrew and maintenance technicians?

Background

Training can take place in many different formats, but the traditional approach has been to deliver the theory portion in a classroom setting, and then move onto progressively more advanced training tools before progressing to the actual aircraft (Airbus, 2019). This approach is used for both pilot training, where the training tools are used to simulate an aircraft in flight, and maintenance technician training where the tools are used to simulate the function of the various aircraft systems and to promote understanding of how they function in a maintenance setting.

In many programs the final stage of training takes place in the FFS, which replicates the functions, sounds, vibrations, and motion of the aircraft in a fixed indoor facility. An FFS replicates the responses and handling of a physical aircraft and its systems from a cockpit perspective, thereby reducing costs and freeing up the aircraft to perform revenue flights. The most advanced simulators incorporate motion, sound, and visual effects, and allow pilots and technicians to become rated on the aircraft by using only the FFS; such is their similarity to the real world (Bürki-Cohen et al., 1998). Much like all civil aviation authorities worldwide, the Federal Aviation Administration (FAA) of the United States of America certifies and inspects these simulators on a regular basis, allowing them to be a powerful tool in qualification and recurrent training for flight crews and technicians. Despite these benefits, an FFS can cost as much to acquire as an actual aircraft, requires a dedicated facility to house it, and needs constant maintenance on its systems and computers. It also ties training to a specific location, requires operators to dedicate resources for personnel to attend these facilities, and trainees are forced to compete for a scarce resource (Curnow, personal communication, 2021).

The Regulatory Environment of Aviation Training

Training for licensed and/or certified aviation crewmembers and technicians is governed by the FAA in the United States, the European Aviation Safety Agency (EASA) in the European Union, and other national and pan-national authorities worldwide. Each has its own set of regulations that govern licensing, and in the United States these are drawn from the Code of Federal Regulations (CFR) Part 14.

14 CFR § 61 (Certification: Pilots, Flight Instructors, and Ground Instructors, 1997) deals with licensing requirements for pilots and flight instructors, while 14 CFR § 65 (Certification: Airmen Other Than Flight Crewmembers, 1962) deals with the certification of, “airmen other than flight crewmembers,” which includes Aviation Maintenance Technicians (AMTs). Aviation Maintenance Technician Schools (AMTS) that train technicians are regulated by 14 CFR § 147 (Aviation Maintenance Technician Schools, 2022) to deliver curricula that can result in the award of an AMT Airframe or Powerplant certificate in a general sense, with much more specific training provided depending on the nature of the school (e.g., a student could specialize in helicopter airframes). A technician who holds both Airframe and Powerplant qualifications is commonly referred to as an A&P. These schools are closely regulated by the FAA, and specific approval is given for the facilities that are used for training, the approved devices with which training can be conducted, the specific course content and syllabi that are used to deliver training, and the order in which training events are sequenced (FAA, 2021). If it is desired to deviate from an already approved course, or method of instruction, additional FAA approval is required.

Economics of Simulator Training

Although considerably cheaper to operate than an actual aircraft, an FFS requires considerable investment to install, operate and maintain. The computers to run the simulator, the visual systems that present to the crew a simulated world that is a high-quality replication of the real one, the hydraulics to move the simulator, the electricity to power it, and the infrastructure to house the FFS can make it an investment that approaches the cost of an aircraft itself (Curnow, personal communication, 2021; see Figure 1). Rates for use plus instruction can approach $1,500/hour, and a typical FFS would expect to see an optimal utilization of around 5,000 hrs/yr after taking maintenance and upgrades into account (Curnow, personal communication, 2021). This introduces a real bottleneck into the production of crews for any given operator: a standard pilot training course requires between 36-40 FFS hours per crew of 2, meaning that a single FFS has a theoretical maximum production rate of 125 crews/year (Airbus, 2016). Airlines typically crew 6-8 pilots per aircraft, meaning that a fleet of 100 aircraft would require between 600-800 pilots, and would occupy 3-4 FFS for a year to qualify them (Clark, 2007). This gives an appreciation of the level of utilization of the FFS within the industry just to produce new pilots and does not even begin to account for the ongoing training requirements that operators face to comply with 14 CFR § 121 and 135 operations.

A&P technicians also have a requirement to utilize the FFS to complete initial and recurrent training to run engines and operate aircraft systems. A typical training center, such as the one shown in Figure 1, may contain anywhere from 2-10 simulator bays while still being unable to meet the disparate training requirements of its customers due to the combination of training and requalification requirements that they work under. The operators of training centers are constantly looking to optimize utilization of their FFS, and to shift training to alternate or lower fidelity devices when possible in order to free up FFS hours for dedicated pilot training (Bürki-Cohen et al., 1998).

Figure 1

Modern Full-Motion Flight Simulators

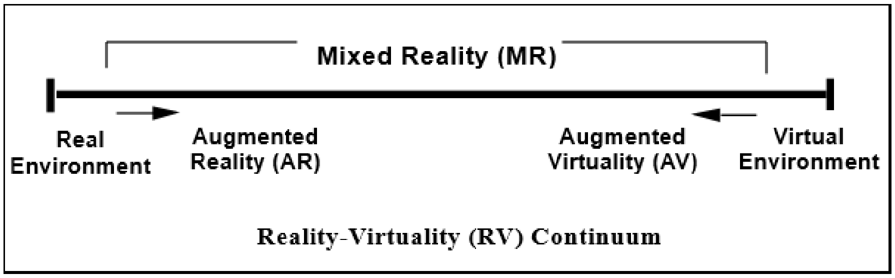
Note. Modern full-motion flight simulators at the Airbus Training Center in Miami, Florida. These simulators can be used to fully qualify pilots to fly the A320 without ever flying the actual aircraft. Copyright 2018 by M. E. McCullins.

Virtual Reality Training

One area that holds immense promise to supplement or replace the use of FFS in training is Virtual Reality (VR). VR is defined as, “the use of computer graphics systems in combination with various display and interface devices to provide the effect of immersion in the interactive 3D computer-generated environment" (Pan et al., 2006). Pure VR is a powerful tool that synthetically generates images, information, sounds, and haptic feedback through its interfaces, and is viewed on a continuum with both VR and the real environment as shown in Figure 2.

Figure 2

Simplified Representation of the Reality-Virtuality (RV) Continuum



Note. Adapted from "Augmented Reality: A Class of Displays on the Reality-Virtuality Continuum by P. Milgram, H. Takemura, A. Utsumi, and F. Kishino, 1995, *Proceedings SPIE: Telemanipulator and Telepresence Technologies*, p. 283 (https://doi.org/10.1117/12.197321). Copyright 1995 by SPIE.

The goal of VR is to produce a synthetic environment in which the user can interact, explore, and influence. Through the use of vision systems (VR head-mounted devices) and various Human Interface Devices (HIDs), the user can interact with the virtual environment as if it were real. The quality of the virtual environment can vary due to factors such as the resolution of the graphics, detail programmed into the virtual world, the speed of the processor that controls the simulation, and feedback available to the user through the HIDs used. When compared to part-task trainers and FTDs, VR is considered an immersive simulation technology and is actively being explored for use in maintenance and pilot training (Li, 2023; Macchiarella et al., 2005).

Major airline training providers, the airlines themselves, and other third-party training providers are currently examining the use of VR environments to reduce their overall dependence on the FFS, and its associated infrastructure, for the training of pilots and maintenance personnel (Airbus, 2019). A sample system, designed specifically for training maintenance personnel in engine run techniques, is shown in Figure 3, and incorporates a visual display that allows the user to see both the virtual cockpit environment and another trainee sitting in an adjacent seat. Hand controllers are provided as HIDs to allow the user to interface with the cockpit systems, which respond as they would in the FFS. This VR system does not allow for full interactivity with the virtual environment and trainees must follow a pre-defined script which is based on specific learning objectives and sequences; however, it does allow the user to see the logical consequences of actions taken within the script and to see the real time functioning of aircraft systems based on user input (Airbus, 2019).

Figure 3

VR Training System

Note. Copyright 2019 by M. E. McCullins

A VR system such as this that lets the trainee interact with the aircraft, in much the same manner as in the FFS, has the potential to revolutionize airline training. It would both reduce costs and improve access to training devices due to the much smaller footprint and infrastructure required for the VR system (Curnow, 2021). Portions of initial or recurrent training could be conducted using a VR platform, thereby reducing the demand on the FFS, and improving trainee throughput by optimizing the time spent in the FFS (Airbus, 2019).

It is not being suggested that the trainee could learn to fly or maintain an aircraft using a system as shown in Figure 3, but rather that those portions of the training that involve interacting with and manipulating aircraft systems could be learned and practiced in the VR environment rather than the FFS, which could then be used only for those sequences requiring sound, motion, and physical manipulation of the aircraft. VR training could be particularly effective in teaching maintenance technicians how to perform sequences in the flight deck, such as running the aircraft's engines, and could replace the FFS entirely for recurrent training requirements (Jerald, 2016). Recurrent engine-run training does not involve use of the FFS motion capabilities, so it is highly desirable to find an alternative means of delivering this training (Airbus, 2019). VR tools that have been successfully used in the medical field to train surgeons on procedures and techniques, particularly for robotic surgery, are beginning to be adapted for use in education, aviation, and industry. Their use is increasing as shown in Table 1, which lists a selection of more recent works that studied the use of VR in various learning and training environments. Renganayagalu et al. (2021) reviewed studies on the effectiveness of VR training over the last 30 years and extracted a total of 60 studies to analyze. They found a total of 30 articles published in the period spanning from 1988-2013, with a further 30 articles published from 2013-2018. Interest in the field of VR training is expanding; however, "over the last three decades there have been limited reviews covering the effects of VR on training" (Abich et al., 2021).

Table 1

Current Research in the Field of VR Training in Various Environments

Environment	Research Type	Context	Limitation of study	Reference
Aviation	Study	Evaluating the efficacy of VR devices for pilot training	Comparative study of PC based and VR training	Guthridge & Clinton-Lisell, 2023
Aviation	Study	VR part-task trainer (PTT) development for cockpit familiarization	Analysis, design, and development of PTT for military pilots	Sikorski et al., 2017
Aviation education	Study	TAM for AR use in maintenance training	Original TAM constructs, AR not VR	Wang et al., 2018
Aviation	Study	VR PTT for checklist training	Usability and validation of PTT for military pilots; did not use TAM	Palla et al., 2018
Education / Aviation	Quasi - Experiment	Effectiveness of VR training for ab initio civilian pilot training	Small sample, no random assignment	Hight et al., 2022
Education / Training	Review	Review of studies of VR use in education and training	Literature review	Jensen & Konradsen, 2018
Industrial	Review	Evidence for training effectiveness with VR technology	Literature review	Abich et al., 2021
Industrial	Review	A review of the effectiveness of VR head-mounted displays in professional training	Literature Review	Renganayagalu et al., 2021
Maintenance	Study	Design and development of aviation aircraft maintenance training platform	Conceptual design and analysis	Li, 2023
Maintenance	System development	Developing VR training systems for industrial training	System development	Yuviler- Gavish et al., 2013
Medicine	Review	Review of VR training for improving operating room performance	Literature review of medical studies	Seymour, 2008
Medicine and Gaming	Study	VR gaming for the rehabilitation of stroke	Gamification, medical rehabilitation	Saposnik et al., 2010
Nuclear	Pilot Study	Simulation-based VR training	Small sample, TAM based	Masiello et al., 2022

Note. Adapted from *Determinants of Aviation Student's Intentions to use Virtual Reality for Flight Training* (p.8), by S.G. Fussell, 2020, Scholarly Commons (https://doi.org/10.15394/ijaaa.2020.1504). Copyright 2020 by S.G. Fussell.

Problem Statement

Within the United States, the FAA must approve every part of a certified training program. This means that when an operator or school wants to adopt a new way of teaching, or a new delivery method, no matter how promising, they must get certification of their program by the FAA (2021). Aviation training is of necessity closely regulated for both quality and safety reasons, which makes a lot of sense from a public safety perspective, but it also makes it difficult to quickly adopt new technologies and methods of teaching. To introduce a VR system for training in an airline environment, the training must be proven to be at least as good as the training provided by the current certified systems in order to provide an equivalent level of safety (FAA, 2021). The use of immersive simulation outside the FFS has seen limited study, and limited studies of the effectiveness of VR training in an aviation environment have been found (Airbus, 2019; Renganayagalu et al., 2021). This is significant given the public perception that aviation is at the forefront of using simulation technologies such as the FFS. Fussell (2020) also found limited studies on the effectiveness of VR in aviation training, although a number of studies in other fields were found and listed in Table 1.

Although VR systems have been successfully used in other fields to deliver technical training at the same level demanded from pilots and AMTs, the FAA still requires that any VR based system be specifically approved for use in delivering a program of studies (FAA, 2021). In practical terms this means that a new system using VR must be designed, perfected, validated, and then certified by the FAA prior to use. This required certification is a significant barrier to the use of VR in training, and one that must be overcome before this technology sees significant use in the industry. To provide the FAA, or any similar certifying authority, evidence of the suitability of a VR training suite would require a study of the effectiveness of training conducted under a VR environment, compared to training conducted in an FFS. The is no currently accepted methodology for proving the Equivalent Level of Safety (ELOS) required by the FAA, and other national regulatory authorities, to implement an expanded use of VR in training. A method for assessing this ELOS is required that could serve as a template on which future validation studies of VR environments could be based.

Purpose of Study

The current study assessed the effectiveness of VR delivered training as compared to traditional training as currently delivered in an FFS. Its purpose was to provide a quantitative study, using robust experimental methods and an adequate sample size, to examine the causal relationship between the use of VR training and subsequent task performance of the subjects. A&P technicians, certificated by the FAA to conduct engine runs, were assessed on selected tasks post-training in order to allow for a quantitative assessment of the effectiveness of VR training as compared to training delivered in the FFS (Airbus, 2019). The group that underwent VR training was further sub-divided into two sub-groups depending on whether or not the subjects had previous exposure to commercial VR systems in order to examine if prior exposure to VR had a measurable effect on task performance. The subsequent comparison between the subjects who received only FFS training, and the two sub-groups who received VR training, allowed conclusions to be drawn as to the relative performance of the groups trained via VR versus the group that received training as it is currently delivered.

Significance of Study

In order to assure an ELOS in training that would allow for the certification and use of more VR systems by the FAA, an established methodology for proving their effectiveness is needed. This study developed an instrument that combines high validity and reliability in measuring VR training effectiveness in a multivariate approach that can be replicated and further developed by future researchers. The competencies for A&P engine–run training are not significantly different between A&Ps and pilots interfacing with the same aircraft systems and performing similar tasks (i.e., starting engines, using checklists, and reacting to abnormal situations), which allows for the generalization of results beyond maintenance technicians to any other training involving identical competencies using a VR environment to replace an FFS. This generalization can be expanded beyond the FAA regulatory environment through comparison and analysis of compatible international standards such as those put in place by EASA, or more generically recommended by ICAO. The operational use of a multivariate method of assessing VR training performance is an important contribution to the aviation body of knowledge and is now available for future research using more complex scenarios and VR environments.

An accepted VR training solution for recurrent training of certificated FAA Maintenance Technicians would demonstrate that it is possible to replace a significant portion of the training, currently delivered in an FFS with its associated infrastructure, by a VR system that is smaller, cheaper, and more readily scalable. To do this, the FAA requires quantitative evidence that the performance of technicians who undergo VR training is equal to or better than those who receive training under the current system.

Each group of technicians who undergo recurrent training occupies one 4-hour simulator slot. The freeing up of a total of eight of these slots would allow for an additional crew to receive a type rating qualification on the aircraft and is a more efficient use of the FFS resource (Airbus, 2016). As discussed in the prior section on the economics of simulator training, this would provide a significant reduction in the fixed costs associated with A&P engine-run recurrent training as the FFS would not be used, thereby opening up the possibility of conducting this training remotely through the use of a networked VR system that would allow the instructor to interface in real-time with the students from a different facility.

Research Question

The research question of this study was: Does VR delivered A&P engine-run recurrent training produce equivalent test performance when compared to training in the FFS, when we control for the subject's level of experience both as an A&P and in conducting engine runs?

Hypotheses

The present study compared the training results of three main groups: one that received engine run recurrent training as currently delivered in an FFS, and two that received training delivered in a VR environment. The two VR trained sub-groups consisted of one containing subjects who had never used a VR system before (VR), and the other with subjects who had exposure to commercial VR systems (VR Exp). The results across the three groups were then compared to determine if there was a statistically significant difference between them. The null hypothesis for this comparison was that there was no difference between the groups, and the alternate hypothesis was that a statistically significant difference was observed. Table 2 shows the global variables that were compared in this study, with the independent variable being the practical training delivery, and the dependent variables being the results achieved on the evaluation. These variables are examined in Chapter III, Methodology, to control for varying levels of experience within the independent variable groups, and differing measures of performance that were obtained from the assessment tool. A multivariate analysis of means technique was used to quantify the differences between the groups.

Table 2

Comparison Factors and Grouping

	Test Performance	
Practical Training	**Score**	**Time to Completion**
FFS	X_A	X_B
VR	X_C	X_D
VR Exp	X_E	X_F

As there were three distinct groups to be compared (FFS, VR and VR Exp), and two dependent variables per group (Score and Time to Completion) that were further broken down, a MANCOVA analysis was used. The hypotheses were:

HA_0

There is no collective statistically significant difference in Test Performance between the groups when controlled for A&P and engine-run experience.

HA_a

There is a collective statistically significant difference in Test Performance between the groups when controlled for A&P and engine-run experience.

If the null hypothesis were to be rejected, post-hoc tests would be performed to determine which groups demonstrated superior Test Performance and on which metric.

The MANCOVA analysis was carried out to examine if either the experience levels of the A&P technicians (Experience AandP – measured in years holding an A&P license), or the number of years that they have been performing engine runs (Experience ERun – measured in years as a qualified engine runner), influenced the dependent variables.

As a part of this analyses the dependent variables were also considered individually, and for each dependent variable, Score (S), or Time (T), the hypotheses were:

HS_0

There is no statistically significant difference in Score between the groups when controlled for A&P and engine-run experience.

HS_a

There is a statistically significant difference in Score between the groups when controlled for A&P and engine-run experience.

and,

HT_0

There is no statistically significant difference in Time between the groups when controlled for A&P and engine-run experience.

HT_a

There is a statistically significant difference in Time between the groups. Should the null hypothesis be rejected, post-hoc tests would be performed to determine which groups demonstrated superior Test Performance and on which metric when controlled for experience.

Delimitations

The results of the current study are directly applicable to already qualified and certificated FAA A&P technicians undergoing A320 engine-run requalification under an approved FAA syllabus. No simulator motion was used in the FFS, and no motion was modelled in the VR training. The evaluation used an established training system that is approved under 14 CFR § 147 and taught using a competency-based training system (Airbus, 2019). The results are generalizable to any other A320 competency-based training course that uses the same competencies being evaluated.

Limitations and Assumptions

Limitations

Only FAA qualified and certificated A&P technicians participated in this study, which limits the applicability of the immediate results to AMTs certificated under the applicable FAA 14 CFR § 147 training and licensing program. The technicians were already qualified to perform engine-run maintenance tasks on the A320 aircraft and met all requirements for recency and currency to undertake annual re-qualification training; this limits the results of this study to recurrent training scenarios where the goal is to refresh and update knowledge, rather than present knowledge and systems for the first time for the purposes of an initial qualification.

Assumptions

The subjects of this study were all certificated FAA A&P maintenance technicians who qualified for engine-run recurrent training under existing training center policies and procedures (14 CFR § 147). Their experience and qualifications were vetted by the training center, but no outside validation was possible, and the qualifications as presented by their airlines were accepted at face value and assumed to be representative of the subject's true status. They were asked to indicate on the participant information form if they had previously used a VR system for any purpose, and told that they should answer "yes" for any interactive use of a VR system, or any passive use beyond a quick demonstration. It was assumed that this question was answered truthfully and accurately.

The FFS was assumed to be equivalent to an actual aircraft for the purposes of evaluation of learning in this study. This is well supported both by regulation and FAA practice, and the low-risk nature of the tasks being performed and evaluated eliminates any possibility of a simulator effect (i.e., a behavioral change on the part of the subject due to a perceived difference in risk level between the simulation and the real world (Wilde, 1998) .

Summary

The case for the use of VR in aviation training is compelling from an economic and efficiency standpoint and must now be validated from a technical standpoint to ensure that the current level of safety inherent in the current licensing and training system is not compromised, and also to comply with the applicable FAA regulations that govern the implementation of a new training method. This study isolated a subset of aviation workers and used an existing technology and platform to perform a direct comparison of learning that occurred in a VR environment, as compared to learning that occurred under existing FFS technology. The results of this study allowed conclusions to be drawn regarding the relative effectiveness of VR delivered training, and its potential to supplement or replace already existing training methods.

Chapter II of the current study examines the methods of teaching and learning that are currently applied in modern aviation training courses, and discusses the theories and methodologies used to evaluate students and assess their learning. The extensions of these methods from the physical to the virtual environment are then discussed, as are current methods of evaluating learning that have taken place in a virtual environment.

Chapter III combines the understanding of evaluating learning developed in Chapter II, with the methods of evaluating learning that have taken place in a VR environment and outlines an experimental method whereby students were evaluated on learning that has taken place either in a VR or a physical setting. This method provides a direct comparison between the control and VR trained students, thus meeting this study's stated aim of evaluating the relative effectiveness of VR training in an aviation environment. Results will be presented in Chapter IV and conclusions and discussion in Chapter V.

List of Acronyms

AMT	Aviation Maintenance Technician
AMTS	Aviation Maintenance Technician School
A&P	Airframe and Powerplant certificated technician
AR	Augmented Reality

CBT	Competency Based Training (CBT has also been used to refer to “Computer Based Training” in other publications. In this dissertation CBT will always refer to Competency Based Training with any reference to computer-based training fully written out.)
CFR	Code of Federal Regulations (USA)
EBT	Evidence Based Training
EASA	European Aviation Safety Agency
ECAM	Electronic Central Aircraft Monitoring
ELOS	Equivalent Level Of Safety
ERAU	Embry-Riddle Aeronautical University
EXP	Experienced
FAA	Federal Aviation Administration
FFS	Full Flight Simulator
FTD	Flight Training Device
GEARS	Global Evaluative Assessment of Robotic Skills
HID	Human Interface Device
ICAO	International Civil Aviation Organization
IRB	Institutional Review Board
MRO	Maintenance and Repair Organization
PTT	Part-Task Trainer
SPSS™	Statistical Package for Social Science (IBM)
TAM	Technology Acceptance Model

VR	Virtual Reality

Chapter II: Review of the Relevant Literature

The use of VR trainers in the field of aviation is a new and emerging domain. A survey of the literature on their use in aviation, standards of application for VR systems, and evaluation of their results over the last thirty years, turns up a surprisingly limited list (Renganayagalu et al., 2021). Fussell (2020) found the same issue when conducting a study of student's intentions to use VR for flight training: little attention has been paid to the subject using objective measures that focus on the ability of this technology to support specific training outcomes (Abich et al., 2021).

VR applications in aviation are either too new to have been extensively studied, or in those instances where they have been used the corporate or security environment precludes a public discussion of their effectiveness. Renganayagalu et al. (2021) found in their survey of VR studies that, "many training effectiveness studies reviewed lack experimental robustness due to limited study participants and questionable assessment methods" (p. 999). A search of other fields is therefore in order, and where VR training has been used or studied the methodologies employed should be analyzed and applied to the aviation training environment.

To do this we must first examine the state of the art in aviation training. Once we determine the current best practices for delivering and evaluating aviation training we can then compare and contrast how aviation training and evaluation compares to the training and evaluation conducted in other fields that currently use VR training environments. The comparison between aviation and similar fields allows for the use of established theories of learning and evaluation to produce a validated instrument for use in evaluating VR training in an aviation environment. Understanding how learning occurs, and how it is evaluated, allows for the proposal of an experimental design to properly evaluate the effectiveness of a VR training system in an aviation training environment.

Trends in Aviation Training

It is expected that VR will rapidly permeate the aviation training environment, with end-to-end virtualization becoming the norm for future training environments (Jenab et al., 2016). The increased use of VR can make sense from a cost and student engagement point of view, and also to leverage the advantages of VR from an andragogic perspective. To gain those advantages, we must first consider the learning models used for adults in the aviation environment and how those models can shape and influence how we implement and assess VR learning. The next step is then to examine how to best evaluate if a VR training solution is proving effective in the field, and if the training effect that is being observed is worth the cost and effort of its implementation.

Carl (2018) found that on many occasions, a lower-fidelity training approach, such as a VR headset as opposed to an FFS, could produce acceptable training results and be just as effective as higher fidelity simulators while having a wider reach across space and time. The actual training devices are critical to the delivery of the content, but the model of learning that is being used must also be well understood. In order to evaluate learning, and thereby training effectiveness, we must understand the model that is being used to deliver the material, and how its results can be assessed (Jerald, 2016).

Competency Based Training in Aviation

Aviation training at both the professional flight school and the airline level has moved towards a Competency Based Training (CBT) model. CBT has its roots in behaviorism, as represented in the works of experimental psychologists like Watson, Pavlov, Thorndike, and Skinner, whose legacy has led to a focus on observable behaviors (Morcke et al., 2013). Behaviorism focuses on observed behavior as an indicator of learning, and CBT seeks to use observed skills and behaviors to both teach and assess competency (Hodge, 2007). Rutherford (2009) contends that competency is measured by assessing the successful application of learned skills and behaviors towards the satisfactory completion of a task. Systems theory is also an important component of CBT, as it is necessary to construct a learning system that breaks down each component skill that is necessary to complete a given task, and then assign desired behaviors to it that can be quantified, taught, and measured (Hodge, 2007).

Consider the task of landing an aircraft. This task could be deconstructed into behaviors such as maintaining a set descent rate (e.g., the approach glide path), maintaining tracking (e.g., keeping the aircraft centered on the runway), and flaring the aircraft at the correct place and rate. These three behaviors can be taught and then combined into a landing task for a student pilot. Competency is then judged by observing the correct application of those behaviors, and if a hard landing occurred it could be traced to a misapplication of an observed behavior (e.g., a hard landing could be due to a late flare, which is linked to the 'flare the aircraft' behavior) (Franks et al., 2014). A student pilot would be considered to be competent at the landing task when all three behaviors were correctly applied, and an acceptable landing was observed when the behaviors were combined.

A criticism of behaviorism is that it teaches specific behaviors that, when correctly emulated, result in competency being assessed; there is no inherent requirement for assessing understanding (Dakers, 2005). Behaviorism contrasts with constructionism, which was identified by Ibáñez and Delgado-Kloos (2018) as a model that utilizes a scientific discovery, or constructionist strategy to engage learners. It requires learners to construct knowledge for themselves, rather than being instructed, which is felt by its proponents to stimulate critical thinking and support overarching metacognitive skills (Black & McClintock, 2017). These different approaches to training require different means of evaluation, making it critical to understand the model of teaching and learning being applied prior to attempting to evaluate training effectiveness (Dakers, 2005).

Beyond the ICAO (2013) commitment to use Evidence Based Training (EBT) and CBT as vehicles to improve pilot training and reduce accident rates, the adoption of CBT is seen by some educators as an enhancement of the classroom experience which makes it more productive. Franks et al. (2014) note that:

> a move towards competency-based training (CBT) – sometimes referred to as evidence-based training (EBT) – is a step towards creating more valuable classroom hours. With CBT, the knowledge, skills and attitudes required for professional competence are identified and organised [sic] into a series of 'competency statements,' which themselves become training objectives to be measured against. (p.137)

The theory is that students have well defined goals, are told what they need to learn to meet those goals, are presented the material required, and are then assessed against them. The CBT model is a pragmatic real-world approach, which focuses on the desired behavior and the skills required to specifically accomplish identified goals. It recognizes that while segregating subjects into different topics may be easier to teach, students eventually end up having to learn how to use all their different skills together.

Instead of segregating tasks CBT advocates ‘whole task training’ (Competency-based training: The future of the aviation industry?, 2017). This is clearly applicable to technical pilot tasks such as navigation, systems operation, and trajectory control, as well as to maintenance actions that use multiple aircraft systems. A big advantage of assessing a CBT course is that one can focus on observable behaviors and make easily quantifiable observations. This results in a de-coupling of the instructional vehicle from an assessment of its effectiveness. Provided that the delivery of the content is based on CBT principles, and that the evaluated competency can be clearly observed and judged, the system will be agnostic to the means of delivery or instruction of the content. This fact makes it well suited to an experimental study as the normal evaluation of skills taught in a CBT environment is through observation, and those same observations can be quantified and tested against various hypotheses (Marzano et al., 2017, p.2).

The FAA advocates the use of Bloom’s Taxonomy to describe and evaluate different stages of learning throughout flight instruction (FAA, 2020). Bloom’s taxonomy classifies educational goals into three separate and individually observable domains: affective, psychomotor, and cognitive. Rupasinghe et al. (2010) describe these as follows:

> The affective domain describes how people react to educational objectives which induce awareness, attitudes, emotions, and feeling. The psychomotor domain identifies and categorizes objectives in terms of physical manipulation of tools or instruments. The cognitive domain has been widely researched in many disciplines and the hierarchy of sub levels includes knowledge, comprehension, application, analysis, synthesis and evaluation. (p. 3)

Rupasinghe et al. (2010) studied the use of Bloom's taxonomy in the development and delivery of courses to A&P technicians, thereby extending its use beyond flying courses and into the realm of aviation maintenance. Their positive results indicate that it is equally applicable to technical training as well as flying, and that its use in developing and evaluating curricula in both areas is strongly supported, especially in the cognitive and psychomotor domains (Rupasinghe et al., 2010). The three areas of Bloom's Taxonomy, however, are not equally suited to observation in the evaluation of a CBT skill. Psychomotor, and elements of the cognitive domain can be readily assessed through observation, while the affective domain requires different methods or longer-term observations to evaluate. The answers to questions surrounding the affective domain would need to come from the students themselves, suggesting that a survey or interview method would be more appropriate to studying this domain (Vogt et al., 2012).

CBT is the current state of the art in aviation training, but it is not without its critics (Kearns et al., 2016, p.1). Dakers (2005) sees a clear need to move beyond CBT by leveraging newly developed technological opportunities. His paper, "The Hegemonic Behaviorist Cycle," states this clearly:

> A learning environment that will enable the technologist of the future to shape our world for the better will be one in which risk taking and creativity is encouraged. This can occur, I would argue, only when the current authoritarian transmission model of instruction has been replaced with one in which the formation of values relating to the technologically mediated world we inhabit, is allowed to occur in a manner which encourages debate as opposed to one in which values are imposed. (p.124)

As our tools for instruction evolve and change, so also should our models and methods of instruction. Brady et al. (2001) have identified that aviation students employ different learning styles from traditional adult or child learners; they tend to combine both pedagogic and andragogic styles of learning in a manner unique to aviation students, while tending towards a more andragogic style. Current emerging technologies give us an excellent opportunity to address the points made by both Dakers (2005) and Brady et al. (2001) by specifically adapting content delivery to the learning styles most often employed by students in the aviation field. What remains to be seen is whether or not the results obtained using these methods differ appreciably from those previously obtained. Before this evolution can take place, the potential of the tools that could enable it must also be assessed, and their usage must evolve from the laboratory into the mainstream.

The Use of VR in Education

The use of both VR and AR in education has evolved slowly over the last decade, and a series of studies have been conducted across various fields to assess their usefulness as teaching aids and platforms. The meta-analyses conducted by Renganayagalu et al. (2021), Wang et al. (2018) and Ibáñez and Delgado-Kloos (2018) listed many benefits associated with the use of AR/VR in education and training. Benefits were found in the areas of learning outcomes, pedagogical contributions and interactions, and visualization of abstract concepts. VR has the capability to detach advanced and complicated learning from its associated laboratories and facilities, but Wang et al. (2018) note that it is still in its infancy; more work remains to be done to determine the true effects of using virtual or semi-virtual environments for the learning of skills that require a specific manual or psychomotor component. Studying the effectiveness of VR training in a recurrent training environment has the advantage of using subjects who have already demonstrated mastery of the manual skills required to perform the required tasks; this allows researchers to focus on the effectiveness of the training medium in transferring knowledge, independently from the acquisition of manual skills by the trainees. Such a study performed with AMTs, training on flight deck systems, is an important first step in determining VR training effectiveness. The results have the potential to move us to fundamentally rethink current learning practices (Carl, 2018).

Assessing VR Learning Environments

D.L. Kirkpatrick developed a four-level model to evaluate training in 1959. It has been used extensively in the training industry ever since and was updated in 2016 by W.K. Kirkpatrick and J.D. Kirkpatrick. The model provides a framework for measuring training effectiveness by looking at four key levels of the training experience (Kirkpatrick & Kirkpatrick, 2016, Ch. 2, Table 2-1):

1. Reactions: the degree to which the participants find the training favorable
2. Learning: the degree to which the participants acquire the intended knowledge, skills attitude, confidence, and commitment
3. Behavior: the degree to which participants apply what they learned during training when they are back on the job
4. Results: the degree to which targeted outcomes occur as a result of the training and the support and accountability package

To assess the comparative effectiveness of VR delivered recurrent training it is necessary to observe, in some measure, each of these dimensions and compare virtual training to training delivered in the physical environment. This requires a drawn-out and comprehensive study to complete an assessment across all four points, as subjects would need to be baselined and monitored over time to determine the long-term effects of a switch to a VR platform.

Fussell (2020) addressed point one, student's reactions to the training, through the application of the Technology Acceptance Model (TAM) to measure student's intentions to accept and use VR platforms. Extension of this research would naturally address point two, to assess the actual learning that takes place when a VR platform is used. Valimont et al. (2007) note that the most efficient transfer of knowledge and training occurs whenever the similarity between the training, the training environment, the task, and the task environment is maximized. This supports the use of the FFS, which is considered to be the equivalent of the aircraft due to its stringent certification process. The use of a CBT syllabus, combined with either the FFS or VR training environment, allows for a direct assessment of point two above, while point one could be measured subsequently through post-training surveys. An assessment of Kirkpatrick's levels three and four would require a long-term evaluation of a program and are recommended for future research (Kirkpatrick & Kirkpatrick, 2016).

Measurement of Students' Success

In order to accurately assess real world performance and training results, it is insufficient to use univariate measures of success. Results can be best measured as a combination of both time and errors, as demonstrated by Chittaro et al. (2018) and supported by Ahmedyanova (2017). This addresses the main measures of success in the FFS environment, which include both accuracy and timeliness of actions; it also supports evaluation of the results level of Kirkpatrick's 2016 framework, which looks beyond the learning and behavior dimensions which can be measured readily with a simple scoring metric. Development of a multivariate assessment model for VR effectiveness is critical to continued research in this domain as, “Current standards of research have failed to provide a standard of measure against which virtual reality flight training can be compared” (Whitson, 2019, p.2). This multivariate approach, which includes time as a measure of success, is appropriate in fields where the professional technical competence of the subjects strongly predisposes them to both correct and timely completion of a task, such as in the case of licensed aviators and mechanics, or of medical personnel. In addition to surgical technique, time to task completion is a critical measure of surgical competence as it accounts for the time window in which a patient is exposed to the risk of being under anesthetic, surgical discomfort, and exposure to infection (Hoogenes et al., 2018, p.112). In aviation maintenance mechanical competence is vital, and the time element relates directly to the economics of the operation; an A&P who can successfully complete three tasks in the time that it takes another to complete two is worth more to an airline. As such, there is a continuous push towards efficiency in A&P training, and the first order measurement of this in training is the time to task completion.

Assessment Tools for VR Platforms

The medical field has struggled with similar training issues as aviation and has attempted to advance the study of the effectiveness of VR training through a number of experiments using VR training devices to help surgeons manipulate robots to perform laparoscopic procedures. Chen et al. (2019) performed a meta-analysis of all current methods of objectively assessing robotic surgical technique and concluded that, “No universally accepted robotic skills assessment currently exists” (p. 461). They found that assessment techniques generally fall into two broad categories: those that use automatic means of evaluation, provided through the training devices themselves; or those that rely on manual assessment and use some form of structured evaluation (Chen et al., 2019). Hoogenes et al. (2018) used a validated manual assessment tool, Global Evaluative Assessment of Robotic Skills (GEARS), to conduct a randomized comparison of two robotic VR simulators, and an evaluation of the trainees' skills transfer to a simulated robotic urethrovesical anastomosis task. Figure 4 shows a representative GEARS tool, developed for use in assessing a robotic surgical procedure. Hoogenes et al.'s (2018) experiment involved 39 medically qualified participants who underwent a VR based training session on one of two different VR training devices and were then asked to perform a simulated surgical procedure. The GEARS evaluation assessed trainee performance on a technical level, and trainees were also measured on their task time to completion. The GEARS rating tool, "consists of a 5-point anchored Likert scale across 6 domains that deconstruct the fundamental elements of robotic surgical procedures" (Hoogenes et al., 2018, p. 112). A direct comparison of surgical performance and task time to completion between the two differently trained groups was then possible, and conclusions were drawn about the relative effectiveness of the VR training devices.

Figure 4

GEARS Scale Adapted for Robotic Surgery

Depth Perception

1	2	3	4	5
Constantly exceeds the target, large movements, fixes slowly.		Some failures in making the goal but corrected quickly.		Directs the instruments in the correct plane to the target.

Bimanual Skill

1	2	3	4	5
Uses only one hand, ignores the non-dominant hand, poor coordination between the two.		Use both hands, but the interaction between them is not optimal.		Use both hands in a complementary manner for optimal exposure,

Efficiency

1	2	3	4	5
Many tentative movements, frequent changes in the thing to do, no progress.		Slow movements, but reasonable and organized.		Confident, efficient, remains focused on the goal.

Force Control

1	2	3	4	5
Jerking, tearing the tissue, damage to structures. Frequent breaking of the suture.		Reasonable handling of tissues, less damage occurs. Occasional rupture of the suture.		Proper handling of tissues, proper traction thereof. Without breaking the suture.

Autonomy

1	2	3	4	5
Unable to complete the procedure.		The individual is able to complete the task safely, with some guidance tutor.		Able to complete the task alone, without a guide.

Robot control

1	2	3	4	5
Does not optimize the position of the hands on the console, frequent collision. Vision is not optimal.		Occasional collision of hands. Vision is sometimes not optimum.		Adequate control of the camera. Optimal hand position without collision.

Note. Adapted from Robotic surgery training: construct validity of Global Evaluative Assessment of Robotic Skills (GEARS) (p. 229), by Sánchez et.al, 2016.

Schulz et al. (2018) conducted a non-randomized evaluation of the use of VR training for surgical procedures. It used self-evaluation questionnaires that covered both the technical aspects of the surgery, and the speed with which the surgeons were able to perform the tasks. While this study centered more on the surgeon's confidence to perform these procedures, it also found strong evidence of the effectiveness of VR training in preparing surgeons for specific tasks (Schulz et al., 2018). Neumann et al. (2019) conducted a variation of the two previous studies, and assessed the difference in effectiveness between a group of medical students who underwent VR training on a specific procedure, and another group who viewed a traditional video tutorial by an expert surgeon. Their evaluation included procedure time, and a series of technical elements, which were recorded automatically by the simulator (p. 909). Schmidt et al. (2018) also studied the evaluation of laparoscopic VR training and concluded that objective feedback, in the form of single parameters, made overall evaluation of trainees' performance difficult, and that an expert-based composite scoring system such as GEARS was needed.

Strong parallels exist between the teaching and employment of robotics in the field of medicine and the teaching and use of highly automated systems in the latest generation of airliners. Dr. K. Abbott, an FAA researcher in the field of Artificial Intelligence and Complex Systems, likened the modern airliner to a flying robot (2017). Robotic surgery and modern airline operations use highly skilled and knowledgeable experts in their respective tasks (as evidenced by their qualifications and certificates), each requires the manipulation of complex automatic systems through specific user inputs, and both provide an output that is influenced by the specific user input received.

Mindell (2015) reminds us that it is the human who remains at the core of the process, and that is why the training and education that they receive must be scrutinized, understood, and assessed; it is through learning to interact with a robotic interface that these professionals learn how to do their jobs. The elements of the evaluation of VR training environments in the robotic medical field can be easily adapted to the aviation environment and can serve as a valuable tool with which to assess the effectiveness of VR training environments in aviation.

Validity of the GEARS Construct

Sánchez et al. (2016) performed a cross sectional study to directly assess the construct validity of the GEARS tool in differentiating between varying skill levels of subjects performing robotic surgical procedures. In addition to the GEARS tool, time to complete a procedure was also used as a discriminator between groups assessed as having the following degrees of experience in robotic surgery: expert, intermediate, and novice (p. 228). They found that the GEARS tool had a high reliability, with an inter-observer coefficient of $r = 0.96$; all fields of the tool were found to provide excellent discrimination between the groups with the exception of "depth perception," which was found to be equal between all groups by virtue of the outstanding qualities of the robotic system being used (p. 231). This suggests that GEARS is an appropriate tool to adapt to the evaluation of A&P procedures learned on a VR system; however, care must be taken in choosing the fields that are being evaluated lest the quality of the systems being used and tested compensate for a student's deficiencies that may otherwise be present.

Gaps in the Literature

Renganayagalu et al. (2021) reviewed studies on the effectiveness of VR training over the last 30 years and extracted a total of 60 studies to analyze. They found a total of 30 articles published in the period spanning 1988-2013, with a further 30 articles published 2013-2018. In their survey they noted a lack of experimental robustness in existing studies, small sample sizes, and questionable assessment methods. Abich et al. (2021) conducted a similar review and noted that most current research focuses on hardware and software development, and not on the ability of VR to deliver appropriate learning outcomes. While there are a number of VR systems in use in the private and military sector, a lack of accessible reporting on them seems to indicate a level of classification or proprietary information present that prevents assessment and study of their usefulness. Both Abich et al. (2021) and Renganayagalu et al. (2021) identify a need for further experimental study of VR learning outcomes, properly specified and controlled, with adequate sample sizes.

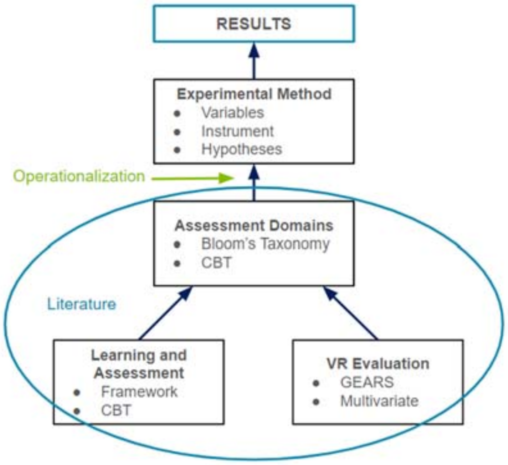
As discussed in Chapter I, the FAA sets a high bar for the certification of a system to be used in the training of aviation technicians, and these certification requirements present a barrier to the introduction of VR training in the commercial sector that is yet to be overcome. Adapting and using the GEARS tool for the evaluation of a VR system for training A&P technicians accomplishes two things: firstly, it provides direct comparative data to assess the suitability of a VR tool to deliver recurrent training; and secondly, it delivers a tool adapted for the aviation training environment that can be used in future applications to streamline the certification of other VR platforms in this domain. This study effectively provides a validation of an existing tool in a controlled experimental setting, while also providing a method that could be used with other VR systems to facilitate their certification.

Theoretical Foundation

The theoretical foundation for this study is based on the use of Bloom's Taxonomy to modify an existing GEARS tool for use in measuring training effectiveness in an aviation environment. Figure 5 shows how Kirkpatrick's model, CBT theory, and established VR assessment methods were combined with Bloom's Taxonomy to adapt the GEARS tool for use in an aviation environment. The use of the GEARS tool in a CBT learning environment has been validated in the field of urology and robotic surgery (Sánchez et al., 2016), and the use of Bloom's Taxonomy to adapt it to aviation use ensures that the items measured align with the theoretical basis for currently delivered aviation training courses approved by the FAA (FAA, 2020; Jensen & Konradsen, 2018).

Figure 5

Theoretical Framework and Research Model



Note. Copyright 2024 by M. E. McCullins

The current study then operationalized this instrument and associated variables to construct a pure experimental method to assess and study the differences between two groups: one that was trained to complete a task using a VR system, and one that was trained using current methods. The VR trained group was then divided into two sub-groups based on whether or not the subject had prior experience with VR systems. The combination of a CBT training syllabus, and the GEARS evaluation tool, provided the theoretical foundation for the selection of dependent variables that can be observed per CBT principles, and evaluated using the GEARS scale with a high degree of validity. Both groups came from a common pool of A&P technicians undergoing recurrent training, and the random selection of the subjects for each group and blinding of the evaluators was possible thereby meeting the key criteria for an experiment (Vogt et al., 2012). The skill chosen for training and evaluation was drawn from a CBT syllabus, and the training results were directly observable by the evaluator; this link between education and observable behavior in a CBT environment was crucial to ensure the integrity of the experimental method. The modified GEARS scale contained both cognitive and psychomotor elements, as was appropriate for the assessment of a robotic interaction, and time to task completion was measured to allow for a multivariate analysis of differences between groups.

Research Model

CBT methods focus on observable behavior (Hodge, 2007). This makes the use of a CBT syllabus well suited to an experimental method, where results are observed, quantified, and conclusions are drawn based solely on what is seen (Vogt et al., 2012). This study quantified differences in performance between two groups, who were trained on a common task drawn from a CBT syllabus but delivered using a different method (either VR or traditional delivery). The assessment of the subjects’ performance was measured using a modified GEARS tool that was adapted to the aviation environment, and on the time required for task completion. The high degree of validity observed in the use of GEARS gives excellent confidence that if there is any difference between the performance of the VR group and non-VR group it would be detected given an appropriate selection of effect and sample size. The research model for this study combined CBT, which focuses on observable behavior, with GEARS, a tool that has a high degree of construct validity and a proven history of quantifying observed behavior in high-tech environments. GEARS has been used to assess VR learning in medicine for the last 8 years with an excellent track record as reported in the literature and is an appropriate tool to extend into other areas that use a combination of CBT and VR learning.

Hypotheses and Support

The main hypotheses presented in Chapter I were:

***HA*$_0$**

There is no collective statistically significant difference in Test Performance between the groups (VR trained and FFS) when controlled for A&P and engine-run experience.

***HA*$_a$**

There is a collective statistically significant difference in Test Performance between the groups when controlled for A&P and engine-run experience.

The literature review for this study has demonstrated that it is appropriate to use an experimental method to observe the quantitative performance differences between these groups based on the CBT syllabus that is used to instruct them. The tool that will be used for assessment is drawn from a series of studies that assessed the effectiveness of VR delivered training in an interactive environment that required both cognitive and psychomotor skills; this is exactly what is required in interacting with a modern transport–category aircraft. Should the results show that the performance of the VR group is similar to that of the group taught using a traditional method, then a path to the certification of that system and its use in aviation training can now be defined through establishment of an equivalent level of safety (ELOS). Regardless of the result, the use of the GEARS tool and multivariate data analysis provides the training industry with a pathway to assess the effectiveness of VR training in aviation and offers a clear roadmap to assessing the suitability of future systems.

Summary

CBT is a model of learning that is well suited to application in the aviation environment and allows for direct assessment of learning through direct observation of behaviors. A multivariate analysis of these behaviors is necessary to give a true picture of the effectiveness of a VR learning environment, and this measurement will help establish standards against which the success of VR training can be judged. There have been a number of studies in the medical field that have looked to validate the use of VR training for surgeons and medical students, and their methodologies can be readily adapted to the evaluation of VR training in aviation.

Chapter III discusses the operationalization of the GEARS tool for use in this study and demonstrates its application to a representative A&P task to be performed by both groups in an FFS following a training session. The results from GEARS and task time to completion are analyzed using a multivariate means (MANCOVA) in Chapter IV and reported alongside full demographic data for the sample to either validate or reject the null hypothesis. These results frame the discussion in Chapter V surrounding the effectiveness of the VR system used to deliver the training, as well as the overall suitability of the research model that combines elements of CBT and GEARS to evaluate the use of a VR system.

Chapter III: Methodology

To explore the effectiveness of a VR training system relative to an existing means of training, a direct comparison of the results of that training is desirable. As outlined in Chapter II, this is possible by focusing on observed behaviors related to the training delivered, and by comparing those behaviors between groups that were trained using a VR system against groups that were provided equivalent training in a traditional environment. This chapter outlines the experiment that allowed for that comparison to take place and discusses the sampling process for the evaluation.

Research Method Selection

A quantitative study of training effectiveness is necessary in order to produce the required data to satisfy the regulatory requirement for an ELOS. This data must be collected in the field from actual certificated operators and then compared with existing training systems. This data is best collected by using an experimental method that allows for direct comparison between groups. In order to assess if VR delivered engine-run recurrent training, delivered to already qualified and certificated FAA Maintenance Technicians, produces equivalent training results to a course delivered in an FFS, a between-groups experimental study was performed to directly compare training results between FFS and VR trainees. This was a quantitative study that used a pure experimental method to directly compare observed results between the experimental (VR) and control (FFS) groups. The VR group was then divided into two sub-groups, one that had no prior experience with VR systems and one that had previously used a VR device. Prior exposure was determined from the participant information form (see

Appendix C) and was considered to be any previous use of a VR system or VR technology reported by the subjects.

Data Collection Process

The study was conducted using a 3 x 4 experimental design with three between-subject group independent variables (i.e., training method: FFS or VR, with the VR group being subdivided into groups with prior exposure to VR systems or those with none) and four measures of learning effectiveness (overall GEARS score, cognitive GEARS score, psychomotor GEARS score, and time to completion). This yielded three total groups for comparison across four measures each. No pre-test is permissible as this would have provided additional refresher training beyond what a student would normally get and would therefore have rendered the results of the evaluation non-representative of a normal course (Vogt et al., 2012).

Following the prescribed training, assessment was conducted in an FAA Certified Level D FFS, which is considered as equivalent to an aircraft for training and licensing purposes. Students were quantitatively assessed during an engine-run scenario in two areas: procedural accuracy using the GEARS tool, and time to completion. In order to account for varying levels of experience of the technicians within the sample, these main groups were controlled based on the years of experience as a certificated A&P technician (Experience AandP) and the number of years that they have been performing engine runs (Experience ERun), reported in the training entry information forms that are required for all students undergoing training. This is a quantitative study that draws conclusions based on statistical differences in training results measured between experimental groups that underwent training using a VR training system, and a control group that was trained solely in the FFS using an existing training system. This study was performed at the Airbus Miami Training Center using an Airbus developed VR platform as shown in Figure 3; this platform is based on a commercial Oculus Rift VR headset, hand controllers, the Microsoft XR platform, and Airbus-developed VR training software as described in Appendix G.

Design and Procedures

Students for the engine-run requalification course were trained in groups of three, per an FAA approved syllabus. Once students were checked into the training facility, they underwent a block of classroom training to cover all required topics for the engine-run requalification. Following this training, trainees were briefed on the nature of this study and offered a chance to participate. Those who chose to participate were then guided through the informed consent process, and each trainee was randomly assigned to either the VR or FFS group. Students assigned to the VR group completed a recurrent training sequence delivered by the VR platform and were then evaluated in the FFS. These students then completed the traditional recurrent FFS training in order to ensure that they met the FAA requirements for annual requalification and had not been disadvantaged by participating in the VR training study. The FFS students completed the regular recurrent training program in the FFS, and then underwent the same evaluation.

The training schedule was arranged so that both the VR and FFS students completed their evaluations having received approximately four total hours of training, including the activities listed in Table 3, thus ensuring that fatigue was equivalent for each student and was also representative of what could happen with a normal training schedule. As the experiment was embedded in normal training course activities this timeline could vary slightly, but the classroom portion of the training was always completed prior to the FFS evaluation session and was generally programmed to take up a morning, while the FFS session was scheduled in the afternoon.

Table 3

Sample Experimental Timeline

Activity	**Time (Minutes)**
Presentation	5
Questionnaire	5
VR or FFS Introduction	10
VR or FFS Training	15
Break/Lunch	30
FFS Evaluation	15
Total time	80

Table 3 outlines the schedule for a subject undergoing training in VR and evaluation in the FFS; the schedule for traditional FFS training and evaluation would be similar but might have varied occasionally due to FFS availability and course programming. The schedule was designed to ensure that all subjects received an equivalent break between the training and evaluation activities.

Evaluation Scenario

The scenario that was delivered to subjects in both traditional and VR format, and was evaluated in the FFS, was a normal engine start procedure in the A320 aircraft that terminated with the engine exceeding engine start temperature limitations (a Start Valve that failed open). This was a moderately complex scenario that required students to follow established procedures, manipulate both checklists and aircraft systems, interpret information that was given by the aircraft, and act correctly in accordance with trained procedures when the scenario did not progress as expected. The length and complexity of the scenario was appropriate for evaluation with the GEARS tool and gave the evaluators multiple opportunities to observe each dimension of behavior that was evaluated by the tool.

The procedure for both a normal engine start and a failed start valve are found in Appendix D, and the instructional package guided the subjects through the procedures that they must follow to complete this scenario. The VR system simulated the action of the second crewmember during the training to ensure that each student received an equivalent level of prompting and assistance throughout the training and intervened and corrected the student if they were not progressing as required, as would be expected from a human instructor. This scenario combined elements of both the cognitive and psychomotor domains wherein the subjects had to manipulate the aircraft controls and switches correctly, as well as apply procedures, interpret instruments and readings, and decide on a correct course of action. It included observed procedural knowledge, and problem-solving competencies that certificated A&P technicians who are engine-run qualified would already have mastered, but due to the malfunction presented would not be routinely practicing (hence its inclusion in the refresher training syllabus). A list of training competencies, along with those competencies specifically observed in this scenario, are listed in Appendix D.

Apparatus and Materials

In the instructional phase of the experiment the VR trainer, pictured in Figure 3 and described in Appendix G, was used to deliver training on the scenario that would later be evaluated in the FFS. Training on the operation of the VR trainer was given, followed by a scripted training scenario that was to be delivered by the VR trainer. Students in the control group received the regular classroom instruction on the scenario, and then were then given the opportunity to practice it in the FFS as is currently done prior to the evaluation. Details of the VR training system can be found in Appendix G, and an example certification of the FFS that was used is in Appendix H.

Sources of the Data

All experimental data was gathered from an evaluation in the FFS, using the modified GEARS tool in Appendix F. The evaluation was a timed event, with a timer running from the evaluator's clearance to begin until the termination of the Hot Start procedure. The GEARS scale in Appendix F has been modified from that shown in Figure 4 to provide aviation related guidance to the evaluators but retains the cognitive and psychomotor elements of the original as well as the rating guidance elements. Evaluators were selected from a group of experienced and standardized instructors from the Airbus training center, and as a part of the pilot study were specifically trained on the use of the experimental instrument through a standardization process. All evaluators held A&P Certificates and were both qualified instructors and evaluators under the Training

Center’s Quality Manual which is FAR Part 14 CFR § 147 compliant. They all had received specific training and qualification in the use of CBT methods, assessments, and evaluations as part of their normal employment as instructors.

Pilot Study

A pilot study is desirable whenever possible when using an experimental design in order to validate the test instrument being used in the study, and to allow the entire experimental procedure to be assessed prior to its use for data gathering (Vogt et al., 2012). A pilot study was run using a sample size of 10 technicians and pilots in order to ensure that the test protocols were effective, and that the data being gathered was appropriate. It also provided the opportunity to train a small group of evaluators on the assessment instrument and to ensure that the group was using a uniform standard. These evaluators were experienced A&P engine-run instructors and assessors who were trained and certified in administering CBT courses and evaluations. There was no contact between the pilot study group and the main sampled group that could influence the results; the results of the pilot study were held separately from those of the main experiment.

The pilot study followed the form of the main experiment in that FAA certificated A&P technicians were selected to undergo a VR training session similar to the one proposed in the main study and were then evaluated on the performance of the task during the routine performance of their duties. This study took place at an Airbus training center where there was a population of certificated A&P technicians who currently perform engine-run duties and who could be observed in the field during the course of their normal duties for the purposes of testing the data gathering instrument. This served to train the experimenters on the effective use of the VR trainer, uncover any issues with its use, and allowed further refinement of the documentation that was used to deliver the training content and to report the results of the study. While the training device was designed to be intuitive and easy to use, it was essential that the evaluators were very familiar with it to ensure its correct usage.

Following the VR training session, the technicians were evaluated using the assessment instrument during performance of engine-run duties in the FFS. The main purpose of these evaluation sessions was to further refine and critique the assessment instrument, and to develop protocols to standardize its use. At the conclusion of the pilot study there was a trained cadre of three experimenters who were competent in using the VR trainer to deliver the prescribed training session, and who could subsequently evaluate the subjects in the FFS using the assessment tool.

Population/Sample

The population for this study was comprised solely of FAA certificated A&P technicians who hold an engine-run qualification on Airbus aircraft. There are currently in excess of 2,000 Airbus aircraft operating in North America, and this number is expected to exceed 5,900 in the next twenty years (Airbus, 2020). Each aircraft requires multiple certificated technicians to keep it maintained and flying in airline service, and in turn this population requires annual recurrency training that can be delivered at any appropriately certificated training facility using an FAA approved syllabus. Many airlines will contract out this training and certificated facilities provide pricing that is highly dependent on other business that is being conducted at the training facility, overall volume, and the operator’s relationship with the training provider. Operators will often change training providers based on pricing, availability, and convenience factors (Curnow, 2021). This inter-changeability bears testament to the uniformity of training given by the various training centers, which is due in a large part to the regulatory environment under which they operate. The sample for this study was drawn from certificated A&P technicians undergoing engine-run recurrent training at the training facility of a major aircraft manufacturer in the United States. These technicians were employees of both major airlines and third-line maintenance facilities who are required to perform engine-runs as part of their job function, and whose company has contracted for this training.

Sampling Frame and Size

Prior to commencing recurrent training, all participants were pre-screened by the training center to ensure that they held the appropriate certificate and had suitable recent experience to allow them to undertake engine-run refresher training under the regulations established by the FAA. The sample for this study was drawn from those students entering the Airbus Miami Training Center to undergo an FAA approved engine-run recurrent training course, and who were deemed qualified to undergo recurrent training. This screening ensured a uniform level of knowledge and qualification between all subjects, leaving only relative experience in the engine-run role as the prime differentiator. This study was run over a fourteen-month period and all students who entered the training center during this time were offered the opportunity to participate. A total of 100 students were assessed over this fourteen-month period.

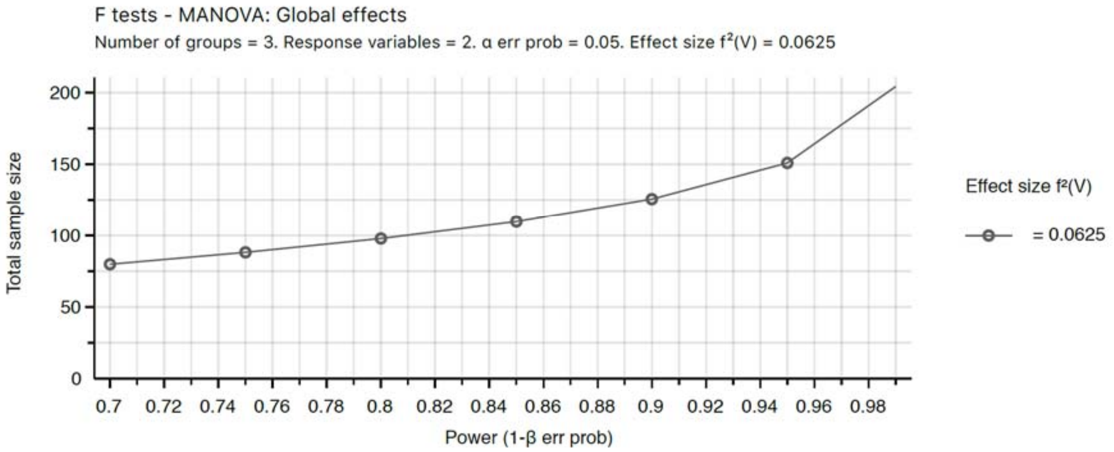
To determine the required sample size, a power analysis was performed using G*Power statistical software. Using a medium effect size of 0.25 for MANCOVA

(Cohen, 1988), and a confidence interval of 0.95, it was determined that the minimum sample size needed was 99 technicians, with random distribution between the experimental groups (Morris, 2007). Figure 6 shows the G*Power plot of Sample size vs Power used to determine this number (Faul et al., 2009). Training sessions were sampled continuously until the desired number of participants was reached, yielding approximately 50 technicians in each group (VR and FFS). Group assignment was random from within each course, thus meeting the requirement for both random assignment and independent observations. The breakdown of subjects between the VR and VR (Exp) groups was assigned after the evaluation based on the responses received in the participant information sheets, and no attempt was made to influence the number of subjects in those groups. Prior exposure to VR was considered to be any previous use of a VR system or VR technology for any purpose that was reported by the subjects.

Oberhauser and Dreyer (2017) note that, “In aviation, flight crews, engineers, and technicians are highly trained professionals with expertise in their specific domains. Their performance, appraisal, and feedback during experiments are more profound” (p. 264). They go on to state that this leads to valuable feedback being received on the operational implications of a proposed solution, even with a small sample size (Oberhauser & Dreyer, 2017).

Figure 6

*G*Power Sample Size Plot*



Sampling Strategy

All students who entered the training center over a fourteen-month period were offered a chance to participate in the experiment, with selection into the experimental or control group assigned at random. Demographic data was collected on each group to include age, education level, overall A&P experience (Experience AandP), and engine-run experience (Experience ERun). As overall experience can contribute to an individual's performance, both length of time as an A&P and length of time performing engine-run tasks were recorded and assessed as potential covariates during data analysis.

Ethical Considerations

An Institutional Review Board (IRB) application is required by Embry-Riddle Aeronautical University, which operates as a research institution under United States law. This is mandatory for research involving human subjects, and consideration must be given to participant privacy, consent, and possible harm. IRB approval, without which no research or gathering of data is permitted, is found in Appendix A. The informed consent form provided to all participants in found in Appendix B. Participation in the research was voluntary and all participants were provided with appropriate informed consent documentation. Privacy was safeguarded by de-identifying all participant data in the experimental record and assigning unique participant IDs to each subject using the form found in Appendix C. In addition to measures taken by the researcher, the participant data is also protected by the Airbus Data Protection policy under both US and European Union law owing to the location of the research institution in the United States, and the fact that Airbus, as a European-based entity, is likewise bound by European Data Protection law.

The risk to participants in the FFS group was no different than they would otherwise experience while undergoing normal refresher training. There are two specific risks that were addressed for the VR group:

1. Physical Safety while using the VR Tool: As the VR tool was completely absorbing, participants were not aware of their real-world surroundings while using it. The area in which they worked had to be free from obstructions and objects that could pose a hazard. Additionally, participants had to be appropriately restrained to protect against loss of balance. This risk was addressed by providing a dedicated room for this study where the participants were unable to physically contact items in the real world through their full range of movement, and an appropriate chair was provided that was secured from movement, fully supported the participants in a normal posture, and had sides that would prevent a participant from overbalancing or falling. The VR tool itself incorporated a setup procedure

that was followed to ensure that the usage area was clear of any hazards or obstructions, and this tool was used whenever the surroundings were altered.

2. Simulator Sickness: The possibility of disorientation and nausea from the use of the VR tool had to be addressed (Wu, 2018). Participants were briefed on this risk, and protocols were put in place to discontinue the experiment if any discomfort or disorientation took place. Participants who felt any discomfort or motion sickness were briefed to remove the VR goggles, remain seated, and inform the researcher. The experiment would have been terminated at this point and normal procedures to deal with a student who felt unwell in the FFS would have been followed. The risk of disorientation was no different than with any of the commercial VR gaming systems that are currently being sold and used throughout the world.

All information for this study was anonymous; there was no requirement for participants to be identified by name in order to participate. Following the initial collection of demographic data relating to age, education, and relevant experience, each participant was assigned a numerical ID for the duration of the experiment. A list of the names associated with the experimental IDs was stored separately from any experiential data, along with the informed consent forms, and was not available to those conducting the study or collecting data; this data was stored and encrypted on a secure Airbus server in compliance with all US and European Data Compliance requirements, and the named data custodian was not involved in the research project. The assessment data and demographic data collected was only referenced to the participant ID and was stored on a separate server at a different facility.

In accordance with US and European data protection laws all study data is being kept on secure and encrypted servers owned and operated by Airbus. Informed consent forms and participant names will be stored for three years after study completion; following this period the electronic files will be deleted from primary storage and backups in accordance with Airbus procedures and in compliance with US and European data protection laws. This will be supervised by a named data custodian in accordance with Airbus policies and procedures. The data gathered during the study (demographic and assessment) will be deleted one year following acceptance of the manuscript for the granting of a degree or publishing following the same procedures.

Measurement Instrument

The primary experimental measurement instrument was the adapted GEARS scale found in Appendix F, alongside a measure of the time to task completion in minutes and seconds.

Variables and Scales

The adapted GEARS tool that was used for evaluating the scenario in the FFS can be found in Appendix F and is a 5-point Likert scale consisting of 6 rating items. It provides an interval scale with values ranging from 6 – 30; the cognitive and psychomotor sub-domains are interval scales ranging from range from 3-15. Time to task completion was also recorded on this form and was measured in minutes and seconds. Definitions of the variables on the adapted GEARS instrument are provided on the scales themselves to aid with ratings, and rating items have been sub-categorized as either cognitive or psychomotor. These categories were used to further create sub-groups for comparison. Definitions of each variable along with the scoring criteria are found on the GEARS scale itself, and are expanded in Appendix F.

Data Analysis Approach

The GEARS assessment in combination with the time to completion recorded for each subject allowed for a multivariate comparison of results between the control group and the experimental group. The GEARS scores were further separated into cognitive and psychomotor segments to assess for differences in performance in either of these domains. A MANCOVA analysis was performed using GEARS scores and Time and was then expanded to analyze differences between the variables shown in Table 4.

Table 4

Comparison Factors and Grouping – Variables for Analysis

	Test Performance (DV)			
Practical Training Group (IV)	**GEARS Score**	**Cognitive Score**	**Psychomotor Score**	**Time to Completion**
FFS	X_1	X_2	X_3	X_4
VR	X_5	X_6	X_7	X_8
VR (Exp)	X_9	X_{10}	X_{11}	X_{12}

Technician experience (Experience AandP and Experience ERun) was used as controlling variables, or covariates, and their degree of correlation with the IVs was assessed. A univariate ANCOVA analysis was also conducted on each dependent variable to individually examine the effect of the experimental manipulation.

Reliability Assessment Method

The pre-screening performed by the training center ensured that the subjects had sufficient knowledge to perform the engine-run task, and that they were certificated, competent, and able to understand all instructions and training delivered. This uniform entry standard, alongside a common academic preparation, and random selection between groups, ensured maximized validity of the between groups experimental structure. Reliability was addressed by using a scripted training flow, prompting and interventions from the VR tool, and a pool of specially trained evaluators who had undergone standardization training during the pilot study to conduct the assessment. Evaluator scores were tracked and compared to assess inter-rater reliability.

Validity Assessment Method

Sánchez et al. (2016) performed a cross sectional study to directly assess the construct validity of the GEARS tool in differentiating between varying skill levels of subjects performing robotic procedures. In addition to the GEARS tool, time to complete a procedure was also used as a discriminating factor between groups assessed as having the following degrees of experience in robotic surgery: expert, intermediate, and novice (p. 228). They found that the GEARS tool had a high reliability, with an inter-observer coefficient of $r = 0.96$; all fields of the tool were found to provide excellent discrimination between the groups with the exception of “depth perception,” which was found to be equal between all groups due to the outstanding qualities of the robotic system being used (p. 231). This suggests that GEARS is an appropriate tool to adapt for the evaluation of A&P procedures learned on a VR system; however, care was taken in choosing the fields that were evaluated lest the quality of the systems being used and tested compensate for a subject's deficiencies that might otherwise be present.

Data Analysis Process/Hypothesis Testing

Following completion of the experiment the total GEARS scores, individual GEARS categories and time to completion of the task were loaded into SPSS for analysis. The data was sorted by the Practical Training Received (VR or FFS), and descriptive statistics were generated to examine the suitability of all groups. Measures of central tendency, dispersion, and distribution were calculated for the aggregate GEARS Score, individual GEARS elements (Cognitive and Psychomotor), and Time variables. Outliers were examined as described in Chapter IV, and the dataset was checked for missing values or errors; the odds of any missing values were extremely low due to the nature of the experiment.

It was considered highly likely that the assessment scores would be strongly negatively skewed due to both the high passing grade and the tendency for students at this level to perform well. Accordingly, following the generation of descriptive statistics for each group, the data was transformed, where indicated, prior to running additional statistical tests. Transformations, such as logarithmic, inverse, and square root were tested as applicable and examined for any improvements in the data distributions (Hair et al., 2010). A MANCOVA test was then performed on the variables, as outlined in Table 4, using Experience AandP and Experience ERun as controlling variables. In the case that a difference had been found between the two groups (i.e., the null hypothesis was rejected), further post-hoc tests would have been performed to determine which group had the higher level of achievement.

Nine requirements concerning the data have to be met in order to ensure the validity of the MANCOVA test results (Grande, 2015a; Laerd Statistics, 2019). They were:

1. The dependent variables were measured at the interval or ratio level. This was the case in this study as the dependent variables of Test Score and Time Taken are both interval values.
2. The independent variable should be two or more categorical groups. Practical Training Received is the independent variable and is categorical.
3. Independence of observations was required. The design of this study ensured that each observation was completely independent; the groups themselves were independent.
4. Adequate sample size was required. A total of 100 participants were assessed which met the G*Power minimum calculated sample size of 49 per group. The experimental duration was extended from six to thirteen months to meet these numbers.
5. There should be no univariate or multivariate outliers. Following the generation of descriptive statistics the data was examined for outliers and those observations would have been dealt with. A Mahalanobis distance measure was generated using SPSS to check for multivariate outliers.
6. Multivariate normality was required. This was tested using the Shapiro-Wilk test of normality.

7. A requirement for a linear relationship between each pair of dependent variables for each group of the independent variable was tested in SPSS by generating Q-Q plots and examining them.
8. The requirement for homogeneity of variance-covariance matrices was tested in SPSS using Box's M test of equality of covariance.
9. There was no multicollinearity. This was screened for using SPSS.

Requirements one through four were met through the research design, while requirements five through nine were tested using SPSS. A rigorous application of this methodology ensured valid MANCOVA results.

The SPSS MANCOVA test produced both a multivariate test result using a set of test statistics, and also separate univariate tests for each Dependent Variable (DV). The main purpose of using the MANCOVA test was to reduce the overall possibility of an inflated Type I error by combining multiple univariate tests and while the univariate tests gave us some idea of the effect of the Independent Variable (IV) on each DV, the MANCOVA test answered the empirical question surrounding the particular multivariate research question that was being examined (Warne, 2014). Warne demonstrates that when examining a multivariate problem using univariate tests, it is possible that certain combinations of univariate tests can cause the researcher to keep the null hypothesis when it should be rejected (p. 2).

The MANCOVA test in SPSS used four separate test statistics, which were converted to an F-statistic in order to calculate the p-value which was used to keep or reject the null hypothesis (Laerd Statistics, 2019); the criteria for this study was $p < .05$. If the null hypothesis would have been rejected then a post-hoc Descriptive Discriminant

Analysis (DDA) would have been performed using the discriminant tool in SPSS in order to calculate the functions that distinguished the IVs from each other using the DV scores, and also to avoid reliance on univariate tests for post-hoc analysis (Warne, 2014). These functions would show the relative effect of the IV on the collective DV, and thus allow conclusions to be drawn on the reasons for the differences between groups. The MANCOVA test was carried out to examine if the Experience AandP and Experience ERun variables substantively influenced the results.

Summary

This chapter presented a research methodology to assess the effectiveness of a VR device in delivering recurrent engine-run training to FAA certificated A&P technicians. The experimental design, sampling techniques, data gathering, and reduction were discussed, alongside the ethical considerations that would allow this study to proceed with the use of human subjects. This methodology produced high quality data using a well-established research tool, thereby allowing direct comparison of training performed in a virtual environment and training performed in a real-world environment. Chapter IV will present those results, which allow the conclusions drawn in Chapter V to be made on the relative effectiveness of the VR tool that was used.

Chapter IV: Results

Experimental data was collected over a fourteen-month period from FAA certificated A&P technicians undergoing engine-run recurrent training. Prior to the main experiment, a pilot study was conducted with a small group in order to assess the experimental procedure and to train the research assistants who performed the GEARS assessments. This chapter reports the findings from the pilot study and the main study, including sample demographics, experimental results, and hypothesis testing. It answers the research questions based on these results, and concludes with a chapter summary.

Pilot Study

A pilot study was conducted to validate the test instrument being used in the study, and to allow the entire experimental procedure to be assessed prior to its use for data gathering. It was conducted over a four-week period and included a total of 10 test subjects who were qualified to perform the procedure being evaluated. As the purpose of this study was to examine the procedure and the GEARS instrument, and to train research assistants in the use of the GEARS tool, all pilot study subjects were trained on the VR trainer and then evaluated in the FFS. The VR tool proved to be robust and easy to use, and the research assistants adapted quickly to the GEARS tool. As a result of their feedback, the scoring notes were added to the assessment tool (see Appendix F) to clarify when intermediate scores of *2* or *4* should be used; this wording was not added to the cells of the GEARS scale being used to avoid clutter. The overall experimental procedure was assessed to be practical and easily executable, and the supporting paperwork and tools were revised slightly for clarity and ease of use. No results are reported from the pilot study due to its small sample size and lack of control group.

Demographics Results

The participant information form found in Appendix C was used to gather demographic data on the participants in the study. All participants had been pre-screened by the training center to ensure that they held the appropriate licenses and ratings to allow them to participate in engine-run recurrent training, and therefore, this study. Participants were employed by major airlines, smaller air carriers (both passenger and cargo), and FAA certified Maintenance and Repair Organizations (MROs) that are based within the United States and Latin America and have regulatory oversight from the FAA. The total sample consisted of 100 participants who were split into three groups: those who underwent traditional FFS training, those who underwent VR training but had no prior experience with VR systems, and those who underwent VR training and had used VR systems previously. The demographic breakdown of participant age and both A&P and engine-run experience is shown in Table 5. Four out of 100 participants were female, comprising 4.0% of the sample, which is comparable to the total FAA population of A&P certificated mechanics, of which 2.7% are female (FAA, 2023).

Table 5

Demographic Characteristics of Participants

		Age		A&P Experience		Engine-Run Experience	
Group	Participants	Mean	*SD*	Mean	*SD*	Mean	*SD*
Control	52	40.65	11.89	14.62	10.80	8.71	7.94
VR	37	41.65	10.80	15.54	10.03	9.30	7.74
VR with Experience	11	34.73	11.31	13.09	9.82	9.45	9.91

Figure 7 shows the scatter of participant ages by experimental group. Ages ranged from 22 to 72 years old, with an mean age among all groups of 40.37 (*SD* = 11.50). An ANOVA between the three experimental groups showed that there was not a statistically significant difference in participant age, *F*(2,97) = 1.586, *p* = .210.

Figure 7

Participant Ages by Experimental Group

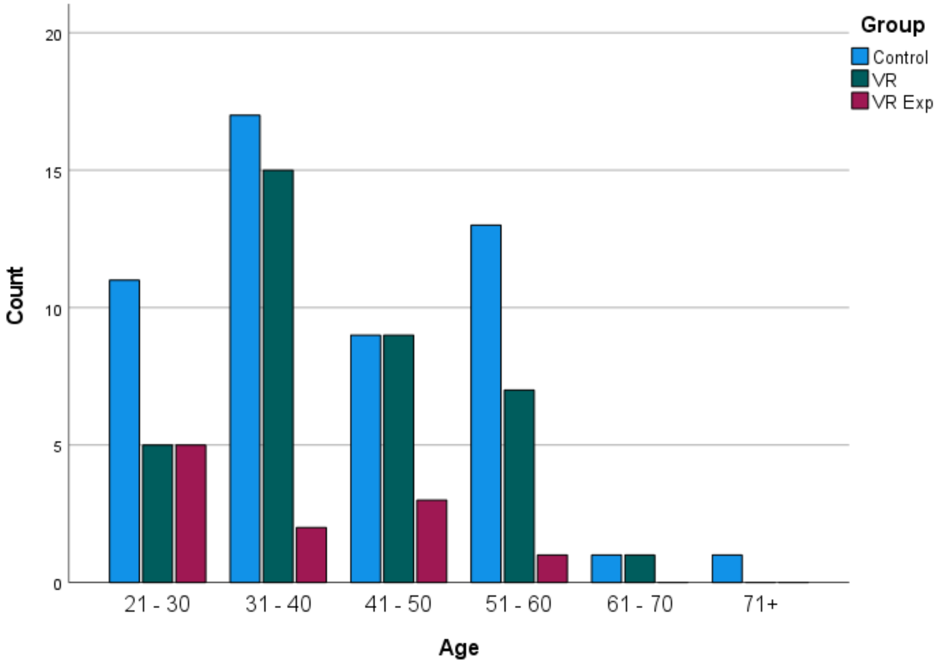


Figure 8 shows the distribution of A&P experience by experimental group. A&P experience ranged from 1 to 42 years, with a mean experience among all groups of 14.79 years (*SD* = 10.34). An ANOVA between the three experimental groups showed that there was not a statistically significant difference in participant A&P experience, *F*(2, 97) = .249, *p* = .780.

Figure 8

Participant A&P Experience by Experimental Group

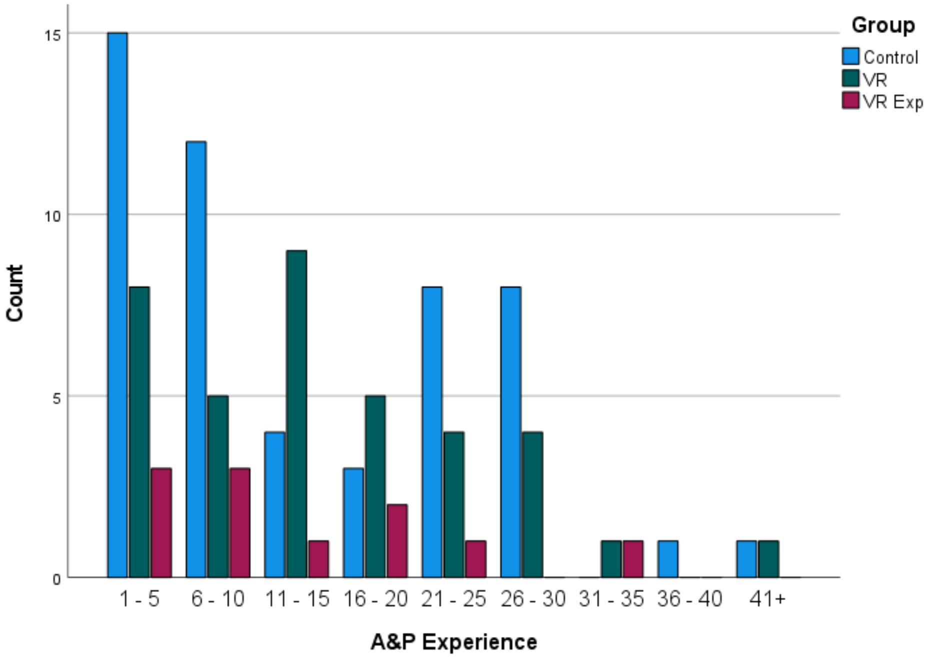
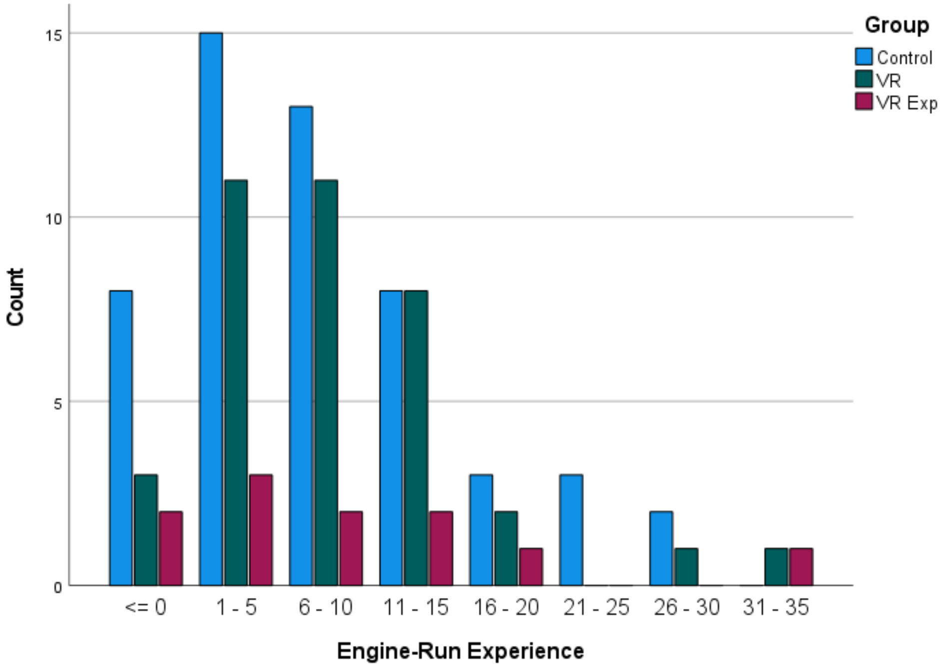


Figure 9 shows the total engine-run experience of all participants by experimental group. Engine-run experience ranged from 0 to 35 years, with a mean experience among all groups of 9.01 (*SD* = 8.02) years. An ANOVA between the three experimental groups showed that there was not a statistically significant difference in participant engine-run experience, *F*(2, 97) = .075, *p* = .928.

Figure 9

Participant Engine-Run Experience by Experimental Group



Descriptive Statistics

Post-training assessment was conducted using the modified GEARS scale during a timed event. A total GEARS score ranging from 6-30 was possible, with each of the individual rating components being scored from 1-5. This can be broken down into a possible sub-score ranging from 3-15 for both Psychomotor and Cognitive aspects of the learning evaluation which each comprised three of the six total rating components. Time to task completion was also reported in seconds. Table 6 summarizes the results by experimental group. GEARS scores assigned from all three evaluators were compared, and an ANOVA test comparing the results showed that there was not a statistically significant difference between the scores assigned by each evaluator, *F*(2, 97) = 1.117, *p* = .331.

Table 6

Evaluation Results by Experimental Group

	GEARS Total		GEARS Psychomotor		GEARS Cognitive		Time (S)	
Group	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Control	23.33	3.75	11.71	1.88	11.56	2.04	718	82
VR	23.22	3.95	11.62	2.07	11.59	2.13	738	118
VR with Experience	23.45	3.24	11.91	1.76	11.55	1.86	677	116

Figure 10 shows the distribution of total GEARS scores by experimental group. Total GEARS scores recorded ranged from 13 to 30, with a mean score of 23.30 (*SD* = 3.74). An ANOVA performed between the three experimental groups showed that there was not a statistically significant difference in participants' total GEARS scores, *F*(2, 97) = .075, *p* = .928.

Figure 10

Total GEARS Score by Experimental Group

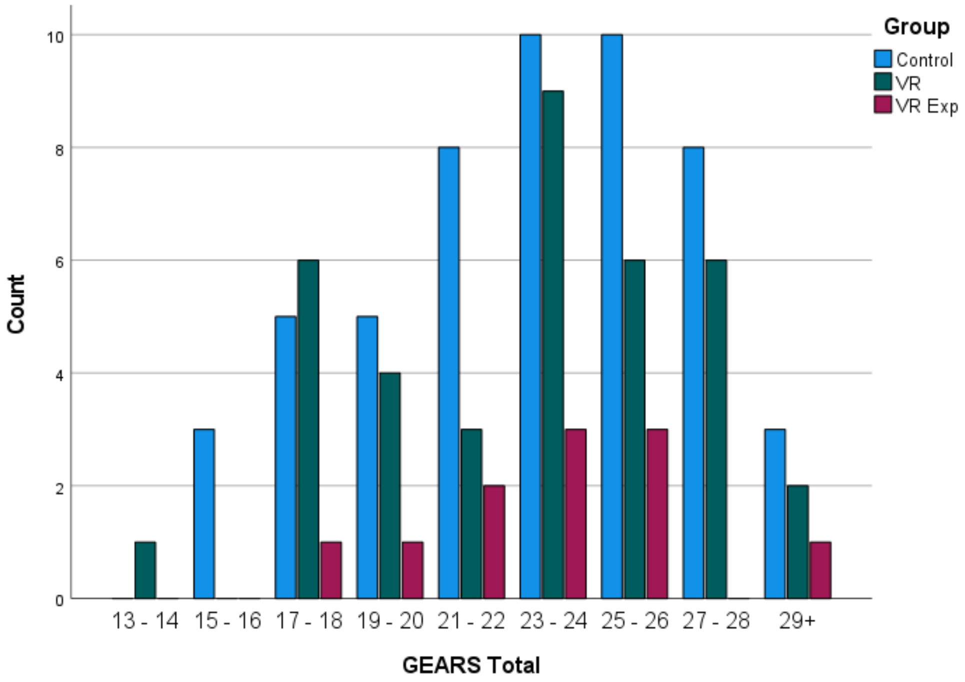


Figure 11 presents a boxplot comparison of the three experimental groups showing a visual comparison of the scores and means.

Figure 11

Boxplot of Total GEARS Score by Experimental Group

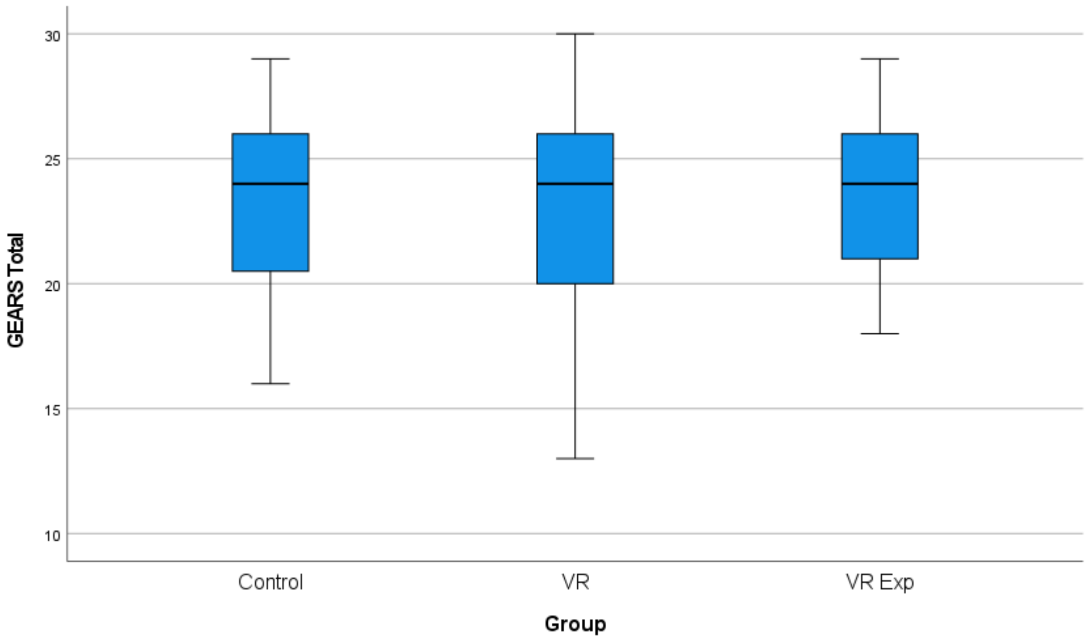


Figure 12 shows the distribution of GEARS Psychomotor scores by experimental group. GEARS Psychomotor scores ranged ranged from 7 to 15, with a mean score among all groups of 11.70 (*SD* = 1.93). An ANOVA between the three experimental groups showed that there was no statistically significant difference in participant GEARS Psychomotor scores, *F*(2, 97) = .095, *p* = .910. Figure 13 presents a boxplot comparison of the three experimental groups.

Figure 12

GEARS Psychomotor Score by Experimental Group

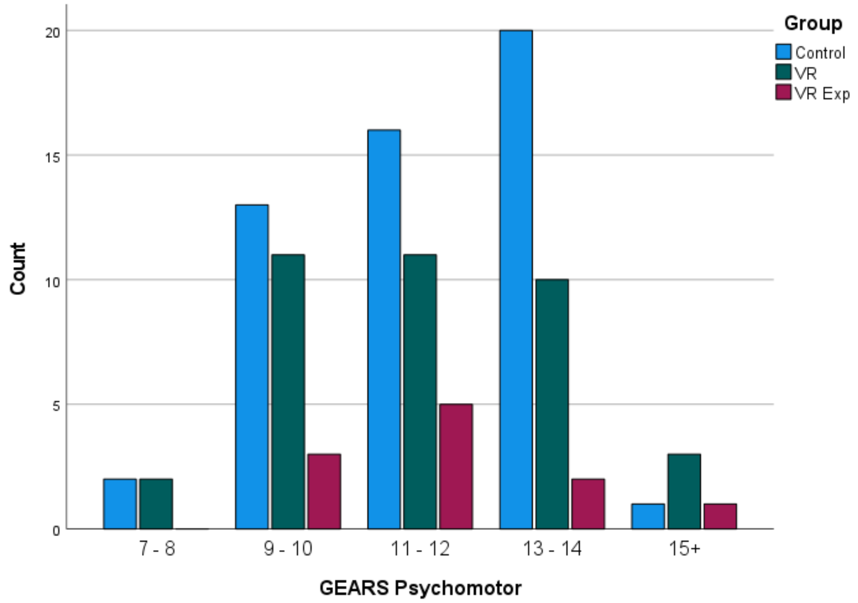


Figure 13

Boxplot of GEARS Psychomotor Score by Experimental Group

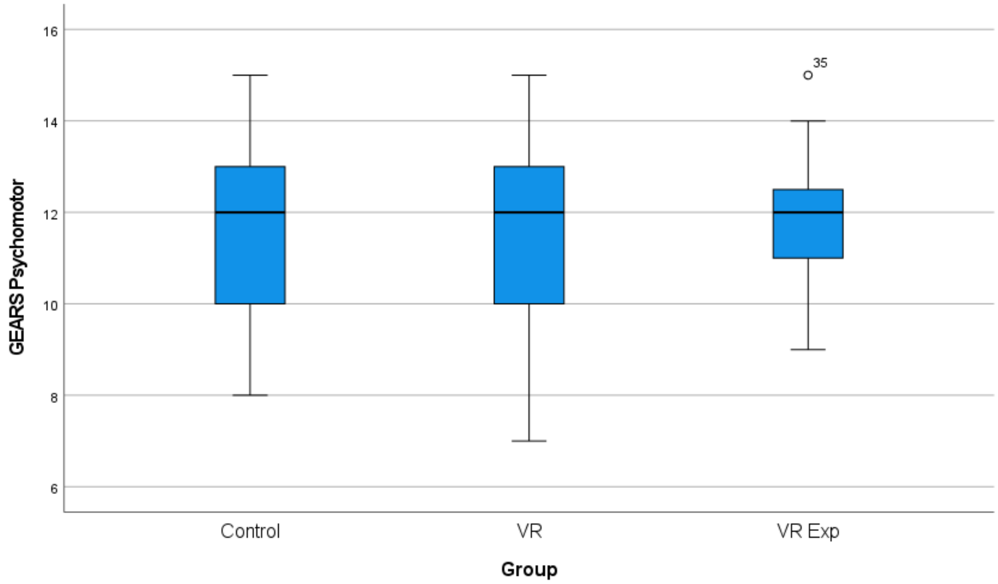


Figure 14 shows the distribution of GEARS Cognitive scores by experimental group. GEARS Cognitive scores ranged ranged from 6 to 15, with a mean score among all groups of 11.57 (*SD* = 2.04). An ANOVA between the three experimental groups showed that there was not a statistically significant difference in participants' GEARS Cognitive scores, *F*(2, 97) = .004, *p* = .996. Figure 15 presents a boxplot comparison of the three experimental groups.

Figure 14

GEARS Cognitive Score by Experimental Group

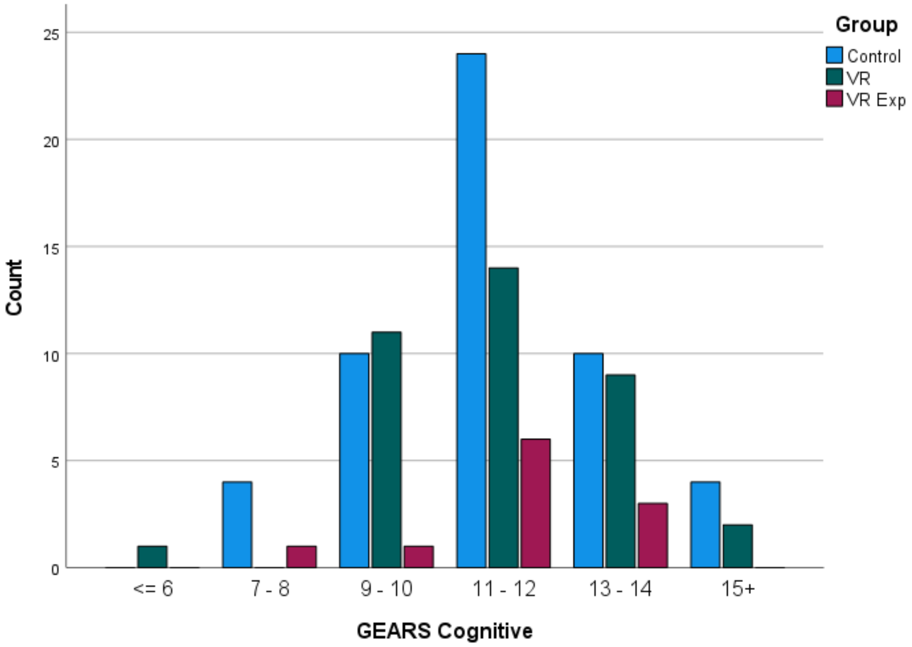


Figure 15

Boxplot of GEARS Cognitive Score by Experimental Group

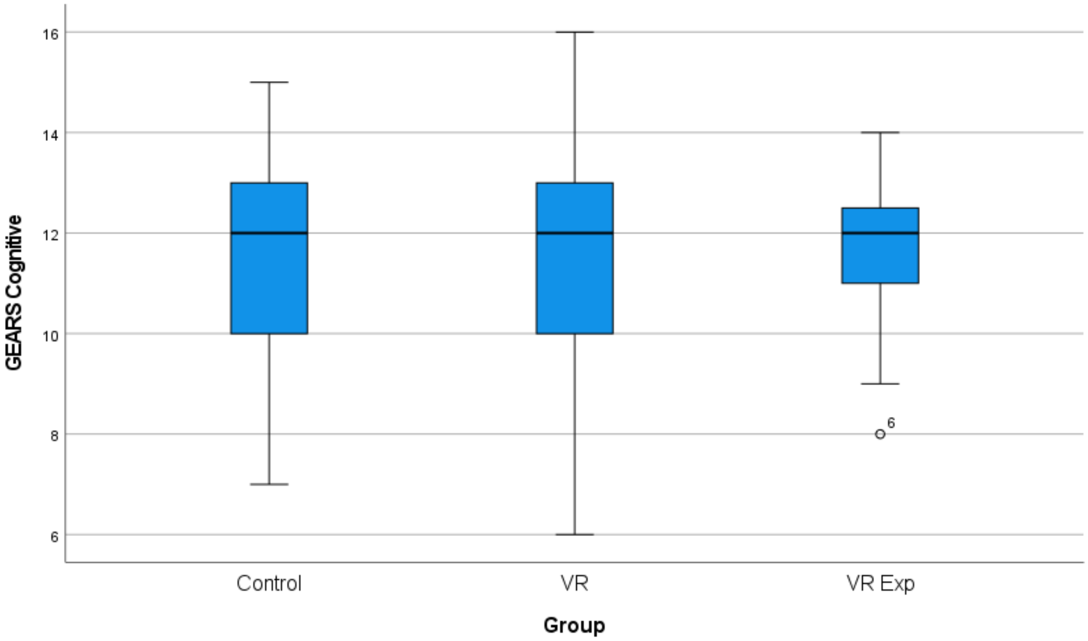


Figure 16 shows the distribution of Time to Completion values by experimental group. Time values ranged ranged from 590 to 1050 seconds, with a mean value among all groups of 721 seconds (*SD* = 101). An ANOVA between the three experimental groups showed that there was not a statistically significant difference in participant times, *F*(2, 97) = 1.562, *p* = .215. Figure 17 presents a boxplot comparison of the three experimental groups.

Figure 16

Time by Experimental Group

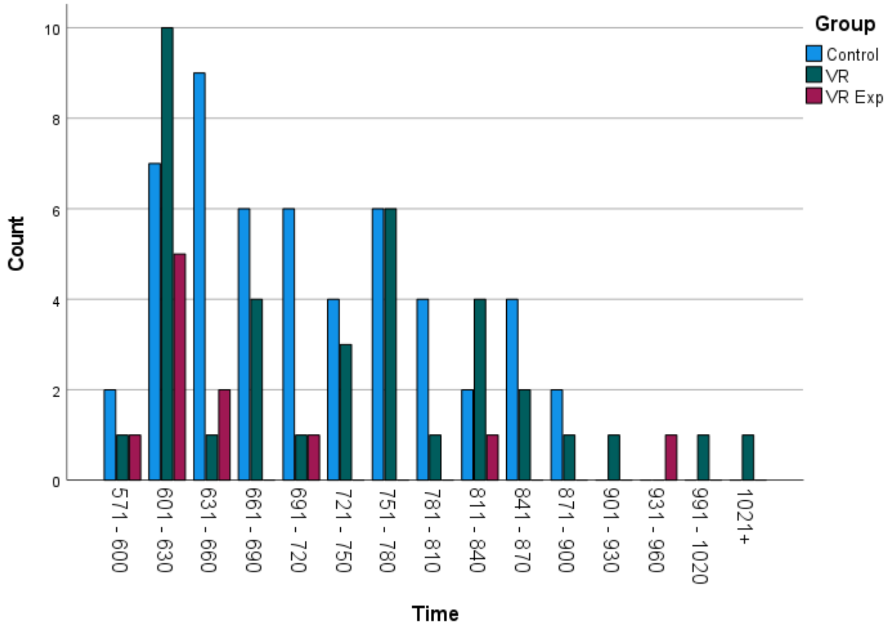
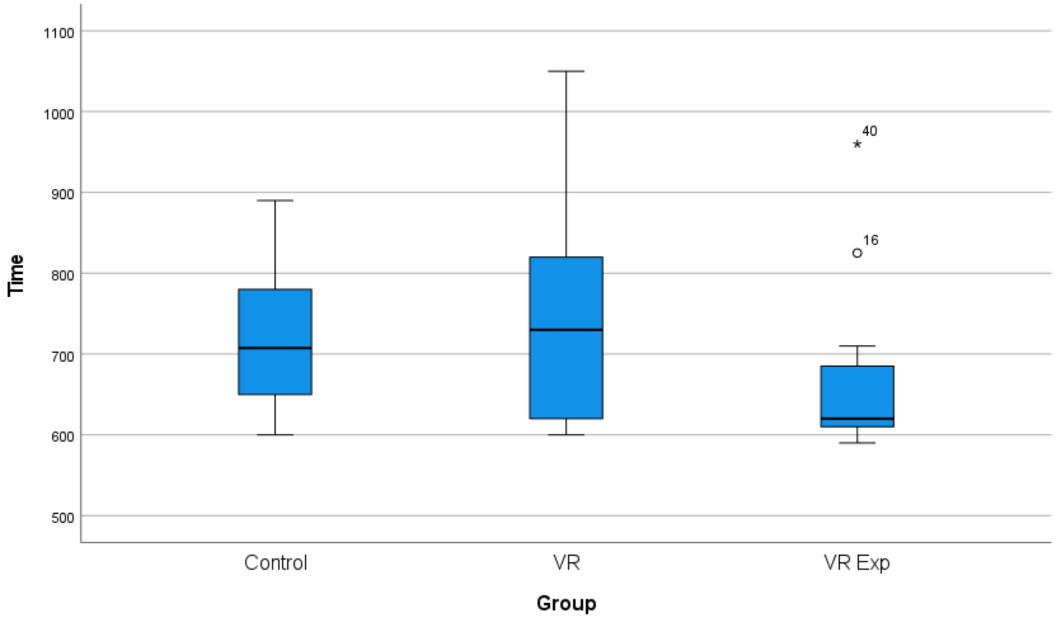


Figure 17

Boxplot of Time by Experimental Group



Reliability and Validity Testing Results

Reliability testing for the GEARS test was performed using a Cronbach's Alpha test. A Cronbach's Alpha score of greater that .7 is desired to demonstrate the relaibility of the data collection device (Tavakol & Dennick, 2011). A computed score of .911 for the GEARS test indicates that the test had acceptable internal reliability. A score in excess of .9 indicates an overall high level of reliability for the GEARS test. Inter-rater relaibility was acceptable, and an ANOVA test comparing the results showed that there was not a statistically significant difference between the GEARS scores assigned by each evaluator, $F(2, 97) = 1.117$, $p = .331$.

Hypothesis Testing Results

The main experimental hypotheses presented in Chapter I were:

HA_0

There is no collective statistically significant difference in Test Performance between the groups (VR trained and FFS) when controlled for experience.

HA_a

There is a collective statistically significant difference in Test Performance between the groups when controlled for experience.

Test performance was measured as a multivariate combination of GEARS score and time to completion for a defined task. A MANCOVA analysis was conducted using GEARS Score and Time as dependent variables, with training provided as the independent variable. The group that received VR training was further broken down into those who had prior experience using any VR system and those who had never used one.

Additional analysis was conducted by decomposing the GEARS score into its Cognitive and Psychomotor elements, and conducting independent MANCOVA analyses along with time in order to explore if the independent variable had a more noticeable effect on one specific domain of learning. The requirements for conducting a MANCOVA analysis were met as indicated in the following sections.

Measurement of Dependent Variables

The dependent variables were measured at the interval or ratio level; both dependent variables of Test Score and Time Taken are interval values reported as whole integers.

Categorical Independent Variable

The independent variable was comprised of three categorical groups. In this case the groups were Control, who received traditional training, VR who received VR training but had no prior VR experience, and Exp, who received VR training and had reported that they had previously used a VR system. Group assignment was random between the Control and either VR group, with assignment to VR or Exp done based on the participant questionnaire.

Independence of Observations

Independence of observations was observed. In this study each observation was completely independent by design, and the groups themselves were independent.

Sample Size

Adequate sample size was met. A total of 100 participants were tested to meet the G*Power calculated sample size. Group size was met by extending the experimental duration to 14 months.

Testing for Outliers

There should be no univariate or multivariate outliers. Following generation of descriptive statistics, the data was examined for outliers. Several potential outliers were identified in the VR Exp group due to the smaller size of the group relative to the others and can be seen on the associated boxplots; the scores were judged to be valid as they fell well within the bounds of all participants who received VR training and were therefore retained in the analysis. Notes from the experiment indicate that these participants were highly experienced A&P Technicians who performed exceptionally well during the evaluation. To check for multivariate outliers a Mahalanobis distance measure (MAH_1) was generated for all subjects by SPSS using GEARS Total Score and Time (Grande, 2015b). These were then compared to the chi-square distribution with the same degrees of freedom to check for multivariate outliers. The formula used for the calculation with SPSS was Prob_MD = 1-CDF.CHISQ(MAH_1,2). Any records with a Prob_MD < .001 would have been considered a multivariate outlier. No multivariate outliers were identified using this method.

Multivariate Normality

Multivariate normality is required. This was tested by examining the Skewness and Kurtosis of the results as shown in Table 7, and by examining the Q-Q plots for each variable.

Table 7

Skewness and Kurtosis for GEARS Scores and Time

	GEARS Score	Time
Kurtosis	-.607	.232
Skewness	-.401	.805

The totality of GEARS Scores and Time results follow a sufficiently normal distribution for the MANCOVA test, with the absolute skewness values of less than 2, and absolute kurtosis of less than 4.

The MANCOVA analysis is resistant to deviations from normality, and no significant improvement in the data was noted using log, power, and inverse transformations. Multivariate normality is demonstrated in Figures 18 – 21 that show Q – Q plots of each dependent variable, along with Cognitive and Psychomotor sub-groups. These plots demonstrate that the results display sufficient multivariate normality to continue with a MANCOVA analysis.

Figure 18

Normal Q-Q Plot of GEARS Total Score

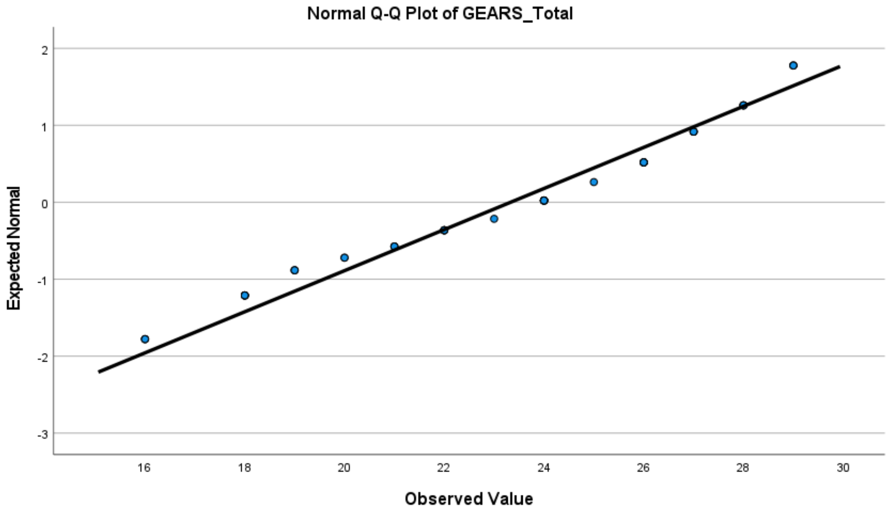


Figure 19

Normal Q-Q Plot of GEARS Psychomotor Score

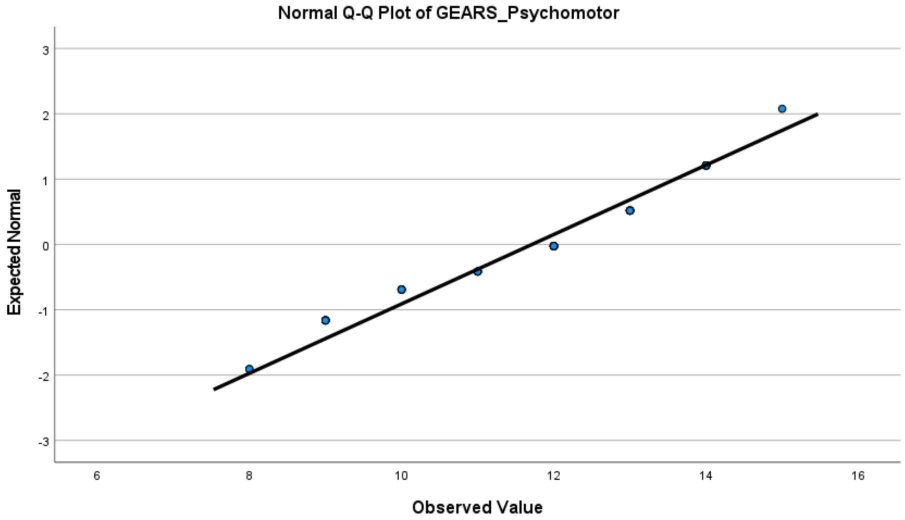


Figure 20

Normal Q-Q Plot of GEARS Cognitive Score

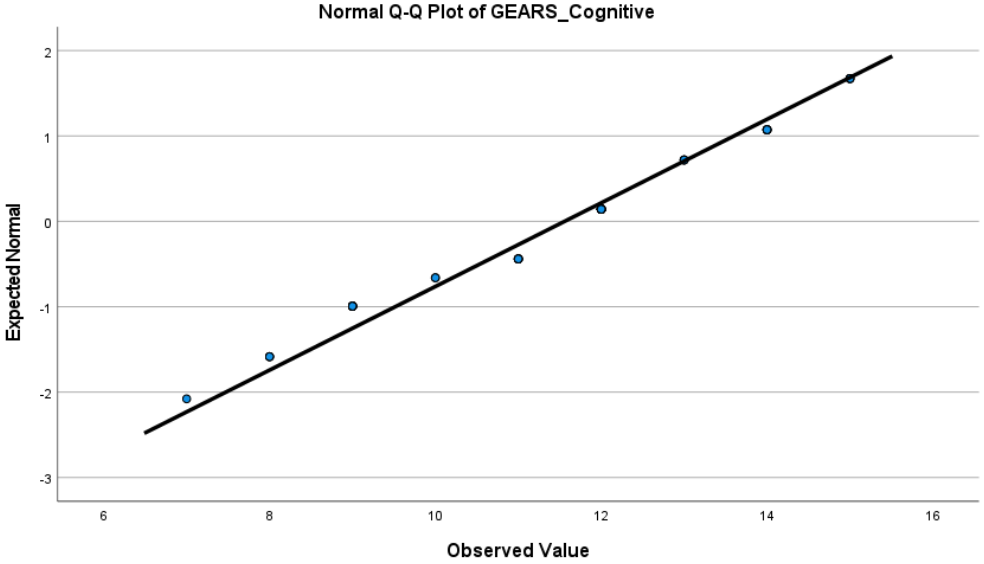
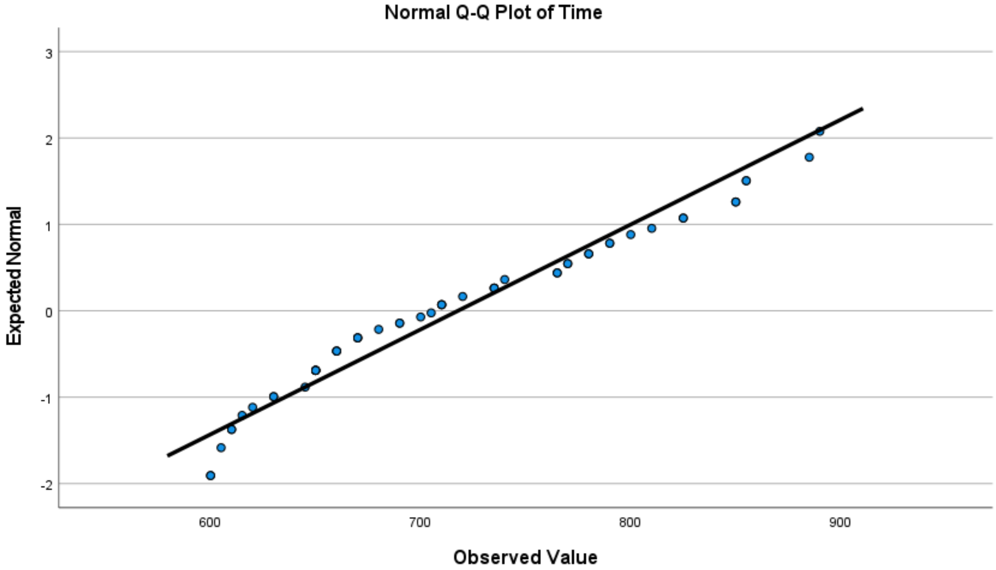


Figure 21

Normal Q-Q Plot of Time Score

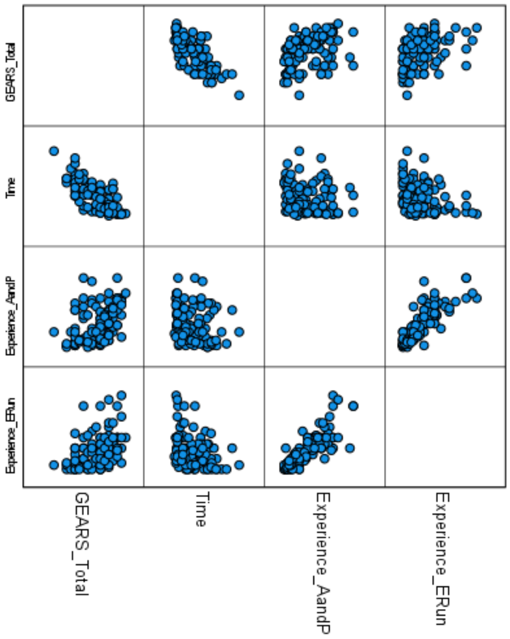


Linear Relationship Between Variables and Covariates

A MANCOVA analysis requires that there be a linear relationship between the dependent variables and covariates for each group of the independent variable. Figure 22 shows that for GEARS Total and Time, plotted against A&P Experience and Engine-run Experience, the scatterplot follows a roughly elliptical shape originating in the lower left for GEARS Total which displays a positive correlation, and lower right for Time which displays a negative correlation. This is sufficient to conclude that there is an adequate linear relationship between the dependent variables and covariates to proceed with the MANCOVA analysis. Full size scatterplots for each group can be found in Appendix I for each group.

Figure 22

ScatterPlot of Dependant Variables and Covariates



Homogeneity of Variance-Covariance Matrices

There is homogeneity of variance-covariance matrices. This was tested using Box's M test of equality of covariance with none of the results being significant, as shown in Table 8, indicating that this condition is met.

Table 8

Box's Test of Equality of Covariance Matrices

	GEARS Total / Time	GEARS Psychomotor / Time	GEARS Cognitive / Time
Box's M	11.361	9.095	10.533
F	1.798	1.439	1.667
df1	6	6	6
df2	6488	6488	6488
Sig.	.095	.195	.125

Multicollinearity

Multicollinearity was screened for in SPSS by examining the Pearson Correlation Coefficients in Table 9. As the GEARS Psychomotor and Cognitive groups are extracted from the GEARS total score, they are expected to be highly correlated as demonstrated by the high correlation values between those three variables; all three of these variables have a negative correlation with Time that is stronger than -0.8, indicating that the level of multicollinearity is acceptable to conduct a MANCOVA analysis. In order to examine the relationship between the groups it was necessary to conduct three separate MANCOVA analyses, using each of GEARS Total, GEARS Psychomotor, and GEARS Cognitive along with Time as the multivariate dependent variables.

Table 9

Correlation Coefficients

		GEARS Total	GEARS Psychomotor	GEARS Cognitive	Time
GEARS Total	Pearson Correlation	1	.936	.940	-.758
	Sig. (2-tailed)		<.001	<.001	<.001
	N	100	100	100	100
GEARS Psychomotor	Pearson Correlation	.936	1	.763	-.745
	Sig. (2-tailed)	<.001		<.001	<.001
	N	100	100	100	100
GEARS Cognitive	Pearson Correlation	.940	.763	1	-.675
	Sig. (2-tailed)	<.001	<.001		<.001
	N	100	100	100	100
Time	Pearson Correlation	-.758	-.745	-.675	1
	Sig. (2-tailed)	<.001	<.001	<.001	
	N	100	100	100	100

Main Hypothesis

The MANCOVA test for the Experimental Hypothesis, using Group as the independent variable, GEARS Total and Time as the dependent variables, and controlling for A&P Experience and Engine-Run Experience is shown in Table 10. There was no significant difference in training effectiveness (GEARS Total Score and Time) based on Training Group (Control, VR, and VR with Experience), $F(4,190) = 1.307$, $p = .269$; Wilk’s lambda = .946.

Furthermore, there was no significant effect of training Group on GEARS Total Score, $F(2, 95) = .069$, $p = .934$. There was no significant effect of training Group on Time, $F(2, 95) = 1.611$, $p = .205$.

Table 11 shows that A&P Experience was not shown to significantly influence GEARS Total Score, $F(1,95) = .714$, $p = .400$, nor Time $F(1,95) = .788$, $p = .377$.

Engine-Run Experience was not significant in influencing GEARS Total Score, *F*(1,95) = 3.751, *p* = .056, but was significant in influencing Time *F*(1,95) = 9.346, *p* = .003, partial eta squared = .090.

The null hypothesis is retained and it is concluded that there is no difference in training effectiveness based on the training received. Engine-Run Experience was shown to have a significant influence on the results.

Table 10

MANCOVA Test for Total GEARS Score and Time

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Intercept	Pillai's Trace	.991	5373.584	2.000	94.000	<.001	.991
	Wilks' Lambda	.009	5373.584	2.000	94.000	<.001	.991
	Hotelling's Trace	114.332	5373.584	2.000	94.000	<.001	.991
	Roy's Largest Root	114.332	5373.584	2.000	94.000	<.001	.991
Experience AandP	Pillai's Trace	.056	2.807	2.000	94.000	.065	.056
	Wilks' Lambda	.944	2.807	2.000	94.000	.065	.056
	Hotelling's Trace	.060	2.807	2.000	94.000	.065	.056
	Roy's Largest Root	.060	2.807	2.000	94.000	.065	.056
Experience ERun	Pillai's Trace	.091	4.728	2.000	94.000	.011	.091
	Wilks' Lambda	.909	4.728	2.000	94.000	.011	.091
	Hotelling's Trace	.101	4.728	2.000	94.000	.011	.091
	Roy's Largest Root	.101	4.728	2.000	94.000	.011	.091
Group	Pillai's Trace	.054	1.307	4.000	190.000	.269	.027
	Wilks' Lambda	.946	1.311	4.000	188.000	.268	.027
	Hotelling's Trace	.057	1.314	4.000	186.000	.266	.027
	Roy's Largest Root	.056	2.671	2.000	95.000	.074	.053

Table 11

Tests of Between-Subjects Effects

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	GEARS Total	257.346	4	64.336	5.430	<.001	.186
	Time	183665.917	4	45916.479	5.219	<.001	.180
Intercept	GEARS Total	12961.184	1	12961.184	1093.864	<.001	.920
	Time	15995709.377	1	15995709.377	1818.086	<.001	.950
Experience AandP	GEARS Total	8.455	1	8.455	.714	.400	.007
	Time	6935.605	1	6935.605	.788	.377	.008
Experience ERun	GEARS Total	44.446	1	44.446	3.751	.056	.038
	Time	82226.906	1	82226.906	9.346	.003	.090
Group	GEARS Total	1.623	2	.812	.069	.934	.001
	Time	28352.600	2	14176.300	1.611	.205	.033
Error	GEARS Total	1125.654	95	11.849			
	Time	835819.643	95	8798.102			
Total	GEARS Total	55672.000	100				
	Time	52948804.000	100				
Corrected Total	GEARS Total	1383.000	99				
	Time	1019485.560	99				

Hypothesis using Psychomotor Subgroup

The MANCOVA test for a subset of the Experimental Hypothesis, using Group as the independent variable, GEARS Psychomotor and Time as the dependent variables, and controlling for A&P Experience and Engine-Run Experience is shown in Table 12. There was no significant difference in training effectiveness in the Psychomotor domain (GEARS Psychomotor Score and Time) based on Training Group (Control, VR, and VR with Experience), *F*(4, 190) = 1.039, *p* = .389; Wilk's lambda = .957.

Furthermore, there was no significant effect of training Group on GEARS Psychomotor Score, *F*(2, 95) = .164, *p* = .849. There was no significant effect of training Group on Time, *F*(2, 95) = 1.611, *p* = .205, as in the previous analysis.

Table 13 shows that A&P Experience was not shown to significantly influence GEARS Psychomotor Score, *F*(1, 95) = .287, *p* = .593, nor Time *F*(1,95) = .788, *p* = .377. Engine-Run Experience was significant in influencing GEARS Psychomotor Score, *F*(1, 95) = 5.732, *p* = .019, partial eta squared = .057, and was also significant in influencing Time *F*(1, 95) = 9.346, *p* = .003, partial eta squared = .090.

We fail to reject the null hypothesis and conclude that there is no difference in training effectiveness based on the training received. Engine-Run experience was shown to have a significant influence on the results.

Table 12

MANCOVA Test for GEARS Psychomotor Score and Time

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Intercept	Pillai's Trace	.990	4828.121	2.000	94.000	<.001	.990
	Wilks' Lambda	.010	4828.121	2.000	94.000	<.001	.990
	Hotelling's Trace	102.726	4828.121	2.000	94.000	<.001	.990
	Roy's Largest Root	102.726	4828.121	2.000	94.000	<.001	.990
Experience AandP	Pillai's Trace	.036	1.740	2.000	94.000	.181	.036
	Wilks' Lambda	.964	1.740	2.000	94.000	.181	.036
	Hotelling's Trace	.037	1.740	2.000	94.000	.181	.036
	Roy's Largest Root	.037	1.740	2.000	94.000	.181	.036
Experience ERun	Pillai's Trace	.090	4.675	2.000	94.000	.012	.090
	Wilks' Lambda	.910	4.675	2.000	94.000	.012	.090
	Hotelling's Trace	.099	4.675	2.000	94.000	.012	.090
	Roy's Largest Root	.099	4.675	2.000	94.000	.012	.090
Group	Pillai's Trace	.043	1.039	4.000	190.000	.389	.021
	Wilks' Lambda	.957	1.039	4.000	188.000	.388	.022
	Hotelling's Trace	.045	1.039	4.000	186.000	.388	.022
	Roy's Largest Root	.045	2.117	2.000	95.000	.126	.043

Table 13

Tests of Between-Subjects Effects - Psychomotor

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	GEARS Psychomotor	76.140	4	19.035	6.217	<.001	.207
	Time	183665.917	4	45916.479	5.219	<.001	.180
Intercept	GEARS Psychomotor	3262.373	1	3262.373	1065.547	<.001	.918
	Time	15995709.377	1	15995709.377	1818.086	<.001	.950
Experience AandP	GEARS Psychomotor	.879	1	.879	.287	.593	.003
	Time	6935.605	1	6935.605	.788	.377	.008
Experience ERun	GEARS Psychomotor	17.550	1	17.550	5.732	.019	.057
	Time	82226.906	1	82226.906	9.346	.003	.090
Group	GEARS Psychomotor	1.002	2	.501	.164	.849	.003
	Time	28352.600	2	14176.300	1.611	.205	.033
Error	GEARS Psychomotor	290.860	95	3.062			
	Time	835819.643	95	8798.102			
Total	GEARS Psychomotor	14056.000	100				
	Time	52948804.000	100				
Corrected Total	GEARS Psychomotor	367.000	99				
	Time	1019485.560	99				

Hypothesis Using Cognitive Subgroup

The MANCOVA test for a subset of the Experimental Hypothesis, using Group as the independent variable, GEARS Cognitive Score and Time as the dependent variables, and controlling for A&P Experience and Engine-Run Experience is shown in Table 14. There was no significant difference in training effectiveness in the Psychomotor domain (GEARS Cognitive Score and Time) based on Training Group (Control, VR, and VR Exp), $F(4, 190) = 1.333$, $p = .259$; Wilk's lambda = .945.

Furthermore, there was no significant effect of training Group on GEARS Cognitive Score, $F(2, 95) = .002$, $p = .998$. There was no significant effect of training Group on Time, $F(2, 95) = 1.611$, $p = .205$, as in the previous analysis.

Table 15 shows that A&P Experience was not shown to significantly influence GEARS Cognitive Score, $F(1, 95) = 1.082$, $p = .301$, nor Time $F(1, 95) = .788$, $p = .377$. Engine-Run Experience was not significant in influencing GEARS Cognitive Score, $F(1, 95) = 1.283$, $p = .260$, but was significant in influencing Time $F(1, 95) = 9.346$, $p = .003$, partial eta squared = .090 per the previous analysis.

The null hypothesis is retained and it is concluded that there is no difference in training effectiveness based on the training received. Engine-Run experience was shown to have a significant influence on the results.

Table 14

MANCOVA Test for GEARS Cognitive Score and Time

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
Intercept	Pillai's Trace	.988	3764.016	2.000	94.000	<.001	.988
	Wilks' Lambda	.012	3764.016	2.000	94.000	<.001	.988
	Hotelling's Trace	80.085	3764.016	2.000	94.000	<.001	.988
	Roy's Largest Root	80.085	3764.016	2.000	94.000	<.001	.988
Experience AandP	Pillai's Trace	.054	2.698	2.000	94.000	.073	.054
	Wilks' Lambda	.946	2.698	2.000	94.000	.073	.054
	Hotelling's Trace	.057	2.698	2.000	94.000	.073	.054
	Roy's Largest Root	.057	2.698	2.000	94.000	.073	.054
Experience ERun	Pillai's Trace	.101	5.302	2.000	94.000	.007	.101
	Wilks' Lambda	.899	5.302	2.000	94.000	.007	.101
	Hotelling's Trace	.113	5.302	2.000	94.000	.007	.101
	Roy's Largest Root	.113	5.302	2.000	94.000	.007	.101
Group	Pillai's Trace	.055	1.333	4.000	190.000	.259	.027
	Wilks' Lambda	.945	1.338[b]	4.000	188.000	.258	.028
	Hotelling's Trace	.058	1.342	4.000	186.000	.256	.028
	Roy's Largest Root	.058	2.742	2.000	95.000	.070	.055

Table 15

Tests of Between-Subjects Effects - Cognitive

Source	Dependent Variable	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	GEARS Cognitive	49.540	4	12.385	3.259	.015	.121
	Time	183665.917	4	45916.479	5.219	<.001	.180
Intercept	GEARS Cognitive	3230.797	1	3230.797	850.280	<.001	.900
	Time	15995709.377	1	15995709.377	1818.086	<.001	.950
Experience AandP	GEARS Cognitive	4.111	1	4.111	1.082	.301	.011
	Time	6935.605	1	6935.605	.788	.377	.008
Experience ERun	GEARS Cognitive	4.876	1	4.876	1.283	.260	.013
	Time	82226.906	1	82226.906	9.346	.003	.090
Group	GEARS Cognitive	.013	2	.007	.002	.998	.000
	Time	28352.600	2	14176.300	1.611	.205	.033
Error	GEARS Cognitive	360.970	95	3.800			
	Time	835819.643	95	8798.102			
Total	GEARS Cognitive	13797.000	100				
	Time	52948804.000	100				
Corrected Total	GEARS Cognitive	410.510	99				
	Time	1019485.560	99				

Analysis of Covariates

The main hypothesis tested was that there was no collective statistically significant difference in Test Performance between the groups (VR trained and FFS) when controlled for experience. The covariates that comprised experience were a technician's total number of years of experience as an A&P (Experience_A&P) and their number of years of experience as a qualified engine runner (Experience_ERun). The analysis in Table 11 shows that A&P Experience was not shown to significantly influence GEARS Total Score, *F*(1 ,95) = .714, *p* = .400, nor Time *F*(1, 95) = .788, *p* = .377. Engine-Run Experience, however, was significant in influencing both GEARS Psychomotor Score, *F*(1, 95) = 5.732, *p* = .019, partial eta squared = .057, and was also significant in influencing Time *F*(1, 95) = 9.346, *p* = .003, partial eta squared = .090. Engine-Run Experience was a significant covariate in this study, while overall A&P experience was not.

Reportable Safety and Physiological Incidents

A safety reporting process was put in place for this study that required reporting of any physiological incidents during VR training, and an immediate stop to the study until any such events were analyzed. Subjects were advised before beginning a VR session that they could stop the session at any time and were asked to inform the evaluators of any physiological discomfort that might occur. No safety reports were filed during the period of the study, and no sessions were stopped or interrupted at the request of the subject. The evaluators reported no negative comments throughout and noted that the VR training apparatus was well tolerated by the subjects.

Summary

Chapter IV presented the quantitative results of an experimental study to determine if there was a statistically significant difference in Test Performance between two experimental groups, one VR trained and the other trained in the FFS, when controlled for experience. A pilot study was conducted to analyze the experimental methodology, and to train the evaluators on the use of a modified GEARS scale as an evaluation instrument. A total sample of 98 was desired, and the experiment was conducted over 14 months to reach a final sample size of 100 participants. Descriptive statistics were gathered to allow comparison of the sample with the total population of FAA certified A&P technicians with respect to age and gender.

Analysis of the experimental results found that for the total sample, as well as each sub-group (experimental and control), the demographic data did not vary significantly from the total A&P population. A MANCOVA analysis was conducted comparing overall test results between the control group (FFS), and two VR trained sub-groups (VR and VR_Exp), while controlling for A&P and Engine-run Experience and found that the main hypothesis was not supported. The null hypothesis that there was no statistically significant difference between the groups was retained, and years of engine-run experience was found to be a significant covariate. The same result was found when the total GEARS scores were decomposed into Psychomotor and Cognitive sub-groups, and a MANCOVA analysis was run using those scores. The Engine-Run Experience of the subjects was found to be a significant covariate and influenced both psychomotor scores and time. In Chapter V these results are discussed, conclusions drawn, and recommendations for future research given. The contribution of these results to the aviation body of knowledge is also discussed, alongside the theoretical underpinnings of this study.

Chapter V: Discussion, Conclusions, and Recommendations

This study assessed whether a VR delivered engine-run recurrent training package could produce equivalent training results as the same training delivered in a traditional FFS setting. Chapter IV presented the results of the study which included the sample demographics, experimental results, and hypothesis testing, and then concluded by addressing the research question. Chapter V discusses the significance of these results, presents conclusions, and offers recommendations for future research.

Discussion

Characteristics of the Participants

Demographic information was collected from the subjects and compared to the population of FAA Licensed A&Ps (FAA, 2019). The current mean age of FAA Certified A&P mechanics is approximately 40 years, and 2.7% of certificate holders are female (FAA, 2023). Table 5 shows the mean ages and standard deviation for each group, and all are within one standard deviation of the population mean. Female A&P mechanics represented 4.0% of the sample, which is slightly greater than in the population. The sample demographics are representative of the general population and cause no concern in generalizing the sample results to the population within the bounds of certainty of this study.

Discussion of the Research Findings

The research question for this study was, Does VR delivered A&P engine-run recurrent training produce equivalent test performance when compared to training in the FFS, when we control for the subject’s level of experience both as an A&P and in conducting engine runs? The measure of equivalent performance was assessed through a combination of GEARS scores and Time to task completion, and the level of experience was measured through years of experience both as a certificated A&P technician (Experience A&P), and years of experience conducting engine-runs (Experience ERun). The null hypothesis used in the MANCOVA analysis was that there is no collective statistically significant difference in Test Performance between the groups when controlled for experience, and this hypothesis was retained. When looking at the individual components of the GEARS scores (Cognitive and Psychomotor) and testing using a MANCOVA with the same null hypothesis, the null hypothesis was also retained. In all measures, the VR training system produced a statistically equivalent test performance to the traditional FFS training.

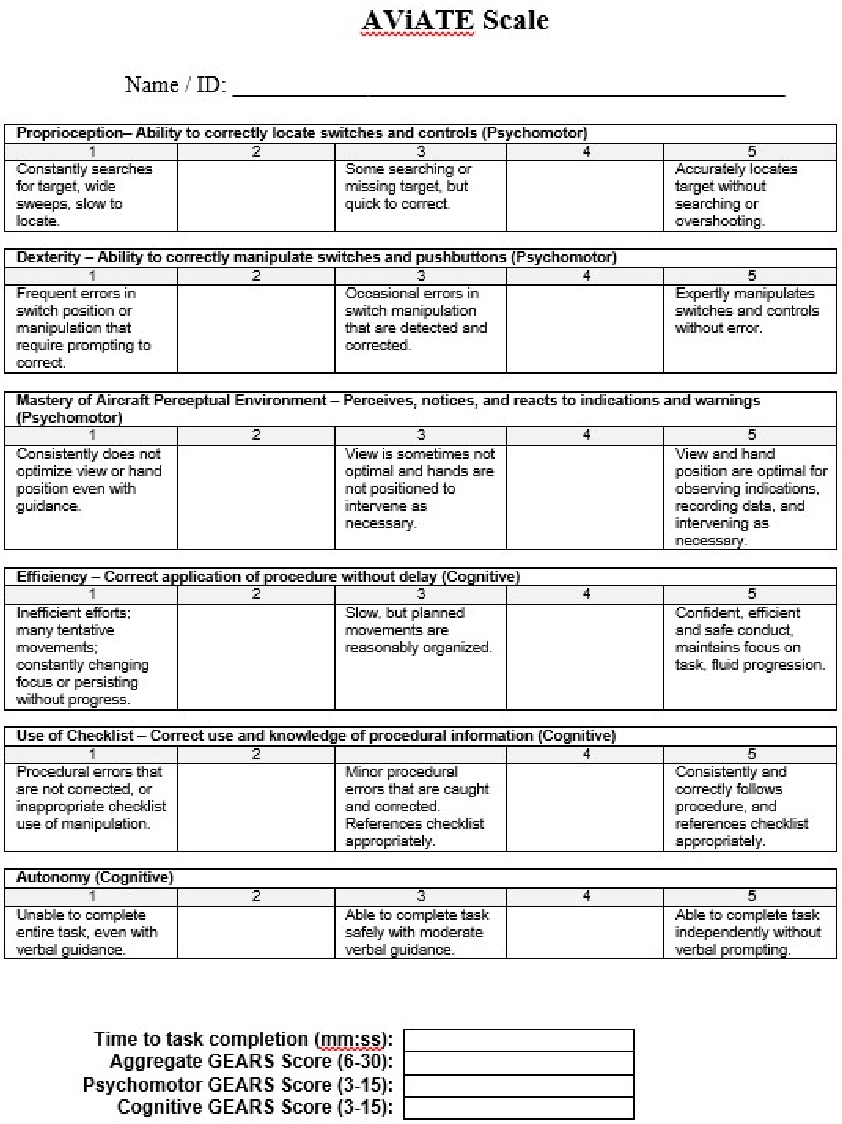
In practical terms the current study has demonstrated, with robust experimental controls and an adequate sample size that the VR trainer used produced equivalent training results to the FFS. This is a highly significant finding with potentially enormous economic and training significance. The FFS could now be replaced by a suite of VR devices for procedural training and be properly dedicated to teaching those sequences that require aircraft motion. Aircraft operators in remote locations could conduct procedural training using VR over an internet connection, giving them access to the same tools and instructors that larger companies build into their training centers. The ability to bring a virtual world to trainees, instead of bringing trainees to real-world infrastructure, will be truly transformative.

The Modified GEARS Scale: Aerospace Virtual Reality Assessment of Training Effectiveness (AViATE)

The modified GEARS scale that was discussed in Chapter III and is presented in Appendix F proved an effective tool in assessing the subjects' performance. It provided an easy format in which to score both cognitive and psychomotor elements of the task alongside the time used. The adaptation of GEARS using Bloom's Taxonomy and the Airbus CBT competencies was easily understood and utilized by the evaluators and was observed to work extremely well in the FFS environment. Inter-rater reliability was demonstrated to be acceptable and consistent with previous studies of the GEARS instrument reported in Chapter II. Figure 23 presents the scale that was used for the current study re-titled as the Aerospace Virtual Reality Assessment of Training Effectiveness (AViATE) Scale to recognize the specific adaptations incorporated to allow it to be easily used in the flight deck environment. It supports standardized grading, clearly identifies the task elements to be assessed, and incorporates time as a measure of task performance to facilitate a multivariate analysis of overall performance. The new naming scheme clearly demonstrates its adaptation to the field of aviation, and more clearly identifies that the overall learning outcome and training effectiveness are being assessed as opposed to the apparatus that has delivered the training. This paper will continue to use the GEARS acronym for clarity and consistency, but it should be read as synonymous with AViATE for future use and publication.

Figure 23

Aerospace Virtual Reality Assessment of Training Effectiveness Scale



AViATE Scale

Name / ID: ____

Proprioception– Ability to correctly locate switches and controls (Psychomotor)				
1	2	3	4	5
Constantly searches for target, wide sweeps, slow to locate.		Some searching or missing target, but quick to correct.		Accurately locates target without searching or overshooting.

Dexterity – Ability to correctly manipulate switches and pushbuttons (Psychomotor)				
1	2	3	4	5
Frequent errors in switch position or manipulation that require prompting to correct.		Occasional errors in switch manipulation that are detected and corrected.		Expertly manipulates switches and controls without error.

Mastery of Aircraft Perceptual Environment – Perceives, notices, and reacts to indications and warnings (Psychomotor)				
1	2	3	4	5
Consistently does not optimize view or hand position even with guidance.		View is sometimes not optimal and hands are not positioned to intervene as necessary.		View and hand position are optimal for observing indications, recording data, and intervening as necessary.

Efficiency – Correct application of procedure without delay (Cognitive)				
1	2	3	4	5
Inefficient efforts; many tentative movements; constantly changing focus or persisting without progress.		Slow, but planned movements are reasonably organized.		Confident, efficient and safe conduct, maintains focus on task, fluid progression.

Use of Checklist – Correct use and knowledge of procedural information (Cognitive)				
1	2		4	5
Procedural errors that are not corrected, or inappropriate checklist use of manipulation.		Minor procedural errors that are caught and corrected. References checklist appropriately.		Consistently and correctly follows procedure, and references checklist appropriately.

Autonomy (Cognitive)				
1	2	3	4	5
Unable to complete entire task, even with verbal guidance.		Able to complete task safely with moderate verbal guidance.		Able to complete task independently without verbal prompting.

Time to task completion (mm:ss):
Aggregate GEARS Score (6-30):
Psychomotor GEARS Score (3-15):
Cognitive GEARS Score (3-15):

Effect Size

In determining the sample size for this study, a medium effect size was assumed (Cohen’s d = .25) and was used to arrive at the total desired sample of 99; if a medium effect were present we would expect to observe it given the calculated sample size. A medium effect size was an appropriate metric on which to base this study as a small effect would have an equally small bearing on test performance, and therefore on quality and safety, given the aviation-based requirement for 100% accurate task completion in a timely manner. While it would have been possible to expand the sample to test for a smaller effect, it would have made little practical difference in this domain and would have significantly lengthened the time and cost of the study. The partial eta squared values for Group in Table 11 indicate small or lower effect sizes for both GEARS total and Time, which also serves to demonstrate an equivalence in training effect from the VR system.

Use of Time as a Measure of Test Performance

In Chapter III the use of time as a variable was discussed, and it was stated that

> a multivariate approach, that includes time as a measure of success, is appropriate in fields where the professional technical competence of the subjects strongly predisposes them to both correct and timely completion of a task, such as in the case of licensed aviators and mechanics, or of medical personnel. (Hoogenes et al., 2018, p.112)

A&P mechanics working to a task are required to complete it to 100% accuracy, and it is expected that, as they become more proficient, that they will complete the task more quickly. Table 9 shows a significant negative Pearson Correlation of -.737 between GEARS Total Scores and Time, thus demonstrating that as Total GEARS scores improved (indicating increased proficiency and performance) the time for completion decreased. This is a phenomenon that may be counter-intuitive to non-aviation researchers, making it a significant finding that the experimental results support the linking of time to measures of competency, and also demonstrating the link between a more rapid completion of a task and superior results.

Safety and Physiological Factors

There were no reported safety or physiological incidents during this study, and the evaluators noted that the VR apparatus was well tolerated by the subjects. Training sessions were programmed to last for 15 minutes. The physiological effects of extended use of VR were not examined during this study, but it is significant that there were zero reported ill effects induced by the use of a VR system on a representative sample of the total A&P population conducted over 14 months. This suggests that, as would be expected from a commercial system, training times can be extended from what was examined during this study with little concern. Future studies should continue to track and report any observed ill-effects of using a VR system as part of their experimental methodology in order to build upon the body of knowledge of VR physiological effects.

Applicability of the Use of the Tested VR Training Device

The VR trainer used for the delivery of the recurrent engine-run training produced a similar training outcome to the traditional training conducted in the FFS, and the use of a virtual environment for training allowed the subjects to effectively perform in the flight deck environment. This is the first use of a structured evaluation tool such as GEARS in an aviation application, and in addition to validating the training which is being delivered by an existing VR system, it allows for the possibility of much more comprehensive evaluation of virtual aviation applications in the future.

Between Groups Observations

The Boxplots shown in Figures 11, 13, and 15 all have a similar pattern. They support the analysis that there is no statistically significant difference between groups but seem to indicate that there may be a benefit gained by using the VR tool. In each figure the median VR scores are slightly lower than for the FFS group, while the VR Exp group had slightly higher median scores than the FFS group. While not statistically significant, this does suggest that a useful area for further research would be to investigate if there is an additional performance benefit to dedicated VR training and an ongoing use of the device, and if testing for a smaller effect size may be necessary to fully expose and understand any existing correlation.

Conclusions

The purpose of this study was to evaluate the effectiveness of VR learning in an aviation environment. While there are certainly a large number of emerging VR tools appearing on the market, limited research with sufficent robustness to draw statistically significant conclusions has been conducted to assess their overall effectiveness in transferring knowledge and skills to the students who use them. This study has provided, in a controlled experimental setting with an adequate sample size, both a method to evaluate the effectiveness of VR learning in the cognitive and psychomotor domains in an aviation setting, and a method to assess actual learning on a commercial system that is ready for employment. It has adapted an instrument with a high degree of validity to the aviation environment, and used it to conclude that the VR system and environment used to deliver engine-run recurrent training is capable of producing results that are at least equivalent to the traditional instruction in an FFS.

Theoretical Contributions

This study contributes to the literature in three key ways. Firstly, it provides a structured method for evaluating the effectiveness of a VR learning system that is based on existing CBT and instructional theory where no such method existed previously. It contributes to aviation training by developing this methodology and by establishing the linkage between CBT and a structured evaluation method. Future researchers will be able to use either the GEARS (AViATE) scale to conduct similar evaluations, or they can use the existing methodology to adapt other scales that may be better suited to a specific purpose. Those searching for a methodology to evaluate the effectiveness of VR learning can now base their research on the current study and develop it further.

Secondly, the results show that in developing methodologies for evaluating VR systems that a period of adaptation to the VR system should be taken into account. This is also a recommendation for future research, and an important addition to the theory of evaluating VR systems. Much like how the initial GEARS testing on the DaVinci™ robot tracked the performance of the student as they learned robotic surgery, the addition of an evaluative component to the virtual environment allows for continuous assessment of a student's progress by the software and training device. The mere suggestion that an adaptation period exists is critical to the design of future studies looking for incrementally smaller effect sizes.

Thirdly, it provides an adapted scale (AViATE) for the evaluation of VR delivered training in an aviation environment, and in combination with Time to task completion has demonstrated both an ease of use and consistency of results. It has further demonstrated that existing CBT evaluations can be combined into a single instrument to allow for multivariate analysis of the effectiveness of VR Training. Used in combination with existing CBT methods, the AViATE Scale will provide a basis for evaluating future VR systems and studying their effectiveness in differing applications and duration of use.

The theoretical aviation body of knowledge has been extended through the experimental application of adapted methodology, a better understanding of the use of a VR system gained through experimentation, and the development and validation of a multivariate means of assessing VR systems. These three contributions address notable gaps in the existing literature and suggest a direction for several future studies.

Practical Contributions

This study focused on a VR tool that provided recurrent training to certificated A&P technicians, but did so under a CBT framework which leveraged an existing training platform and methodology that are recognized as industry state of the art. CBT tasks in aviation are no different whether performed by a technician, systems operator, or pilot, and the findings of this study are significant and applicable to any program of study that uses this methodology.

Primarily, the results provide not only a validation of an existing VR training environment, but also a tool and methodology for evaluating future VR applications in aviation training. Although the study was conducted in a recurrent training environment for A&P technicians, any CBT task being taught in this environment could be similarly evaluated. Systems training for pilots and system operators falls under this umbrella, and the ability to use this study's methodology immediately in other forms of CBT evaluation is significant.

It has particular application to policymakers and regulators, training development centers, and those who wish to exploit similar applications in the field of aviation. The FAA gives specific approval for the facilities that are used for aviation training, the approved devices with which training can be conducted, the specific course content and syllabi that are used to deliver training, and the order in which training events are sequenced (FAA, 2021). This study should serve to validate the existing Airbus VR system as an effective learning platform for recurrent A&P training and should be delivered to the FAA through the standard approval process to support the validation of the platform for use. In following the normal application process for approval, it should also be recommended that this evaluation serve as a template for future evaluations of VR systems, and that a standard template for the validation of virtual learning environments be adopted. The courseware for this study, while applicable to an FAA training course, was also designed to be usable for training certified by EASA as well. It is further recommended to use this validation study to apply for EASA approval, and to use it for future EASA certifications as well.

Courseware designers who wish to migrate their existing courseware to virtual environments can use the methodology from this study to conduct their own independent validation studies, which can then be used to apply for regulatory approval of other VR training products. Close adherence to the experimental methodology, and the establishment of a control group early in the planned transition to VR learning, allows for rapid authoring, editing, and certification of new virtual tools. VR training times should be extended in these studies and a reporting system established to track and analyze any observed or reported physiological effects.

Limitations of the Findings

The present study has four main limitations which, while they constrain its results, can also serve to bound future studies and to serve as a basis for their design.

Firstly, a VR environment that was developed and adapted for recurrent training was utilized in this experiment. In such a recurrent training environment the students have already learned how to do the required tasks and have real-world experience in performing them. The recurrent training environment places emphasis on reinforcing existing knowledge, rather than building new knowledge, which may have a different focus. It provided predictable and stable learning requirements, and also allowed for a uniform and well controlled standard of students who presented as subjects. This was highly desirable for an initial study such as this as it eliminated a large degree of variability in subjects, but this must be acknowledged as both a limitation on its results and a stepping-off point for future research.

Secondly, no study of the affective domain of learning was performed. As this study focused on the physical effectiveness of learning using a CBT system, it did not analyze the elements of affective learning. Affective learning is an important domain to understand, and the combination of psychomotor, cognitive, and affective domains gives a complete picture of learning under Bloom's Taxonomy. Practical constraints on time and resources did not allow for this to be studied in combination with cognitive and psychomotor factors, and this should be addressed in future dedicated or mixed-methods studies.

Thirdly, the results of this study provide a single snapshot of learning effectiveness after a single VR training session. The evolution of skills over time, when a

VR system is used continuously, will be an important factor in designing future recurrent training programs, and will need to be well understood. A longitudinal study that examines the effect of VR training on a population or sample over time is recommended and could be implemented alongside the operational implementation of a VR training program.

Finally, this study used FAA Certificated A&P technicians as subjects and a structured VR training program that allowed for a limited amount of deviation from the script. Generalization of the findings is possible through any CBT program that uses the same entry standard as the FAA program, but a direct comparison with other programs and regulatory authorities was not done as part of this study. The structured nature of the VR training program was appropriate for a recurrent training environment where the students have a baseline degree of knowledge of how to complete tasks; it may not be appropriate for initial learning where students may deviate significantly from established procedures in the process of learning them, or where a large degree of free play is desired to let students fully explore the capabilities and functions of their complete environment.

Recommendations

The results of this study provide a starting point for a deeper and more profound understanding as to how VR tools can be leveraged to enhance aviation training. It leads to several specific recommendations for the target population, which could be implemented quickly and for immediate benefit and effect. It can also serve as a starting point for future studies by providing a methodology and research instrument that is validated and specifically adapted to aviation. Finally, it leads to five specific recommendations for future research that would further expand understanding of the effectiveness of VR training in aviation, and the benefits that it may have for more effective and targeted education of future technicians, system operators, and pilots.

Recommendations for the Target Population

The results of this study show that the use of a VR based recurrent training solution produced results that were equivalent to those for a group using the FFS. These results should be leveraged to expand the use of VR training in A&P training for both efficiency and cost reasons. This will serve to immediately open FFS slots for other training uses, and also to vastly increase the accessibility of A&P recurrent training by de-coupling it from access to an FFS.

The existing VR system as used in this study proved easy to use for the students and effective in imparting the knowledge and skills required for their recurrent training. It is available for immediate use, and it is recommended that it be operationally employed to examine how it could address the totality of recurrent training requirements for FAA certified requalification programs. The feedback loop that this would generate from post-training surveys and additional experimentation would provide vital information to allow for the further development and growth of this and similar systems. The FAA certificated A&P population would then benefit from a more rapid development of VR technologies that would be progressively more adapted to their needs. Additionally, the decoupling of training from the FFS infrastructure has the potential to radically change the accessibility of this training and to facilitate a fundamental restructuring of how it is managed.

The utilization of a VR system for A&P training comes with it a need to examine the effect of its use over time. Alongside the implementation of VR training programs for the A&P population, a longitudinal study should be conducted to ensure that the effects of the training on the overall competency of the population is maintained over time. This could be done concurrently with implementation and should form part of an overall monitoring and standardization program.

Recommendations for Future Research Methodology

The present study used trained evaluators to assess subject performance using the adapted GEARS scale. One of the strengths of new VR systems is that they can independently monitor trainee performance in both the physiological and cognitive domains and automate the scoring of a GEARS assessment. The automation of this assessment would increase accuracy of the tool, as well as remove any human evaluator failings such as bias and inattention. An equivalent study could be conducted using an automated scoring system both virtually and in the FFS using a mixed-reality version of the software to score both. An AI agent such as Chat GPT or its equivalent could then conduct a short verbal survey to study the affective domain of learning, and reporting could be completely automated.

Future research methodologies should also account for an adaptation period to the virtual environment, and in so doing could monitor the subjects' performance to assess their level of adaptation to the VR system. The results of this study suggest that performance can improve with any exposure to or familiarity with VR systems and it is highly desirable for future researchers to understand this effect within their study.

Recommendations for Future Research

There are five recommendations for further research to more fully understand and explore the effectiveness of VR learning in an aviation environment. This study's methodology can serve as a starting point for any of them, and each would expand these findings to allow for a more complete understanding of how we may best utilize virtual learning.

First, a study of the affective domain of learning using VR systems should be conducted in order to form a complete understanding of VR learning using Bloom's taxonomy. Understanding the affective domain will allow researchers to fully comprehend how students feel about using VR systems; this will in turn drive improvements in the overall learning experience. This would then further extend the body of knowledge by taking research already conducted on student's intentions to use VR training, and this study on its effectiveness, and expanding them to post-learning attitudes and impressions to yield a truly end-to-end modeling and understanding of VR's potential.

Second, additional research on an expanded range of tasks, both initial and recurrent learning, is recommended. The particular task selected for this study was well quantified in terms of cognitive and psychomotor elements, and this range of elements should be expanded to include all tasks contained within the competencies listed in Appendix E. Once a complete understanding of the effectiveness of VR learning on all associated competencies has been evaluated, it will be possible to generalize those results across the entire range of training that uses the same CBT tasks.

Third, this study did not examine the effect of personal interactions in the VR environment. The VR system used had the capability to display avatars of others using the virtual trainer, and real-time voice communication was possible. In order to standardize the scenario being used, and to eliminate the variability of receiving instructions and responses from another person also under training, this study used the computer generated second crew member with its defined set of responses and actions. A follow-on study should be conducted using the same methodology as in this study, but with the further addition of interpersonal interactions within the virtual environment to assess if there is an effect. This study could be combined with a study on affective learning which should include this aspect of the virtual environment.

Fourth, it is recommended that the effect of the continued use of VR systems over time be further studied. A longitudinal study conducted over multiple recurrent training periods is needed to assess whether there is any degradation of skills, in particular psychomotor, that is observed by continuous use of a virtual environment in place of the real world. This study could form part of a continuous quality monitoring program which would be a vital part of any aviation training program, and whose results would help identify if any training currency in the FFS or real aircraft was required. As regulations on the use of VR training evolve, and the effects of the virtual environment for learning are better understood, knowledge of the effect of VR on learners over time will be critical to effectively employ these systems.

Fifth, aircraft motion was neither simulated nor studied. For technicians who taxi the aircraft on the ground, or pilots who fly it, the perceived motion of the aircraft and the physical sensations provide a continuous source of information that is processed at the subconscious level and provide a powerful tool in learning and reacting to events. The FFS can simulate this for an entire crew, and with a VR system it is possible to simulate this for individuals using specially motorized chairs that can produce the same range of motion as the FFS at a small fraction of the acquisition and operating costs. It is recommended to study the effectiveness of a VR system, combined with a basic motion system, to assess the effectiveness of extending existing CBT training sequences that require aircraft motion onto a system that can provide it. The ability to add motion cues to a fully capable VR system and virtual world has the potential to supplant a good deal of FFS training, and this should be studied and understood due to the enormous potential economic benefits.

Appendix J provides a sample of costs that could be expected to be incurred in conducting similar research using a commercial FFS, and is provided to help future researchers plan and secure funding of the correct order of magnitude.

Summary

This study assessed whether a VR delivered engine-run recurrent training package could produce equivalent training results to the same training delivered in a traditional FFS setting. It evaluated a sample of FAA Certificated A&P technicians, taught with a dedicated VR recurrent training, to assess if they demonstrated equivalent task performance when compared to a control group taught in the FFS. Results were evaluated using a modified GEARS scale and time to task completion and were controlled for technician experience. The results found that those trained using the VR system demonstrated equivalent task performance when compared to those trained in the FFS. Those results were valid across both the psychomotor and cognitive domains, and the technician's total engine-run experience was found to have a significant effect on observed performance; all main hypotheses were rejected, and the null hypotheses retained. No safety or physiological incidents were reported. The results of this study validate the use of the tested VR system for A&P recurrent training, and its methodology may be used to expand both the depth and breadth of aviation knowledge on the effectiveness of VR training. Further research via longitudinal studies is recommended to better understand both continued use of VR in education, and to better understand an expanded role of VR in substituting a virtual environment for learning in a wider range of areas.

References

Abbott, K. (2017). Will a computer fly you on your vacation in the future? - Artificial intelligence, automation and autonomy in aviation [Paper presentation]. Society of Experimental Test Pilots 2017 Annual Symposium, Anaheim, CA, United States.

Abich, J., Parker, J., Murphy, J. S., & Eudy, M. (2021). A review of the evidence for training effectiveness with virtual reality technology. *Virtual Reality: The Journal of the Virtual Reality Society*, *25*(4), 919–933. https://doi.org/10.1007/s10055-020-00498-8

Ahmedyanova, G. (2017). Simulator as a tool of training to modern equipment management. *MATEC Web of Conferences, 129*. https://doi.org/10.1051/matecconf/201712906019

Airbus. (2016). A320 type rating course syllabus (Std 1.9).

Airbus. (2018). Commercial Aircraft Orders and Deliveries.

Airbus. (2019). Engine run-up recurrent course syllabus (TP Rev 7).

Airbus. (2020, July 24). Airbus reaches milestone number of aircraft flying in North America.Retreived from https://www.airbus.com/newsroom/press-releases/en/2017/09/airbus-reaches-milestone-number-of-aircraft-flying-in-north-amer.html

Airbus. (2021). Flight course training program: Competency based training and assessment/Evidence based training pilot handbook.

Aviation Maintenance Technician Schools, 14 C.F.R § 147 (2022). https://www.ecfr.gov/current/title-14/chapter-I/subchapter-H/part-147\

Black, J. B., & McClintock, R. O. (2017*). Interpretation Construction (ICON) design model*. SUNY Cortland. Retrieved 26 June 2017, from http://web.cortland.edu/frieda/id/IDtheories/9.html

Brady, T., Stolzer, A., Muller, B., & Schaum, D. (2001). A comparison of the learning styles of aviation and non-aviation college students. *Journal of Aviation/Aerospace Education & Research*, *11*(1). https://doi.org/10.15394/jaaer.2001.1286

Brainy Quote. (2019). Richard Branson quote. (2019). Retrieved 16 October 2019, from https://www.brainyquote.com/quotes/richard_branson_452106

Bürki-Cohen, J., Soja, N. N., & Longridge, T. (1998). Simulator platform motion-The need revisited. *The International Journal of Aviation Psychology*, *8*(3), 293–317. https://doi.org/10.1207/s15327108ijap0803_8

Cao, K., Zhang, Y., & Feng, J. (2024). Prediction of training cost and difficulty for aircraft-type transition based on similarity assessment. *Aerospace*, *11*(2), 166-. https://doi.org/10.3390/aerospace11020166

Carl, D. R. (2018). The shifting realities of performance improvement: VR, AR, MR. *Performance Improvement, 57*(4), 6-9. https://doi.org/10.1002/pfi.21774

Certification: Airmen Other Than Flight Crewmembers, 14 C.F.R § 65 (1962). https://www.ecfr.gov/current/title-14/chapter-I/subchapter-D/part-65

Certification: Pilots, Flight Instructors, and Ground Instructors, 14 C.F.R. § 61 (1997). https://www.ecfr.gov/current/title-14/chapter-I/subchapter-D/part-61

Chen, J., Cheng, N., Cacciamani, G., Oh, P., Lin-Brande, M., & Remulla, D., Gill, I., & Hung, A. J. (2019). Objective assessment of robotic surgical technical skill: A systematic review. *Journal of Urology*, *201*(3), 461-469. https://doi.org/10.1016/j.juro.2018.06.078

Chittaro, L., Corbett, C. L., McLean, G. A., & Zangrando, N. (2018). Safety knowledge transfer through mobile virtual reality: A study of aviation life preserver donning. *Safety Science, 102*, 159-168. https://doi.org/10.1016/j.ssci.2017.10.012

Clark, P. (2007). *Buying the big jets: Fleet planning for airlines* (2nd ed.). Ashgate.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.).Lawrence Erlbaum Associates, Publishers.

Competency-based training: The future of the aviation industry? (2017). Retrieved from https://www.civilaviation.training/pilot/competency-based-training-future-aviation/

Dakers, J. R. (2005). The hegemonic behaviorist cycle. *International Journal of Technology and Design Education, 15*(2), 111-126. https://doi.org/10.1007/s10798-005-8275-3

Fafard, A. (2020). Civil Simulator Census 2020. Retrieved 11 February 2021, from https://www.flightglobal.com/flight-international/civil-simulator-census-2020/138981.article

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149-1160. https://doi.org/10.3758/BRM.41.4.1149

Federal Aviation Administration. (2019). *Airmen certification.* United States Department of Transportation. https://www.faa.gov/licenses_certificates/airmen_certification/releasable_airmen_download/

Federal Aviation Administration. (2020). *Aviation instructor's handbook* (2020 ed.). Aviation Supplies & Academics, Inc.

Federal Aviation Administration. (2021). *Aviation Maintenance Technician Schools (AMTS).* United States Department of Transportation. https://www.faa.gov/licenses_certificates/airline_certification/amts/

Federal Aviation Administration. (2023). *U.S. civil airmen statistics*. United States Department of Transportation.

Franks, P., Hay, S., & Mavin, T. (2014). Can competency-based training fly?: An overview of key issues for ab initio pilot training. *International Journal of Training Research, 12*(2), 132-147. https://doi.org/10.1080/14480220.2014.11082036

Fussell, S. (2020). *Determinants of aviation students' intentions to use virtual reality for flight training* [Ph.D. dissertation, Embry-Riddle Aeronautical University]. Scholarly Commons. https://commons.erau.edu/edt/542/

Grande, T. (2015a, August 31). *MANCOVA in SPSS with the testing of assumptions* [Video]. YouTube.

Grande, T. (2015b, September 16). *Identifying multivariate outliers with Mahalanobis distance in SPSS* [Video]. YouTube.

Guthridge, R., & Clinton-Lisell, V. (2023). Evaluating the efficacy of virtual reality (VR) training devices for pilot training. *Journal of Aviation Technology and Engineering*, *12*(2). https://doi.org/10.7771/2159-6670.1286

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). Multivariate data analysis. Prentice Hall.

Hight, M. P., Fussell, S. G., Kurkchubasche, M. A., & Hummell, I. J. (2022). Effectiveness of virtual reality simulations for civilian, ab initio pilot raining. *Journal of Aviation/Aerospace Education & Research, 31*(1). https://doi.org/10.15394/jaaer.2022.1903

Hodge, S. (2007). The origins of competency-based training. *Australian Journal of Adult Learning*, *47*(2), 179-209.

Hoogenes, J., Wong, N., Al-Harbi, B., Kim, K. S., Vij, S., Bolognone, E., Quantz, M., Guo, Y., Shayegan, B., & Matsumoto, E. D. (2018). A randomized comparison of 2 robotic virtual reality simulators and evaluation of trainees' skills transfer to a simulated robotic urethrovesical anastomosis task. *Urology, 111*, 110-115. https://doi.org/10.1016/j.urology.2017.09.023

Ibáñez, M., & Delgado-Kloos, C. (2018). Augmented reality for STEM learning: A systematic review. *Computers & Education*, *123*, 109-123. https://doi.org/10.1016/j.compedu.2018.05.002

International Civil Aviation Organization. (2013). *Manual of evidence-based training* (1st ed.).

Jenab, K., Moslehpour, S., & Khoury, S. (2016). Virtual maintenance, reality, and systems: A review. *International Journal of Electrical and Computer Engineering, 6*(6), 2698-2707. https://doi.org/10.11591/ijece.v6i6.pp2698-2707

Jensen, L., & Konradsen, F. (2018). A review of the use of virtual reality head-mounted displays in education and training. Education and Information Technologies, 23(4), 1515-1529. doi:10.1007/s10639-017-9676-0

Jerald, J. (2016). *The VR book: Human-centered design for virtual reality*. Association for Computing Machinery. https://doi.org/10.1145/2897826.2927320

Kearns, S. K., Hodge, S., & Mavin, T. J. (2016). *Competency-based education in aviation: Exploring alternate training pathways*. Routledge. https://doi.org/10.4324/9781315563220

Kirkpatrick, D. L. (1959). Techniques for evaluating training programs: Pt.1. Reactions. *Journal of the American Society for Training and Development*, 13(11), 3–9.

Kirkpatrick, D. L., & Kirkpatrick, J. D. (2006). *Evaluating training programs: The four levels* (3rd ed.). Berrett-Koehler.

Kirkpatrick, J. D., & Kirkpatrick, W. K. (2016). *Kirkpatrick's four levels of training evaluation*. Association for Talent Development.

Laerd Statistics. (2019). One-way MANOVA in SPSS statistics - Step-by-step procedure with screenshots. https://statistics.laerd.com/spss-tutorials/one-way-manova-using-spss-statistics.php

Lampert, A., Freed, J., & Shepardson, D. (2021). Airlines scour the world for scarce 737 MAX simulators. *Reuters*. Retrieved 11 February 2021, from https://www.reuters.com/article/us-boeing-737max-training/airlines-scour-the-world-for-scarce-737-max-simulators-idUSKBN1ZL0EH

Li, S. (2023). Design and development of aviation aircraft maintenance training platform based on VR technology. *Procedia Computer Science*, *228*, 898–906. https://doi.org/10.1016/j.procs.2023.11.118

Macchiarella, N. D., Brady, T., & Lyon, B. S. (2008). An application of high fidelity FTDs for ab initio pilot training: The way ahead. *Collegiate Aviation Review*, *26*(1). https://doi.org/10.22488/okstate.18.100367

Macchiarella, N. D., Liu, D., Gangadharan, S. N., Vincenzi, D. A., & Majoros, A. E. (2005). Augmented reality as a training medium for aviation/aerospace application. Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting, 2174-2178. https://doi.org/10.1177/154193120504902512

Marzano, R. J., Norford, J.S., Finn, M., & Finn, D., III. (2017). *A handbook for personalized competency-based education*. Marzano Research.

Masiello, I., Herault, R., Mansfeld, M., & Skogqvist, M. (2022). Simulation-Based VR training for the nuclear sector—A pilot study. *Sustainability*, *14*(13). https://doi.org/10.3390/su14137984

McCullins, M. E. (2019). The evolution of pilot training. *Air Pilot*, *36,* 27-32.

Milgram, P., Takemura, H., Utsumi, A., & Kishino, F. (1995). Augmented reality: A class of displays on the reality-virtuality continuum. In H. Das (Ed.), *Proceedings of the Photonics for Industrial Applications Conference* , *2351*(1). SPIE Digital Library. https://doi.org/10.1117/12.197321

Mindell, D. (2015). *Our robots, ourselves*. Viking.

Morcke, A. M., Dornan, T., & Eika, B. (2013). Outcome (competency) based education: An exploration of its origins, theoretical basis, and empirical evidence. *Advances in Health Sciences Education, 18*(4), 851-863. https://doi.org/10.1007/s10459-012-9405-9

Morris, S. B. (2007). Estimating effect sizes from pre-test – post-test control group designs. *Organizational Research Methods, 11(2),* 364-386. https://doi.org/10.1177/1094428106291059

Neumann, E., Mayer, J., Russo, G., Amend, B., Rausch, S., Deininger, S., Harland, N., Anselmo da Costa, I., Hennenlotter, J., Stenzl, A., Kruck, S., & Bedke, J. (2019). Transurethral resection of bladder tumors: Next-Generation virtual reality training for surgeons. *European Urology Focus, 5*(5), 906-911. https://doi.org/10.1016/j.euf.2018.04.011

Oberhauser, M., & Dreyer, D. (2017). A virtual reality flight simulator for human factors engineering. *Cognition, Technology & Work, 19*(2), 263-277. https://doi.org/10.1007/s10111-017-0421-7

Pan, Z., Cheok, A., Yang, H., Zhu, J., & Shi, J. (2006). Virtual reality and mixed reality for virtual learning environments. *Computers & Graphics*, *30*(1), 20-28. http://dx.doi.org/10.1016/j.cag.2005.10.004

Preston, J. (2017). *Competence based education and training (CBET) and the end of human learning*. Springer. https://doi.org/10.1007/978-3-319-55110-4

Renganayagalu, S. K., Mallam, S. C., & Nazir, S. (2021). Effectiveness of VR head mounted displays in professional training: A systematic review. *Technology, Knowledge and Learning*, *26*(4), 999–1041. https://doi.org/10.1007/s10758-020-09489-9

Sánchez, R., Rodríguez, O., Rosciano, J., Vegas, L., Bond, V., Rojas, A., & Sanchez-Ismayel, A. (2016). Robotic surgery training: construct validity of Global Evaluative Assessment of Robotic Skills (GEARS). *Journal of Robotic Surgery*, *10*(3), 227–231. https://doi.org/10.1007/s11701-016-0572-1

Saposnik, G., Teasell, R., Mamdani, M., Hall, J., McIlroy, W., Cheung, D., & Bayley, M. (2010). Effectiveness of virtual reality using Wii gaming technology in stroke rehabilitation: a pilot randomized clinical trial and proof of principle. Stroke, 41(7), 1477-1484. Retrieved from https://www.ahajournals.org/doi/full/10.1161/strokeaha.110.584979

Schmidt, M. W., Kowalewski, K.-F., Schmidt, M. L., Wennberg, E., Garrow, C. R., Paik, S., Benner, L., Schijven, M. P., Müller-Stich, B. P., & Nickel, F. (2019). The Heidelberg VR score: Development and validation of a composite score for laparoscopic virtual reality training. *Surgical Endoscopy, 33*(7), 2093-2103. https://doi.org/10.1007/s00464-018-6480-x

Schulz, G. B., Grimm, T., Buchner, A., Jokisch, F., Casuscelli, J., Kretschmer, A., Mumm, J., Ziegelmuller, B., Stief, C., & Karl, A. (2019). Validation of a high-end virtual reality simulator for training transurethral resection of bladder tumors. *Journal of Surgical Education, 76*(2), 568-577. https://doi.org/10.1016/j.jsurg.2018.08.001

Seymour, N. E. (2008). VR to OR: A review of the evidence that virtual reality simulation improves operating room performance. World Journal of Surgery, 32(2), 182-188.

Sikorski, E., Palla, A., & Brent, L. (2017). Developing an immersive virtual reality aircrew training capability. In Proceedings of the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC).

Rupasinghe, T. D., Kurz, M. E., Gramopadhye, A. K., & Washburn, C. (2010). Improving aircraft maintenance technology education: Bloom's taxonomy perspective. *IIE Annual Conference. Proceedings,* 1.

Rutherford, P. (2009). Competencies undergo a review. *Training and Development in Australia, 36*(5), 25-28. ProQuest.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach’s alpha. *International Journal of Medical Education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Valimont, R. B., Gangadharan, S. N., Vincenzi, D. A., & Majoros, A. E. (2007). The effectiveness of augmented reality as a facilitator of information acquisition in aviation maintenance applications. *Journal of Aviation/Aerospace Education & Research, 16*(2). https://doi.org/10.15394/jaaer.2007.1478

Vogt, W. P., Gardner, D. C., Haeffele, L. M., & Ebrary, I. (2012). *When to use what research design* (1st ed.). Guilford Press.

Wang, M., Callaghan, V., Bernhardt, J., White, K., & Peña-Rios, A. (2018). Augmented reality in education and training: Pedagogical approaches and illustrative case studies. *Journal of Ambient Intelligence and Humanized Computing, 9*(5), 1391-1402. https://doi.org/10.1007/s12652-017-0547-8

Warne, R. (2014). A primer on multivariate analysis of variance (MANOVA) for behavioral scientists. *Practical Assessment, Research, and Evaluation, 19(17).*

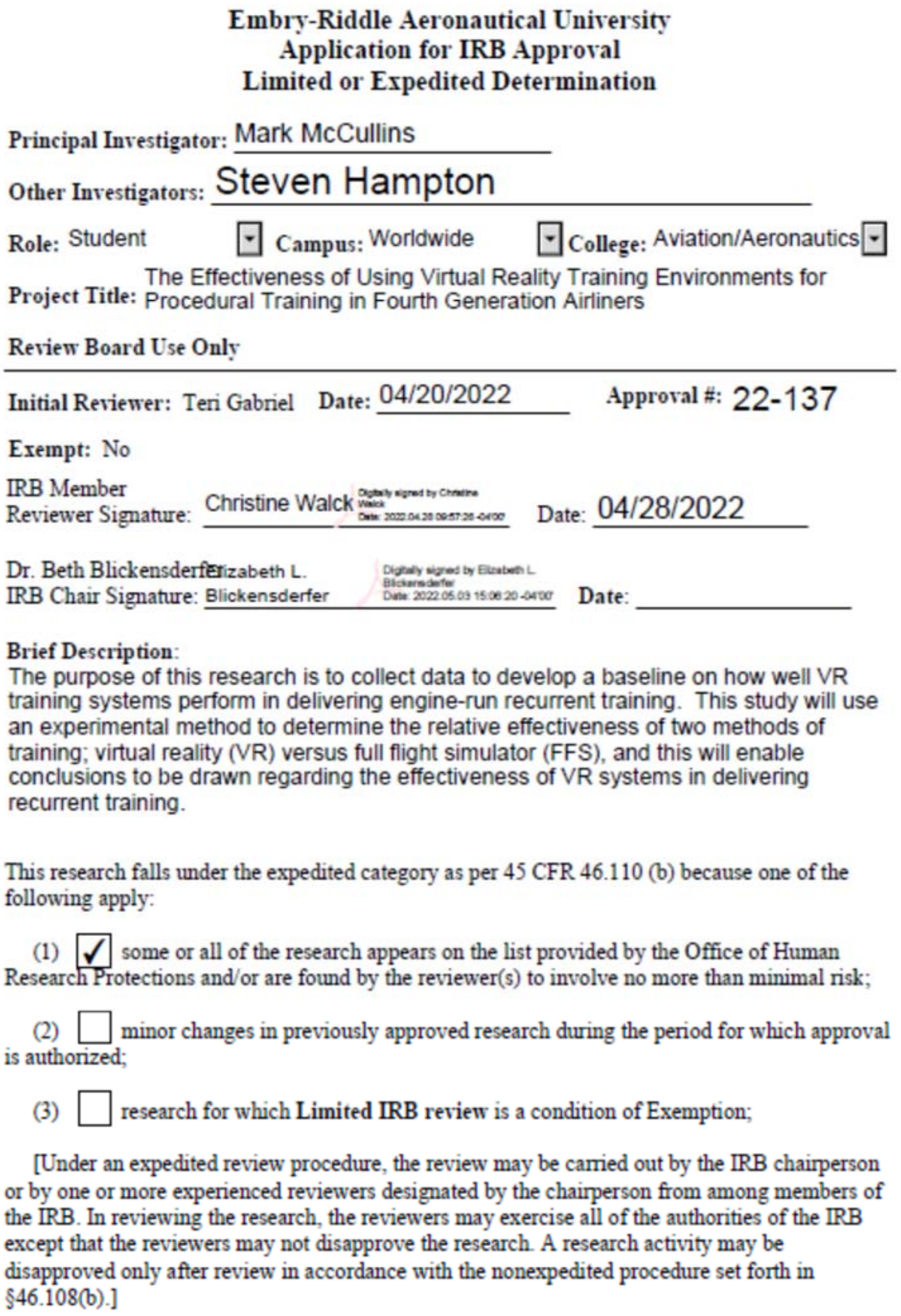
Wilde, G. J. S. (1998). Risk homeostasis theory: An overview. *Injury Prevention, 4*(2), 89-91. https://doi.org/10.1136/ip.4.2.89

Whitson, R. (2019). *Training in a modern age* [Unpublished master’s thesis]. Arizona State University.

Wu, S. (2018). *Psychological presence in immersive virtual environments.* [Master’s Thesis, San José State University]. https://doi.org/10.31979/etd.wcss-8s4p

Yuviler-Gavish, N., Krupenia, S., & Gopher, D. (2013). Task analysis for developing maintenance and assembly VR training simulators. Ergonomics in Design, 21(1), 12–19. https://doi.org/10.1177/1064804612463214

Appendix A: Permission to Conduct Research



Embry-Riddle Aeronautical University
Application for IRB Approval
Limited or Expedited Determination

Principal Investigator: Mark McCullins

Other Investigators: Steven Hampton

Role: Student **Campus:** Worldwide **College:** Aviation/Aeronautics

Project Title: The Effectiveness of Using Virtual Reality Training Environments for Procedural Training in Fourth Generation Airliners

Review Board Use Only

Initial Reviewer: Teri Gabriel **Date:** 04/20/2022 **Approval #:** 22-137

Exempt: No

IRB Member Reviewer Signature: Christine Walck Digitally signed by Christine Walck Date: 2022.04.28 09:57:26 -04'00' Date: 04/28/2022

Dr. Beth Blickensderfer IRB Chair Signature: Elizabeth L. Blickensderfer Digitally signed by Elizabeth L. Blickensderfer Date: 2022.05.03 15:06:20 -04'00' **Date:**

Brief Description:
The purpose of this research is to collect data to develop a baseline on how well VR training systems perform in delivering engine-run recurrent training. This study will use an experimental method to determine the relative effectiveness of two methods of training; virtual reality (VR) versus full flight simulator (FFS), and this will enable conclusions to be drawn regarding the effectiveness of VR systems in delivering recurrent training.

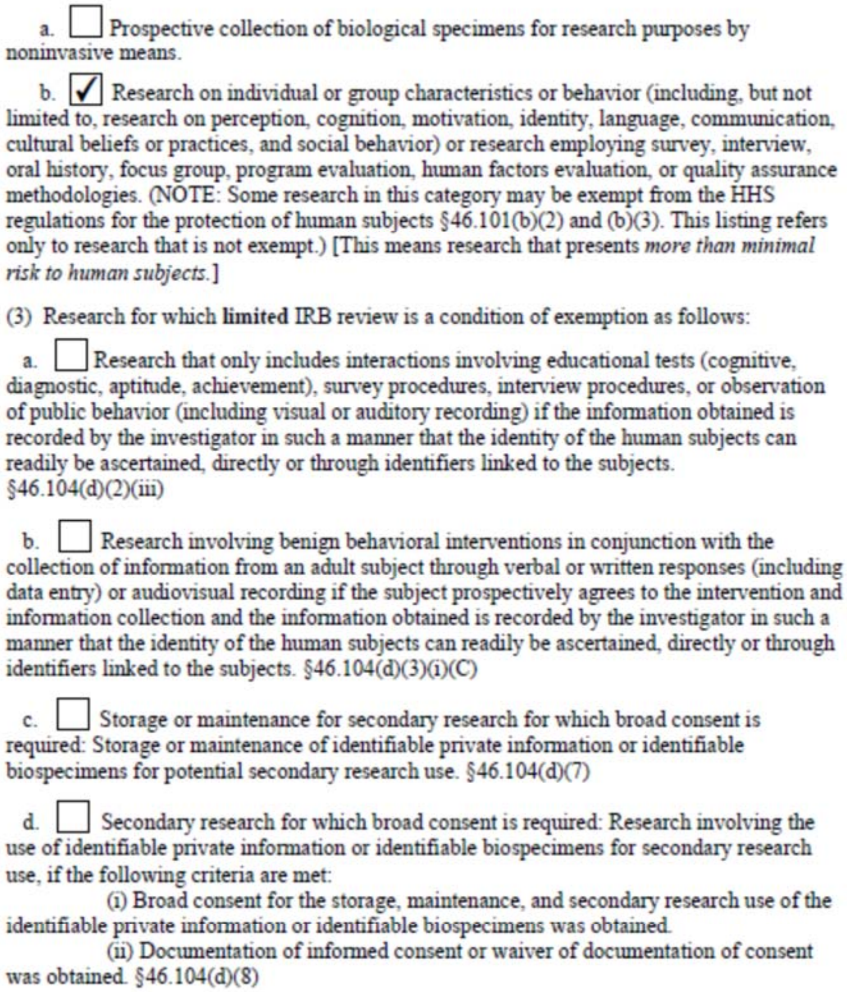
This research falls under the expedited category as per 45 CFR 46.110 (b) because one of the following apply:

(1) ☑ some or all of the research appears on the list provided by the Office of Human Research Protections and/or are found by the reviewer(s) to involve no more than minimal risk;

(2) ☐ minor changes in previously approved research during the period for which approval is authorized;

(3) ☐ research for which **Limited IRB review** is a condition of Exemption;

[Under an expedited review procedure, the review may be carried out by the IRB chairperson or by one or more experienced reviewers designated by the chairperson from among members of the IRB. In reviewing the research, the reviewers may exercise all of the authorities of the IRB except that the reviewers may not disapprove the research. A research activity may be disapproved only after review in accordance with the nonexpedited procedure set forth in §46.108(b).]



a. ☐ Prospective collection of biological specimens for research purposes by noninvasive means.

b. ☑ Research on individual or group characteristics or behavior (including, but not limited to, research on perception, cognition, motivation, identity, language, communication, cultural beliefs or practices, and social behavior) or research employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies. (NOTE: Some research in this category may be exempt from the HHS regulations for the protection of human subjects §46.101(b)(2) and (b)(3). This listing refers only to research that is not exempt.) [This means research that presents *more than minimal risk to human subjects.*]

(3) Research for which **limited** IRB review is a condition of exemption as follows:

a. ☐ Research that only includes interactions involving educational tests (cognitive, diagnostic, aptitude, achievement), survey procedures, interview procedures, or observation of public behavior (including visual or auditory recording) if the information obtained is recorded by the investigator in such a manner that the identity of the human subjects can readily be ascertained, directly or through identifiers linked to the subjects. §46.104(d)(2)(iii)

b. ☐ Research involving benign behavioral interventions in conjunction with the collection of information from an adult subject through verbal or written responses (including data entry) or audiovisual recording if the subject prospectively agrees to the intervention and information collection and the information obtained is recorded by the investigator in such a manner that the identity of the human subjects can readily be ascertained, directly or through identifiers linked to the subjects. §46.104(d)(3)(i)(C)

c. ☐ Storage or maintenance for secondary research for which broad consent is required: Storage or maintenance of identifiable private information or identifiable biospecimens for potential secondary research use. §46.104(d)(7)

d. ☐ Secondary research for which broad consent is required: Research involving the use of identifiable private information or identifiable biospecimens for secondary research use, if the following criteria are met:

(i) Broad consent for the storage, maintenance, and secondary research use of the identifiable private information or identifiable biospecimens was obtained.

(ii) Documentation of informed consent or waiver of documentation of consent was obtained. §46.104(d)(8)

Appendix B: Informed Consent Form

The Effectiveness of Virtual Reality Training Environments in Aviation

Purpose of this Research: We are asking you to take part in a research project for the purpose of collecting data to develop a baseline on how well Virtual Reality (VR) training systems perform in delivering engine-run recurrent training. You will be asked to complete a short training session using a VR trainer that consists of VR goggles and hand controllers. Following this, you will be asked to perform the task that you were trained on in the full flight simulator, and your performance will be observed and recorded on an evaluation form. The total time of your participation is estimated to be about 30 minutes.

Risks or discomforts: The risks of participating in this study are no greater than what is experienced in using commercial VR goggles. This can include eye strain, headaches, and motion sickness. If you experience any discomfort you may pause the training at any time or discontinue the session.

Benefits: Participation in this research will provide you with approximately 20 minutes of additional training in engine-run procedures and techniques using an advanced training system, and approximately 10 additional minutes of Full Flight simulator usage.

Confidentiality of records: Your individual information will be protected in all data resulting from this study. While the members of the research team will have access to your personal information, publication of the data will not include any identifying information. You will be assigned a number; the key code will be stored separately from the data. Information collected as part of this research ***will not be used*** or ***distributed*** for future research studies.

Contact: If you have any questions or would like additional information about this study, please contact the principal researcher, Mark McCullins, mark.m.mccullins@airbus.com, or the faculty member overseeing this project, Dr. S. Hampton, hamptons@erau.edu. For any concerns or questions as a participant in this research, contact the Institutional Review Board (IRB) at 386-226-7179 or via email teri.gabriel@erau.edu.

Voluntary Participation: Your participation in this study is completely voluntary. You may discontinue your participation at any time without penalty or loss of benefits to which you are otherwise entitled. Should you wish to discontinue the research at any time, no information collected will be used.

CONSENT. By signing below, I certify that I am an FAA certificated A&P technician, a resident of the U.S. and I am 18 years of age or older. I further verify that I understand the information on this form, that the researcher has answered any and all questions I have about this study, and I voluntarily agree to participate in the study.

Signature of Participant____Date: _________

Printed Name of Participant ____

Appendix C: VR Study Participant Information Form

Name (Last, First): ______________________________

ID Number: ______________________________

DETACH FORM HERE ONCE COMPLETED AND ID NUMBER ASSIGNED

DO NOT STORE WITH PARTICIPANT NAME

ID Number:

Age: ______________________________

Gender: ______________________________

Engine run experience (years): ______________________________

Aircraft types qualified on: ______________________________

A&P experience (years): ______________________________

Previous VR experience? Yes / No (Circle one)

Appendix D: A320 Normal Start Procedure and Start Fault Checklist

NORMAL ENGINE START PROCEDURE (Approx 8 minutes)

The FFS or VR trainer will be set up for a normal engine start. The subjects will verify the settings per applicable Airbus SOPs and will use the real or virtual iPad to run the following checklist. During the second engine start a HOT START fault will be injected, and the subjects will switch to and complete the **ENG 2 EGT OVERLIMIT** ECAM procedure. This will add another 4 minutes to the sequence. Upon successful securing of the second engine the scenario will be considered complete.

Ident.: PRO-NOR-SOP-08-00010162.0060001 / 04 MAR 14

Use the automatic engine start procedure in most circumstances. However, if the start aborts due to insufficient starter inlet air pressure (e.g., on high airfields, or in case of low pressure from an external pneumatic power group), it is recommended to use the manual start procedure, instead the automatic procedure.
If, during the engine start, the ground crew reports a fuel leak from the engine drain mast, run the engine at idle for 5 min. If the leak disappears during these 5 min, the aircraft can be dispatched without maintenance action. If the leak is still present after 5 min, maintenance action may be required before the flight.

ENG MODE selector.. IGN/START

The lower ECAM displays the ENG SD page.

START ENGINE 2...ANNOUNCE

Engine 2 is usually started first. It powers the yellow hydraulic system that pressurizes the parking brake.

ENG 2 MASTER sw..ON

- Do not turn the ENG 2 MASTER sw ON before all amber crosses and messages have disappeared on the engine parameters (upper ECAM display).
- Parameter callouts are not mandatory.
- In case the electrical power supply is interrupted during the start sequence (indicated by the loss of ECAM DUs), abort the start by switching OFF the ENG 2 MASTER sw. Then, perform a 30 s dry crank.

ON ECAM UPPER DISPLAY ON ECAM LOWER DISPLAY
N2 increases Corresponding start valve in line.
Bleed pressure indication green.
Oil pressure increases.
At 16 % N2 Indication of the active igniter (A or B).
At 22 % N2
  - FF increases
15 s (maximum) after fuel is on
  - EGT increases
  - N1 increases
At 50 % N2 Start valve starts closing. (It is fully closed between 50 % and 56 % N2).
Igniter indication off.

When idle is reached (AVAIL indication is displayed):

ENG IDLE PARAMETERS.. CHECK

At ISA sea level : N1 about 19.5 %
N2 about 58.5 %
EGT about 390 °C
FF about 275 kg/h (600 lb/h)
Grey background on N2 indication disappears.

START ENGINE 1... ANNOUNCE

ENG 1 MASTER sw.. ON

Same procedure as for engine 2.
Both pack valves reopen with 30 s delay after the second engine N2 is above 50 %.
Note: A PTU FAULT is triggered, if the second engine is started within 40 s following the end of the cargo doors operation.

ENG 1(2) START FAULT

Ident.: PRO-ABN-70-AD-NG01333

ANNUNCIATIONS

Triggering Conditions:

This alert triggers when start fault due to:

- No light up, or
- Engine stall, or
- **Engine overtemperature (above 725 °C),** or
- Starter time exceeded, or
- Low start air pressure, or
- Thrust lever not at idle.

ENG 1(2) IGNITION FAULT:

The engine does not start within the 18 s that follow the ignition start.

On ground (auto start):

If the engine does not start, the FADEC can attempt an additional engine restart. After any start attempt that is not successful, a dry crank phase automatically occurs. The ECAM displays the following messages:

L1 AUTO CRANK IN PROGRESS
NEW START IN PROGRESS

When the final dry crank process is completed:

ENG MASTER (AFFECTED)..OFF

After the starter cools, and for any subsequent attempt to start the engine, the flight crew must decide whether to attempt auto or manual engine start, or must report the "no start condition" to maintenance for appropriate action.

ENG 1(2) STALL, ENG 1(2) EGT OVERLIMIT:

On ground (auto start):

If the FADEC detects a stall or a potential EGT overheat, the FADEC will reduce the fuel schedule in stages, if necessary, to achieve a normal condition. The following message will be displayed on the ECAM:

NEW START IN PROGRESS

If restart not possible:

If normal conditions cannot be achieved, the FADEC shuts off fuel and turn off ignition. Then a dry crank phase automatically occurs. The ECAM displays the following message:

AUTO CRANK IN PROGRESS
ENG MASTER (AFFECTED)...OFF

- The fuel metering valve and starter air valve are automatically closed. Both igniters are turned off
- Setting ENG MASTER to OFF confirms automatic start abort
- In case of ENG STALL, consider making a X BLEED start, if pressure is low. flight crew must decide whether to attempt auto or manual engine start, or must report the "no start condition" to maintenance for appropriate action.

STARTER TIME EXCEEDED:

MAN START (IF MANUAL START IS PERFORMED)...OFF

ENG MASTER (AFFECTED)...OFF

LO START AIR PRESS:
BLEED AIR SUPPLY.. CHECK

Appendix E: Airbus Competency List and Definitions (Airbus, 2021)

Application of knowledge (KNO)	
Description:	Demonstrates knowledge and understanding of relevant information, operating instructions, aircraft systems and the operating environment
OB 0.1	Demonstrates practical and applicable knowledge of limitations and systems and their interaction
OB 0.2	Demonstrates required knowledge of published operating instructions
OB 0.3	Demonstrates knowledge of the physical environment, the air traffic environment including routings, weather, airports and the operational infrastructure
OB 0.4	Demonstrates appropriate knowledge of applicable legislation.
OB 0.5	Knows where to source required information
OB 0.6	Demonstrates a positive interest in acquiring knowledge
OB 0.7	Is able to apply knowledge effectively

Application of procedures and compliance with regulations (PRO)	
Description:	Identifies and applies appropriate procedures in accordance with published operating instructions and applicable regulations
OB 1.1	Identifies where to find procedures and regulations
OB 1.2	Applies relevant operating instructions, procedures and techniques in a timely manner
OB 1.3	Follows SOPs unless a higher degree of safety dictates an appropriate deviation
OB 1.4	Operates aircraft systems and associated equipment correctly
OB 1.5	Monitors aircraft systems status
OB 1.6	Complies with applicable regulations
OB 1.7	Applies relevant procedural knowledge

Communication (COM)	
Description:	Communicates through appropriate means in the operational environment, in both normal and non-normal situations
OB 2.1	Determines that the recipient is ready and able to receive information
OB 2.2	Selects appropriately what, when, how and with whom to communicate
OB 2.3	Conveys messages clearly, accurately and concisely
OB 2.4	Confirms that the recipient demonstrates understanding of important information
OB 2.5	Listens actively and demonstrates understanding when receiving information
OB 2.6	Asks relevant and effective questions
OB 2.7	Uses appropriate escalation in communication to resolve identified deviations
OB 2.8	Uses and interprets non-verbal communication in a manner appropriate to the organisational and social culture
OB 2.9	Adheres to standard radiotelephone phraseology and procedures
OB 2.10	Accurately reads, interprets, constructs and responds to datalink messages in English

Aeroplane flight path management — automation (FPA)	
Description:	Controls the flight path through automation
OB 3.1	Uses appropriate flight management, guidance systems and automation, as installed and applicable to the conditions
OB 3.2	Monitors and detects deviations from the intended flight path and takes appropriate action
OB 3.3	Manages the flight path to achieve optimum operational performance
OB 3.4	Maintains the intended flight path during flight using automation whilst managing other tasks and distractions
OB 3.5	Selects appropriate level and mode of automation in a timely manner considering phase of flight and workload
OB 3.6	Effectively monitors automation, including engagement and automatic mode transitions

Aeroplane flight path management — manual control (FPM)	
Description:	Controls the flight path through manual control
OB 4.1	Controls the aircraft manually with accuracy and smoothness as appropriate to the situation
OB 4.2	Monitors and detects deviations from the intended flight path and takes appropriate action
OB 4.3	Manually controls the aeroplane using the relationship between aeroplane attitude, speed and thrust, and navigation signals or visual information
OB 4.4	Manages the flight path to achieve optimum operational performance
OB 4.5	Maintains the intended flight path during manual flight whilst managing other tasks and distractions
OB 4.6	Uses appropriate flight management and guidance systems, as installed and applicable to the conditions
OB 4.7	Effectively monitors flight guidance systems including engagement and automatic mode transitions

Leadership & teamwork (LTW)	
Description:	Influences others to contribute to a shared purpose. Collaborates to accomplish the goals of the team
OB 5.1	Encourages team participation and open communication
OB 5.2	Demonstrates initiative and provides direction when required
OB 5.3	Engages others in planning
OB 5.4	Considers inputs from others
OB 5.5	Gives and receives feedback constructively
OB 5.6	Addresses and resolves conflicts and disagreements in a constructive manner
OB 5.7	Exercises decisive leadership when required
OB 5.8	Accepts responsibility for decisions and actions
OB 5.9	Carries out instructions when directed
OB 5.10	Applies effective intervention strategies to resolve identified deviations
OB 5.11	Manages cultural and language challenges, as applicable

Problem-solving — decision-making (PSD)	
Description:	Identifies precursors, mitigates problems, and makes decisions
OB 6.1	Identifies, assesses and manages threats and errors in a timely manner
OB 6.2	Seeks accurate and adequate information from appropriate sources
OB 6.3	Identifies and verifies what and why things have gone wrong, if appropriate
OB 6.4	Perseveres in working through problems whilst prioritising safety
OB 6.5	Identifies and considers appropriate options
OB 6.6	Applies appropriate and timely decision-making techniques
OB 6.7	Monitors, reviews and adapts decisions as required
OB 6.8	Adapts when faced with situations where no guidance or procedure exists
OB 6.9	Demonstrates resilience when encountering an unexpected event

Situation awareness and management of information (SAW)	
Description:	Perceives, comprehends and manages information and anticipates its effect on the operation
OB 7.1	Monitors and assesses the state of the aeroplane and its systems
OB 7.2	Monitors and assesses the aeroplane’s energy state, and its anticipated flight path
OB 7.3	Monitors and assesses the general environment as it may affect the operation
OB 7.4	Validates the accuracy of information and checks for gross errors
OB 7.5	Maintains awareness of the people involved in or affected by the operation and their capacity to perform as expected
OB 7.6	Develops effective contingency plans based upon potential risks associated with threats and errors
OB 7.7	Responds to indications of reduced situation awareness

Workload management (WLM)	
Description:	Maintains available workload capacity by prioritising and distributing tasks using appropriate resources
OB 8.1	Exercises self-control in all situations
OB 8.2	Plans, prioritises and schedules appropriate tasks effectively
OB 8.3	Manages time efficiently when carrying out tasks
OB 8.4	Offers and gives assistance
OB 8.5	Delegates tasks
OB 8.6	Seeks and accepts assistance, when appropriate
OB 8.7	Monitors, reviews and cross-checks actions conscientiously
OB 8.8	Verifies that tasks are completed to the expected outcome
OB 8.9	Manages and recovers from interruptions, distractions, variations and failures effectively while performing tasks

Appendix F: Data Collection Device

GEARS SCALE ADAPTED FOR FFS USE – Participant ID: ____________

Proprioception– Ability to correctly locate switches and controls (Psychomotor)

1	2	3	4	5
Constantly searches for target, wide sweeps, slow to locate.		Some searching or missing target, but quick to correct.		Accurately locates target without searching or overshooting.

OB 1.4
OB 4.1

Dexterity – Ability to correctly manipulate switches and pushbuttons (Psychomotor)

1	2	3	4	5
Frequent errors in switch position or manipulation that require prompting to correct.		Occasional errors in switch manipulation that are detected and corrected.		Expertly manipulates switches and controls without error.

OB 1.4
OB 1.5
OB 3.1
OB 4.1

Mastery of Aircraft Perceptual Environment – Perceives, notices, and reacts to indications and warnings (Psychomotor)

1	2	3	4	5
Consistently does not optimize view or hand position even with guidance.		View is sometimes not optimal, and hands are not positioned to intervene as necessary.		View and hand position are optimal for observing indications, recording data, and intervening as necessary.

OB 0.1
OB 1.4
OB 1.5
OB 1.7
OB 2.3
OB 3.1
OB 4.1

Efficiency – Correct application of procedure without delay (Cognitive)

1	2	3	4	5
Inefficient efforts; many tentative movements; constantly changing focus or persisting without progress.		Slow, but planned movements are reasonably organized.		Confident, efficient and safe conduct, maintains focus on task, fluid progression.

OB 0.1
OB 0.2
OB 0.7
OB 1.3
OB 1.4
OB 1.5
OB 1.7
OB 2.3
OB 3.1
OB 4.1
OB 6.5
OB 7.1

Use of Checklist – Correct use and knowledge of procedural information (Cognitive)

1	2		4	5
Procedural errors that are not corrected, or inappropriate checklist use of manipulation.		Minor procedural errors that are caught and corrected. References checklist appropriately.		Consistently and correctly follows procedure, and references checklist appropriately.

OB 0.1
OB 0.2
OB 0.7
OB 1.3
OB 1.4
OB 1.5
OB 1.7
OB 2.3
OB 3.1
OB 4.1
OB 6.5
OB 7.1

Autonomy (Cognitive)

1	2	3	4	5
Unable to complete entire task, even with verbal guidance.		Able to complete task safely with moderate verbal guidance.		Able to complete task independently without verbal prompting.

OB 0.1
OB 0.7
OB 1.7
OB 6.5
OB 8.8

Time to task completion (mm:ss):
Aggregate GEARS Score (6-30):
Psychomotor GEARS Score (3-15):
Cognitive GEARS Score (3-15):

GEARS Scoring Notes

All six dimensions of the GEARS scale are designed to be scored in the same manner. Each score must be a whole number; half numbers are not allowed. Scores of 5 are indicative of an expert performance, as one would expect from a trained professional, and scores of 1 indicate that the subject is essentially unable to perform the task. A score of three represents the minimum desired performance. Intermediate scores such as a 2 or 4, are used to indicate that the observed behavior falls in between the definitions provided on the scale.

A score of five indicates a close to 100% success rate in that dimension, as indicated by words such as accurately, expertly, consistently, and confidently. A score of three would indicate that the minimum standard of performance was met, but not necessarily smoothly as indicated by language that refers to slow speed, searching, and sub-optimal placement of hands or eyes. The third column uses descriptors such as occasional and some to quantify the number of errors and specifies that all critical errors must be caught and corrected. A score of one would be given in the case of many errors, lack of knowledge, inefficiency that interferes with task completion, and an inability to complete the task safely.

The score of four is meant to be used when several errors are observed, but not so many as to meet the intent of the wording, ‘some’ or ‘occasional.’ It could also be used to indicate slight inconsistencies in an otherwise expert performance. A score of two would indicate some degree of capacity to complete the task beyond what a score of one would indicate, but the performance would not yet meet the minimum standard; it could also come from a performance that would have merited a three if it were not for one or several errors that were uncorrected.

Appendix G: Experimental Apparatus

Two variations of equipment were used in this study, both of which connected to the same VR system and provided equivalent VR performance. A desktop computer system was used for the pilot study subjects and was replaced by a laptop run system for the main study subjects for obsolescence management and portability issues.

The initial desktop setup included a Desktop HP Z4 computer, using an i7-7800X 7th Gen Intel CPU with 16GB DDR4 RAM and a GTX 1080 GPU. The Laptop that replaced this was an MSI Vector GP76 computer, using an i7 12700H 12th Gen Intel CPU with 32GB DDR5 RAM and an RTX 3070Ti GPU. Both setups ran the same VR headset, which was an Acer AH101 Windows Mixed Reality Headset, with 1440x1440 resolution per eye, and a maximum refresh rate of 90Hz. Figure 24 shows the final experimental apparatus.

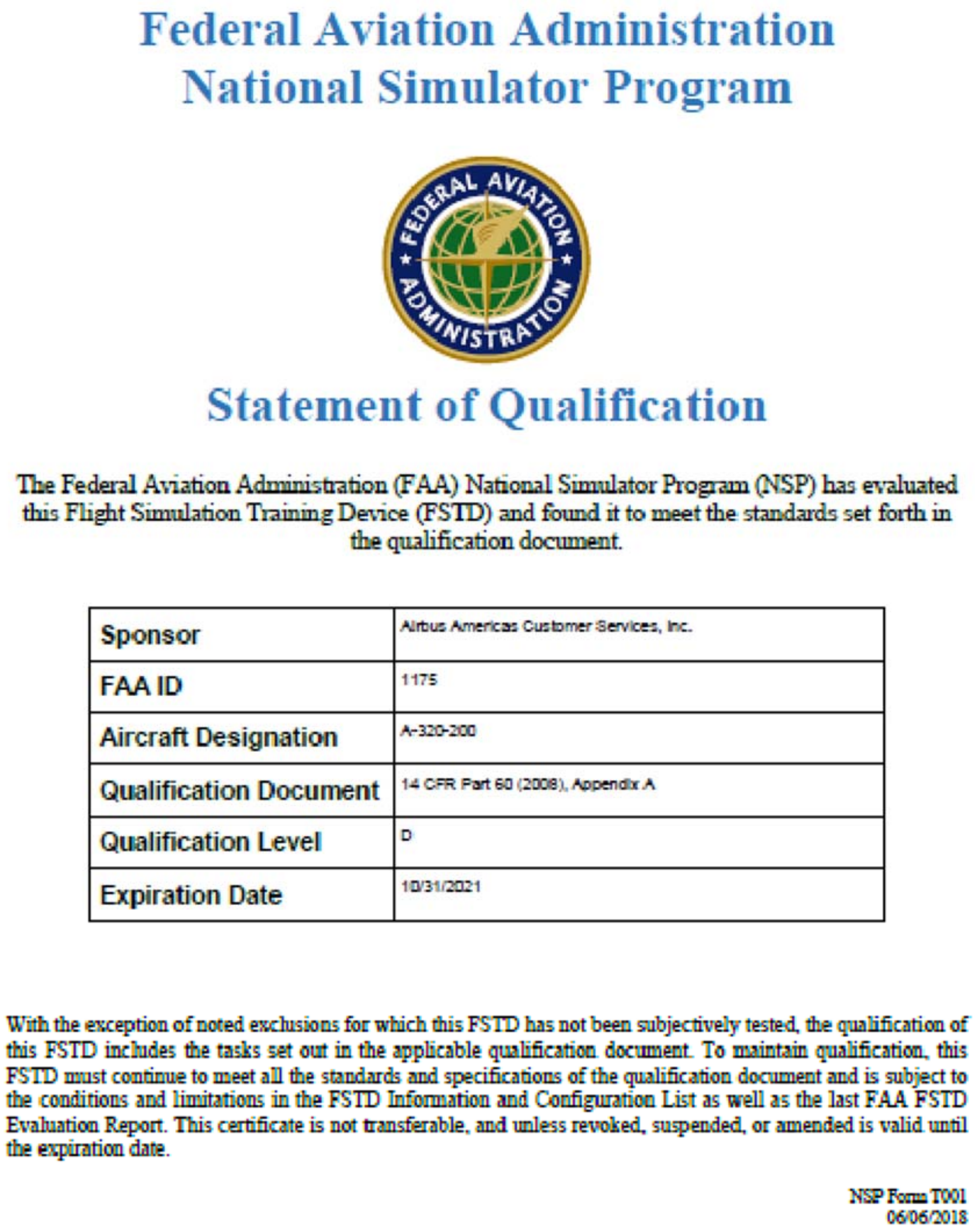
The GPU Load on the initial GTX 1080 was never limiting, as the VR Software being used was not particularly demanding on the GPU. Either system could have adequately presented the training software for the purposes of this study. The system change to the laptop did not have any effect on the refresh rate and quality since the headset was limited to 90Hz regardless of what GPU is being used. CPU and RAM differences did not change the VR system performance, and the experience of the subjects was assessed to be identical between the two systems.

Figure 24

Experimental Apparatus with Victor Liriano and Erik Marrero

Note. Copyright 2024 by M. E. McCullins

Appendix H: Full Flight Simulator Certification Certificate



Federal Aviation Administration
National Simulator Program

Statement of Qualification

The Federal Aviation Administration (FAA) National Simulator Program (NSP) has evaluated this Flight Simulation Training Device (FSTD) and found it to meet the standards set forth in the qualification document.

Sponsor	Airbus Americas Customer Services, Inc.
FAA ID	1175
Aircraft Designation	A-320-200
Qualification Document	14 CFR Part 60 (2008), Appendix A
Qualification Level	D
Expiration Date	10/31/2021

With the exception of noted exclusions for which this FSTD has not been subjectively tested, the qualification of this FSTD includes the tasks set out in the applicable qualification document. To maintain qualification, this FSTD must continue to meet all the standards and specifications of the qualification document and is subject to the conditions and limitations in the FSTD Information and Configuration List as well as the last FAA FSTD Evaluation Report. This certificate is not transferable, and unless revoked, suspended, or amended is valid until the expiration date.

NSP Form T001
06/06/2018

Appendix I: Scatterplots of Dependent Variables and Covariates

A scatterplot for each DV and covariate was generated using SPSS and presented in Chapter IV. These figures provide a larger view of each DV individually plotted against the Covariates (Experience AandP and Experience ERun).

Figure 25

ScatterPlot of Total GEARS Score and Covariates

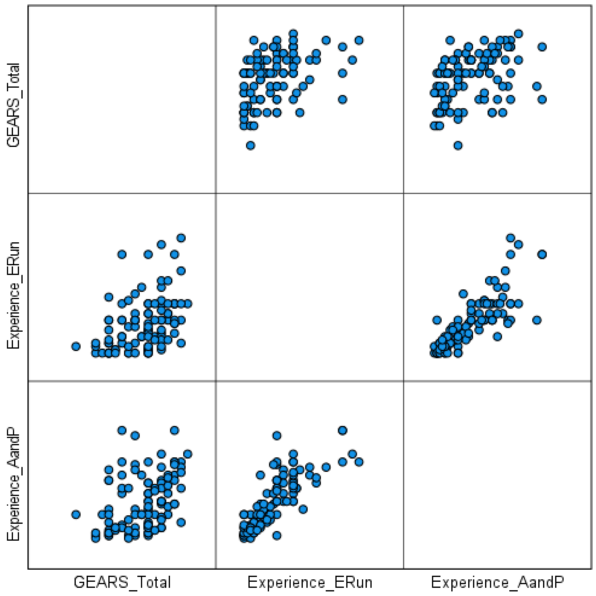


Figure 26

ScatterPlot of GEARS Cognitive Score and Covariates

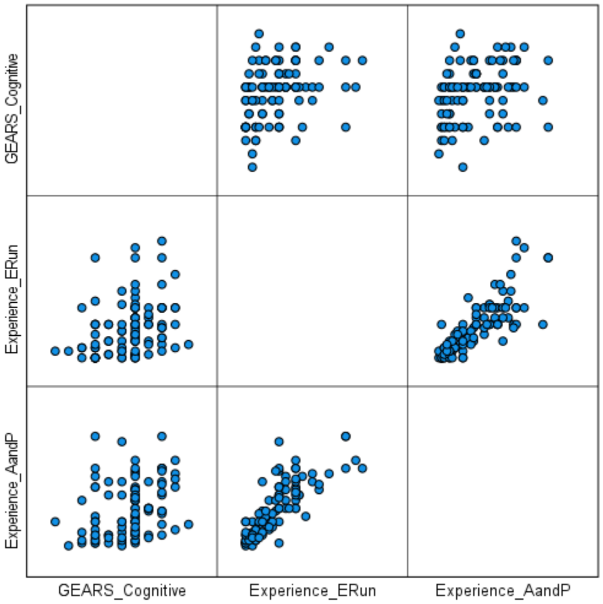


Figure 27

ScatterPlot of GEARS Psychomotor Score and Covariates

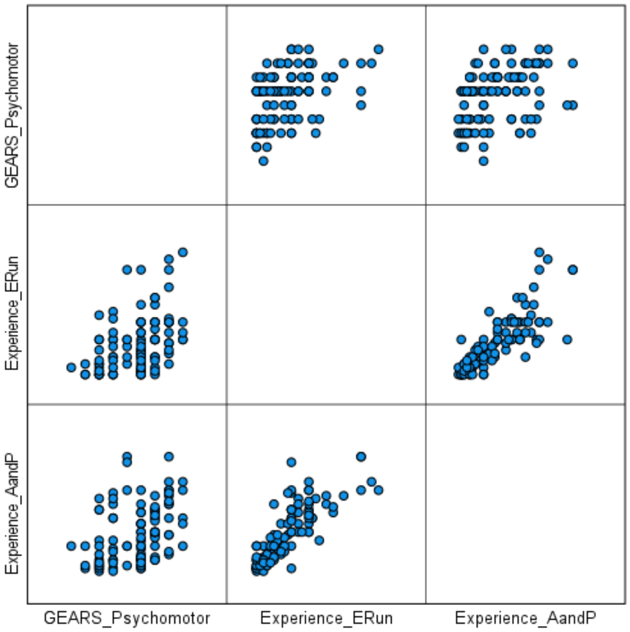
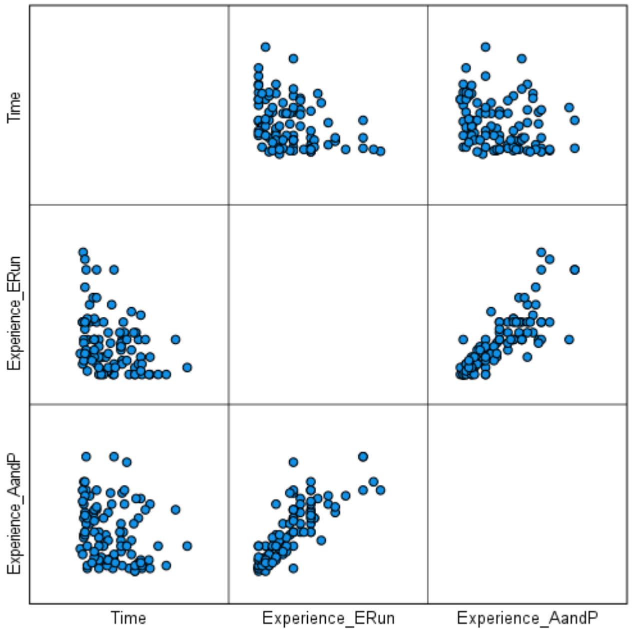


Figure 28

ScatterPlot of Time and Covariates



Appendix J: Study Cost

For those who would contemplate conducting an experiment such as this one, it may be useful to have some idea on the commercial costs involved and the magnitude of effort. These figures are approximate, and intended to be of some use in the event that grants, or funding are sought for future research. Extensive use of existing infrastructure kept costs for this study down, as did the synergies gained from using real-life training. Researchers should keep in mind the need to account for factors such as these when planning studies remotely from existing research facilities.

Each participant required approximately two hours of time to process and complete paperwork, and it was possible to process two subjects simultaneously. This means that a minimum of 100 hours was required for the actual experimental study by the researchers, outside of training and reporting which required another 20 hours. While this study gained synergies from using real students and their instructors, a stand-alone study would need to budget at least 120 hours of work from a single research assistant.

The laptop computers and VR systems are commercially available, and a comparable system to the one used could be built for approximately $US 10,000. The hourly rate for the FFS varies, but a figure of approximately $US 1,500 can be used to estimate the cost for the simulator and operator. Approximately 34 FFS hours were used in total for evaluations.

Table 16 summarizes the basic costs for this study, with a final total of $US 64,400 for the basic equipment and facilities. This should serve as a template for those seeking funding to perform similar research in the future.

Table 16

Cost Summary

Item	Hours	Cost per Hour ($US)	Total ($US)
Research Assistant	120	20	2400
FFS	34	1500	51000
Supplies	N/A	N/A	1000
VR Laptops and Equipment	N/A	N/A	10000
Total			64400