The Aeronautical Journal (2025), 1–20
doi:10.1017/aer.2025.10086

RESEARCH ARTICLE

The effectiveness of using virtual reality training environments for procedural training in fourth-generation airliners

M.E. McCullins[1], S. Hampton[1], S.G. Fussell[2], K. Kiernan[3] and J. Thropp[1]

[1]College of Aviation, Embry Riddle Aeronautical University, Daytona Beach, FL, USA
[2]Aptima Inc. Training, Learning, and Readiness Division, Fairborn, OH, USA
[3]Boeing Center for Aviation and Aerospace Safety, Embry Riddle Aeronautical University, Daytona Beach, FL, USA
Corresponding author: M.E. McCullins; Email: mccullim@erau.edu

Received: 9 June 2025; **Revised:** 28 September 2025; **Accepted:** 30 September 2025

Keywords: flight simulation; virtual reality; VR; aviation training

Abstract

This article examines the effectiveness of using virtual reality training environments for procedural training in fourth generation airliners. It is based on a study that assessed whether the training outcomes from a current recurrent training course for FAA certificated airframe and power plant 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 programme and was assessed in the FFS, while the second group was trained using a VR device and was subsequently assessed in the FFS. Training effectiveness was assessed using a modified Global Evaluative Assessment of Robotic Skills (GEARS) tool that measured both cognitive and psychomotor aspects of learning alongside the time to successful completion of the assessed task was also measured. The population sampled for the study were all Federal Aviation Administration (FAA) certificated airframe and power plant technicians who were engine-run qualified; a total sample of 100 was used to achieve a 95% confidence interval ($p < 0.05$). The hypothesis under test was that there is no difference in test performance between the three groups. A multivariate analysis of covariance (MANCOVA) analysis was performed using the GEARS scores and time to completion as variables, and the null hypothesis was retained. The VR system, as tested, was found to provide equivalent task performance to the traditional training method. Recommendations for future research and ongoing application of the specific experimental methodology were provided.

Nomenclature

AMT	aviation maintenance technician
A&P	airframe and powerplant certificated technician
AR	augmented reality
AViATE	Aerospace Virtual reality Assessment of Training Effectiveness
CBT	competency-based training
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
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 organisation
PTT	part-task trainer
SPSS™	Statistical Package for Social Science (IBM – Software package V28.0.0.0)
TAM	technology acceptance model
VR	virtual reality

1.0 Introduction

Aircrew and aircraft maintenance technician 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. 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 programmes the final stage of training takes place in the full flight simulator (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 qualified by using only the FFS; such is their similarity to the real world [4].

1.1 The regulatory environment of aviation training

Training for licensed and/or certified aviation crewmembers and technicians is governed by the Federal Aviation Administration (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. Training organisations are closely regulated, 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 [12]. If it is desired to deviate from an already approved course, or method of instruction, additional approval is required.

1.2 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' [27]. 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 Fig. 1.

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. When compared to part-task trainers and flight training devices (FTDs), VR is considered an immersive simulation technology and is actively being explored for use in maintenance and pilot training [20].

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. A VR system that lets the trainee interact with the aircraft, in much the same manner as in the FFS, has the potential to revolutionise

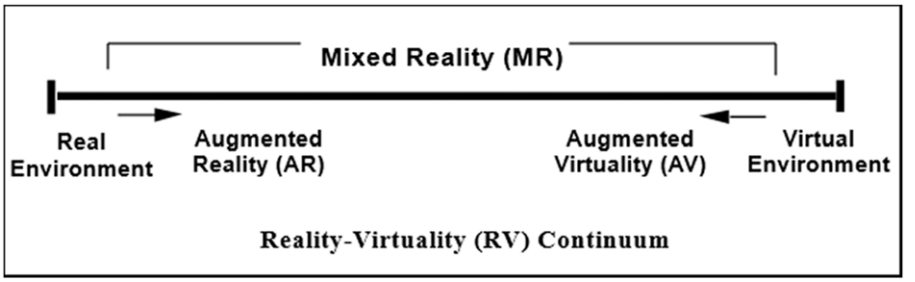

Figure 1. *Simplified representation of the reality-virtuality (RV) continuum.*
Note: Adapted from 'Augmented Reality: A Class of Displays on the Reality-Virtuality Continuum' by P. Milgram, H. Takemura, A. Utsumi, and F. Kishino, 1995, Proceedings SPIE: Telemanipulator and Telepresence Technologies, p. 283 (https://doi.org/10.1117/12.197321). Copyright 1995 by SPIE.

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. 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 optimising the time spent in the FFS. 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 [19]. 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 that would free up currently used FFS slots. VR tools that have been successfully used in the medical field to train surgeons on procedures and techniques, particularly for robotic surgery, are only beginning to be adapted for use in education, aviation and industry. 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' [2].

1.3 Problem statement

Certifying agencies must approve every part of a certified training programme. 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 for their programme. 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 (ELOS) [12]. 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 [15,28]. This is significant given the public perception that aviation is at the forefront of using simulation technologies such as the FFS. In practical terms this means that a new system using VR must be designed, perfected, validated and then certified by the certifying authority 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 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. There is no currently accepted methodology for proving the ELOS required by the FAA, and other national regulatory authorities, to implement an expanded use of VR in training. This study proposes a method for assessing this ELOS and evaluates an existing VR training solution using the proposed method.

1.4 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. Airframe andpowerplant (A&P) technicians, certificated by the FAA to conduct engine runs, were assessed on selected tasks post-training to allow for a quantitative assessment of the effectiveness of VR training as compared to training delivered in the FFS. 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 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. 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 generalisation of results beyond maintenance technicians to any other training involving identical competencies using a VR environment to replace an FFS.

1.5 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?

1.6 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 alternative hypothesis was that a statistically significant difference was observed. 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:

$\mathbf{HA_{0:}}$ There is no collective statistically significant difference in test performance between the groups when controlled for A&P and engine-run experience.

$\mathbf{HA_{\mathit{a}:}}$ There is a collective statistically significant difference in test performance between the groups when controlled for A&P and engine-run experience.

Covariates were included in the analysis to assess if either the experience levels of the A&P technicians (measured in years holding an A&P license), or the number of years that they had been performing engine runs, influenced the dependent variables.

1.7 Limitations and assumptions

Only FAA qualified and certificated A&P technicians participated in this study, which limits the applicability of the immediate results to aviation maintenance technicians (AMTs) certificated under the applicable FAA 14 CFR § 147 training and licensing programme. The technicians were already qualified to perform engine-run maintenance tasks on 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.

2.0 A review of 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 30 years, turns up a surprisingly limited list [28]. Fussell [15] 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 [2]. 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. [28] 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'. A search of other fields is therefore in order, and where VR training has been used or studied the methodologies employed should be analysed and applied to the aviation training environment.

2.1 Competency-based training

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 behaviourism, as represented in the works of experimental psychologists like Watson, Pavlov, Thorndike and Skinner, whose legacy has led to a focus on observable behaviours [24]. Behaviourism focuses on observed behaviour as an indicator of learning, and CBT seeks to use observed skills and behaviours to both teach and assess competency [16]. Rutherford [32] contends that competency is measured by assessing the successful application of learned skills and behaviours towards the satisfactory completion of a task.

Consider the task of landing an aircraft. This task can be deconstructed into behaviours such as maintaining a set descent rate (e.g. the approach glide path), maintaining tracking (e.g. keeping the aircraft centred on the runway) and flaring the aircraft at the correct place and rate. These three behaviours 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 behaviours, and if a hard landing occurred it could be traced to a misapplication of an observed behaviour (e.g. a hard landing could be due to a late flare, which is linked to the 'flare the aircraft' behaviour) [14]. A student pilot would be considered to be competent at the landing task when all three behaviours were correctly applied, and an acceptable landing was observed when the behaviours were combined. 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 [21].

The FAA advocates the use of Bloom's taxonomy to describe and evaluate different stages of learning throughout flight instruction [11]. Bloom's taxonomy classifies educational goals into three separate and individually observable domains: affective, psychomotor and cognitive. 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 [34].

2.2 The use of VR in education

The use of both VR and augmented reality (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. [28], Wang et al. [37] and Ibáñez and Delgado-Kloos [18] 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 visualisation of abstract concepts. VR has the capability to detach advanced and complicated learning from its associated laboratories and facilities, but Wang et al. [35] 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 A&Ps, 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 [5].

2.3 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. [7] and supported by Ahmedyanova [3]. This addresses the main measures of success in the FFS environment, which include both accuracy and timeliness of actions. 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' [36]) 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 [17]. 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.

2.4 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. [6] 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'. 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 [6]. Hoogenes et al. [17] used a validated manual assessment tool, Global Evaluative Assessment of Robotic Skills (GEARS), to conduct a randomised comparison of two robotic VR simulators, and an evaluation of the trainees' skills transfer to a simulated robotic urethrovesical anastomosis task. Figure 2 shows a representative GEARS tool, developed for use in assessing a robotic surgical procedure. Hoogenes et al. [17] 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

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.

Figure 2. *GEARS scale adapted for robotic surgery.*
Note: Adapted from Robotic surgery training: construct validity of Global Evaluative Assessment of Robotic Skills (GEARS) (p. 229), by Ref. [29].

tool 'consists of a 5-point anchored Likert scale across 6 domains that deconstruct the fundamental elements of robotic surgical procedures' [17]. 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.

Schulz et al. [31] conducted a non-randomised 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 centred 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 [31]. Neumann et al. [26] 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. Schmidt et al. [30] 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. Abbott [1], FRAeS, an FAA researcher in the field of artificial intelligence and complex systems, likened the modern airliner to a flying robot. 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 [23] 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 scrutinised, understood and assessed. It is through this training that effective interaction with a robotic interface is enabled, and that these professionals learn how to do their jobs in a safe and efficient manner. 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.

2.5 Validity of the GEARS construct

Sánchez et al. [29] 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. 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. 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.

2.6 Gaps in the literature

Renganayagalu et al. [28] reviewed studies on the effectiveness of VR training over the last 30 years and extracted a total of 60 studies to analyse. 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. [2] 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. [2] and Renganayagalu et al. [28] identify a need for further experimental study of VR learning outcomes, properly specified and controlled, with adequate sample sizes. Cross and Ryley [9] studied the effectiveness of VR systems in providing collaborative VR learning using competency-based training and assessment (CBTA) methods, but also acknowledge that their study has an increased likelihood of Type 2 errors due to the small sample size.

As discussed in the introduction, certifying authorities set 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.

3.0 Methodolgy

3.1 Research method selection

A quantitative study of training effectiveness was selected in order to produce the required data to satisfy the regulatory requirement for an ELOS. The data was collected in the field from actual certificated operators and then compared with existing training systems. In order to assess if VR delivered engine-run recurrent training, delivered to already qualified and certificated FAA maintenance technicians, produced 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 quantitative study 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 and was considered to be any previous use of a VR system or VR technology reported by the subjects.

3.2 Data collection process

The study was conducted using a 3×4 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 was 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 [34].

Following the prescribed training, assessment was conducted in an EASA and 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 and the number of years that they have been performing engine runs, reported in the training entry information forms that are required for all students undergoing training.

3.2.1 Design and procedures

This study was performed using a full VR platform based on a commercial Oculus Rift VR headset, compatible hand controllers, using the Microsoft XR platform. The VR environment was a commercial simulation of a fourth-generation airliner that is currently in service, and the flight deck environment was fully VR modelled and interfaced with using the VR device and hand controllers. Students for the engine-run requalification course were trained in groups of three, per an 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. The FFS students completed the regular recurrent training programme 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, thus ensuring that fatigue was equivalent for each student and was also representative of what could happen with a normal training schedule. All training and evaluation was individually performed.

3.2.2 Modified GEARS scale: aerospace virtual reality assessment of training effectiveness (AViATE)

The GEARS scale was modified to use language and task divisions that were appropriate for an aviation environment as shown in Fig. 3. This provided an easy format with which to score both cognitive and psychomotor elements of the task alongside the time used. The adaptation of GEARS using Bloom's taxonomy and CBT competencies was easily understood and utilised by the evaluators and was observed to work extremely well in the FFS environment. It supports standardised 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.

3.2.3 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 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. It required the students to manipulate each different type of switch and control on the flight deck, and to change the focus of their attention multiple times. The length and complexity of the scenario was appropriate for evaluation with the AViATE tool and gave the evaluators multiple opportunities to observe each dimension of behaviour that was evaluated by the tool. 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 the tool intervened and corrected the student if they were not progressing as required, as would be expected from a human instructor. This was done to ensure that each student being evaluated would have the same directions and prompts given an identical scenario, thus removing the variability of a second human as either instructor or second student in the evaluation. 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).

3.2.4 Sources of the data

All experimental data was gathered from an evaluation in the FFS, using the AViATE scale shown in Fig. 3. 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. Evaluators were selected from a group of experienced and standardised instructors from the training centre, and as a part of the pilot study they were specifically trained on the use of the experimental instrument through a standardisation 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.

AViATE Scale

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.

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):	

Figure 3. *Aerospace virtual reality assessment of training effectiveness scale (AViATE).*

3.3 Population/sample

The population for this study consisted solely of FAA certificated A&P technicians who hold an engine-run qualification on transport aircraft. 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. 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. 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.

3.4 Sampling frame and size

Prior to commencing recurrent training, all participants were pre-screened by the training centre to ensure that they held the appropriate certificate and had suitable recent experience to allow them to

undertake engine-run refresher training under established regulations. The sample for this study was drawn from those students entering the training centre 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 14-month period and all students who entered the training centre during this time were offered the opportunity to participate. A total of 100 students were assessed over this 14-month period.

To determine the required sample size, a power analysis was performed using G∗Power statistical software [10]. Using a medium effect size of 0.25 for MANCOVA [8], 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 [25]. 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.

3.5 Measurement instrument

The primary experimental measurement instrument that was used for evaluating the scenario in the FFS was the AViATE scale, alongside a measure of the time to task completion in minutes and seconds. The AViATE scale is a 5-point Likert scale consisting of six rating items. It provides an interval scale with possible total scores ranging from 6 to 30 measuring both cognitive and psychomotor elements; the cognitive and psychomotor sub-domains are also interval scales with total possible scores ranging from 3 to 15. Time to task completion was also recorded on this form and was measured in minutes and seconds. Definitions of the variables on the AViATE scale 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 analysis.

3.6 Data analysis approach

The AViATE 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 AViATE 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 AViATE scores and time and was then expanded to analyse differences between the variables. Technician experience and engine-run experience were used as controlling variables, or covariates, and their degree of correlation with the IVs was assessed. A univariate analysis of covariance (ANCOVA) analysis was also conducted on each dependent variable to individually examine the effect of the experimental manipulation.

3.6.1 Reliability assessment method

The pre-screening performed by the training centre 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 maximised 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 standardisation training during the pilot study to conduct the assessment. Evaluator scores were tracked and compared to assess inter-rater reliability.

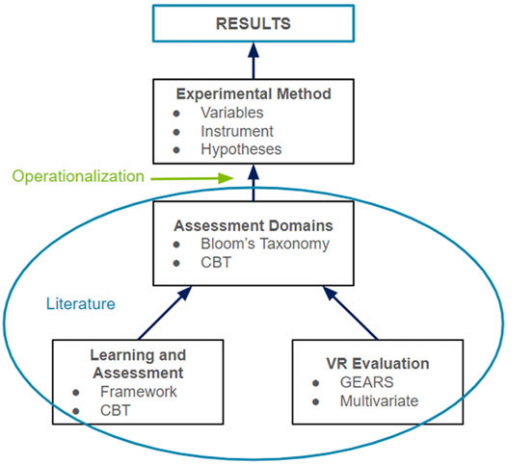

Figure 4. *Theoretical framework and research model.*

3.6.2 Validity assessment method

The existing construct validity of the GEARS tool as studied by Sánchez et al. [29] was leveraged to ensure a high degree of reliability in assessing students' interaction with the FFS, and the addition of aviation specific language and terms did not change the basic construction or usage. In addition to the modified GEARS tool (AViATE), time to complete the scenario was used as a discriminating factor. Training on the use of the AViATE scale was conducted during the pilot study, and a continuous comparison of results between evaluators was conducted as an experimental control. Figure 4 shows the building block approach taken to build the research model and ensure that the validity of adapted measures and tools was retained.

3.6.3 Data analysis process/hypothesis testing

Following completion of the experiment the total AViATE scores, individual AViATE categories and time to completion of the task were analysed using a MANCOVA performed using IBM's SPSS™ software (V28.0.0.0). 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 AViATE Score, individual AViATE elements (Cognitive and Psychomotor), and Time variables. Outliers were examined and the dataset was checked for missing values or errors.

4. Results

4.1 Demographics results

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. 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 [13]. Ages ranged from 22 to 72 years old, with a 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, $F(2,97) = 1.586$, $p = 0.210$.

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, $F(2, 97) = 0.249$, $p = 0.780$. 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, $F(2, 97) = .075$, $p = 0.928$.

4.2 Descriptive statistics

Post-training assessment was conducted using the AViATE scale during a timed event. A total AViATE score ranging from 6 to 30 was possible, with each of the individual rating components being scored from 1 to 5. This can be broken down into a possible sub-score ranging from 3 to 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. AViATE 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, $F(2, 97) = 1.117$, $p = 0.331$.

4.3 Reliability and validity testing results

Reliability testing for the AViATE test was performed using Cronbach's Alpha test. A Cronbach's Alpha score of greater than .7 is desired to demonstrate the reliability of the data collection device [33]. A computed score of .911 for the AViATE test indicates that the test had acceptable internal reliability. A score exceeding .9 indicates an overall high level of reliability for the AViATE test. Inter-rater reliability was acceptable, and an ANOVA test comparing the results showed that there was not a statistically significant difference between the AViATE scores assigned by each evaluator, $F(2, 97) = 1.117$, $p = 0.331$.

4.4 Hypothesis testing results

Test performance was measured as a multivariate combination of AViATE score and time to completion for a defined task. A MANCOVA analysis was conducted using AViATE 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 AViATE 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.

4.4.1 Main hypothesis

The MANCOVA test for the Experimental Hypothesis, using Group as the independent variable, AViATE Total and Time as the dependent variables, and controlling for A&P Experience and Engine-Run Experience was not significant. There was no significant difference in training effectiveness (AViATE Total Score and Time) based on Training Group (Control, VR, and VR with Experience), $F(4,190) = 1.307$, $p = 0.269$; Wilk's lambda = 0.946.

Furthermore, there was no significant effect of Training Group on AViATE Total Score, $F(2, 95) = 0.069$, $p = 0.934$, nor on Time, $F(2, 95) = 1.611$, $p = 0.205$.

A&P Experience was not shown to significantly influence AViATE Total Score, $F(1,95) = 0.714$, $p = 0.400$, nor Time $F(1,95) = 0.788$, $p = 0.377$. Engine-Run Experience was not significant in influencing AViATE Total Score, $F(1,95) = 3.751$, $p = 0.056$, but was significant in influencing Time $F(1,95) = 9.346$, $p = 0.003$, partial eta squared = 0.090.

It is concluded that there was no difference in training effectiveness based on the training received. Engine-Run Experience was shown to have a significant influence on the time to task completion results.

4.4.2 Hypothesis using psychomotor subgroup

The MANCOVA test for a subset of the Experimental Hypothesis, using Group as the independent variable, AViATE Psychomotor and Time as the dependent variables, and controlling for A&P Experience

and Engine-Run Experience was not significant. There was no significant difference in training effectiveness in the Psychomotor domain (AViATE Psychomotor Score and Time) based on Training Group (Control, VR, and VR with Experience), $F(4, 190) = 1.039$, $p = 0.389$; Wilk's lambda = 0.957.

Furthermore, there was no significant effect of Training Group on AViATE Psychomotor Score, $F(2, 95) = 0.164$, $p = 0.849$, nor on Time, $F(2, 95) = 1.611$, $p = 0.205$, as in the previous analysis.

A&P Experience was not shown to significantly influence AViATE Psychomotor Score, $F(1, 95) = 0.287$, $p = 0.593$, nor Time $F(1,95) = 0.788$, $p = 0.377$. Engine-Run Experience was significant in influencing AViATE Psychomotor Score, $F(1, 95) = 5.732$, $p = 0.019$, partial eta squared = 0.057, and was also significant in influencing Time $F(1, 95) = 9.346$, $p = 0.003$, partial eta squared = 0.090.

We conclude that there was no difference in psychomotor training effectiveness based on the training received. Engine-Run experience was shown to have a significant influence on the time to task completion and Psychomotor scores.

4.4.3 Hypothesis using cognitive subgroup

The MANCOVA test for a subset of the Experimental Hypothesis, using Group as the independent variable, AViATE Cognitive Score and Time as the dependent variables, and controlling for A&P Experience and Engine-Run Experience was not significant. There was no significant difference in training effectiveness in the Cognitive domain (AViATE Cognitive Score and Time) based on Training Group (Control, VR, and VR Exp), $F(4, 190) = 1.333$, $p = 0.259$; Wilk's lambda = 0.945.

Furthermore, there was no significant effect of training Group on AViATE Cognitive Score, $F(2, 95) = 0.002$, $p = 0.998$. There was no significant effect of training Group on Time, $F(2, 95) = 1.611$, $p = 0.205$, as in the previous analysis.

A&P Experience was not shown to significantly influence AViATE Cognitive Score, $F(1, 95) = 1.082$, $p = 0.301$, nor Time $F(1, 95) = 0.788$, $p = 0.377$. Engine-Run Experience was not significant in influencing AViATE Cognitive Score, $F(1, 95) = 1.283$, $p = 0.260$, but was significant in influencing Time $F(1, 95) = 9.346$, $p = 0.003$, partial eta squared = 0.090 per the previous analysis.

It is concluded that there was no difference in cognitive training effectiveness based on the training received. Engine-Run experience was shown to have a significant influence on the time to task completion.

4.4.4 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 and their number of years of experience as a qualified engine runner. The analysis in Table 1 shows that A&P Experience was not shown to significantly influence AViATE Total Score, $F(1,95) = 0.714$, $p = 0.400$, nor Time $F(1, 95) = 0.788$, $p = 0.377$. Engine-Run Experience, however, was significant in influencing both AViATE Psychomotor Score, $F(1, 95) = 5.732$, $p = 0.019$, partial eta squared = 0.057, and was also significant in influencing Time $F(1, 95) = 9.346$, $p = 0.003$, partial eta squared = 0.090. Engine-Run Experience was a significant covariate in this study, while overall A&P experience was not.

4.4.5 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 analysed. 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.

Table 1. *Evaluation results by experimental group*

		AViATE Total		AViATE Psychomotor		AViATE Cognitive		Time (S)	
Group	n	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
Control	52	23.33	3.75	11.71	1.88	11.56	2.04	718	82
VR	37	23.22	3.95	11.62	2.07	11.59	2.13	738	118
VR Exp[1]	11	23.45	3.24	11.91	1.76	11.55	1.86	677	116

Note: 1. VR Exp = VR Subjects who indicated prior use of or experience with VR systems.

5.0 Discussion, conclusions and recommendations

5.1 Discussion

5.1.1 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 AViATE scores and time to task completion, and the level of experience was measured through years of experience both as a certificated A&P technician, and years of experience conducting engine-runs. It used a medium effect size of 0.25 for MANCOVA [8], a confidence interval of 0.95, and a total sample size of 100 subjects to achieve a statistical power of 0.8 [10]. The null hypothesis used in the MANCOVA 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 AViATE 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.

5.1.2 Effect size

In determining the sample size for this study, a medium effect size was assumed (Cohen's $d = 0.25$) 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 each group obtained 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.

5.1.3 Use of time as a measure of test performance

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 [17].

A&P mechanics and pilots working to a task are required to complete it to 100% accuracy, and it is expected that, as they become more proficient, they will complete the task more quickly. 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.

5.1.4 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 outcome was reached using a real-world training course with qualified technicians undergoing recurrent training, giving a powerful endorsement of both the specific training tool and the medium of instruction. The VR system as tested was both effective in transferring knowledge and was easy to use. The use of a structured VR evaluation tool, such as AViATE, in an aviation VR environment allowed for a comprehensive evaluation of the VR tool's training effectiveness.

5.2 Conclusions

While there are certainly a large number of emerging VR tools appearing on the market, limited research with sufficient 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. The instrument was adapted with a high degree of validity to the aviation environment, and may be used 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.

This study contributes to the aviation body of knowledge by providing 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. This method utilised an adapted scale (AViATE) for the evaluation of VR delivered training in an aviation environment in combination with time to task completion. It demonstrated both ease of use for the evaluators and consistency of results.

5.2.1 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:

1. A VR environment that was developed and adapted for recurrent training was utilised in this experiment. The recurrent training environment places emphasis on reinforcing existing knowledge, rather than building new knowledge, which may have a different focus.
2. No study of the affective domain of learning was performed. 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.
3. 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 programmes, and will need to be well understood.
4. This study used FAA certificated A&P technicians as subjects and a structured VR training programme that allowed for a limited amount of deviation from the script. Generalisation of the findings is possible through any CBT programme that uses the same entry standard as the FAA programme, but a direct comparison with other programmes and regulatory authorities was not done as part of this study.

5.3 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 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, and 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:

1. 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. It is recommended that future studies explore using an automated scoring system both virtually and in the FFS using a mixed-reality version of the software to score both. 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.
2. 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 as a framework. 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.
3. 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 standardise 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.
4. 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 programmeme which would be a vital part of any aviation training programme, 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.
5. 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 motorised platforms 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 due to the large potential economic benefits.

5.4 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.

Each engine-run course occupies the FFS for a single 4-hour session without the need for simulator motion. Moving seven of these courses to a VR platform would allow a training centre to add another full pilot qualification course to their programme, thereby increasing throughput and making better use of resources. An approved and certified VR trainer would improve financial results by allowing for additional training of both pilots and technicians without the need for additional costly FFS infrastructure.

Acknowledgements. The authors would like to gratefully acknowledge the work and assistance of Victor Liriano, Erik Marrero, Jimmy Torres, Dennis Zubizarreta and Jose Barro in both data gathering and study administration.

References

[1] Abbott, K. Will a computer fly you on your vacation in the future? – Artificial intelligence, automation and autonomy in aviation, In *Presentation, Society of Experimental Test Pilots 2017 Annual Symposium*, Anaheim, CA, 2017.

[2] Abich, J., Parker, J., Murphy, J.S. and Eudy, M. A review of the evidence for training effectiveness with virtual reality technology, *Virt. Real. J. Virt. Real. Soc.*, 2021, **25**, (4), pp 919–933. https://doi.org/10.1007/s10055-020-00498-8

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

[4] Bürki-Cohen, J., Soja, N.N. and Longridge, T. Simulator platform motion-the need revisited, *Int. J. Aviat. Psychol.*, 1998, **8**, (3), pp 293–317. https://doi.org/10.1207/s15327108ijap0803_8

[5] Carl, D.R. The shifting realities of performance improvement: VR, AR, MR, *Perform. Improv.*, 2018, **57**, (4), pp 6–9. https://doi.org/10.1002/pfi.21774

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

[7] Chittaro, L., Corbett, C.L., McLean, G.A. and Zangrando, N. Safety knowledge transfer through mobile virtual reality: A study of aviation life preserver donning, *Saf. Sci.*, 2018, **102**, 159–168. https://doi.org/10.1016/j.ssci.2017.10.012

[8] Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers, 1988.

[9] Cross, J.I. and Ryley, T. Assessing evidence-based training in a collaborative virtual reality flight simulator, *Aeronaut. J.*, 2025, **129**, (1332), pp 261–281.

[10] Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. Statistical power analyses using G∗Power 3.1: Tests for correlation and regression analyses, *Behav. Res. Methods*, 2009, **41**, pp 1149–1160.

[11] Federal Aviation Administration (FAA). *Aviation Instructor's Handbook* (2020th ed.). Aviation Supplies & Academics, Inc., 2020.

[12] Federal Aviation Administration (FAA). *Aviation Maintenance Technician Schools (AMTS)*, 2021. Available from Federal Aviation Administration: https://www.faa.gov/licenses_certificates/airline_certification/amts/

[13] Federal Aviation Administration (FAA). Airmen Certification Database, 2023. Available from Federal Aviation Administration: https://www.faa.gov/licenses_certificates/airmen_certification/releasable_airmen_download/

[14] Franks, P., Hay, S. and Mavin, T. Can competency-based training fly?: An overview of key issues for ab initio pilot training, *Int. J. Train. Res.*, 2014, **12**, (2), pp 132–147. https://doi.org/10.1080/14480220.2014.11082036

[15] Fussell, S. *Determinants of aviation students' intentions to use virtual reality for flight training*. PhD Dissertations and Master's Theses, 542, 2020. https://commons.erau.edu/edt/542

[16] Hodge, S. The origins of competency-based training, *Austr. J. Adult Learn.*, 2007, **47**, (2), pp 179–209.
[17] Hoogenes, J., Wong, N., Al-Harbi, B., Kim, K.S., Vij, S., Bolognone, E., Quantz, M., Guo, Y., Shayegan, B. and Matsumoto, E.D. A randomized comparison of 2 robotic virtual reality simulators and evaluation of trainees' skills transfer to a simulated robotic urethrovesical anastomosis task, *Urology*, 2018, **111**, pp 110–115. https://doi.org/10.1016/j.urology.2017.09.023
[18] Ibáñez, M., & Delgado-Kloos, C. Augmented reality for STEM learning: A systematic review, *Comput. Educ.*, 2018, **123**, pp 109–123. https://doi.org/10.1016/j.compedu.2018.05.002
[19] Jerald, J. (2016). *The VR Book: Human-Centered Design for Virtual Reality*. New York, NY: Association for Computing Machinery.
[20] Macchiarella, N.D., Liu, D., Gangadharan, S.N., Vincenzi, D.A. and Majoros, A.E. Augmented reality as a training medium for aviation/aerospace application, In *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting*, pp 2174–2178. Orlando: Hunan Factors and Ergonomics Society, 2005.
[21] Marzano, R.J., Norford, J.S., Finn, M. and Finn III, D. A Handbook for Personalized Competency-Based Education. Marzano Research, 2017.
[22] Milgram, P., Takemura, H., Utsumi, A. and Kishino, F. Augmented reality: A class of displays on the reality-virtuality continuum, In *Paper presented at the SPIE,* vol. 2351, (1), 1995. https://doi.org/10.1117/12.197321
[23] Mindell, D. *Our Robots, Ourselves*. New York: Viking, 2015.
[24] Morcke, A.M., Dornan, T. and Eika, B. Outcome (competency) based education: An exploration of its origins, theoretical basis, and empirical evidence, *Adv. Health Sci. Educ.*, 2013, **18**, (4), pp 851–863. https://doi.org/10.1007/s10459-012-9405-9
[25] Morris, S.B Estimating effect sizes from pre-test – post-test control group designs. *Org. Res. Methods*, 2007, **11**, (2), pp 364–386. https://doi.org/10.1177/1094428106291059
[26] Neumann, E., Mayer, J., Russo, G., Amend, B., Rausch, S., Deininger, S., Harland, N., Anselmo da Costa, I., Hennenlotter, J., Stenzl, A., Kruck, S. and Bedke, J. Transurethral resection of bladder tumors: next-generation virtual reality training for surgeons, *Eur. Urol. Focus*, 2019, **5**, (5), pp 906–911. https://doi.org/10.1016/j.euf.2018.04.011.
[27] Pan, Z., Cheok, A.D., Yang, H., Zhu, J., & Shi, J. Virtual reality and mixed reality for virtual learning environments, *Comput. Graph.*, 2006, **30**, (1), pp 20–28. https://doi.org/10.1016/j.cag.2005.10.004
[28] Renganayagalu, S.K., Mallam, S.C. and Nazir, S. Effectiveness of VR head mounted displays in professional training: A systematic review, *Technol. Knowl. Learn.*, 2021, **26**, (4), pp 999–1041. https://doi.org/10.1007/s10758-020-09489-9
[29] Sánchez, R., Rodríguez, O., Rosciano, J., Vegas, L., Bond, V., Rojas, A. and Sanchez-Ismayel, A. Robotic surgery training: construct validity of Global Evaluative Assessment of Robotic Skills (GEARS), *J. Robot. Surg.*, 2016, **10**, (3), pp 227–231. https://doi.org/10.1007/s11701-016-0572-1
[30] Schmidt, M.W., Kowalewski, K., Schmidt, M.L., Wennberg, E., Garrow, C.R., Paik, S. and Nickel, F. The Heidelberg VR score: Development and validation of a composite score for laparoscopic virtual reality training, *Surg. Endosc.*, 2019, **33**, (7), pp 2093–2103. https://doi.org/10.1007/s00464-018-6480-x
[31] Schulz, G.B., Grimm, T., Buchner, A., Jokisch, F., Casuscelli, J., Kretschmer, A., Mumm, J., Ziegelmuller, B., Stief, C. and Karl, A. Validation of a high-end virtual reality simulator for training transurethral resection of bladder tumors, *J. Surg. Educ.*, 2019, **76**, (2), pp 568–577. https://doi.org/10.1016/j.jsurg.2018.08.001
[32] Rutherford, P. Competencies undergo a review, *Train. Develop. Austr.*, 2009, **36**, (5), pp 25–28. http://search.proquest.com.ezproxy.libproxy.db.erau.edu/docview/208559908?accountid=27203
[33] Tavakol, M. and Dennick, R. Making sense of Cronbach's alpha, *Int. J. Med. Educ.*, 2011, **2**, pp 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd
[34] Vogt, W.P., Gardner, D.C., Haeffele, L.M. and Ebrary, I. *When to Use What Research Design* (1st ed.). New York, N.Y: Guilford Press, 2012.
[35] Wang, M., Callaghan, V., Bernhardt, J., White, K. and Peña-Rios, A. Augmented reality in education and training: Pedagogical approaches and illustrative case studies, *J. Amb. Intell. Human. Comput.*, 2018, **9**, (5), pp 1391–1402. https://doi.org/10.1007/s12652-017-0547-8
[36] Whitson, R. Training in a Modern Age. (Master's Thesis). Arizona State University, Tempe, AZ, 2019.
[37] Wu, S. *Psychological Presence in Immersive Virtual Environments*. (Master's Thesis), San José State University, San José, CA, 2018.

Cite this article: McCullins M.E., Hampton S., Fussell S.G., Kiernan K. and Thropp J. The effectiveness of using virtual reality training environments for procedural training in fourth-generation airliners. *The Aeronautical Journal*, https://doi.org/10.1017/aer.2025.10086