Dr. Jamie Cross[1] and Christine C Boag-Hodgson[1]

[1]Affiliation not available

March 14, 2024

A Collaborative Virtual Reality Flight Simulator: Efficacy, Challenges and Potential

Jamie I. Cross, Christine C. Boag-Hodgson

***Abstract*—The incorporation of immersive technologies into student pilot training has been hindered by a lack of empirical evidence to support their efficacy. Existing research on virtual reality flight simulators is limited in scope, predominantly focused on single-users in small, piston-engine aircraft, with little concern for its application to commercial pilot operations. This paper initiates the process to evaluate a virtual reality flight simulator to train ab-initio pilots in a multi-crew environment using a complex jet aircraft.**

Proof of concept was demonstrated by satisfactory usability, fidelity, and collaboration in a two-pilot virtual reality flight simulator. Participants in the virtual environment exhibited reduced workload (effort) in comparison to a desktop flight simulator, indicating a potential decrease in cognitive processing. This in turn suggests enhanced spatial memory, corroborated by measures of heightened team situational awareness in the virtual environment. The benefits of these findings are numerous, including the potential for a virtual reality flight simulator to supplement traditional pilot training methods.

***Index Terms*— Collaboration, Flight simulation, Pilot training, Virtual reality**

I. Introduction

VARIOUS levels of simulation have been used for decades to enable both ab-initio and commercial pilots to acquire and maintain skills. However, the process is costly, time consuming and inflexible. Recent advances in immersive technologies have piqued interest in virtual reality (VR) for pilot training since it may offer supplementary options to traditional desktop and full flight simulators. The effectiveness of simulator training is undeniably beneficial and is embedded into current pilot training regimes [1]. However, there are several limitations to these simulators [2],[3], including their cost, size, and in the case of full flight simulators, the bottleneck imposed by only being able to accommodate two trainees at a time. This latter restriction is particularly pertinent since airlines are currently short of flight crew [4],[5]. This may exacerbate pilots being trained with the minimum number of regulated hours, and yet the validity of equaling flight hours to sufficient experience is questionable [6]–[8].

Due to these shortfalls and training demand, using virtual reality flight simulators (VRFSs) to supplement traditional training methods is gaining an increasing amount of attention. However, there is a great deal of anecdotal evidence, uninformed commentary, conflicting terminology, and unrealistic expectations that has prevented its wide-spread introduction into mainstream pilot training. Immersive technologies have advanced significantly over recent years, and are a capable mechanism used to deliver training in industries such as healthcare, tourism, and education [3]. The main reasons delaying the implementation of VRFSs is the dearth of empirical evidence required by regulators to support their efficacy, coupled with a lack of human factors design principals.

A. Collaboration and Multi-Crew Cooperation in Virtual Environments

What little empirical evidence that has been provided on the usefulness of VRFSs has been limited to single-user simulators (e.g., [3],[9]-[11]). This may be of benefit if single-pilot commercial operations are introduced, which may commence in the early 2030s [12]. However, within most current commercial pilot operations, the flight crew comprise of at least two collaborating pilots. Collaboration within a team is at the center of any multi-crew operation [13]-[15]. Within the pilot training environment, collaboration is termed multi-crew cooperation (MCC). MCC training is compulsory for all pilots and is typically combined with crew resource management training (CRM), which incorporates human factors and non-technical skills training. It is therefore essential to research MCC within a virtual environment (VE).

B. Situational Awareness in a Collaborative Environment

Alongside presence (i.e. "being there"), sufficient situational awareness (SA) in a VE is important [3]. Although there is no universally accepted model of SA [16], at a very simplistic level, SA is an appropriate awareness of a situation [17]. Most critical to this current paper is that measures of SA are used to provide evidence of the useability of a VRFS in a collaborative environment.

Since the study described in this paper involved participants working together in a collaborative VE, their collaborative SA must be considered. This is termed team situational awareness

This paragraph of the first footnote will contain the date on which you submitted your paper for review, which is populated by IEEE. It is IEEE style to display support information, including sponsor and financial support acknowledgment, here and not in an acknowledgment section at the end of the article.

Jamie. I. Cross is a senior lecturer and program director with the School of Engineering & Built Environment (Aviation) at Griffith University, 170 Kessels Rd, Nathan QLD 4111, Australia (e-mail: jamie.cross@griffith.edu.au).

Christine. C. Boag-Hodgson is an adjunct with the School of Engineering & Built Environment (Aviation) at Griffith University (e-mail: c.boag-hodgson@griffith.edu.au).

(TSA), which, for the purposes of this paper, is analogous to shared situational awareness. One definition of TSA is "the degree to which every team member possesses the SA required for his or her responsibilities" [18, p. 39]. Like SA, TSA is considered a critical and influencing factor in collaborative task performance [19]. Many researchers have measured TSA by aggregating the SA of the individual team members. This approach involves measuring individual SA in order to infer a shared mental model [20]-[24]. The basis of this type of measure is that individual SA includes the 'shared SA requirements' such that each team member is aware of their independent task requirements as well as aspects of their overlapping requirements [25]. Given that Endsley's individual model of SA [18] has been successfully extended to the team environment (e.g., [26],[27]), such aggregate measures of TSA are generally acceptable. Furthermore, research found that direct measures of individual SA give a relatively accurate measure of TSA [28]. Therefore, individual SA was measured in this study, which can be considered as a reliable indicator for TSA.

C. Workload and Situational Awareness

SA is intricately entwined with mental workload [29],[30]. This relationship was explained using an example of a pilot becoming aware of an instrument failure on approach to landing [30]. Diagnosis of the problem increases the pilot's workload, resulting in the pilot losing some SA and not noticing that the aircraft is veering away from the required track. Therefore, workload is a construct that represents "the cost incurred by human operators to achieve a specific level of performance" [31, p. 140]. Consequently, it is sensible to measure workload to provide potential explanations for any variations found in SA.

D. Aims

Since most ab-initio student pilots will inevitably train in a MCC environment involving at least two pilots, it is essential to examine the interaction of those pilots in a VE. The challenge is therefore to provide evidence of the usability of two-pilot VRFSs. For this reason, a study was designed to evaluate the novel concept of assessing pairs of participants collaborating in a VE. To further align the research to commercial pilot operations, the study simulated a multi-crewed Boeing 737-800 (B738) aircraft. Therefore, this paper describes a study to investigate the affordances of two-pilot VRFSs. Specifically, it (1) provides a proof of concept for collaborative VRFSs, (2) measures SA and workload in a VRFS, and (3) provides a comparison of a VRFS to a desktop simulator. Demonstrating collaboration in a VE has the potential to revolutionize both ab-initio and commercial pilot training, and may establish that cheaper, more flexible VRFSs can be used to supplement or replace aspects of traditional non-immersive training.

II. Literature Review

There are few academic papers on immersive technology pilot training since the publication of a systematic quantitative literature review (SQLR) [2]. The papers discussed in the SQLR, and the limited number of published papers since the SQLR (e.g., [32]) are all related to single-user, non-collaborative simulators with the exception of [33] who demonstrated evidence-based training in a collaborative VRFS. Similarly, there is little evidence of collaborative VR for education and training in non-aviation contexts, although a few examples are [34] who concluded that there are some use cases in higher education to complement conventional methods and [35],[36] who found that training utilising immersive technologies created greater levels of presence and engagement compared to other mediums (e.g., paper, desktop computers).

Ab-initio pilot training is largely restricted to non-immersive simulators (e.g., [37], [38]). However, in commercial pilot training, there have been attempts to introduce VR over the last couple of years (e.g., [39]). In spite of the SQLR reporting that "Many major airlines have, or are in the process of, rolling out VRFSs for specific parts of aircrew training" (p. 6), there has not been any progress on the regulatory certification of VR for aircrew training in fixed-wing aircraft.

Recently, some major aircraft manufactures have teamed up with leading airlines, for example, Airbus with Lufthansa, to accelerate the certification process of training pilots using VR [39],[40]. These partnerships have made significant advancements in the development of VRFSs. For example, the Airbus virtual procedure trainer reported in the SQLR [41] compared to more recent reports of the same trainer [39],[40] incorporates an advanced artificial intelligence (AI) capability that enables trainees to practice by themselves but with the aid of a co-pilot avatar, in addition to other "training enhancements [that] will enable further use cases as well as regulatory acceptance" [39, p. 1].

III. Method

A. Methodology

Due to the novel concepts of measuring SA and workload in a VR complex aircraft in the study (i.e., the B738), a design-based research methodology was initially adopted to identify and fix any methodological issues. Parameters that were investigated included varying the VR simulator setup to explore the level of fidelity and ensure that the VRFS replicates a real operational environment for collaborating participants as accurately as possible while maintaining minimal latency. This step was particularly important since it is known that hardware performance issues can severely impact user experience and generate unreliable data [2]. Two different scenarios were also evaluated with their respective supporting material, as well as assessing different head mounted displays (HMDs) (a Hewlett Packard Reverb G2 and an Oculus Quest 1). A secondary focus of the study was to capture measures of SA and workload (discussed below). Completion times for each phase of flight (i.e., pre-flight, taxi, etc.) were also recorded. Finally, the study was configured to allow evaluation of different combinations of VR and desktop flight simulators in a multi-crew environment.

B. Participants

After ethical clearance was obtained from the authors' institution, participants for the study were sourced from the Griffith University aviation programs and the local flight school utilized by the university for flight training. A pre-requisite placed on participants was that some amount of flight simulator experience was preferred, but no previous real-world flying experience was necessary. The aviation participants were highly representative of potential beneficiaries of a VRFS since they were either enrolled in an undergraduate degree in aviation or had completed the degree and were enrolled in a post-graduate commercial pilot training program. One exception was a participant who already had some experience as a commercial pilot.

Data was collected from 24 participants comprising of males (n=21, 88%) and females (n=3, 12%) between the age of 18 and 42 (M=21.1, SD=4.8). Most participants were studying the first year of an aviation program (n=12; 50%), three participants (12%) were studying the second year of an aviation program, four participants (17%) were studying the third year of an aviation program, four participants (17%) were enrolled in the institution's commercial pilot training program, and also the one participant (4%) who was employed as a commercial pilot. Post-experiment debriefs revealed that the majority of participant-pairs (n=16, 67%) did not know each other before the study.

Familiarity with desktop simulators and VRFSs is shown in in Fig. 1. All participants complied with the experiment pre-requisite since they reported having at least some familiarity with desktop flight simulators, while 30% of participants reported being not at all familiar with VR in flight simulators. The most recent time desktop simulators and VRFSs were used by each participant is shown in Fig. 2.

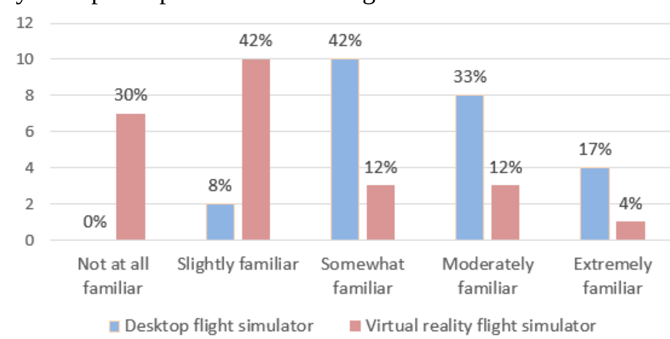

Fig. 1. Number of participants familiar with desktop simulators and virtual reality flight simulators.

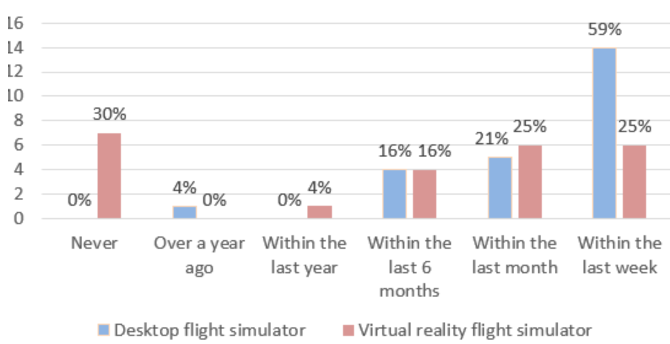

Fig. 2. The last time participants used desktop simulators and virtual reality flight simulators.

Three participants (12%) had no real-world flying experience, while the majority of participants did have some real-world flying experience (n=21, 88%). The only participant with multi-engine turbine experience (200 hours) was employed as a commercial pilot. Familiarity with a B738 and MCC is shown in Fig. 3. The B738 familiarity was revealed in a post-experiment debrief to be recreational simulator usage, that is, no participants had any real-world B738 experience. The MCC familiarity was revealed in the debrief to be the commercial pilot who had 200 hours of multi engine turbine experience.

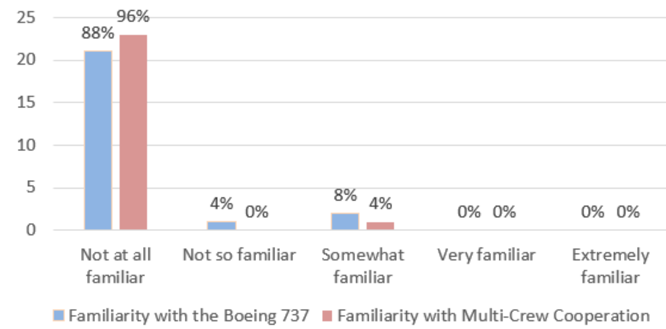

Fig. 3. The number of participants familiar with the B738 and multi-crew cooperation

C. *Measuring Team Situational Awareness*

Since the reliability of individual SA as an indicator of TSA has already been established, this section will discuss the measurement of individual SA. Previous research [3] reviewed various methods for measuring SA and subsequently SART was identified as best suited to the present study. SART uses three domains to measure operator SA: Attentional Demand, Attentional Supply and Understanding [42]. The internal consistency measures for SART during this study yielded alphas [43] of 0.71 in the VE and 0.89 in the real-world. This is comparable to alphas of between 0.71 and 0.85 in a study that investigated the validity of SART for cognitively complex human–machine work [44]. SART uses 7-point Likert scale where only the end anchors are labelled (e.g., 1: low, 7: high).

D. *Measuring Workload*

The National Aeronautics and Space Administration Task Load Index (NASA-TLX) [31] is a subjective, multidimensional assessment tool that rates perceived workload to estimate task effectiveness or performance. At its inception, the NASA-TLX measure and its subscales were validated and found to sufficiently represent sources of cognitive workload among different tasks. Since then, the NASA-TLX has achieved a certain venerability [45], with independent studies finding it to be a valid measure of subjective workload (e.g., [46],[47]). However, the NASA-TLX has also received some criticism, suggesting that it measures perceived task difficulty rather than mental workload [48]. Given the extensive use of the NASA-TLX within the aviation community, it was used to measure subjective workload in the present study.

The NASA-TLX is divided into six subjective subscales: Mental Demand, Physical Demand, Temporal Demand, Performance, Frustration Level and Effort [31]. An overall workload score is calculated based on a weighted average of ratings on these subscales. The internal consistency measures for NASA-TLX during this study yielded alphas [43] of 0.80 in the VE and 0.91 in the real world. This is comparable to an alpha of more than 0.80 in an internal consistency test of the NASA-TLX by [47], and alphas of between 0.83 and 0.86 in a study that investigated the validity of the NASA-TLX [44]. NASA-TLX uses 20-point scale where the end anchors are labelled (e.g., 1: very low/perfect, 20: very high/failure).

E. Scenario Development

Based on the International Civil Aviation Organization (ICAO) core-competency framework [49]-[52] and aligned to typical initial First Officer airline training programs [51],[53], two different scenarios were developed, as follows:

1) **Scenario 1 (Aircraft Handling and Normal Procedures)**

Heathrow Airport (LHR/EGLL) was used as the base of the first scenario because it is a suitably large airport and one of which is highly detailed and authentically represented within the flight simulator software. For scenario 1, the aircraft was positioned at a gate during daylight, in a heavy rainstorm. To increase realism, eight other active aircraft were operational in the scenario, controlled by AI. Note that integration of AI into training scenarios is an identified benefit of VEs [2].

With engines already running, a summary of scenario 1 is as follows:

- Pushback
- Taxi to an assigned runway
- Takeoff
- Perform a simple circuit
- Land
- Taxi and park at an assigned gate.

2) **Scenario 2 (Non-Normal and Emergency Procedures)**

Sydney Airport (SYD/YSSY) was used as the base of the second scenario because it is a suitably large airport and one of which is highly detailed and authentically represented within the flight simulator software. In scenario 2, the aircraft was positioned at a gate during daylight, in CAVOK (i.e., fair weather). As per scenario 1, eight other active AI aircraft were operational in the scenario. A surprise critical event was included in this scenario, since including such events promotes a sense of presence and SA in a VE [3].

From cold and dark (i.e., no engines running, and no power), a summary scenario 2 is as follows:

- Start the Auxiliary Power Unit (APU)
- Pushback
- Start both engines and shutdown the APU
- Taxi to an assigned runway
- Takeoff
- Manage a critical event: an engine fire that automatically triggered at 100 kts during the takeoff run (i.e., before V1, so the takeoff should be aborted).

F. Pilot Flying and Pilot Not Flying Description and Roles

On a typical two-person flight crew, the aircraft commander, who is appropriately qualified to hold the rank of Captain, would normally occupy the left seat and the First Officer or Co-Pilot would normally occupy the right seat. Before each flight, the commander designates one of them as Pilot Flying (PF) and the other as Pilot Not Flying or Pilot Monitoring (PNF, PM). Every airline has an operations manual which fully describes the responsibilities of the Captain and the First Officer, and the roles of the PF and PNF, during all phases of flight.

To align with commercial flight operations, a simplified concept of the PF and PNF was utilized in this study. In summary:

- The PF occupied the left seat. The PF operated the flight controls of the aircraft and was responsible for all the activities which directly affected flight path management (i.e., taxiing and flying the aircraft). The PF was also responsible for confirming the actions of the PNF.
- The PNF occupied the right seat. The PNF monitored the course of the flight and was responsible for reading and actioning the checklists, navigation (on the ground and in the air), all radio communication, landing gear and flap operation, and generally assisting the PF as necessary.

G. Description of the System

1) **Hardware and Software**

Two computers were set up alongside each other, both running the X-Plane Flight Simulator. The left computer was for the PF (as viewed from behind the PF, or from the flight deck door) and the right computer for the PNF. The two computers were networked and set up as 'master' (for the PF) and 'slave' (for the PNF) within X-Plane to allow independent views of the same simulation; that is, the PF controlled their own view, and the PNF controlled their own view. See Fig. 4.

The PF controlled the aircraft using a yoke, throttle quadrant and rudder pedals connected to the master computer. The PNF operated the cockpit switches, buttons and levers using a mouse, also connected to the master computer. VR hand controllers were evaluated during the design stage but were not utilized in study because they were found to be too cumbersome to operate switches and levers. Any simulation flight data (e.g., location, speed, altitude) or movement of switches was fed from the master to the slave computer. This enabled both PF and PNF to be in the same simulation and cockpit environment. A researcher sat behind both the PF and the PNF and could monitor their screens, and also follow the simulation on a tablet.

2) **Setup Combinations**

Three setup combinations of the real-world and the VE were used in the study, as shown in Table I, and Fig. 5, Fig. 6, and Fig 7. Participants immersed in the 'Dual Real-World' setup can be seen in Fig. 8.

TABLE I
SETUP COMBINATIONS

Setup Combination	Description
Dual Real-World	Both the PF and PNF operated in the real-world. See Fig. 5.
Mixed VE / Real-World	The PF operated in the VE and PNF operated in the real-world. See Fig. 6.
Dual VE	Both the PF and PNF operated in the VE. See Fig. 7 and Fig. 8.

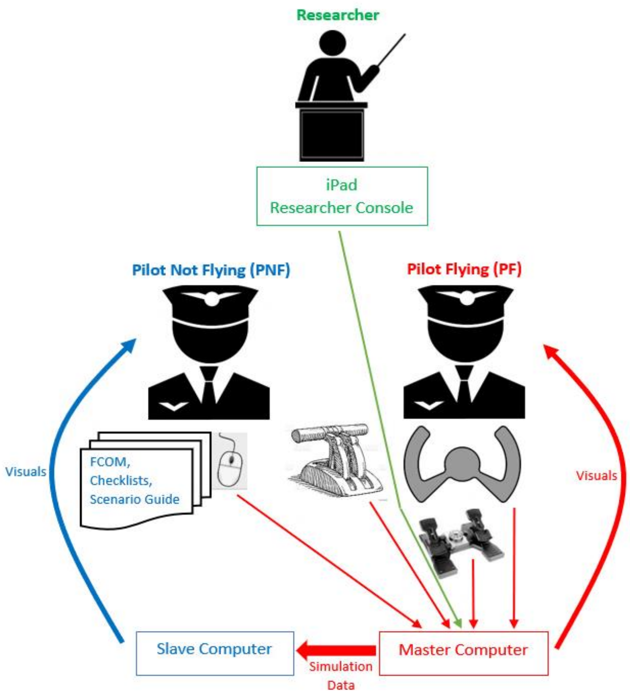

Fig. 4. Hardware setup.

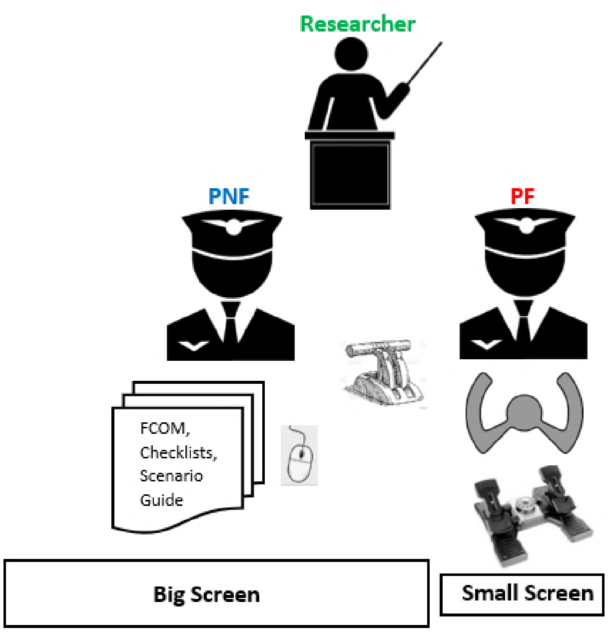

Fig. 5. 'Dual Real-World' setup.

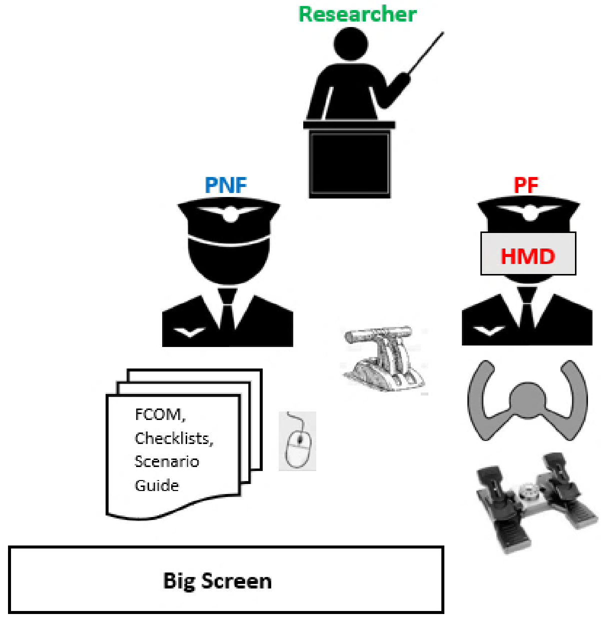

Fig. 6. 'Mixed VE / Real-World' setup.

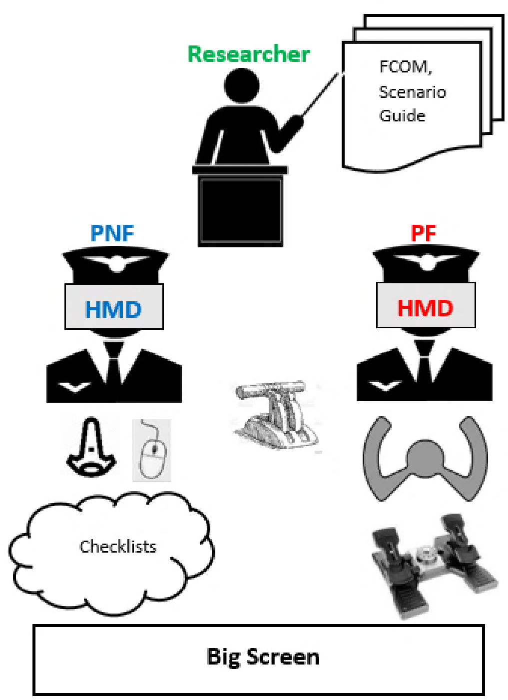

Fig. 7. 'Dual VE' setup.

Fig. 8. Participants in the 'Dual VE' environment.

H. Materials Developed for the Study

Substantial supporting materials were developed for the study, summarized below. In the real-world, participants had access to hardcopy documentation including the checklists. In the VE, participants were restricted to virtual checklists, and reliant on the researcher to provide additional guidance where necessary (e.g., taxi guidance).

- Participant briefing PowerPoint slides used pre-experiment.
- Flight crew operating manual (FCOM), based on a typical commercial airline's B738 FCOM, although significantly simplified.
- Scenario guides (objectives and instructions), one for each for scenario.
- Apron and aerodrome ground movement charts, obtained for both Heathrow and Sydney airports.
- Checklists, based on a real B738 checklist.

I. Experimental Procedure

A within-participants design was utilized which involved four stages: pairing and scenario assignment, pre-simulation, flight simulation, and post-simulation, as follows:

1) **Pairing and scenario assignment**

Participants were randomly paired together, and within the pair, randomly assigned to act as the PF or the PNF. Participant-pairs were assigned to either the first or the second scenario and allocated to one of the three setup combinations in a cyclic fashion. Therefore, there was an even distribution of participants across the two scenarios and the three setup combinations. In total, 12 participants acted as PF and 12 acted as PNF. Also, in total, 12 participants operated in the real-world, and 12 operated in the VE.

2) **Pre-simulation**

Participants-pairs were asked to read an information sheet and provide their consent. They then completed a pre-simulation demographic survey, after which they were given a short presentation which was completed at a spare workstation to allow practice with the flight controls. Participants indicated competence with the simulator and HMD (if used) after about ten minutes, which established a baseline of experience. This reduced confounds or unrelated performance errors due to the effects of familiarity, apprehension, or novelty bias.

3) **Flight simulation**

Participants pairs were taken to the two flight simulator computers where use of the flight controls was demonstrated once again, and the scenario objectives were restated. If the PNF was operating in the real-world, that participant was provided with the necessary materials (e.g., scenario guide, checklists, charts, and paper/pen). If either or both of the PF and/or the PNF were assigned to operate in the VE and use VR, they donned the HMD and were shown how to adjust the strap fittings and the interpupillary distance. Participants were advised to stop the simulation if they felt any simulator sickness and then asked to complete their designated scenario. During the experiment, the researcher acted as air traffic control and provided some cues, although unnecessary interference was avoided, and conformed to ICAO instructor training guidelines [49, p. I-7-2, para.7.4.3]. Completion times for each phase of flight were recorded. The researcher also took observational notes.

4) **Post-simulation**

After completion of the scenario, participant-pairs were asked to complete the SART and NASA-TLX surveys. Immediately after completing the surveys, participant pairs were asked to take part in a debrief.

IV. Results

A. Proof of Concept

Once the various network and simulator settings had been configured during the design phase, the study ran smoothly, and all participant-pairs successfully completed their scenarios. The settings were configured before any participants used the simulator, so that all participants were subject to the same environment. Many settings were manipulated to achieve the highest level of fidelity balanced against minimal latency. The Hewlett Packard Reverb G2 HMDs were used in favor of the Oculus Quest 1 HMD which was found to produce slightly blurred images. This was particularly apparent when attempting to read the text on or around switches in the cockpit or taxiway signage which was used to navigate to the correct runway (for takeoff) or gate (after landing).

B. Completion Times

Completion times for each scenario, broken down by phase of flight, for each setup combination, are shown in Table II. It is not appropriate to perform any statistical analysis on this data due to the small sample size (i.e., four participants within any setup combination). In addition, it was not possible to assess the impact of real-world flying experience since participants were randomly paired together with a variety of experience.

TABLE II

AVERAGE COMPLETION TIMES IN MINUTES FOR THE THREE SETUP COMBINATIONS FOR BOTH SCENARIOS BY PHASE OF FLIGHT

		Setup Combination 1 'Dual Real-World'	Setup Combination 2 'Mixed VE / Real-World'	Setup Combination 3 'Dual VE'
Scenario 1 (Heathrow)	Pre-Flight, Before Start, Pushback	9	14	9
	Taxi out	4	5	4
	Takeoff, Initial Climb	5	5	5
	Further Climb, Cruise, Approach, Landing	9	10	9
	Taxi in, Shutdown	9	11	8
	Overall Completion Time (Scenario 1)	36	45	35
Scenario 2 (Sydney)	Pre-Flight, Before Start, Pushback, Engine Start	13	19	12
	Taxi out	6	6	6
	Takeoff (Engine fire event requiring engine shutdown)	5	7	5
	Taxi in (using one engine), Shutdown	8	10	7
	Overall Completion Time (Scenario 2)	32	42	30

However, a summary of the completion times is as follows:

- Due to the design of the two scenarios, overall completion times for scenario 1 (Heathrow) took longer than scenario 2 (Sydney) as the former involved some flying.
- Completion times for the different phases of flight for the 'Dual Real-World' and 'Dual VE' setup combinations within the same scenario were similar. For example, 'Dual Real-World' scenario 1 (Heathrow) Pre-Flight of 9 minutes was the same as 'Dual VE' scenario 1 (Heathrow) Pre-Flight of 9 minutes.
- The 'Mixed VE / Real-World' setup took longer than either of the 'Dual Real-World' or 'Dual VE' setup combinations within the same scenario. For example, for scenario 1 (Heathrow), 45 minutes versus 36 and 35 minutes.
- Scenario 1 Pre-Flight took less time than scenario 2 Pre-Flight because the latter involved an engine start with more checklists to complete. For example, 9 minutes versus 13 minutes.

C. *Situational Awareness*

Using an independent samples t-test with a 95% confidence interval, the mean scores of the SART domains for the participants operating in the VE were compared against those operating in the real-world to determine whether there is evidence that the associated population means are significantly different. An overall SA for real-world and VE was also calculated and compared using the formula:

SA = Understanding – (Attentional Demand – Attentional Supply) [42].

The results are shown in Table III. No difference in Attentional Demand was found in the VE compared to the real world, but significantly higher Attentional Supply, Understanding and Overall SA was found in VE compared to the real world.

D. *Workload*

Using an independent samples t-test with a 95% confidence interval, the mean scores of the NASA-TLX subscales for the participants operating in the VE were compared against those operating in the real-world to determine whether there is evidence that the associated population means are significantly different. The subscales scores were also combined to provide an overall task load index which was also compared. Given that the importance of the subscales is a priori known to be approximately of equal interest, no weights were applied to the subscales [54].

The results are shown in Table IV. No difference in Mental Demand, Physical Demand, Temporal Demand, Performance, Frustration Level, or Overall Workload was found in the VE compared to the real world, but less perceived Effort was found in the VE compared to the real world.

TABLE III

COMPARISON OF REAL-WORLD SA VERSUS VE SA

Domain	Real World M (SD)	VE M (SD)	t	df	Sig. (2-tailed)
Attentional Demand	10.88 (5.11)	9.44 (3.58)	-.804	22	p = .430
Attentional Supply	15.88 (5.25)	21.94 (3.89)	3.205	22	p = .004 *
Understanding	11.50 (2.98)	14.56 (3.10)	2.312	22	p = .031 *
Overall SA	16.50 (12.41)	27.06 (7.29)	2.642	22	p = .015 *

* p < .05

TABLE IV
COMPARISON OF REAL-WORLD WORKLOAD VERSUS VE WORKLOAD

Subscale	Real World M (SD)	VE M (SD)	t	df	Sig. (2-tailed)
Mental Demand	11.00 (5.13)	12.38 (4.47)	.677	22	p = .505
Physical Demand	8.63 (4.75)	9.00 (5.56)	.163	22	p = .872
Temporal Demand	10.88 (3.72)	9.69 (3.94)	-.708	22	p = .487
Performance	9.13 (5.22)	8.25 (5.39)	-.379	22	p = .708
Frustration Level	11.13 (4.70)	11.50 (4.95)	.178	22	p = .861
Effort	12.00 (4.93)	5.63 (3.93)	-3.445	22	p = .002 *
Overall Workload	10.46 (3.97)	9.41 (3.36)	-.681	22	p = .503

* $p < .05$

V. DISCUSSION

The study provides an invaluable proof of concept that can be applied to future research on collaborative VRFSs. Successful completion of the scenarios by all participant-pairs within a reasonable amount of time demonstrated the viability of the system, adequate control elements, effectiveness of the supporting material, sufficient SA, and effective communication and collaboration within the VE. For example, without collaboration in this study, the PF and PNF would not have been able to start the engines using the checklists.

Although the weather effects for Scenario 1 (Heathrow) gave the scenario greater realism, the flying and landing phases involved researcher intervention since participants had very little experience operating a B738. Scenario 2 (Sydney) with detailed checklists describing engine and APU start provided the best opportunity to observe competencies such as application of procedures, communication, leadership and teamwork. The critical event (engine fire) in this scenario also provided good opportunities to observe these competencies, as well as added realism.

With regard to communication, it was observed that most participants (and all of the VE participants) naturally adopted a closed communication loop using precise language, albeit with non-standard aviation phraseology. That is, the receiver repeated back an instruction, and the sender acknowledged the read-back. This was particularly beneficial to participants since technology is not currently commercially available that permits collaborators in a VRFS to usefully see each other in a VE. The HP Reverb G2 HMDs used in this research provide ‘passthrough’ functionality where the cameras and sensors mounted on the front of the headset allow real-world data to be combined with the virtual content. This would be a shift from VR to augmented reality (see [2] for a description of the extended reality continuum). Assuming the two users are physically sat next to each other, this would allow them to see each other and follow hand movements. However, this would be of no use, since they would be following hand movements in the real world which would bear no relationship to movements in the VE. This functionality would also allow the user to see all of their real-world surroundings and most likely impact their immersive experience. Observation also revealed that non-verbal communication, especially pointing, was used extensively in the ‘Dual Real-World’ combination and was correctly interpreted. Deictic gestures were also initially used by participants using HMDs until they realized that pointing in a VE does not relate to a real-world object, and therefore, even if they could be seen by their co-pilot, it provides limited if any value.

The inability to receive deictic gestures in a VE might cause loss of message or context. This is because a non-verbal message can be a regulator, which is used to control, maintain or discourage interaction [55], for example, by holding up a hand to discourage communication. There is also a common preference for face-to-face communication among pilots, perhaps instilled from traditional training methods, as demonstrated in a study that examined crew interaction on physically separated pilots [56]. They observed some confusion from pilots not being able to point or exchange paper (e.g., charts), although, most importantly, this did not appear to interfere with the quality of decision making.

The measure of SA revealed that a VRFS generated more overall SA than a desktop simulator. As discussed earlier, it can be concluded that TSA was present since individual SA is a reliable indicator for TSA in collaborative environments. The SART domains that gave rise to this difference in calculated SA were Attentional Supply and Understanding (see Table III). With reference to Table III and the definitions of the SART domains [42], in a VRFS compared to a desktop computer, the SART domains reveal that users:

- had spare mental ability and are ready for new activity (from the Attentional Supply domain);
- could concentrate more on the scenario and were more focused (from the Attentional Supply domain);
- received and understood more knowledge, quite possibly because they received better quality information (from the Understanding domain); and
- feel more familiar with the scenario (from the Understanding domain).

The measure of perceived workload revealed less effort is required in a VRFS compared to a desktop simulator (see Table IV). This is surprising, since overall, the participants had less VRFS experience compared to desktop simulator experience, and it could be expected that workload would be greater in an unfamiliar environment. This finding is substantiated by [57] who suggest that reduced workload in a collaborative VE is attributable to communication behavior. That is, a reduction in deictic gestures and an increase in direct communication in a collaborative VE reduces workload. They also suggest that increased individual spatial memory provided by a VE may also be responsible for reduced workload.

A. *Limitations and Future Research*

The limitations of the study were:

- Scenario 1 (Heathrow) required more researcher intervention than scenario 2 (Sydney) because it involved some flying and a landing, and participants had very little B738 operating experience.
- Design limitations imposed a number of restrictions in the VE, including restrictive checklists (items had to be presented one at a time), inability to take notes, and the unavailability of scenario guides and charts. Although this meant that more researcher guidance was given to participants in the VE and these limitations may have impacted the usability of the simulator in the VE, there was no indication that it had an adverse effect on the type of measures undertaken during the study and the subsequent results.
- The software imposed some limitations, such as four of the switches between the slave and the master computer not being mapped correctly. This was identified pre-experiment which allowed participants to be forewarned and may have impacted the usability of the simulator, although there was no indication that it had an adverse effect on the results. In addition, one participant reported a mismatch between the physical throttle height and that observed in the VE. This was because the desk height did not match the viewed position of the throttle quadrant in the VE when a participant was in the normal seated position, where the throttles were lower, and more forward, as in a real cockpit (for example, see Fig. 8). There was no indication that it had an adverse effect on the results, although it could confound any measures of reaction times.
- The PF and PNF observed some different static aircraft (not the AI controlled aircraft). This could cause confusion when, for example, parking the aircraft and one participant observed an empty gate but the other observed an occupied gate. There was no indication that it had an adverse effect on the results.
- The sample size was small, which could lead to more type II errors (i.e., an increase in false negatives). However, the sample size was deemed appropriate for the aims of the study.

VI. Conclusions

Valuable insights have been gained in terms of creating a reliable environment for further research on collaborative VRFSs. The study has shown that participants can successfully collaborate in a VE when conducting simulator sessions aligned to that of a typical initial First Officer airline training program in a complex commercial aircraft.

Greater understanding and knowledge were acquired in a VE compared to a desktop simulator, quite possibly because of enhanced immersion (immersion is discussed in [2]). Non-verbal collaborative communication was replaced by more succinct verbal communication in a VE, with reduced perceived workload (effort) due to reduced cognitive processing. This in turn most likely increased spatial memory, which was demonstrated by heightened SA in a VE compared to a desktop simulator. This is substantiated by [58], who found that non-verbal communication, in the form of gesture commands, created additional workload (cognitive and physical), and also direct voice communication significantly reduced workload while maximizing SA. A possible explanation to that of increased spatial memory in a VE is that collaborators only have to focus on the task in hand and precise communication, as opposed to interpreting, responding to, and being distracted by non-verbal deictic gestures.

The benefits of these findings are numerous, including the concept that VRFSs provide greater mental capacity over traditional simulators, giving trainees an increased capacity to learn and be more productive in a VE, and trainees will be less fatigued in a VRFS. However, consideration should be given to the fact that pilots operate face-to-face in a real environment. Other benefits of VEs, such as those listed by [2], become readily available. However, much more research needs to be undertaken, especially in the areas of knowledge acquisition, development of procedures and flying skills, transfer of training, and the application of human factors principles.

References

[1] R. T. Hays, J. W. Jacobs, C. Prince, and E. Salas, "Flight simulator training effectiveness: a meta-analysis," *Military Psychology*, vol. 4, no. 2, pp. 63-74, Jun. 1992, doi: 10.1016/s0166-4115(08)62386-9

[2] J. I. Cross, C. C. Boag-Hodgson, T. Ryley, T. J. Mavin, and L. E. Potter, "Using extended reality in flight simulators: a literature review," *IEEE Trans. on Visualization and Computer Graphics*, vol. 29, no. 9, pp. 3961-3975, Sep. 2023, doi: 10.1109/TVCG.2022.3173921

[3] J. I. Cross, C. C. Boag-Hodgson, and T. J. Mavin, "Measuring presence and situational awareness in a virtual reality flight simulator," *Aviation Psychology and Applied Human Factors*, vol. 13, no. 2, pp. 83–94, Jun. 2023, doi: 10.1027/2192-0923/a000250

[4] Airbus. (2023). *Global market forecast 2022-2041*. [Online]. Available: https://www.airbus.com/en/products-services/commercial-aircraft/market/global-market-forecast

[5] Boeing. (2023). *Pilot and technician outlook 2022-2041*. [Online]. Available: https://www.boeing.com/commercial/market/pilot-technician-outlook/

[6] M. A. Todd, and M. J. W. Thomas, "Flight hours and flight crew performance in commercial aviation," *Aviation, Space, and Environmental Medicine*, vol. 83, no. 8, pp. 776–82, Aug. 2012, doi: 10.3357/asem.3271.2012

[7] M. A. Todd, and M. J. W. Thomas, "Pilot experience and performance in an airline environment," *ATSB Transport Safety Report, Aviation Research Investigation, AR 2012-023*, Jul. 2013.

[8] M. A. Todd, and M. J. W. Thomas, "Experience, competence, or syllabus? Influences on flight hours at licensing of commercial pilots," *The International Journal of Aviation Psychology*, vol. 23, no. 2, pp. 169–180, Apr. 2013, doi: 10.1080/10508414.2013.772851

[9] M. Oberhauser, D. Dreyer, R. Braunstingl, and I. Koglbauer, "Pilots' flight performance in a virtual reality flight simulator," in *Proc. of the 32nd Conf. of the European Association of Aviation Psychology*, Groningen, The Netherlands, 2017, pp. 322–335.

[10] T. H. Tran, F. Behrend, N. Funning, and A. Arango, "Single pilot operations with AR-glasses using Microsoft HoloLens", paper presented at the *2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC)*, London, pp. 1-7, doi: 10.1109/DASC.2018.8569261

[11] Aslandere, T. I., Dreyer, D., Pantkratz, F., & Schubotz, R, "A generic virtual reality flight simulator," in *Virtuelle und Erweiterte Realität, 11. Workshop der GI-Fachgruppe VR/AR*, G. Zachmann, R. Weller, and A. Hinkenjann, Eds., Aachen, Germany: Shaker, 2014, pp. 1–13.

[12] D. Harris, "Single-pilot airline operations: designing the aircraft may be the easy part," *The Aeronautical Journal*, vol. 127, no. 1313, pp. 1171–1191, Feb. 2023, doi: 10.1017/aer.2022.110

[13] W. R. Hamman, “Line oriented flight training (LOFT): the intersection of technical and human factor crew resource management (CRM) team skills,” in *Cockpit Resource Management*, B. G. Kanki, R. L. Helmreich, and J. M. Anca, Eds., Burlington: Elsevier, 2010, pp. 233–263, doi: 10.1016/B978-0-12-374946-8.10008-1

[14] R. L Helmreich, A. C. Merritt, and J. A. Wilhelm, “The evolution of crew resource management training in commercial aviation,” *The International Journal of Aviation Psychology*, vol. 9, no. 1, pp. 19–32, Jul. 2017, doi: 10.1207/s15327108ijap0901_2

[15] E. Salas, J. E. Fowlkes, R. J. Stout, D. M. Milanovich, and C. Prince, “Does CRM training improve teamwork skills in the cockpit?: two evaluation studies,” *Human Factors*, vol. 41, no. 2, pp. 326–343, Jun. 1999, doi: 10.1518/001872099779591169

[16] P. Salmon, N. Stanton, G. Walker, and D. Green, “Situation awareness measurement: a review of applicability for C4i environments,” *Applied Ergonomics*, vol. 37, no. 2, pp. 225-238, Mar. 2006, doi: 10.1016/j.apergo.2005.02.001

[17] K. Smith, and P. A. Hancock, “Situation awareness is adaptive, externally directed consciousness,” *Human Factors*, vol. 37, no. 1, pp. 137-148, Mar. 1995, doi: 10.1518/001872095779049444

[18] M. R. Endsley, “Toward a theory of situation awareness in dynamic systems,” *Human Factors*, vol. 37, no. 1, pp. 32–64, Mar. 1995, doi: 10.1518/001872095779049543

[19] M. She, and Z. Li, “Team situation awareness: a review of definitions and conceptual models,” in *Proc. International Conf. on Engineering Psychology and Cognitive Ergonomics*, pp. 406-415, May 2017, doi: 10.1007/978-3-319-58472-0_31

[20] H. Artman, “Situation awareness and co-operation within and between hierarchical units in dynamic decision making,” *Ergonomics*, vol. 42, no. 11, pp. 1404–1417, Nov. 1999, doi: 10.1080/001401399184776

[21] H. Artman, “Team situation assessment and information distribution,” Ergonomics, vol. 43, no. 8, pp. 1111–1128, Sep. 2000, doi: 10.1080/00140130050084905

[22] N. J. Cooke, R. Stout, and E. Salas, “A knowledge elicitation approach to the measurement of team situation awareness,” in *New Trends in Cooperative Activities: Understanding System Dynamics in Complex Environments*, M. McNeese, E. Salas, and M. R. Endsley, Eds, Santa Monica, CA: Human Factors and Ergonomics Society, Jan. 2001, pp. 114–139.

[23] S.W.A. Dekker, “Crew situation awareness in high-tech settings: tactics for research into an ill-defined phenomenon,” *Transportation Human Factors*, vol. 2, no. 1, pp. 49–62, Mar. 2000, doi: 10.1207/STHF0201_7

[24] J. E. Fowlkes, E. Salas, D. P. Baker, J. A. Cannon-Bowers, and R. J. Stout, “The utility of event-based knowledge elicitation,” *Human Factors*, vol. 42, no. 1, pp. 24–35, Feb. 2000, doi: 10.1518/001872000779656615

[25] M. R. Endsley, and W. M. Jones, “Situation awareness, information dominance and information warfare,” *Technical Report 97-01*, US Air Force Armstrong Laboratory, 1997.

[26] M. R. Endsley, and W. M. Jones, “A model of inter and intra-team situation awareness: implications for design, training and measurement,” in *New Trends in Cooperative Activities: Understanding System Dynamics in Complex Environments*, M. McNeese, E. Salas, and M. R. Endsley, Eds., Santa Monica, CA: Human Factors and Ergonomics Society, Jan. 2001, pp. 46–68.

[27] M. R. Endsley, and M. M. Robertson, “Training for situation awareness,” in *Situation Awareness Analysis and Measurement*, M. R. Endsley, and D. J. Garland, Eds., Mahwah, NJ: Erlbaum, Jan 2000, pp. 349–365.

[28] L. Saner, and C. Bolstad, “Measuring and predicting shared situation awareness in teams,” *Journal of Cognitive Engineering and Decision Making*, vol. 3, no. 3, pp. 280–308, Oct. 2009, doi: 10.1518/155534309X474497

[29] M. A. Vidulich, “Mental workload and situation awareness: essential concepts for aviation psychology practice,” in *Principles and Practice of Aviation Psychology*, P. S. Tsang, and M. A. Vidulich, Eds., Mahwah, NJ: Erlbaum, Nov. 2003, pp. 115-146.

[30] C. D. Wickens, “Situation awareness and workload in aviation,” *Current Directions in Psychological Science*, vol. 11, no. 4, pp. 128–133, Aug. 2002, doi: 10.1111/1467-8721.00184

[31] S. G. Hart, and L. E. Staveland, “Development of NASA-TLX (Task Load Index): results of empirical and theoretical research,” *Advances in Psychology*, vol. 52, pp. 139–183, 1988, doi: 10.1016/s0166-4115(08)62386-9

[32] R. Guthridge, R., and V. Clinton-Lisell, “Evaluating the efficacy of virtual reality (VR) training devices for pilot training,” *Journal of Aviation Technology and Engineering*, vol. 12, no. 2, pp. 1-14, Jul. 2023, doi:10.7771/2159-6670.1286

[33] J. I. Cross, and T. Ryley, “Assessing evidence-based training in a collaborative virtual reality flight simulator,” to be published.

[34] T. H. Laine, and W. Lee, W, “Collaborative virtual reality in higher education: students’ perceptions on presence, challenges, affordances, and potential,” *IEEE Trans. on Learning Technologies*, vol. 17, pp. 280-293, Sep. 2023, doi:10.1109/TLT.2023.3319628

[35] F. Buttussi, and L. Chittaro, “Effects of different types of virtual reality display on presence and learning in a safety training scenario,” *IEEE Trans. on Visualization and Computer Graphics*, vol. 24, no. 2, pp. 1063-1076, Feb. 2018, doi:10.1109/TVCG.2017.2653117

[36] F. Buttussi, and L. Chittaro, “A comparison of procedural safety training in three conditions: virtual reality headset, smartphone, and printed materials,” *IEEE Trans. on Learning Technologies*, vol. 14, no. 1, pp. 1-15, Feb. 2021, doi:10.1109/TLT.2020.3033766

[37] N. D. Macchiarella, T. Brady, and B.S. Lyon, “An application of high fidelity FTDs for ab initio pilot training: the way ahead,” *Collegiate Aviation Review*, vol. 26, no. 1, pp. 67-75, Oct. 2018, doi:10.22488/okstate.18.100367

[38] G. M. T. McLean, S. Lambeth, and T. J. Mavin, “The use of simulation in ab initio pilot training,” *The International Journal of Aviation Psychology*, vol. 26, no. 1-2, pp. 36-45, Nov. 2016, doi:10.1080/10508414.2016.1235364

[39] Airbus. (2022, Nov. 8). *Airbus virtual procedure trainer offers an innovative way for pilots to learn procedures using virtual reality - Lufthansa Group becomes launch customer*. [Online]. Available: https://aircraft.airbus.com/en/newsroom/press-releases/2022-11-airbus-virtual-procedure-trainer-offers-an-innovative-way-for

[40] FlightGlobal. (2022, Nov. 9). *Lufthansa launches Airbus virtual-reality cockpit training for A320s*. [Online]. Available: https://www.flightglobal.com/airlines/lufthansa-launches-airbus-virtual-reality-cockpit-training-for-a320s/150899.article

[41] Airbus. (2019, Dec. 4). *Airbus brings cockpit to you with new virtual reality flight trainer.*” [Online]. Available: https://aircraft.airbus.com/en/airbus-brings-cockpit-to-you-with-new-virtual-reality-flight-trainer

[42] R. M. Taylor, “Situational awareness rating technique (SART): the development of a tool for aircrew systems design,” in *Situational Awareness*, R. M. Taylor, Ed., Routledge, 2017, pp. 111–128, doi: 10.4324/9781315087924-8

[43] L. J. Cronbach, “Coefficient alpha and the internal structure of tests,” *Psychometrika*, vol. 16, no. 3, pp. 297–334, Sep. 1951, doi: 10.1007/BF02310555

[44] P. Ø. Braarud, “Investigating the validity of subjective workload rating (NASA-TLX) and subjective situation awareness rating (SART) for cognitively complex human-machine work,” *International Journal of Industrial Ergonomics*, vol. 86, Nov. 2021, doi: 10.1016/j.ergon.2021.103233

[45] S. G. Hart, “NASA-task load index (NASA-TLX); 20 years later,” in *Proc. of the Human Factors and Ergonomics Society Annual Meeting*, vol. 50, no. 9, pp. 904-908, Oct. 2006, doi: 10.1177/154193120605000

[46] S. Rubio, E. Díaz, J. Martín, and J. M. Puente, “Evaluation of subjective mental workload: a comparison of SWAT, NASA-TLX, and workload profile methods,” *Applied Psychology*, vol. 53, no. 1, pp. 61-86, Jan. 2004, doi: 10.1111/j.1464-0597.2004.00161.x

[47] Y. M. Xiao, Z. M. Wang, M. Z. Wang, and Y. J. Lan, “The appraisal of reliability and validity of subjective workload assessment technique and NASA-task load index,” *Chinese Journal of Industrial Hygiene and Occupational Diseases*, vol. 23, no. 3, pp. 178-181, Jun. 2005, doi: 10.3760/CMA.J.ISSN.1001-9391.2005.03.007

[48] R. D. McKendrick, and E. Cherry, “A deeper look at the NASA-TLX and where it falls short”, in *Proc. of the Human Factors and Ergonomics Society Annual Meeting*, vol. 62, no. 1, pp. 44–48. Sep. 2018, doi: 10.1177/1541931218621010

[49] ICAO. (2013). *Manual of evidence-based training (Doc 9995).*1st ed. [Online]. Available: https://store.icao.int/en/manual-of-evidence-based-training-doc-9995

[50] ICAO. (2024, Jan.). *Evidence-based training implementation guide*. 2nd ed. [Online]. Available: https://www.iata.org/contentassets/c0f61fc821dc4f62bb6441d7abedb076/ebt-implementation-guide.pdf

[51] ICAO. (2018, Jul.). *International standards and recommended practices. Annex 1 to the convention on international civil aviation personnel licensing.* 12th ed. [Online]. Available:

https://www.icao.int/APAC/Meetings/2019%20COSCAP%20SEAEASA%20PEL/AN01_cons.2019_compressed.pdf
[52] ICAO. (2023). CBTA Instructional system design. [Online]. Available: https://www.icao.int/safety/OPS/OPS-Normal/Pages/CBTA-Instructional-system-design.aspx
[53] CASA. (2023, Nov. 2). *Getting an aircraft type rating*. [Online]. Available: https://www.casa.gov.au/licences-and-certificates/pilots/ratings-reviews-and-endorsements/ratings-and-endorsements/getting-aircraft-type-rating#Differencestraining
[54] K. Virtanen, H. Mansikka, H. Kontio, and D. Harris, "Weight watchers: NASA-TLX weights revisited," *Theoretical Issues in Ergonomics Science*, vol. 23, no. 6, pp. 725–748, Nov. 2021, doi: 10.1080/1463922X.2021.2000667
[55] S. McLean, *The basics of speech communication*, Boston, MA: Allyn & Bacon, 1st ed, May 2002.
[56] J. Lachter, S. L. Brandt, V. Battiste, S. V. Ligda, M. Matessa, and W. W. Johnson, "Toward single pilot operations: developing a ground station," in *Proc. of the International Conference on Human-Computer Interaction in Aerospace (HCI-Aero '14)*, New York, NY, USA. Article 19, pp. 1–8. Association for Computing Machinery (ACM), Jan. 2014, doi: 10.1145/2669592.2669695
[57] J. Müller, R. Radle, and H. Reiterer, "Virtual objects as spatial cues in collaborative mixed reality environments: how they shape communication behavior and user task load," *in Proc. of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*, San Jose, California, USA, pp. 1245–1249, May 2016, doi: 10.1145/2858036.2858043
[58] W. Li, J. Zhang, S. Court, P. Kearney, and G. Braithwaite, "The influence of augmented reality interaction design on pilot's perceived workload and situation awareness," *International Journal of Industrial Ergonomics*, vol. 92, Nov. 2022, doi: 10.1016/j.ergon.2022.103382

Dr. Jamie I. Cross is a Senior Lecturer and the Program Director of the Graduate Diploma of Flight Management at Griffith University, Australia. Jamie flew for a European airline before relocating to Dubai to join Emirates as an instructor. He then worked as an aviation consultant for several years, engaged by leading flight schools, universities, and major airlines to design and deliver training content. He specializes immersive technologies, pilot training, risk management and advisory boards, and is currently conducting research on the use of virtual reality for ab-initio, commercial and military pilot training.

Dr Christine C. Boag-Hodgson is a Team Leader Airspace Strategy at the Civil Aviation Safety Authority (Australia) and is an endorsed organizational psychologist. Formerly she was the Head of Griffith Aviation within the School of Engineering and Built Environment (Aviation) at Griffith University, QLD, Australia. Her career incorporates over 15 years applied aviation industry experience in senior safety and human factors roles and over 20 years in academic appointments. Her research involves enhancing operational safety and improving performance efficiency, including accident analysis and cognitive complexity.