Self-evaluation vs. objective performance measures: Evaluation of fidelity, presence and training transfer in two helicopter simulator tasks

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Master in Philosophy, by Li Ting Wong.

September, 2014

Table of Contents

Self-evaluation vs. objective performance measures: Evaluation of fidelity, presence and training transfer in two helicopter simulator tasks

Abstract

Simulations are widely used in aviation, medical and military training. Simulation *fidelity* is an important element of simulation training development. This study explores the reliability of self-evaluation of *fidelity* and *training transfer* in comparison to objective performance measures in two experiments. In Experiment 1, fifteen participants (aged 20-32, mean= 23.1) completed a target-tracking task in the HELIFLIGHT simulator at the University of Liverpool. They then underwent training on a desktop-based helicopter simulator with basic visuals and a realistic turbine rotor noise as their only motion cues before being re-tested in HELIFLIGHT. Motion cue fidelity was manipulated to explore effects on subjective post-training ratings of fidelity and self-performance.

In Experiment 2, eleven participants (aged 22-27, mean= 24.2) performed a hover task in HELIFLIGHT. As in Experiment 1, they then trained on a desktop-simulator before being retested in HELIFLIGHT. Again, cue fidelity was manipulated to explore effects on ratings of fidelity and self-performance, but also on sensation of presence.

In both experiments, subjective post-trial ratings were compared with continuously-sampled objective measures. Both experiments showed that participants benefited from transferrable training from desktop simulator to full flight simulator. However, participants could not always reliably evaluate their own performance (Experiment 1). Additionally, participants could not always reliably judge cue fidelity (Experiment 1), and fidelity judgements did not always correspond with objective performance measures (Experiments 1 and 2). Self-evaluation of training also did not reflect objective measures of performance (Experiments 1 and 2), but participants did report greater subjective presence with multisensory motion cues compared to without. These findings contribute to the exploration of suitable metrics for fidelity, presence and performance evaluation.

Chapter 1: Introduction

Simulations are artificial environments that are designed to operationally model the features of the real environment they represent (Wignall, Denstedt, Preminger, Cadeddu, Pearle, Sweet & McDougall, 2008). Nowadays, simulations are widely used in aviation, medical and military training (Bowman & McMahan, 2007). For example, simulations have been used as training tools with the view to transfer of skills into the operating room (Seymour, Gallagher, Roman, O'Brien, Bansal, Andersen & Satava, 2002; Aggarwal, Black, Hance, Darzi & Cheshire, 2006), as well as exploration of aviation handling and manoeuvres (Lee, Sezer-Uzol, Horn & Long, 2005; Casner, Geven & Williams, 2013) and is being invested in by the military as a training tool (Lele, 2013). Simulations can take the form of something as simple as a physical object substituting the real object (for example, a wooden horse in place of a real horse), to an immersive 3D Virtual Reality simulation that affords the sensation of a physical presence in the modelled environment. The flexibility of VR simulations allows the creation of lifelike environments for the purpose of training. However, even VR simulations cannot fully replicate all aspects of the real environment, such as visual details that correspond exactly with the real world, or the stimulation of multiple senses (Gallace, Ngo, Sulaitis & Spence, 2011). Therefore, a key issue is whether the cues present in the simulation can enable transferrable learning between the simulated and real environments (Alexander, Brunyé, Sidman & Weil, 2005).

In the case of virtual reality, *fidelity* can be defined as "the objective degree of exactness with which real-world experiences and effects are reproduced by a computing system" (Gerathewohl, 1969, in McMahan, Bowman, Zielinski & Brady, 2012). It is a multifaceted concept that can be broken down into different varieties (Ferwerda, 2003). Simulation development tends to emphasise the importance of *physical fidelity*, where physical characteristics of the real environment are replicated in the simulation as closely as possible (Liu, Blickensderfer, Macchiarella & Vincenzi, 2009). *Physical fidelity* includes the simulation of not only the visual aspects of the environment, but also the auditory, vestibular and even olfactory sensations of the real world (Alexander et al., 2005). Attempts to create the illusion of *physical fidelity* might involve the recreation of a realistic control set-up in, for example, a driving simulator, with a finely-detailed out-of-window display depicting a realistic driving environment. However, exact replication of a real environment is costly, impractical (Londgridge, Bürki-Cohen, Go & Kendra, 2001), and does not help to identify components of the simulation that most assist training transfer. To emphasise training transfer, simulations should afford high *functional fidelity*, which is the extent to which procedural skills in the virtual environment mimic those in the real environment. With regards to our hypothetical driving simulator, this would mean that a high-fidelity simulation would behave in the same manner as a real car, and that the performance of tasks undertaken in the simulation should be transferrable to the real world.

If *physical* and *functional* fidelity successfully model the real environment, then the features of the simulated environment should cause the user to behave and experience sensations in a manner consistent with the real world. This is known as *psychological fidelity*, where the user's perceptions of and reactions to cues in the simulated environment mimic those within the real environment (Duncan, 2006). This aspect of fidelity emphasises comparable levels of the same stress, arousal and emotional responses provoked by the simulation with what would be induced in the real environment (Alexander et al., 2005).

The psychological and emotional impact of a simulation upon the user is closely linked to the sensation of *presence*, which captures the subjective feelings of 'being there' in the modelled environment (Baños, Botella, Alcañiz, Liaño, Guerrero, & Rey, 2004**;** Pausch, Proffitt & Williams, 1997). When a person feels a sense of *presence* in a simulation, they feel more engaged with the simulation than the physical world around them, and that any of their behaviours are self-perceived as taking place within the simulation rather than the real world (Slater & Wilbur, 1997). Witmer and Singer (1998) postulate that several factors contribute to the level of *presence* a person feels in a simulation: control factors, sensory factors, distraction factors and realism factors. In summary, a person would feel a greater sense of *presence* if they possessed a great degree of control over the simulated environment; if they could perceive a rich variety of sensations from the simulated environment; if their attention were focused on the simulated environment rather than the real world; and if the realism of the simulated environment is consistent with the real world.

Another important aspect of feeling present in simulated environments is *immersion*, which as of yet has no fixed definition (Lackey, Maraj, & Barber, 2014). Some argue that *immersion* should be defined as the technical capability of a system to induce the sensation of presence (Slater & Wilbur, 1997). Others believe that *immersion, like presence* should focus on the individual experience the simulation, where a person is mentally and physically involved in the simulated environment (Sherman & Craig, 2003). Despite there being no consensus on the definition of *immersion*, they show that the determinants contributing to the sensation of presence rely upon the capability of the simulation to produce a high-fidelity, compelling environment; for example, positive correlations have been found between subjective measures of presence and visual realism (Mania & Robinson, 2004; Slater, Khanna, Mortensen & Yu, 2009). Although it could be argued that greater simulation fidelity is not necessary to induce a sensation of *presence* (Zimmons & Panter, 2003; Ooms, 2004), the production of a compelling simulated environment can be achieved through the stimulation of multiple senses (including vision, audition, touch and even taste), regardless of whether or not these sensations are realistic or expected in the environment (Gallace et al., 2011; Dinh, Walker, Hodges, Song & Kobayashi, 1999; Sanchez-Vives & Slater, 2005).

Given that full replication of a real-world environment is not currently possible, it is important to explore which cues are the most important to support functional fidelity and presence. Fidelity and presence may be promoted by the use of congruent multisensory cues; for example, auditory and vestibular as well as visual cues. Soto-Faraco, Kingstone and Spence (2003) reviewed how the use of unimodal and multimodal factors influences motion perception, and the extent to which information presented in one modality affected perception of motion for stimuli in other modalities (visual, auditory and vestibular). Visual motion cues dominated auditory and somatosensory motion cues, but performance was better when congruent motion information was presented in multiple modalities. They suggested that multisensory integration of motion signals occurs early in perceptual processing and contributes an important role in determining performance in tasks. The effect of congruent cue presentation in multiple modalities is known as the Redundant Signal Effect (RSE) (Miller, 1982). RSE explains demonstrable improvements in task performance (e.g. reaction time) due to the presentation of cues that are congruent (moving in the same direction), spatially-aligned (presented from the same spatial location) and temporally aligned (presented at such times that they would be perceived as simultaneous by the observer) (Harrison, Wuerger & Meyer, 2010; Hancock, Mercado, Merlo & Van Erp, 2013). Therefore, it may be of merit to include congruent auditory motion cues to promote task performance and enhance simulator fidelity (Väljamäe, 2007; Väljamäe, Larsson, Västfjäll & Kleiner, 2008), as well as vestibular cues (Berger, Schulte-Pelkum & Bülthoff, 2010; Meyer, Wong, Timson, Perfect & White, 2012). However, the information in these cues should be both accessible and behave in a manner that is expected – i.e. to correspond with participant behaviour and visual feedback. Therefore the limitations of multisensory cue production should be held in mind. For example, physical limitations in current motion platforms mean that it is not possible to produced sustained acceleration cues, and thus the vestibular cues produced could be described insufficiently realistic (Bürki-Cohen & Go, 2005; Bürki-Cohen, Sparko & Go, 2007). Still, it has been argued that lower-physical fidelity simulations are sufficiently effective training tools and help researchers to identify the contribution of individual cues to task performance (Patrick, 1992; Taylor, Lintern, Hulin, Talleur, Emanuel Jr. & Phillips, 1999), because the simulation is not overloaded with numerous (and possibly redundant) sensory cues (Dahlstrom, Dekker, van Winsen & Nyce, 2009).

Synthetic auditory cues are often used in virtual environments in the form of Head-Related Transfer Functions (HRTFs). HRTFs are created by reproducing a sound at the pressure level which would be found in a person's ear canal. This enables use of the HRTF to infer the sound's location in multiple dimensions. This means that HRTFs should ideally be unique to an individual's ears. However, most HRTF measurements are taken from multiple human subjects or mannequins and are therefore more difficult for the individual listener to locate in comparison to real sounds (Nykänen, Zedigh, & Mohlin, 2013). There are especial difficulties in localisation when HRTFs are presented in front of or behind participants, as they often confuse the two locations (Cho, Ovcharenko & Chong, 2006), or if the listener does not keep still. One practical way to overcome this difficulty is to use a head-tracking system with binaural HRTFs so that the spatial information of auditory cues is preserved if the listener moves his or her head (Wightman & Kistler, 1999; Seeber & Fastl, 2003). Utilising multisensory cues in this manner may promote simulation fidelity and the sensation of presence, thereby encouraging training transfer from simulation to the real world (Hale, Stanney & Malone, 2009).

If a simulation affords the ability to train users in a particular task, then one can explore how skills honed in the simulation transfer to the real world. There are a number of ways to evaluate simulator training outcomes. For example, experts subjectively evaluate trainees on their skills (Hyltander, Liljegren, Rhodin & Lönroth, 2002; Watterson, Beiko, Kuan & Denstedt, 2002) and occasionally, trainee performance is compared with expert performance on identical tasks (Judkins, Oleynikov & Stergiou, 2009). Such comparisons are also used in aviation training, where behavioural measures taken from expert and novice pilots during simulated tasks are used as performance metrics to identify pilot strategies (Kasarskis, Stehwein, Hickox, Aretz & Wickens, 2001). On the other hand, subjective ratings of fidelity, presence or performance are useful if participants or examiners can clearly perceive and are aware of the cues or behaviours to be judged (Watterson et al., 2002; Jeon, Kim, Cabrera & Bassett, 2008). However, they may be limited in their use when probing attitudes to less accessible features of a simulation, such as subtle pitch or amplitude changes in auditory motion cues. Additionally, rating measures are vulnerable to typical shortcomings associated with self-evaluation (Wiggins & O'Hare, 2003). For example, unlike continuously sampled objective measures, subjective data acquisition normally has to take place after the task has been completed. Subjective data therefore cannot reveal real-time changes in attitude and whether participants are influenced by the most recently-occurring events (Riva, Davide & IJsselsteijn, 2003). Furthermore, in the cases of both presence and fidelity there are effects of personal bias and expectation during assessment. This can be undesirable because participants may evaluate simulation features on the basis of previous experience as opposed to their appropriateness for the current task. This suggests that qualitative measures taken from novices are a poor tool for functional and physical fidelity evaluation (Schricker, Franceschini & Johnson, 2001).

Objective measures of task performance or participant behaviour may provide robust and sensitive referents of fidelity and presence to support subjective measures (Schricker et al., 2001; Meehan, Insko, Whitton & Brooks Jr., 2002). A referent is defined as: “a formal representation of reality that is intermediate between reality and the simulation” (Schricker et al., 2001). For example, task performance can be used as a measure of how well a simulation models an environment in comparison to the real world – the model that affords the best task performance could be considered as having greater fidelity. Some studies have found correlations between subjective presence measures and task performance in the virtual environment (Witmer & Singer, 1998; Sallnäs, Rassmus-Gröhn & Sjöström, 2000). However, there have also been instances where objective task performance does not correspond with subjective presence measures. For example, Mania and Chalmers (2001) presented a 15-minute seminar across four different levels of reality (real world, 3D desktop, 3D head mounted display and audio-only), and tested participants’ recall and sense of presence. It was found that while subjective ratings of presence were highest in the ‘real world’ condition, this did not correspond with recall performance. Similarly, Durlach, Fowlkes and Metevier (2005) found no systematic relationship between subjective presence ratings and objective performance measures where participants were required to quickly reach out with a virtual hand and accurately touch a virtual target. It could therefore be argued that presence might be better evaluated via subjective assessment rather than via a referent such as task performance.

Although there are methods to subjectively assess the fidelity of flight simulators in relation to task performance (e.g. the Cooper-Harper (1969) Handling Qualities Scale and the Simulation Fidelity Rating Scale (Perfect, Timson, White & Padfield, 2014)), there are no standardised scales similar in format to those produced for subjective presence evaluation (e.g. Witmer & Singer, 1998). The Cooper-Harper (1969) scale is used by test pilots to evaluate the handling qualities of an aircraft, and requires some knowledge of the aircraft and its workings. The Simulation Fidelity Rating Scale (Perfect et al., 2014) involves the ability of a simulation to permit task execution, and how well the simulation promotes task performance and training transfer between the simulation and the real world. This scale uses task performance as an objective referent. These two scales are therefore unlike Witmer and Singer’s Presence Scale (1998), which is a simple Likert scale that can be used by novices, is applicable to almost any simulation, and explores the subjective attitudes of a person toward a simulated environment.

In consideration of this, two experiments were conducted to explore subjective evaluation of simulation fidelity, sensation of presence and self-performance in comparison to objective performance measures. In Experiment One, fifteen naïve participants completed a target-tracking task in the HELIFLIGHT simulator at the University of Liverpool. They then underwent training on a low-fidelity, desktop-based simulator with a realistic turbine rotor noise as their only motion cue. They were re-tested in HELIFLIGHT under the same conditions before training, but with the introduction of an additional ‘substitute’ auditory cue that did not resemble a helicopter turbine. Subjective ratings of fidelity and self-performance were taken after each trial. Continuous objective performance measures were taken throughout: mean of median error and control input. Control input was recorded as control input activity can reveal changes in pilot behaviour not revealed by overall flight path error (Harris, 2011).

Experiment One showed that whilst participants perceived and made use of congruent vestibular cues, they were unable to exploit auditory motion cues to improve task performance. In light of this, Experiment Two explored whether the use of head-tracked HRTFs would enable participants to more effectively localise auditory motion cues. As in Experiment One, Experiment Two further explored the relationship between objective performance data and subjective evaluation of *fidelity* in a simulated flying task. Expanding upon Experiment One, Experiment Two also examined participants’ subjective evaluation of *presence*, and how this affected their task performance. Eleven naïve participants completed a hover task in HELIFLIGHT where they were required to hover in a stationary fashion whilst the aircraft was externally rotated by simulated wind. They were then trained to practise the task on a low-fidelity simulator set-up whilst exposed to an auditory motion cue. Objective measures were taken to evaluate both task performance and participants’ behavioural responses during the task. Post-trial ratings of subjective fidelity, presence and self-performance were also collected.

In consideration of the above literature, the following hypotheses were formed:

1. Subjective measures are a poor tool for fidelity evaluation in comparison to an objective referent such as task performance.
2. Subjective measures are more appropriate for evaluating presence compared to task performance.
3. Low-fidelity simulations are adequate tools for promoting training transfer to higher-fidelity simulations.
4. In the absence of an expert examiner, objective evaluation of training and performance is more suitable than subjective self-evaluation.

Chapter 2: Experiment One – Basket-Tracking Task

Method

Ethics Statement

Both experiments have been approved by the University of Liverpool ethics committee (reference PSYC09100027). Written informed consent was acquired from all participants.

Participants

Fifteen participants obtained via opportunity sampling took part in Experiment One (range 19-32 years, mean= 22.6, twelve males). All participants reported normal or corrected-to-normal vision and normal hearing. Some participants had computer game experience and one had flight simulator experience prior to the experiment, but none had prior experience of the specific simulators or task used during this study.

Apparatus and Materials

Two simulators were used: a high-fidelity, motion-enabled simulator was used in testing (Fig. 1) and a low-fidelity desktop simulator was used in training (Fig. 2). Both simulators shared a common model determining their behaviour. The target moved along a predetermined path which participants were required to follow for 15 minute training blocks. The height variation was defined as a sum of 6 sinusoidal signals, with frequencies ranging from 0.2 to 0.5 rad/s, each with a different amplitude and phase shift.

Fig. 1: Visual display in HELIFLIGHT.

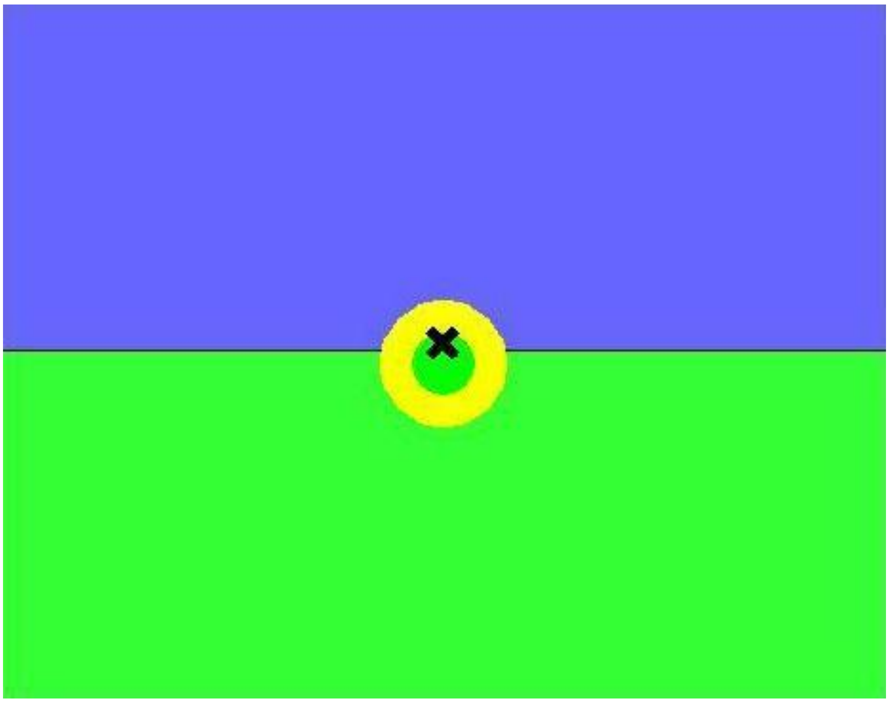

Fig. 2: Visual display in the low-fidelity simulator.

The high-fidelity HELIFLIGHT simulator based at the University of Liverpool's School of Engineering was used during the pre- and post-training tests (Padfield & White, 2003). The visual display consisted of a simulated flight path at 1500 ft above ground and contained a representation of a tanker plane and refuelling basket (Fig. 1). Auditory stimuli were delivered via loudspeakers in the simulator capsule at 87.5 dB(A) while the pilots wore sound attenuating headphones (Flightcom 4DLX (attenuation – 24dB)). The audio signal consisted of two components, the rotor sound and a turbine sound. Both sounds were continuous loops that were generated under control of a Tucker-Davies RM1 real time processor. The turbine signal pitch and amplitude was modulated in direct proportion to the control input. The rotor playback speed (and pitch) was always constant but the rotor signal amplitude covaried with the control input. The overall signal level varied by 3dB (86 – 89 dB(A)). Vestibular cues were delivered via a Maxcue 600 series motion platform. Platform motion was restricted to vertical movements, which were under the control of the participants. An acceleration signal was used to drive the motion platform. A washout filter was used to deliver motion cues within the restricted simulator workspace. Other features of the capsule included a realistic helicopter control set-up, including a collective lever to the left of the pilot's seat which was used for vertical movement of the 'helicopter'. The instrumentation panel was off during all experiments.

The low-fidelity simulator was computer-based in the Visual Perception Laboratory at the University of Liverpool's Psychology Department. Visuals were presented on a single 17" LCD monitor (Fig. 2). Auditory information was presented via two desktop loudspeakers at around 66 dB(A). Control input was delivered through a commercial Thrustmaster T-Flight Hotas X throttle flight stick.

Subjective and Objective Evaluation

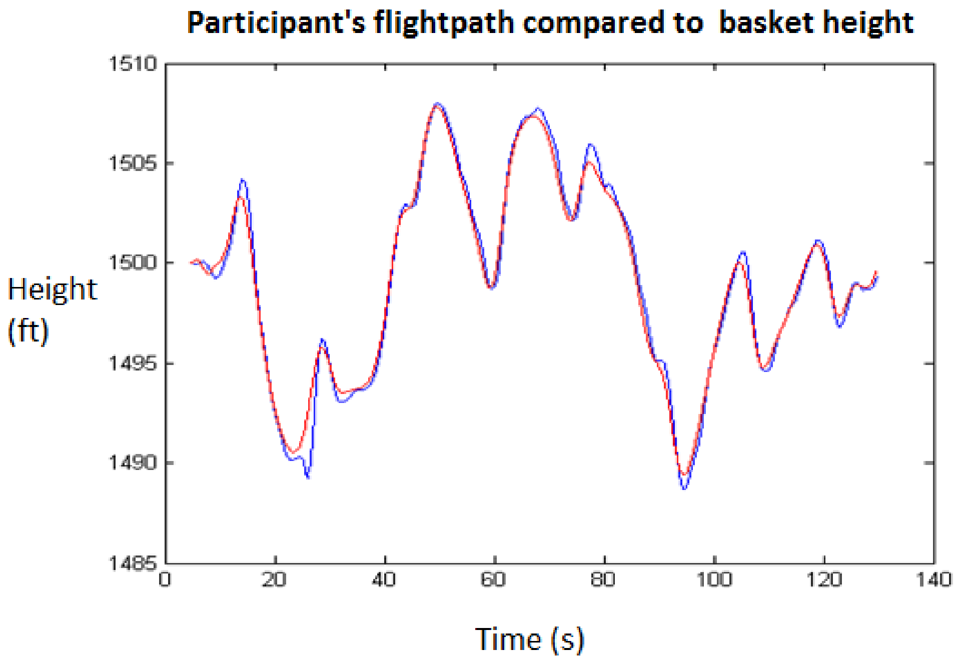

Fig. 3: A participant’s flight path (blue) plotted with the predetermined flight path of the basket (red).

The task was to follow the path of a target or a refuelling basket attached to the tail of a plane. The distance between the crosshair on the visual display and the centre of the target indicated how closely the target was followed. As shown in Figure 3, the vertical trajectory of the target/basket was predetermined by the computer (there was no horizontal movement). Flightpaths for each training and test trial, whilst identical, were designed to be too complex for participants to learn. Data were sampled for each test point at a frame rate of 10ms and a sampling interval of 5 frames. This gave approximately 3000 samples for each 2.5 minute-long test point. Objective performance was quantified as the median absolute difference between the target and the helicopter height, or ‘median error’. Median control input was used as a second objective measure as another insight into participant task behaviour. The flight conditions used for the post-training test were as follows. They were presented in a quasi-random order to each participant (Table 1):

Condition	Acronym	Sound	Difficulty	Motion
1	VEM+	Variable	EASY	ON
2	VEM-	Variable	EASY	OFF
3	VHM+	Variable	HARD	ON
4	VHM-	Variable	HARD	OFF
5	StEM+	Static	EASY	ON
6	StEM-	Static	EASY	OFF
7	StHM+	Static	HARD	ON
8	StHM-	Static	HARD	OFF
9	SubEM+	Substitute	EASY	ON
10	SubEM-	Substitute	EASY	OFF
11	SubHM+	Substitute	HARD	ON
12	SubHM-	Substitute	HARD	OFF

Table 1: Flying conditions used in the post-training task

Subjective ratings were acquired after each trial in the post-training task. Particpants completed a pen-and-paper based 6-item Likert Scale a total of 12 times (the number of trials in the test). This scale was designed to be short enough to complete quickly in between trials, but to contain a few items that would enable assessment of their attitudes to self-performance, simulator handling and simulator fidelity. Responses were given on a scale of 1 to 7, where 1 indicated strong disagreement with the statement and 7 indicated strong agreement. These scale items were always presented in the same order:

Q1) I found the task easy.

Q2) I performed well at this task.

Q3) I felt in control of the helicopter.

Q4) The experience in the simulator seemed real.

Q5) The simulator sounded like a helicopter.

Q6) I feel that training helped me to improve my performance on this task.

Friedman tests were performed on ratings for all scale items. Where Friedman tests were found to be significant ($p \leq .05$), Wilcoxon Signed Ranks tests were used to explore differences between flying conditions. To have a sufficient number of data points to test, questionnaire ratings were pooled across flying conditions (eg. all motion on conditions, all motion off conditions) before paired comparisons.

Results

2.1.1 - Analysis of Objective Performance Measures

Participants completed a tracking task in the HELIFLIGHT simulator, where they were required to control the vertical height of the helicopter in order to refuel a tanker. The tanker moved in a predetermined vertical path which made the task challenging. After two hours' training on a low-fidelity desktop simulator, participants were retested in HELIFLIGHT. The post-training task included additional test trials using a substitute turbine to which participants had not been previously exposed.

A 2×2×2×2 repeated-measures ANOVA was performed on the objective performance measure (median error) across training, auditory cue, difficulty and motion cue. Levels of the factor 'Training' were completion of the task before or after the training session. Levels of the factor 'Auditory cue' included the variable turbine and static turbine. The variable and static turbines were designed to sound like a helicopter, but only the former contained auditory motion information. Easy or hard 'Difficulty' indicated the damping setting on the simulator which eased or impeded control. 'Motion cue' was either present or absence, depending on whether or not the motion platform of the simulator was active.

There was a significant main effect of training, where performance was worse before training compared to after (pre-training mean error= 0.65ft, SE= 0.14; post-training mean error= 0.35ft, SE= 0.05), $F(1,14)= 76.32$, $p< .001$. There was also a significant main effect of motion, where participants performed better with motion (mean error= 0.45ft, SE= 0.09) compared to without motion (mean error= 0.55ft, SE= 0.10), $F(1,14)= 6.77$, $p= .0099$. There was a significant main effect of difficulty, where participants performed better under 'easy' conditions (mean error= 0.44ft, SE= 0.08) compared to 'hard' (mean error= 0.56ft, SE= 0.11), $F(1,14)= 12.01$, $p= .0006$. There was no main effect of auditory cue, and there were no significant interactions.

The ANOVA performed only considered conditions that had been present both before and after training, and therefore excluded conditions using the substitute turbine. Therefore, Bonferroni-corrected t-tests (adjusted alpha= 0.0167) were used to explore differences in post-training performance between the variable, static and substitute turbine conditions. There was no significant difference in performance between variable turbine (mean error= 0.32ft, SD= 0.22) and static turbine conditions (mean error= 0.37ft, SD= 0.29), $t(59)= 2.22$, $p= .031$. There was also no significant difference between the static turbine and substitute turbine conditions (mean error= 0.31ft, SD= 0.17), $t(59)= 2.16$, $p= .035$, or between the variable and substitute turbine conditions, $t(59)= 0.14$, $p= .887$.

A repeated-measures ANOVA was also performed on median control input. There was a significant main effect of training, where control input was greater before training (mean= 0.62in, SE= 0.18) compared to after training (mean= 0.37in, SE= 0.07), $F(1,14)= 35.77$, $p< .001$. There was a significant main effect of motion, where control input was smaller for conditions with vestibular cues (mean= 0.44in, SE= 0.11) compared to without (mean= 0.55in, SE= 0.13), $F(1,14)= 5.86$, $p= .016$. There was also a significant main effect of difficulty, where control input was smaller for 'easy' conditions (mean= 0.45, SE= 0.10) compared to 'hard' (mean= 0.54, SE= 0.13), $F(1,14)= 4.36$, $p= .038$. There was no main effect of auditory cue on control input, and there were no significant interactions.

To explore influences the type of auditory cue may have had on post-training task behaviour, Bonferroni-corrected t-tests (adjusted alpha= 0.0167) were again used to examine post-training median control input under the three auditory conditions. There were no significant differences in control input between the variable turbine (mean = 0.35in, SD= 0.27) and static turbine conditions (mean= 0.38in, SD= 0.29), $t(59)= 1.39$, $p= .169$, the variable turbine and substitute turbine conditions (mean= 0.35in, SD= 0.27), $t(59)= 0.14$, $p= .888$, or the static and substitute turbine conditions, $t(59)= 1.30$, $p= .198$.

2.1.2 – Discussion

Analysis of objective data shows that participants' tracking performance was influenced by training, vestibular cue and task difficulty, but not by auditory cue. Performance significantly improved after training, supporting the idea of positive training transfer from the low-fidelity simulator to the high-fidelity simulator (Tracey & Lathan, 2001; Longridge et al., 2001; Gurusamy, Aggarwal, Palanivelu & Davidson, 2008). In contrast to the findings of Bürki-Cohen and Go (2005) and Bürki-Cohen et al., (2007), participants performed significantly better with simulator motion cues than without. However, it is important to note that improved performance was found in transfer *between* simulators, and it is rarely found that the use of motion platforms in flight simulators transfers to improved performance in *real* aircraft (McCauley, 2006). The physical limitations of current motion platforms means the fidelity of simulated motion cues poorly reflects what pilots would experience in the real world. It should therefore not be expected that improved performance between the low- and high-fidelity simulators automatically indicates training transfer to a real aircraft.

It was expected that the presentation of congruent, temporally- and spatially-aligned motion cues in multiple domains would facilitate task performance (Harrison et al., 2010; Hancock et al., 2013). Participants performed better with vestibular cues compared to without, suggesting that vestibular cues successfully converged with visual tracking information (Chen, DeAngelis & Angelaki, 2011). Although human responses are slower to vestibular-kinaesthetic information alone compared to visual information, small temporal disparities between visual and vestibular motion information do not necessarily prevent cue integration (Barnett-Cowan, Meilinger, Vidal, Teufel, & Bülthoff, 2012; Butler, Campos & Bülthoff, 2015). This may explain why participants were able to use vestibular cues to the advantage of task performance in spite of the in-built time delay in the flight simulator dynamics.

The variable auditory cue also provided motion cueing (through pitch and amplitude changes) consistent with helicopter height, whereas the static turbine contained no motion cues. It would therefore be assumed that, if participants learned its behaviour, the variable turbine cue would promote better tracking performance compared to the non-variable turbine. For instance, the learning effects of exposure to bi-modal cues were demonstrated by Seitz, Kim, van Wassenhove and Shams (2007). Participants were passively exposed to a series of rapidly-presented audio-visual pairs. Tests of the sensory associations made by participants showed that they exhibited more familiarity with audio-visual pairs to which they had previously been exposed in comparison with novel pairs from the same stimulus set.

This finding demonstrated the ability of individuals to form multisensory associations without being aware of the learning process; a concept that was used during training on the low-fidelity simulator, where participants were passively exposed to the variable turbine whilst practicing the tracking task. Although participants were not expressly instructed to attend to the changing pitch and amplitude of the turbine cue, it was thought that with prolonged exposure they would form associations between the visual indication of helicopter height and the behaviour of the auditory motion cue. However, participants did not seem to learn to integrate the changes in pitch and amplitude with visual or vestibular cues enough to influence tracking performance. As the target-tracking task was highly visual, it may be the case that the addition of auditory motion cues conferred little benefit to tracking performance. This finding contrasts with the idea that auditory motion cues can enhance the sensation of self-motion in virtual reality (Väljamäe et al., 2008).

2.2.1 – Subjective Self-performance Measures

Q1) I found the task easy.
Q2) I performed well at this task.
Q3) I felt in control of the helicopter.

Task difficulty was varied during training and testing. Test trials were either 'easy' or 'hard' depending on the damping coefficient used in the simulator model (see Methods). More damping created easy-to-control flight dynamics, whilst less damping made flying more difficult. It was therefore expected that participants would perform better under greater damping conditions. The rating questionnaire used in this study contained items designed to explore participants' ability to self-evaluate their own performance, task difficulty and simulator handling in response to the above difficulty settings. If participants were good at judging these constructs, one would expect to find an inverse correlation between subjective ratings and objective performance error measures. Friedman tests were used to test for significant differences in mean ratings across flying conditions (turbine, motion and difficulty). Where Friedman tests were found to be significant ($p < .05$), Bonferroni-corrected Wilcoxon Signed Ranks tests (adjusted alpha= 0.0167) were used to explore differences within flying conditions.

For Q1, “I found the task easy”, a Friedman test showed no significant differences in ratings across flying conditions, $\chi^2(11) = 16.56$, $p = .121$. For Q2, “I performed well at this task”, there were differences in mean self-performance rating between flying conditions, $\chi^2(11) = 32.31$, $p = .001$, where performance was rated as better in ‘easy’ (mean= 5.63, SD= 1.25) compared to ‘hard’ conditions (mean= 4.91, SD= 1.41), $Z = -3.72$, $p < .001$ (Fig. 4). For Q3, there were no significant differences in perceived control across flying conditions, $\chi^2(11) = 19.02$, $p = .061$.

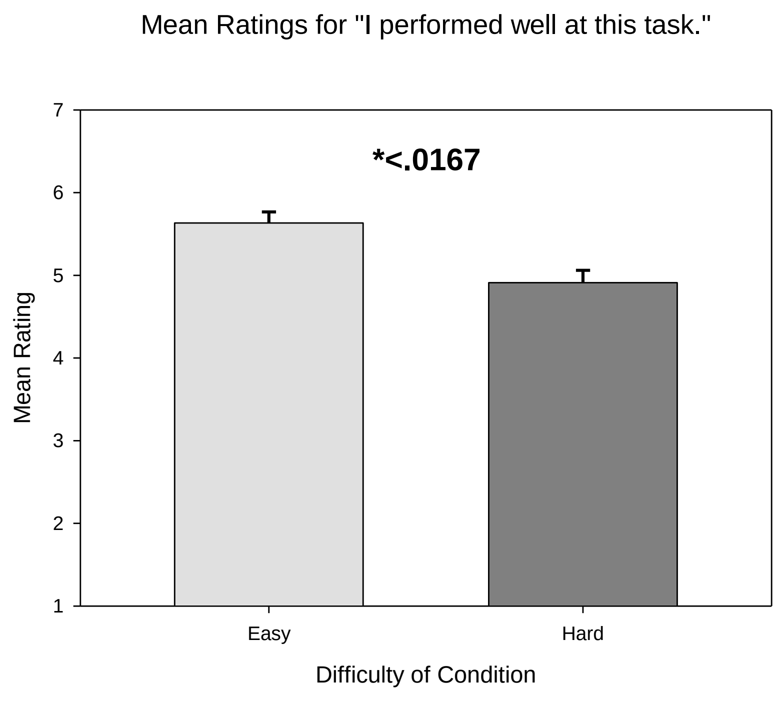

Fig. 4: Although participants did not rate the difficulty of 'easy' and 'hard' conditions any differently, overall they felt that they performed better in easy conditions. Error bars are SEs.

2.2.2 - Discussion

The aim of scale items 1-3 was to explore participants' self-ratings of performance. Analysis of subjective data showed that participants did not rate the difficulty of the tracking task any differently under easy or hard conditions. They also did not report any differences in perceived control across easy or hard conditions. However, objective performance data showed that participants did indeed perform better under easy conditions. When asked to rate their *performance*, participants' responses reflected objective performance data. Variation in ratings for these items suggests that careful wording is required when requesting participants to self-evaluate their performances. Although participants reliably evaluate their tracking performance after task completion, judgement of task difficulty and simulator control could be more consistently assessed by examiners during completion of the task (Watterson et al., 2002; Jeon et al., 2008; Riva et al., 2003).

2.3.1 – Subjective Measures of Fidelity

Q4) The experience in the simulator seemed real.

Q5) The simulator sounded like a helicopter.

In the post-training task, the *physical* and *functional fidelity* of the auditory signal were manipulated to assess any differences in subjective evaluation and objective performance (Ferwerda, 2003). Three types of auditory cue were used: a 'variable' turbine that had both functional fidelity (varied in pitch and amplitude with control input) and physical fidelity (sounded like a real turbine); a 'static' turbine that possessed only physical fidelity; and a 'substitute' saxophone tuning note which possessed functional fidelity but not physical fidelity. In other words, the static turbine sounded like the variable turbine used in training, but did not represent functionally meaningful behaviour. On the other hand, the saxophone note substitute sounded obviously different to the variable turbine, but matched its functional behaviour. Vestibular cues were also manipulated, where they were either present (motion platform switched on) or absent (motion platform off).

Subjective data from two items of the scale were compared with objective performance data to assess participants' ability to evaluate simulator and audio cue fidelity. Since the fidelity of the simulation was modified by systematically manipulating auditory and vestibular cues, it was expected that fidelity ratings would consistently vary with cue type. Q4 referred to the overall fidelity of the helicopter simulation (physical fidelity, visual fidelity and auditory fidelity) whilst Q5 focused only on the auditory fidelity of the simulation. It was expected that exposure to the high-fidelity turbine during training would enable participants to form comparative fidelity judgements of the static and substitute turbines. As for items 1-3, Friedman tests were used to explore significant differences in ratings across flying conditions.

For Q4 - "The experience in the simulator seemed real", a Friedman test found significant differences in ratings across trials, $\chi^2(11) = 44.37$, $p< .001$. Bonferroni-corrected (adjusted alpha= 0.0167) Wilcoxon Signed Ranks tests showed that the simulation was rated as more realistic in conditions with vestibular cues (mean= 5.45, SD= 1.29) compared to without (mean= 4.56, SD= 1.50), $Z= -4.52$, $p< .001$ (Fig. 5). There was no significant difference in ratings between difficulty conditions, $Z= -1.13$, $p= .257$. There was no significant difference in ratings between the variable and static turbine conditions, $Z= -0.86$, $p= .392$. There were also no significant differences in ratings between the variable and substitute turbine conditions, $Z= -1.01$, $p= .313$, or between the static turbine and substitute turbines, $Z= -2.12$, $p= .034$.

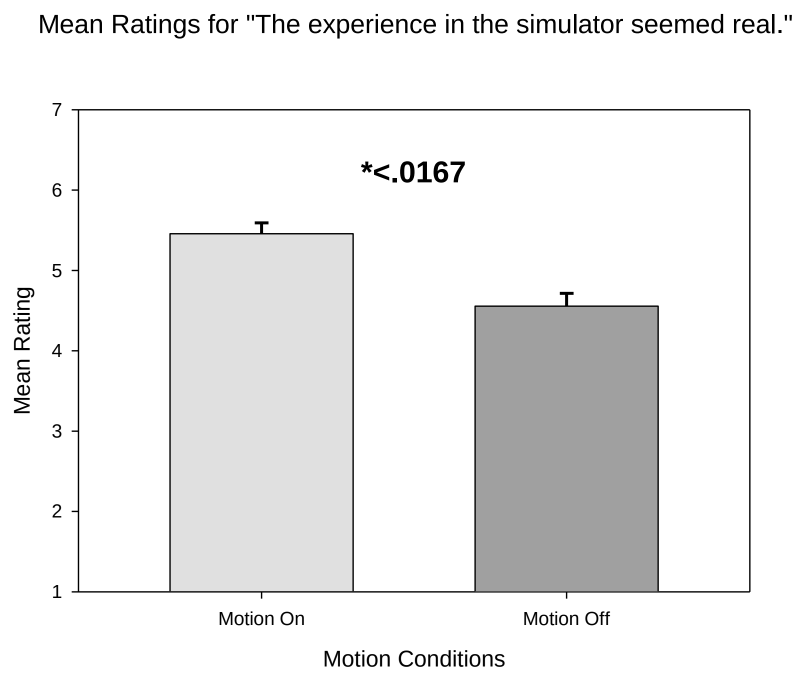

Fig. 5: Participants rated the realism of their experience as being significantly different between motion conditions. Error bars are SEs.

For Q5 - "The simulator sounded like a helicopter", a Friedman test found significant differences across flying conditions, $\chi^2(11)= 20.71$, $p< .036$. Bonferroni-corrected (adjusted alpha= 0.0167) Wilcoxon Signed Ranks tests found significant differences in ratings across turbine condition. Participants rated the variable turbine (mean= 5.15, SD= 1.49) as significantly more realistic than the static turbine (mean= 4.27, SD= 1.80), $Z= -3.75$, $p<.001$. However, there were no significant differences in ratings between the variable and substitute turbines, $Z= -2.16$, $p= .031$, or between the substitute and static turbines, $Z= -0.73$, $p= .463$ (Fig. 6).

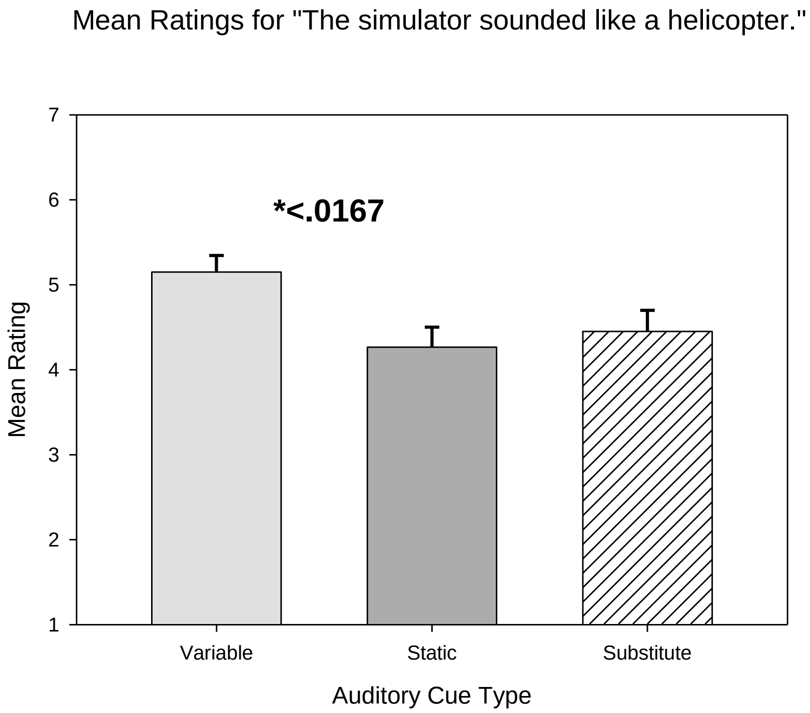

Fig. 6: Mean ratings for turbine fidelity across the three auditory cue conditions. Error bars are SEs.

2.3.2 – Discussion

Scale items exploring subjective fidelity were worded in an attempt to probe attitudes to fidelity without being overly technical and confusing. “The experience in the simulator seemed real” aimed to encompass as many aspects of simulator fidelity as possible, whereas “The simulator sounded like a helicopter” aimed to capture attitudes specifically toward auditory cue fidelity. When rating overall simulator fidelity, participants rated the simulation as being more realistic with the presence of vestibular motion cues. This corresponds with objective performance data, where tracking performance improved with vestibular cueing. Although no motion cues were used during training on the basic simulator, participants were still able to perceive these cues during the post-training tasks, and judged the simulation as being more realistic because of them. The integration of vestibular cue consistent with visual representation of the moving helicopter (through the rising and falling of the tanker) may have enhanced perception of self-motion to benefit both tracking performance and simulator fidelity (Meyer et al., 2012; Berger et al., 2010).

Three different auditory cues were presented throughout the post-training tasks: a rotor-turbine sound combination where rotor amplitude co-varied with control input (variable turbine); the same sound combination but without any variance (static turbine); and a saxophone tuning note whose amplitude co-varied in the same fashion as that of the variable turbine (substitute turbine). The variable turbine was designed to sound like a real helicopter turbine and changed in pitch and amplitude in concordance with helicopter height. The static turbine also sounded like a helicopter turbine, but did not vary with helicopter height. The substitute turbine's saxophone note sounded unlike a helicopter turbine, but its pitch and amplitude, like that of the variable turbine, adjusted with helicopter height. It was expected that trained participants would show significant variation in attitudes to these turbines. These expectations were partly reflected by subjective rating data, where participants rated the variable turbine as more realistic compared to the static turbine. However, not only does this contradict objective tracking data (which showed no significant difference in performance across the turbine conditions), participants did not report the substitute turbine as sounding any less realistic than the others.

If participants considered the *functional fidelity* of the turbine (its auditory motion properties) to be important, one would also expect to find a significant difference in ratings between the static and substitute turbines, but not between the variable and substitute turbines. This is because the functional aspects of the variable and substitute turbines (variations in pitch and amplitude according to helicopter height) were identical. If they considered *physical fidelity* (what it sounded like) to be more important, then both the variable and static turbines would have been rated as more realistic than the substitute turbine, because the variable and static turbines resembled a helicopter turbine whilst the substitute turbine was a saxophone tuning note (Ferwerda, 2003). Given that participants did not clearly indicate which aspects of auditory cue fidelity they could perceive, it appears the inclusion of auditory motion cues made little contribution to perceived fidelity of the simulation and contradicts those who suggest they enhance the realism of a simulation (Väljamäe et al., 2008; Väljamäe, 2009).

2.4.1 - Subjective Measures of Training

Q6) I feel that training helped me to improve my performance on this task.

The training component of this study involved all 15 participants undertaking 2 hours of training on a low-fidelity, desktop-based simulator in a laboratory. The low-fidelity simulator lacked vestibular cues but used identical flight dynamics to those in the HELIFLIGHT simulator. For each participant, the training session was split into 15-minute blocks where only 'realistic', variable turbine was presented. Task difficulty alternated between easy and hard between trials (4 easy trials, 4 hard trials).

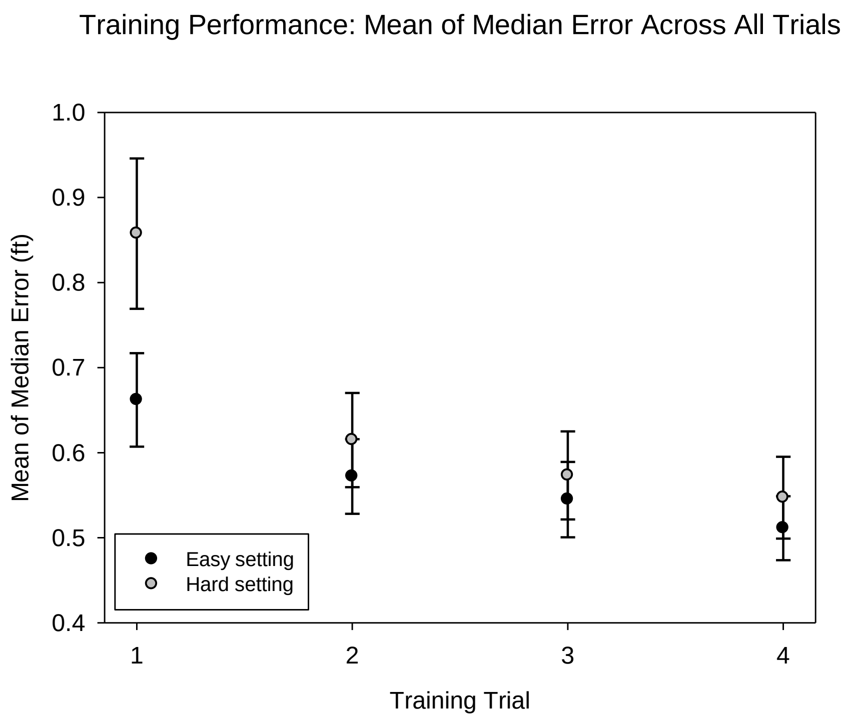

Fig. 7: Training performance of all 15 participants. Error bars are SEs.

Participants' task performance improved over training (Fig. 7). T-tests showed that participants performed significantly better in the fourth hard training trial (mean error= 0.54ft, SD= 0.18) compared to the first (mean error= 0.85ft, SD= 0.33), $t(14)= 5.52$, $p<.001$, and participants also performed significantly better in the fourth easy trial (mean error= 0.51ft, SD= 0.14) compared to the first (mean error= 0.66ft, SD= 0.21), $t(14)= 3.04$, $p= .0087$. However, absolute performance improvement over easy training trials was not a significant predictor of training ratings in equivalent post-training trials, $R^2= .0605$, $p= .377$. Similarly, absolute performance improvement over hard training trials was not a significant predictor of training ratings in equivalent post-training trials, $R^2= .001$, $p= .907$. For the post-training task, a Friedman test showed that there were no significant differences in training ratings across flying condition, $\chi^2(11) = 9.84$, $p= .545$.

2.4.2 - Discussion

Participants were asked to judge the usefulness of training to explore if training with the low-fidelity simulator would transfer to the high-fidelity simulator, regardless of whether post-training conditions matched training conditions. If participants were good at judging training usefulness, one would expect to see higher subjective ratings only for conditions that both included the variable turbine and excluded vestibular cues (which were equivalent to the training conditions). Objective training data showed rapid improvement in task performance. Despite being exposed to the variable turbine throughout training, there was no evidence to suggest that they thought this helped to improve performance. Participants also did not think training was any more useful in 'easy' compared to 'hard' conditions, but consistently performed better in easy conditions post-training. Additionally, ratings of training were not affected by the presence or absence of motion cues, even though tracking performance was significantly better with vestibular cues and participants were trained without a motion platform.

As participants were not receiving feedback on their objective performance, each individual likely judged training using their own personal criteria and expectations after the task had been completed (Riva et al., 2003). Therefore, asking participants to rate training usefulness in relation to the post-training task did not offer consistent evaluation of training. Like with task performance ratings, it may be that objective measures provide a more consistent evaluation of training than subjective ratings (Schricker et al., 2001). For these reasons, it is perhaps expected that perceived usefulness of training would not concur with objective measures of performance.

Summary

Experiment One showed that the presence of congruent vestibular cues afforded improved task performance in a vertical tracking task, but participants were unable to use auditory motion cues to improve their task performance. Their subjective ratings of auditory fidelity also did not reflect the importance of either physical fidelity or functional fidelity in the simulated environment. Improvements in participants' task performance transferred from the basic simulation to high-fidelity simulation, but participants were poor at evaluating the usefulness of training, where ratings were inconsistent with performance across flying conditions.

Publication in PLoS ONE

The work in Experiment One was published in the open-access journal PLoS ONE, and was included in Experiments 2 and 3 of the paper (Meyer et al., 2011). The article explored fidelity measures in the virtual reality tracking task, and described how the contribution of multisensory cues to task performance was aided by training on the low-fidelity simulator. Ten of the best performers during training were reported in the publication, whereas the complete participant pool (N= 15) has been evaluated in this thesis. Additionally, this thesis explored participants' subjective attitudes to fidelity and task performance, which were not included in the published paper.

Chapter 3: Experiment Two – Hover Task

Experiment Two further explored the use of multisensory motion cues in simulated flying tasks. As in Experiment One, auditory and vestibular motion cues were included in the simulated environment. In light of the results of Experiment One, auditory motion cues were this time presented in the form of HRTFs with head tracking, in order to afford each individual participant more ability to localise sounds. As well as exploring the concepts of fidelity and task performance, Experiment Two also considered participants' subjective evaluation of presence and how it related to their task performance.

Method

Participants

Eleven participants obtained via opportunity sampling took part in Experiment Two (range 22-27 years, mean= 24.2, nine males). All participants reported normal or corrected-to-normal vision and normal hearing. Some participants had computer game experience, some had prior experience of the specific simulators used during this study (but not of the activity performed within this experiment), and one was a fully-trained helicopter pilot.

Apparatus and Materials

As in Experiment One, two simulators were used: a low-fidelity simulator was used in training with HELIFLIGHT being used in testing. Again, the dynamic behaviour of both simulators was identical. The low-fidelity simulator was based in the Visual Perception Laboratory at the University of Liverpool's Psychology Department. Visuals were projected on to a 1.5m by 1.6m projection screen on the wall. Auditory cue was presented via Sennheiser HD 435 Vegas headphones, at around 87.5 dB(A). Flight rudder pedals for gaming were used to emulate the foot-and-leg motion required to control the HELIFLIGHT simulator. The visuals of the low-fidelity simulator were identical to those used in Experiment One, but this time participants were required to follow a target moving from side-to-side along a predetermined horizontal path. This horizontal movement was intended to simulate the disturbance of the helicopter along the yaw axis from 0 degrees.

As in Experiment One, the HELIFLIGHT simulator was used during the pre- and post-training tests. The flight dynamics model (see *Appendix*) was restricted to movements along the yaw axis. Control input was delivered via rudder pedals in the cockpit. The input gain was constant in all experiments (N0tr = 4.8). In order to simulate wind disturbance along the yaw axis, a turbulence scale factor of 1.5 was used.

Visual information consisted of white radial lines extending to the horizon on a green 'grass' background with a radial line every 15 degrees. Each of these lines was visually identical to prevent the task from being too easy. At the start of each test point, the helicopter was level at 0 degrees along the yaw axis (Fig. 8).

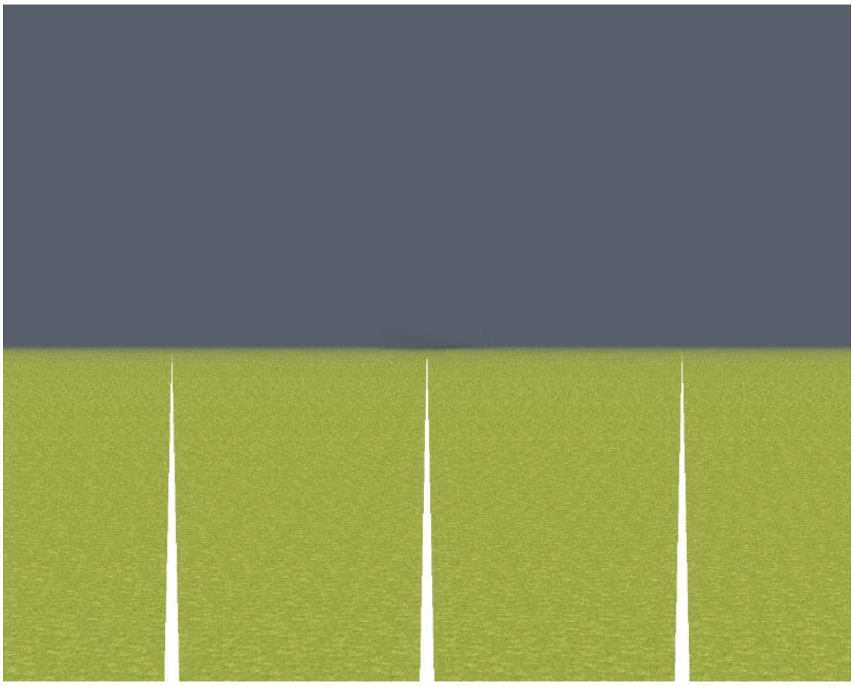

Fig. 8: Visual display for the hover task.

A pulsed white noise whose position varied in horizontal position and amplitude was used as an auditory motion cue. This was delivered via a headset in the simulator capsule at around 87.5 dB(A) while the pilots wore sound attenuating headphones (Flightcom 4DLX (attenuation – 24dB)). The cue was a continuous loop generated by the Tucker-Davis RM1 real time processor. Cue amplitude and horizontal position were modulated in direct proportion to the control input. The overall signal level varied by 3dB (86 – 89 dB(A)). In this experiment, vestibular motion cues were restricted to rotational movements along the yaw axis.

Objective and Subjective Evaluation

The objective of both the training and post-training tasks was to maintain a heading of zero degrees in opposition to simulated wind disturbance along the yaw axis. Participants were required to make use of any available cues (visual, motion or auditory) to determine their position relative to the target heading. As well as a control input-to-response latency of 75ms, each radial line on the display was visually identical, ensuring that the task was complex. The disturbance of the helicopter along the yaw axis was predetermined by the computer.

The simulation of binaural auditory cues using head-related transfer functions (HRTFs) commonly results in difficulties with sound localisation when cues are presented in front of or behind a participant (Cho et al., 2006). This is because the artificial auditory cues usually do not correspond with an individual's own transfer functions. Delivering HRTFs in in correspondence with a head tracking device optimises directional reproduction (Wightman & Kistler, 1999; Seeber & Fastl, 2003). Accurate detection of one's position along the yaw axis should be further enhanced with congruent vestibular cueing. However, a person's sensitivity to movement along the yaw axis is influenced by the frequency of the turn; the thresholds for high-frequency turns (eg. 5Hz) are lower than those for low-frequency turns (eg. 0.05Hz) (Grabherr, Nicoucar, Mast & Merfeld, 2008). Additionally, the aforementioned limitations of current motion platforms mean that the sensation of acceleration cannot be sustained indefinitely before 'washout' occurs.

As in Experiment 1, both task performance and task behaviour measures were taken. The proportion of data points spent within 5 degrees of the central heading was taken as a performance measure, while closed-loop median control input was used as a measure of participants' behavioural response throughout the task. As in Experiment One, data were sampled for each test point at a frame rate of 10ms and a sampling interval of 5 frames, giving around 3000 samples for each 2.5 minute-long test point. The design used for the pre- and post-training tests is as follows. The conditions were presented in a quasi-random order to each participant (Table 2):

Condition	Acronym	Sound	Difficulty	Motion
1	M+V+A+	Variable	EASY	ON
2	M+V+A-	Variable	EASY	OFF
3	M+V-A+	Variable	HARD	ON
4	M+V-A-	Variable	HARD	OFF
5	M-V+A+	Static	EASY	ON
6	M-V+A-	Static	EASY	OFF
7	M-V-A+	Static	HARD	ON
8	M-V-A-	Static	HARD	OFF

Table 2: Flying conditions used in the pre- and post-training tasks of Experiment 2.

As in Experiment 1, subjective ratings were acquired after each trial in the post-training task. Attitudes on presence, functional fidelity and performance were collected. Participants completed a pen-and-paper based 15-item Likert Scale a total of 8 times (the number of trials in the test). Scale items regarding fidelity and performance were adapted from Experiment 1. Items concerning presence were taken from a limited selection from Witmer and Singer's (1998) Presence Scale. These items were chosen because Witmer and Singer's scale has been shown to be robust and reliable measure of presence, and also because they were closely related to the task used in this experiment. The entire scale was not used due to the time constraints of testing. Responses were given on a scale of 1 to 7, where 1 indicated strong negative attitude (eg. “Not at all”) with the statement and 7 indicated strong positive attitude (eg. “Very much”). The items were presented in the same order each time:

Q1) How completely were *all* of your senses engaged?

Q2) How much did the auditory aspects of the environment involve you?

Q3) How aware were you of events occurring in the real world around you?

Q4) How well could you localize sounds?

Q5) How involved were you in the virtual environment experience?

Q6) Did the simulator sound like a helicopter?

Q7) Did your experience seem real?

Q8) How visually realistic was the simulator?

Q9) How realistically did the simulator move?

Q10) How realistic was the environment in which you were flying?

Q11) Did you find training useful for this task?

Q12) Did you perform well in this task?

Q13) Could you easily control the helicopter's movement?

Q14) Did you find the task easy?

Q15) How well could you predict the helicopter's movements?

Friedman tests were performed on ratings for all scale items. Where Friedman tests were found to be significant ($p \leq .05$), Wilcoxon Signed Ranks tests were used to explore differences between groups. As in Experiment 1, item ratings were pooled across flying conditions before t-test comparisons.

Results

3.1.1 – Analysis of Objective Performance Measures

Participants completed a hover task in the HELIFLIGHT simulator, in which they were required to correct their position along the yaw axis towards a central heading whilst their path was 'disturbed' by simulated turbulence (also along the yaw axis). After undergoing an hour's training on the low-fidelity desktop simulator, they were retested in HELIFLIGHT.

To compare performance before and after training, a 2×2×2×2 repeated-measures ANOVA was performed on the proportion of data points spent within 5˚ from the central heading, across training (before or after training), visual cue (present or absent), vestibular cue (present or absent) and auditory cue (variable or static). There was a significant main effect of training, where participants spent a greater proportion of time within 5˚ from the central heading after training (mean= 45.3%, SE= 0.018) compared to before training (mean= 40.3%, SD= 0.029), $F(1,10)= 10.29$, $p= .0094$. There was also a main effect of visual cue, where participants spent a significantly greater proportion of time within 5˚ of the central heading when visuals were on (mean= 77.5%, SD= 0.045) compared to off (mean= 8.1%, SD= 0.003), $F(1,10)= 245.63$, $p<.001$. There were no significant main effects of vestibular cue or auditory cue.

There was a significant interaction between training and visual cue, $F(1,10)= 30.86$, $p<.001$. Before training, performance was better in visuals-on conditions (mean= 72.7%, SE= 0.057) compared to visuals-off (mean= 7.9%, SD= 0.003). After training, performance was also better in visuals-on conditions (mean= 82.3%, SE= 0.036) compared to visuals-off (mean= 8.2%, SE= 0.004).

An ANOVA was also performed on median control input across the same factors to examine participants' task behaviour. There was a main effect of auditory cue, where control input was greater in variable audio conditions (mean= 19.83in, SE= 0.86) compared to static audio conditions (mean= 14.94in, SE= 2.00), $F(1,10)= 5.61$, $p= .039$. There was also a significant main effect of visual cue, where control input was smaller for conditions with visual cues (mean= 13.73in, SE= 0.94) compared to without (mean= 21.04in, SE= 2.00), $F(1,10)= 11.87$, $p= .006$. There were no significant main effects of vestibular cue or training.

There was a significant interaction between vestibular and auditory cues, $F(1,10)= 5.20$, $p= .024$. In motion conditions, control input was greater when the variable auditory cue was used (mean= 18.53in, SE= 1.57) compared to when the static auditory cue was used (mean= 17.04in, SE= 3.34). Likewise, in non-motion conditions control input was greater when variable audio was used (mean= 21.12, SE= 1.16) compared to static audio (mean= 12.85, SE= 1.79).

There was also a significant interaction between visual and auditory cues, $F(1,10)= 10.99$, $p= .001$. In visuals-on conditions, control input was smaller when the variable auditory cue was used (mean= 13.71in, SE= 0.93) compared to when the static cue was used (mean= 13.75in, SE= 0.99). During visuals-off conditions, however, control input was greater when the variable auditory cue was used (mean= 25.94in, SE= 1.65) compared to the static (mean= 16.14in, SE= 3.57).

3.1.2 – Discussion

Participants spent significantly more time within 5˚ of the heading after training compared to before, supporting the notion of positive training transfer between the low-fidelity simulator and HELIFLIGHT (Tracey & Lathan, 2001; Longridge et al., 2001; Gurusamy et al., 2008). Given the difficulty of the task, a main effect of visual cue for both task performance and control input behaviour was expected. As in Experiment 1, there was no main effect of auditory cue on this performance measure. However, analysis of control input data shows that auditory cue had significant effect on participant *behaviour* if not their performance, where more frequent variation in control input was seen in conditions with the auditory motion cue than the static cue. It was noted that even after training, participants tended to spin around in circles at high velocity if they lost control of the simulator. This would have made it difficult to accurately exploit the auditory motion cue to correct position back to the heading. These findings suggest that participants were able to perceive the directional motion cues embedded in the auditory signal, but were unable to use them to improve performance (Väljamäe et al., 2008; Seeber &

Fastl, 2003). Although there was no main effect of vestibular cue, analysis showed an interaction between vestibular and auditory cues. This suggests that motion cues presented in multiple modalities are perceptible enough to influence task behaviour, and supports the merits of motion platforms in training simulations.

3.2.1 – Subjective Measures of Presence

After gaining familiarity with the task and simulator during the pre-training and training sessions, participants' subjective attitudes to experimental cues were evaluated using a 15-item scale. This scale was completed by each participant after every post-training trial to record subjective measures of presence, fidelity and performance.

Q1) How completely were *all* of your senses engaged?

Q2) How much did the auditory aspects of the environment involve you?

Q3) How aware were you of events occurring in the real world around you?

Q4) How well could you localize sounds?

Q5) How involved were you in the virtual environment experience?

A selection of items from Witmer and Singer's (1998) Presence Scale was used to explore participants' feelings of presence, or “being there” (Pausch et al., 1997). The chosen items were included as they were highly related to this experiment, but were limited in number due to the time constraints of the study. For each question, Friedman tests were conducted across flying condition to reveal differences in ratings. Where Friedman tests were found to be significant ($p \leq .05$), Wilcoxon Signed Ranks tests were used to further explore these differences.

Bonferroni-corrected (adjusted alpha= 0.0167) Wilcoxon Signed Ranks tests showed that participants rated their *senses as being engaged* more when visual cues were present (mean= 5.11, SD= 1.37) as opposed to absent (mean= 3.91, SD= 1.06), Z= -2.61, p= .009, and when vestibular cues were present (mean= 4.82, SD= 1.40) as opposed to absent (mean= 4.20, SD= 1.73), Z= -3.29, p= .001. There were no significant differences between auditory cue conditions (Fig. 9). In spite of this, participants felt that the *auditory aspects of the environment involved them* more in auditory motion cue conditions (mean= 4.89, SD= 1.78) than static auditory cue conditions (mean= 3.57, SD= 1.86), Z= -3.18, p= .001 (Fig. 10). Participants rated themselves as having less *awareness of the real world around them* when visual cues were present (mean= 5.13, SD= 2.09) as opposed to absent (mean= 5.84, SD= 1.58), Z= -3.02, p= .003; Bonferroni-adjusted alpha= 0.0167 (Fig. 11). Ratings showed that they felt more able to *localise sounds* when there was an auditory motion cue (mean= 4.34, SD= 1.89) compared to static auditory cue (mean= 3.11, SD= 1.65), Z= -2.87, p= .004 (Fig. 12). All three experimental cues (Bonferroni-adjusted alpha = 0.0167) significantly affected ratings of *feeling involved in the virtual environment experience*, where participants felt more involved with visual cues (mean= 4.95, SD= 1.26) versus without (mean= 4.00, SD= 1.61), Z= -3.23, p= .001, with vestibular cues (mean= 4.75, SD= 1.28) versus without (mean= 4.20, SD= 1.69), Z= -2.74, p= .006, and with auditory motion cues (mean= 4.84, SD= 1.27) versus static auditory cues (mean= 4.11, SD= 1.66), Z= -3.03, p= .002 (Fig. 13).

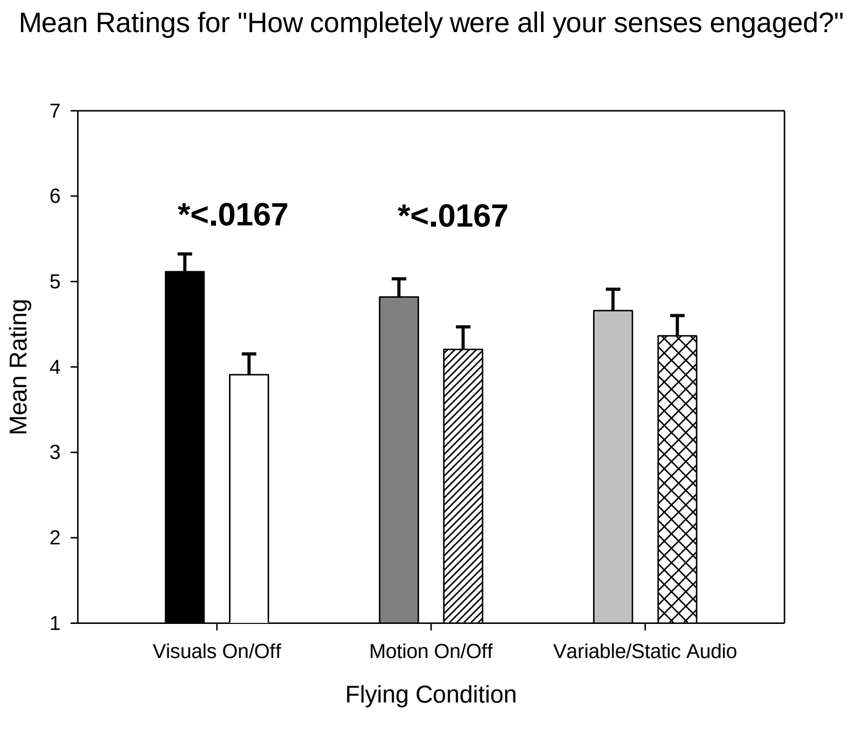

Fig. 9: Mean ratings for sensory engagement. There were significant differences between visual conditions and motion conditions, but not between the auditory conditions. Error bars are SEs.

Mean Ratings for "How much did the auditory aspects of the environment involve you?"

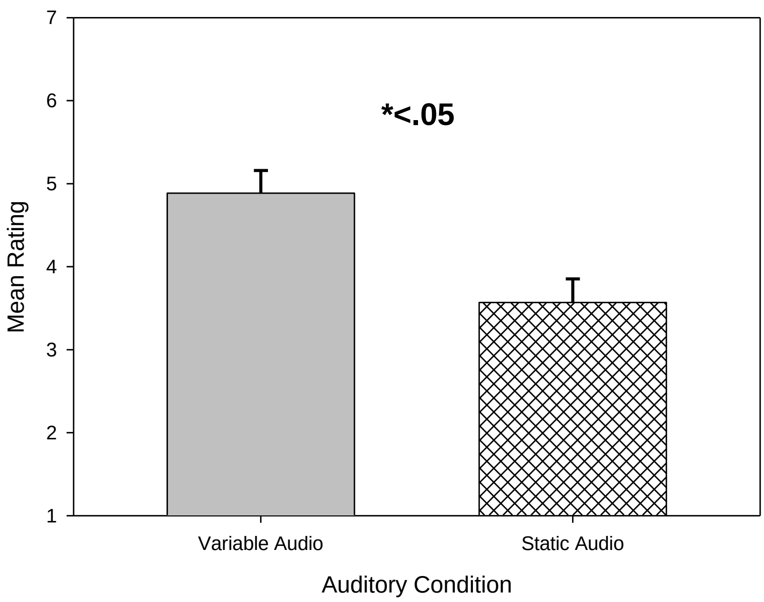

Fig. 10: Mean ratings of engagement for auditory aspects of the environment. Participants found the variable auditory cue to be more engaging compared to static audio. Error bars are SEs.

Mean Ratings for "How aware were you of events occurring in the real world around you?"

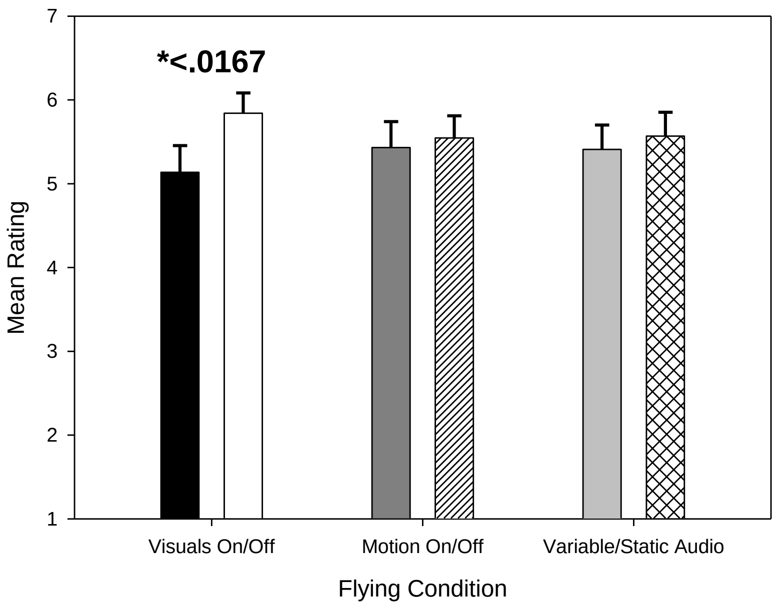

Fig. 11: Mean ratings of awareness of the real world. There was a significant difference between visual conditions. Error bars are SEs.

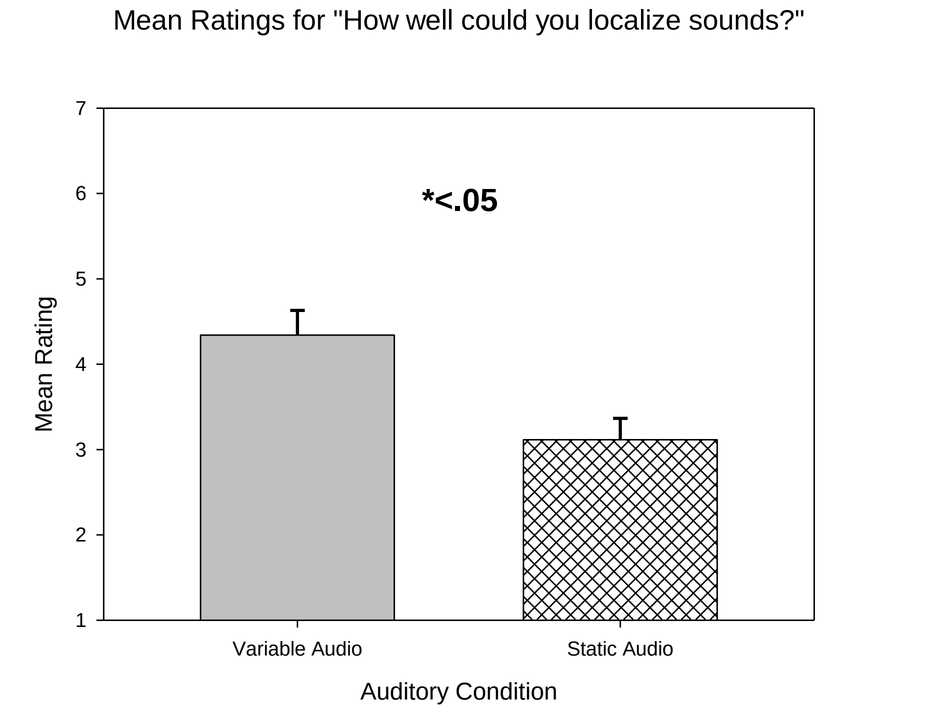

Fig. 12: Participants felt they could localise sounds more in conditions using the variable auditory cue compared to static. Error bars are SEs.

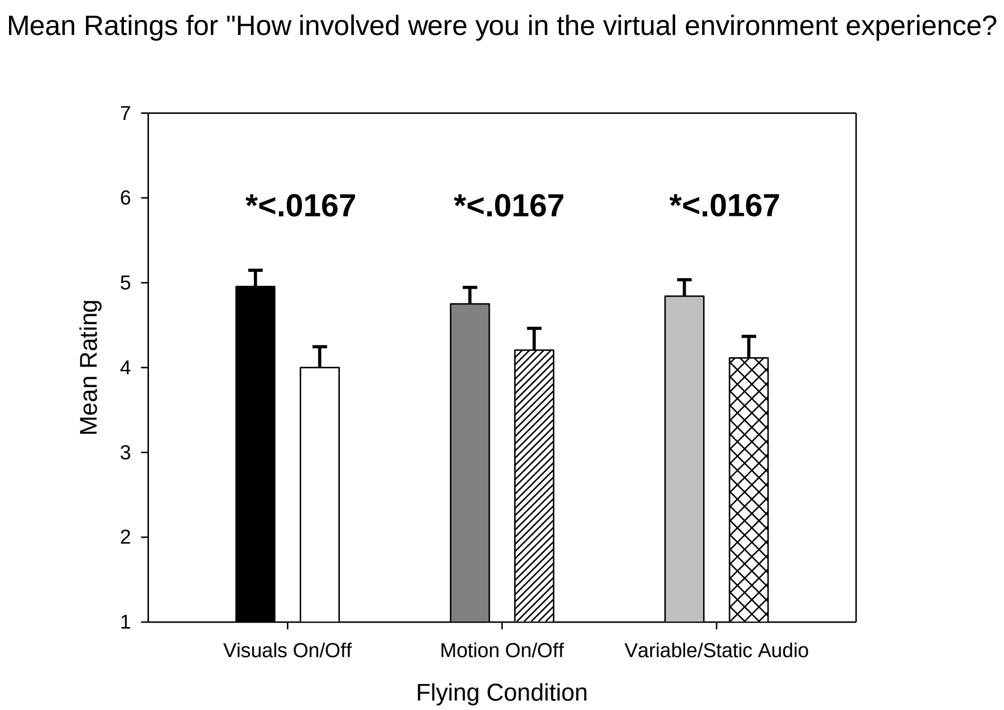

Fig. 13: Ratings for involvement in the VE experience were significantly higher with visual, motion and variable audio cues compared to without. Error bars are SEs.

3.2.2 – Discussion

Items from Witmer and Singer's (1998) Presence scale were used to explore the effect of experimental cues on ratings of presence. Visual and auditory cues were a strong factor in the rating of presence, which is reflected in objective control input behavioural measures. The ratings of presence increased with the number of cues present, supporting the idea that multisensory cues enhance the sensation of presence (Gallace et al., 2011). Given that visual cues were highly important to the task, it was expected that their presence or absence would significantly influence both task performance and presence ratings. Even when visual cues were present, however, they were not realistic enough to mimic the real world. All the same, participants gave higher ratings of presence for conditions with visual cues than without, which suggests that even basic visual representations are sufficient to improve a sensation of presence (Dinh et al., 1999). If presence depended on the use of highly-realistic computer graphics, one would not expect a significant difference in ratings between visual conditions. It may be that participants' responses were made within the context of the experiment rather than comparing their expectations of the real world to the virtual one (Usoh, Catena, Arman & Slater, 2000).

The findings also agree with previous studies that found spatialised auditory cues significantly contribute to the sensation of presence (Hendrix & Barfield, 1996), especially when the auditory cue contained motion information that was relevant to the task (Bormann, 2008). Participants' sensation of presence seemed to be enhanced by the use of binaural HRTFs, but only when the auditory cue acted as a moving sound source (Nykänen, Zedigh & Mohlin, 2013). This corresponds with control input data, which showed that participants' task behaviour significantly differed between variable and static auditory cue conditions. Participants also felt more involved in the virtual environment experience in conditions with vestibular cues, which did not have a significant effect on task performance. This contrasts with the findings in Experiment One, where vestibular cues significantly improved task *performance*, and suggests that while vestibular motion was important to promote the sensation of presence, the low fidelity cues provided by the motion platform did not meaningfully contribute to task performance. On the other hand, an interaction between vestibular and auditory cues for control input measures suggests that vestibular cues influence task *behaviour* and further supports the idea that multisensory sensory cues enhance the sensation of presence.

3.3.1 – Subjective Measures of Fidelity

Items for subjective fidelity measures were adapted from the scale used in Experiment One, reworded to fit the style of items taken from Witmer and Singer's (1998) Presence scale.

Q6) Did the simulator sound like a helicopter?
Q7) Did your experience seem real?
Q8) How visually realistic was the simulator?
Q9) How realistically did the simulator move?
Q10) How realistic was the environment in which you were flying?

Participants significantly rated the simulator as *sounding like a helicopter* more in auditory motion cue conditions (mean= 4.25, SD= 1.60) compared to static auditory cue conditions (mean= 3.68, SD= 1.70), *Z*= -2.86, *p*= .004 (Fig. 14). Only visual cues affected whether participants felt *their experience seemed real,* where they felt it was more real with visual cues (mean= 3.70, SD= 1.27) than without (mean= 2.78, SD= 1.68), *Z*= -2.94, *p*= .003; Bonferroni adjusted alpha= 0.0167 (Fig. 15). They similarly rated the *visual realism of the simulator* as greater with visual cues (mean= 3.88, SD= 1.60) compared to without (mean= 3.27, SD= 2.02), *Z*= -2.36, *p*= .019 (Fig. 16). They also felt that *the simulator moved more realistically* with vestibular cues (mean= 4.25, SD= 1.77) than without (mean= 2.48, SD= 1.91), *Z*= -4.18, *p*< .001 (Fig. 17). Finally, participants felt that the *environment in which they were flying* was more realistic with visual cues (mean= 3.34, SD= 1.83) than without (mean= 2.57, SD= 2.18), *Z*= -2.91, *p*= .004; Bonferroni-adjusted alpha= 0.0167 (Fig. 18).

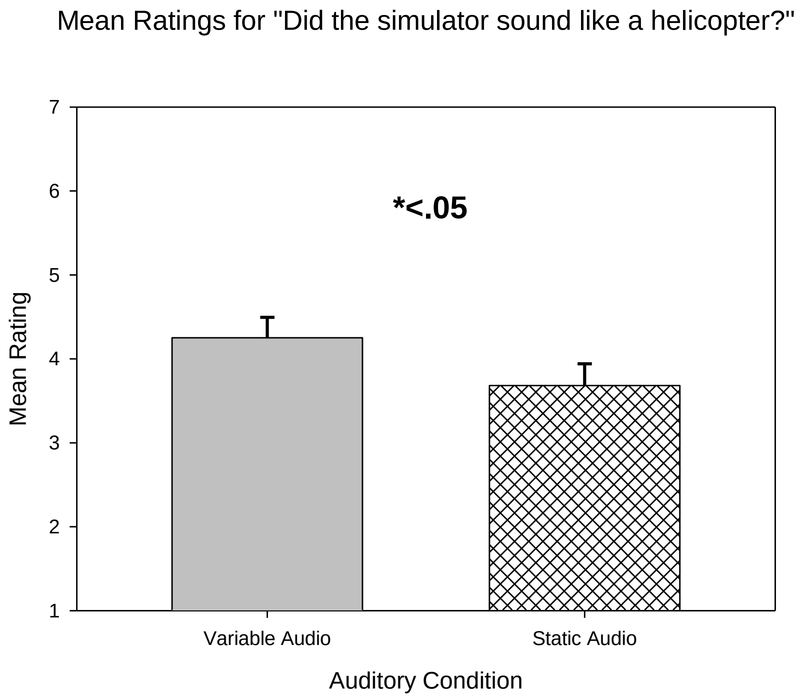

Fig. 14: Ratings for simulator sound fidelity. Although the white pulsed noise used did not resemble a helicopter turbine, participants thought the variable cue sounded more like a helicopter than the static cue. Error bars are SEs.

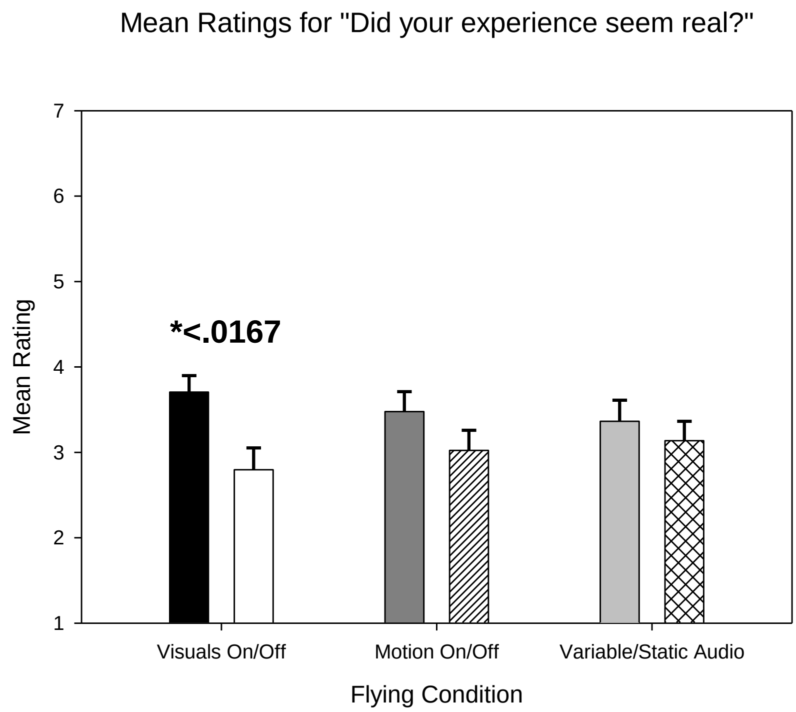

Fig. 15: Ratings for reality of experience. Mean ratings were generally neutral, but there was a significant difference between visual conditions. Error bars are SEs.

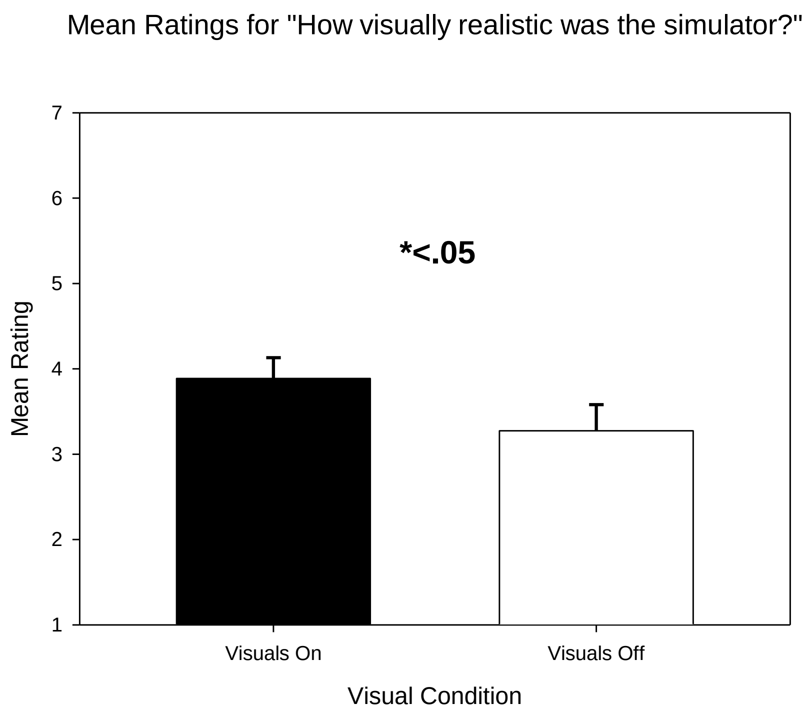

Fig. 16: Ratings for the visual realism of the simulation. Error bars are SEs.

Fig. 17: Ratings for the realism of simulator movement. Error bars are SEs.

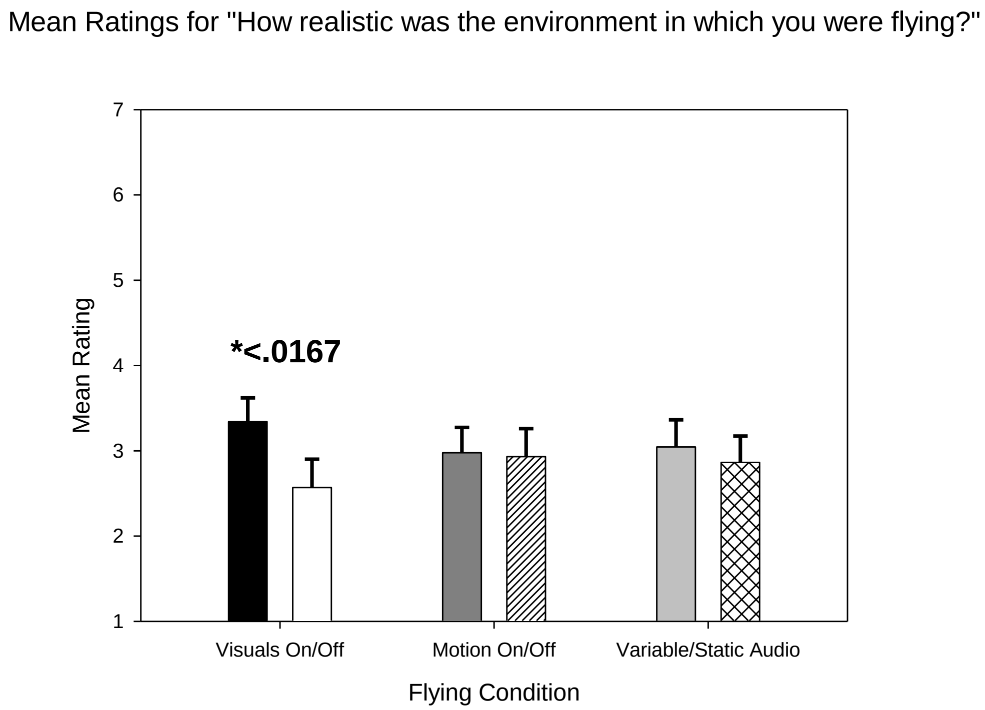

Fig. 18: Ratings for the realism of the environment in the simulator. There was a significant difference in ratings between visual conditions. Error bars are SEs.

3.3.2 – Discussion

For this experiment, a binaural auditory cue (a pulsed white noise) was generated that, through use of a head-tracker, moved consistently relative to the participant and their control input. The cue was not, however, intended to resemble any realistic sound source in a helicopter. Still, when rating the auditory fidelity of the simulator, participants rated the auditory motion cue as being more realistic than the static cue. This could mean that through exposure to the cue in the training sessions, participants expected the sound to move consistently with the environment. Since the *physical fidelity* (what it sounded like) of the cue did not differ throughout the experiment, participants appear to have judged the cue on its *functional* fidelity (how the sound behaved) (Ferwerda, 2003). Although they were unable to exploit the auditory cue to the extent of improving task performance, participants still perceived its motion information enough to modulate their control input. This provides some support to the idea that congruent auditory motion cueing enhances both the sensation of presence and fidelity in a simulation (Hendrix & Barfield, 1996; Nykänen et al., 2013; Väljamäe et al., 2008).

It has been suggested that simulator motion platforms are limited in their capacity to provide congruent acceleration cues (Bürki-Cohen et al., 2007). It could be argued that the auditory and vestibular cues used in this experiment are low in *functional fidelity* as they did not contribute to improvement in task performance after training (Perfect et al., 2014). Nevertheless, participants judged the movement of the simulator as being more realistic when vestibular cues were present. Here, there is a discrepancy between what participants subjectively felt was realistic about the simulated cues, and how the cues affected their task performance. In contrast to arguments made by Schricker et al., (2001), subjective ratings motion fidelity could be a more sensitive measure compared to an objective referent like task performance. On the other hand, this depends on one's definition of fidelity; if the fidelity of a cue was defined as the extent to which it improves task performance, then the vestibular cues used in this experiment would be considered low-fidelity. If the fidelity of a cue was defined by participants' subjective interpretation of what is realistic, then the vestibular cues could be considered high-fidelity.

When rating the realism of their experience, visual realism of the simulator and environmental realism, participants were more affected by the presence or absence of visual cues than auditory or vestibular cues. Compared to the low-fidelity simulator, computer graphics in HELIFLIGHT were more sophisticated but could not be described as photorealistic. Task performance showed training transfer between the low-fidelity simulator and HELIFLIGHT, suggesting that visual cues provided in both simulations were of high enough *functional fidelity* to allow transfer of training to occur (Perfect et al., 2014). Although it is expected that congruent visual cueing would be rated positively, it is surprising that there were no significant effects of auditory motion or vestibular cues on ratings. As in Experiment One, the nature of the task was difficult; even more so without visual cues. It may be that participants prioritised visual cues as being the most important when evaluating the reality of experience.

3.4.1 – Subjective Measures of Performance and Training

As in Experiment One, participants underwent training on a low-fidelity simulator. The objective performance measure used was also mean of median error. As seen in Figure 19, performance rapidly improved after the first trial, with significant improvement in performance between the first (mean= 4.09ft, SD= 4.46) and fourth trials (mean= 0.77ft, SD= 0.43), *t*(10)= 2.43, *p*= .036.

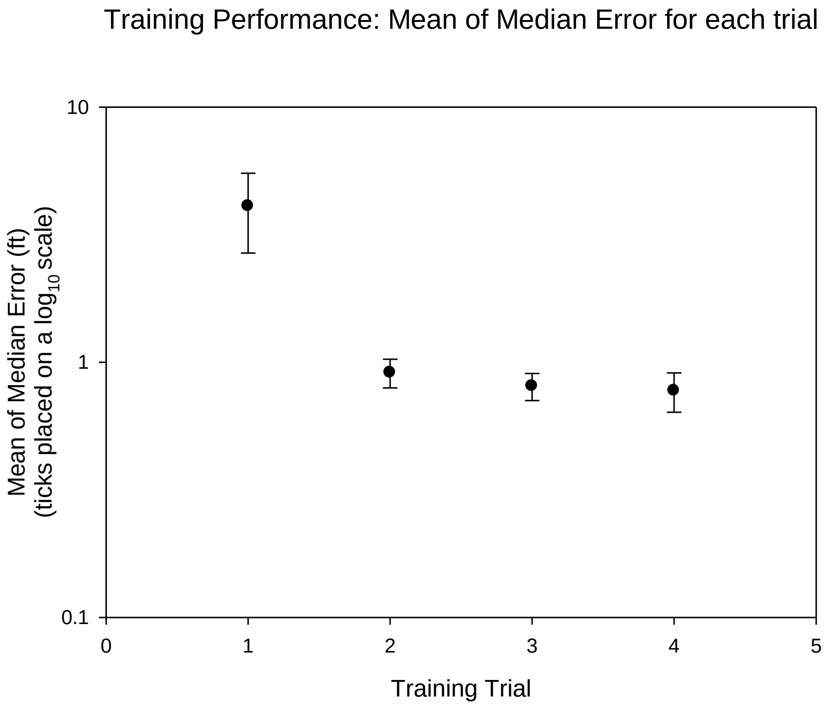

Fig. 19: Training performance of all 11 participants. Error bars are SEs.

Similarly to Experiment One, the following items were used to probe attitudes towards self-performance, task difficulty, simulator control and usefulness of training. Where Friedman tests found a significant difference in ratings, Bonferroni-corrected Wilcoxon Signed Ranks tests were used to confirm where these differences lay:

Q11) Did you find training useful for this task?

Q12) Did you perform well in this task?

Q13) Could you easily control the helicopter's movement?

Q14) Did you find the task easy?

Q15) How well could you predict the helicopter's movements?

Participants rated *training as being more useful* in conditions with visual cues (mean= 5.36, SD= 1.14) than without (mean= 3.31, SD= 2.16), Z= -4.97, p<.001; Bonferroni-adjusted alpha= 0.0167. There were no significant differences in ratings between motion conditions or auditory cue conditions (Fig. 20). Similarly, participants *rated their performance* as being better with visual cues (mean= 4.43, SD= 1.02) than without (mean= 2.23, SD= 1.16), Z= -5.62, p<.001; Bonferroni-adjusted alpha= 0.0167, with no significant differences in ratings between motion conditions or auditory cue conditions (Fig. 21). They also felt that *they could more easily control* the simulator's movement with visual cues (mean= 4.97, SD= 0.88) compared to without (mean= 2.61, SD= 1.48), Z= -5.41, p<.001; Bonferroni-corrected alpha= 0.0167 (Fig. 22). Participants *found the task easier* with visual cues (mean= 4.50, SD= 1.11) compared to without (mean= 2.20, SD= 1.36), Z= -5.50, p<.001 (Fig. 23), and could *predict the helicopter's movements* better with visual cues (mean= 4.61, SD= 1.17) compared to without (mean= 2.07, SD= 1.47), Z= -5.46, p<.001; Bonferroni-corrected alpha= 0.0167 (Fig. 24).

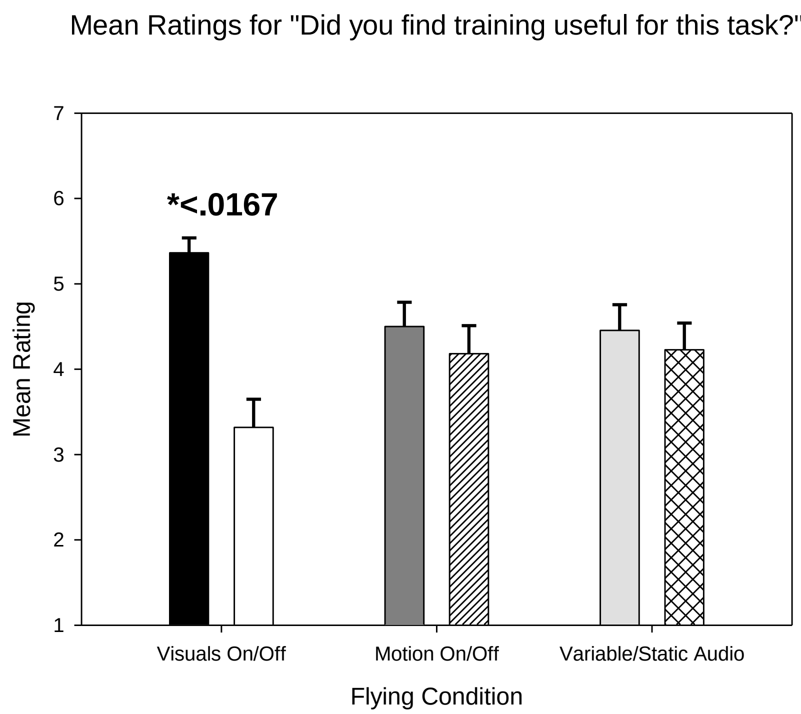

Fig. 20: Mean ratings for training usefulness. There was a significant difference between visual conditions. Error bars are SEs.

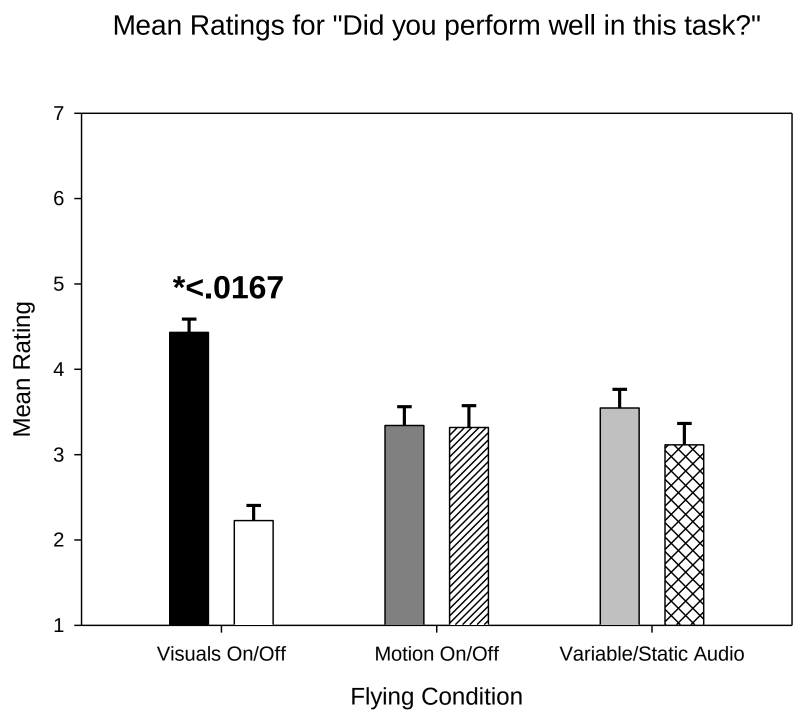

Fig. 21: Mean ratings for self-performance. There was a significant difference between visual conditions. Error bars are SEs.

Mean Ratings for "Could you easily control the helicopter's movement?"

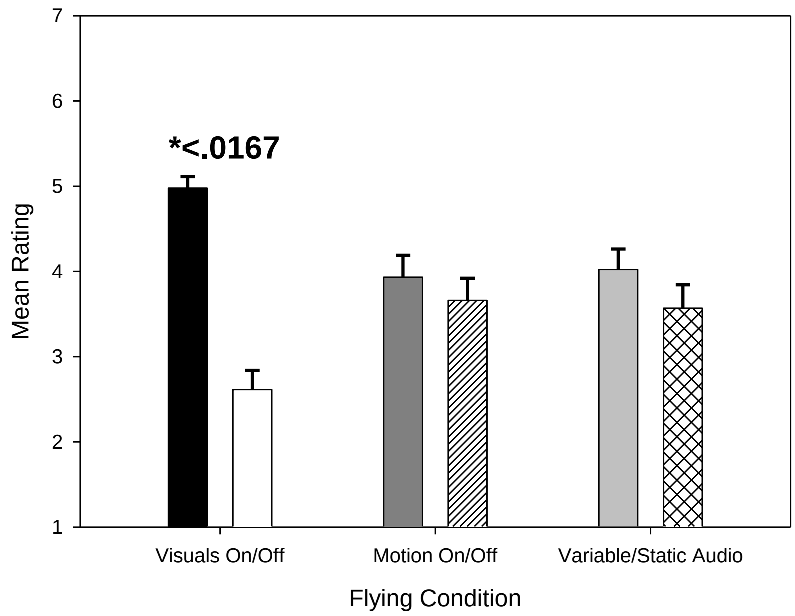

Fig. 22: Mean ratings for how well participants felt they could control the simulator. There was a significant difference between visual conditions. Error bars are SEs.

Mean Ratings for "Did you find the task easy?"

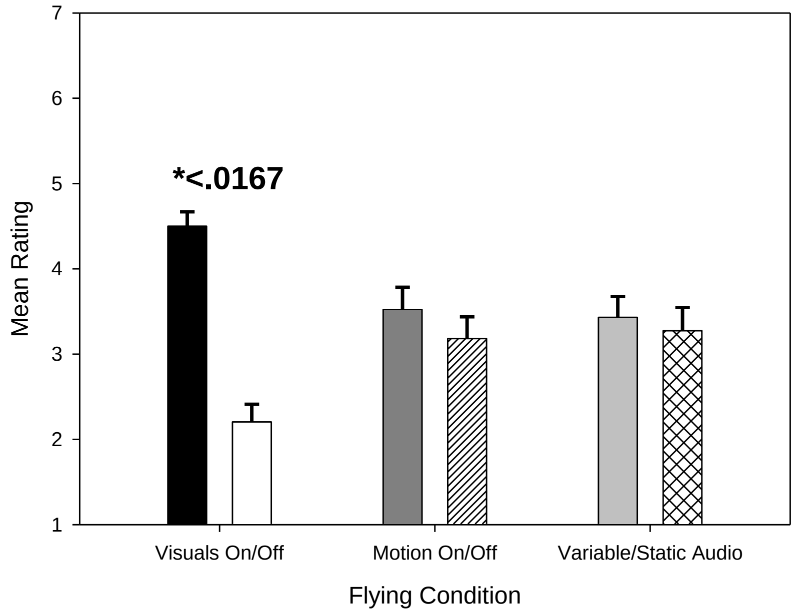

Fig. 23: Mean ratings for how easy participants found the task. There was a significant difference between visual conditions. Error bars are SEs.

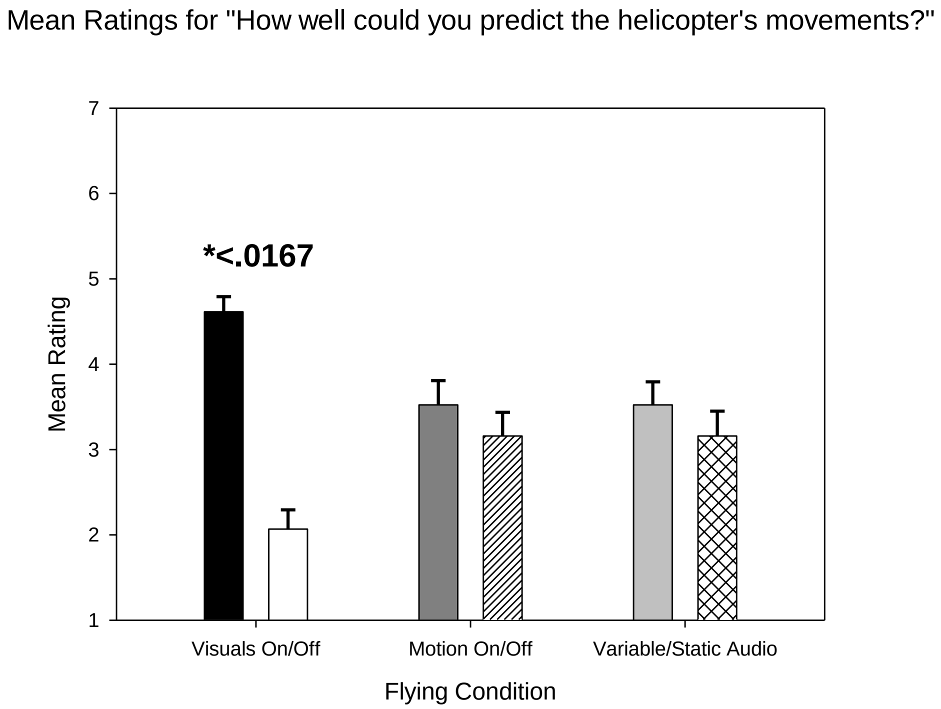

Fig. 24: Mean ratings for how easily participants could control the simulator. There was a significant difference between visual conditions. Error bars are SEs.

3.4.2 – Discussion

As in Experiment One, participants demonstrated rapid performance improvement during training on the low-fidelity simulator. Training transferred from the low-fidelity simulator to HELIFLIGHT (*Section 3.1.1*). Again, participants were asked to judge the usefulness of training to explore whether subject self-ratings co-varied with objective performance data. In contrast to Experiment One, participants reliably judged their self-performance in line with objective performance data. As the most accessible feature of the simulation, it is to be expected that participants judged visual cues as being the most important factor influencing their performance.

In Experiment One, participants were trained with a binaural, vertically-moving auditory cue without the benefit of a head-tracker system. In this experiment, the head-tracker allowed more accurate localisation of self-position and so should have made the auditory motion cue more accessible to participants (Wightman & Kistler, 1999; Seeber & Fastl, 2003). However, participants again did not demonstrably consider training to be useful in conditions with the variable auditory cue (Experiment 1). In spite of this, control input data showed that participants were more responsive to the variable auditory cue in Experiment Two compared to Experiment One (*Section 3.1.1*). Although the auditory motion cue did not offer improvement in performance or influence explicit ratings of performance, participants appear to have implicitly modified their task behaviours in response to the cue (Väljamäe et al., 2008; Väljamäe, 2009).

As in Experiment One, participants were trained without motion cues. Participants did not rate training to be more useful in either motion or no-motion post-training trials. This could be seen as successful self-evaluation on part of the novice participant, because motion cues did not confer any benefits to objective task performance (*Section 3.1.1*). Therefore, there may be some instances where self-evaluation of training can agree objective performance data or evaluations made by expert observers (Morrison & Hammon, 2000).

Chapter 4: Discussion and Conclusions

VR simulations are increasingly used to explore the concepts of fidelity and presence, as well as the effects of these constructs on training transfer from simulations to the real world. One can take steps to painstakingly replicate the physical and visual features of the simulated vehicle, but there is little focus on how realistic the user perceives the simulation to be, which may not be solely influenced by photorealistic replication. Are low-fidelity simulators effective training devices? Also, how do we measure the degree to which the user engages with the simulation captured by the concepts of *fidelity, presence* and *performance?* In light of this, two simulator-based experiments were conducted to explore how post-trial subjective ratings which evaluated the constructs of fidelity, presence and self-performance could be compared with objective, continuously-sampled task performance measures.

Previous literature has claimed that qualitative measures taken from novices are a poor tool for functional and physical fidelity evaluation, and that using an objective metric, such as task performance, as a referent for fidelity is more reliable (Schricker et al., 2001; Meehan et al., 2002). However, this idea has not been wholly supported by the findings of this study. In Experiment 1, there was some agreement between objective task performance and subjective ratings of simulator fidelity. Participants felt that the presence of motion cues made the experience in the simulator more realistic, which corresponded with better task performance with motion cues compared to without. This is an instance where subjective and objective measures of fidelity are correlated. Furthermore, an advantage of subjective evaluation is the relative ease with which data can be collected. These factors support the argument that subjective measures of fidelity may be as valuable as objective measures.

On the other hand, participants rated the variable turbine as being more realistic than the static turbine despite the fact that it had no demonstrable benefit on flight performance. Thus, the *functional fidelity* of the turbine (how it behaved) had little effect on either task performance or subjective fidelity ratings. However, if participant ratings were influenced by *physical fidelity* (what it sounded like), then they should have rated both the variable and static turbines as being more realistic than the substitute turbine, which they did not. In this case, neither objective or subjective measures adequately captured participants' perceptions of physical and functional fidelity, which lends to the idea that objective performance measures may not necessarily be a gold standard metric of fidelity.

Furthermore, in Experiment 2 participants felt that vestibular and variable auditory cues contributed to the fidelity of the simulation. Only the variable auditory cue influenced participant behaviour, but did not significantly improve their task performance. In this instance, subjective measures captured participants' perceived fidelity of the simulation, which objective measures did not. On the whole, these findings disagree with the idea that an objective referent best captures simulation fidelity, and highlights the need for reliable measures in evaluating simulation fidelity.

It could be argued that subjective measures are more appropriate for evaluating presence compared to a referent such as task performance. Despite the simulated sensory cues being unable to correspond exactly with those one would experience in the real world, participants reported increased sensory engagement with visual and motion cues compared to without, and rated the variable auditory cue as being more engaging than the static cue. This supports the idea that high-fidelity cues are not always necessary to induce the sensation of presence (Zimmons & Panter, 2003). These findings also support the assertion that inclusion of multisensory cues increases the sensation of presence for participants in a virtual reality simulation (Dinh et al., 1999; Sanchez-Vives & Slater, 2005; Gallace et al., 2011). Similarly to fidelity ratings, however, there were some discrepancies between presence ratings and objective performance measures. Participants rated themselves as being more able to localize sounds with the variable auditory cue compared to the static, but there was no difference in task performance between these conditions. Likewise, sensory engagement and feelings of involvement were higher in conditions using vestibular cues compared to without, but again there was no effect of vestibular cue on task performance. This supports previous findings in the literature where subjective presence ratings have not correlated with objective performance measures, and suggests that task performance is not a comparable metric for presence evaluation (Mania & Chalmers, 2001; Durlach et al. 2005; Ooms, 2004).

Multimodal motion cues were presented in both experiments in order to examine the effects of cue integration and implicit learning on task performance and training transfer. Extant literature demonstrates the advantageous effects of congruent, temporally-aligned multimodal cue presentation on task performance measures such as reaction time (Harrison et al., 2010; Hancock et al., 2013). In Experiment 1, vestibular and auditory motion cues were presented in tandem with the visual display of helicopter movement. Participants demonstrated sufficient integration of vestibular cues, which benefited task performance in comparison to no vestibular cues (Chen et al., 2011). However, this effect was not seen in Experiment 2, where task performance did not differ between motion and no-motion conditions. Participants had a tendency to make large deviations from the central heading, which often resulted in them turning in circles. The inability of the simulator motion platform to rotate 360˚ mean that vestibular cues could not always remain congruent with visual information. As such, the spatial alignment discrepancy between the visual and vestibular cues may have been responsible for failure to integrate information from the two modalities.

Implicit learning of sensory cues has been demonstrated by those such as Seitz, Kim, et al. (2007), where participants form associations between cues from multiple modalities without explicit instructions to do so. In both experiments of this study, participants were passively exposed to auditory motion cues throughout training, but they did not appear to successfully integrate auditory motion cues with visual or vestibular information. However, a main effect of auditory cue on participant control behaviour was observed in Experiment 2, showing that participants responded to differences between the variable-audio and static-audio conditions. It is possible that some learning of auditory motion cue behaviour had taken place during training, but not to the extent where participants could exploit its information to improve aid task performance.

Despite participants' limited use of sensory cues, computer-based, low-fidelity simulations were demonstrated to be adequate training tools for training transfer to HELIFLIGHT. Experiments 1 and 2 both showed that participant training significantly improved task performance. Improvement of task performance transferred from the low-fidelity desktop simulator to a higher-fidelity simulator, supporting the idea that even basic simulations have their place as a training tool (Patrick, 1992; Taylor et al., 1999) and contradicting the notion that training in a high-fidelity simulation is necessary to improve task performance (Dahlstrom et al., 2009). The findings also contribute to evidence for transfer of training from simulators to the real-world environment in a variety of settings, from surgical training to pilot training (Gurusamy et al., 2008; Longridge et al., 2001).

In the absence of an expert examiner, objective evaluation of training outcomes and task performance was more useful than subjective self-evaluation. There was little evidence from either experiment of participants' ability to introspect on their training experience, and only some evidence of their ability to evaluate self-performance. In Experiment 1, participants gave consistently high ratings of training across all trials regardless of whether or not conditions replicated those of training. This supports the idea that the ratings were driven by some bias or expectation unrelated to actual performance (Wiggins & O'Hare, 2003). In Experiment 2, ratings of training usefulness were only significantly affected by the presence or absence of visual cues. It was expected that training with HRTFs via a head-tracking system would increase participants’ ability to localise their self-position, and therefore influence ratings of training where the variable auditory cue was used. However, this was not the case. These findings support the idea that subjective evaluation of training by experts and objective evaluation of task performance, therefore, may be more reliable than self-assessment (e.g. Hyltander et al. 2002; Watterson et al. 2002; Kasarskis et al. 2001; Morrison & Hammon, 2000).

Limitations of the study are evident in that the scales used to measure fidelity and performance in Experiments 1 and 2 are were formulated for the specific behavioural task at hand and were not tested for internal reliability, as per other usability, presence and fidelity scales. This may go some way to explain why subjective ratings did not generally correspond with objective performance data. In Experiment 2, the difficulty of the hover task (particularly without visual cues) may have been why the variable auditory cue influenced participants’ task behaviour but did not improve their performance. It is also recognised that the small sample sizes used in the experiments may have reduced the chances of detecting effects of sensory cues. The results presented in this thesis need to be interpreted with caution as the validity of the findings depends on the reliability of the data and the degree to which these results would transfer between studies.

This study demonstrated training transfer between the low-fidelity and high fidelity simulators, but this could also have been explained by practice effects. Although the flight paths in both experiments were too complex to learn, it is possible that, as participants were unfamiliar with the experimental task, initial performance was poor with subsequent performance improvement being attributable to increased familiarity with the task. Participants would have gained familiarity with controlling the simulator during the pre-training experiments, because the requirements of each trial were identical (to track a vertically-moving basket in Experiment 1, and return to a central heading in Experiment 2). Training sessions on the low-fidelity simulators were intended to passively expose participants to auditory motion cues before re-testing in the high-fidelity simulator. Although participants demonstrated improved performance during training in both experiments, it may well be the case that familiarity built up with the task during the pre-training phase would have resulted in improved performance without the need for training on the low-fidelity simulators.

Due to a restricted participant pool and time constraints, no participants in this experiment were allocated to a control group. In this experiment, all participants were assigned to training and it therefore cannot be unequivocally claimed that training on a low-fidelity simulator is more superior than prolonged exposure to HELIFLIGHT. The inclusion of a control group in each experiment would afford the opportunity to demonstrate any advantageous effects of training on a low-fidelity simulator on training transfer to a high-fidelity simulator. For example, after the pre-training task participants could be assigned to either train on the low-fidelity simulator or assigned to perform an unrelated task before re-testing in the high-fidelity simulator.

Nonetheless, these experiments contribute to the exploration of suitable metrics for evaluating simulator fidelity, presence and task performance. They showed that subjective ratings were a useful tool for capturing perceived fidelity and presence in a way that objective performance measures could not. They also support the use of low-fidelity simulations as training tools. Although objective performance measures did not appear to be a good referent for fidelity and presence evaluation, it was demonstrated that they were more useful than subjective measures in evaluating performance and training transfer.

In summary, existing literature has suggested that the inclusion of multimodal cues in simulated environments affords better task performance, improves simulation fidelity and enhances subjective feelings of presence (Harrison et al., 2010; Väljamäe et al., 2008; Berger et al., 2010; Gallace et al., 2011). This study only partially supports these claims, as participants utilised vestibular motion cues to aid task performance in Experiment 1, but not in Experiment 2. In both experiments, the inclusion of auditory motion cues did not afford improved task performance. On the other hand, participants' evaluation of fidelity and presence were significantly affected by the presence or absence of motion information in sensory cues, even if they did not successfully use this information to improve task performance. This is consistent with the notion that multisensory simulated environments are more engaging and immersive to the individual (Gallace et al., 2011; Dinh et al., 1999; Sanchez-Vives & Slater, 2005).

The idea that an objective referent, such as task performance, is the ideal measure of fidelity has not been wholly supported by this study (Schricker et al., 2001). This contradiction may be due to participants being unable to successfully integrate multimodal cues, as demonstrated by their inability to use auditory motion cues to aid performance and the inconsistent effects of vestibular cues on task performance. It is possible that if multimodal cues are carefully tailored to the physical limitations of apparatus and are ensured to be meaningful to participants, this would enable objective referents like task performance to be robust and consistent measures of fidelity.

As discussed beforehand, a repetition of this study with the inclusion of control groups is necessary to fully support the idea that low-fidelity simulations are suitable training devices and aid transfer of skills to higher-fidelity environments. A further avenue of exploration of participants' prior experiences would be to examine its effects upon task learning. For example, it would be of interest to investigate the differences in the rate of learning and degree of training transfer between experienced pilots and novice participants, or between participants who play videogames and those who do not. If training on a low-fidelity simulator demonstrated little-to-no advantage on task performance for experienced pilots, this would contradict the idea that simulator training is beneficial for pilot training. If video gamers are quicker to learn a task compared to non-gamers, this suggests that the skills possessed by gamers may translate to better task performance and training transfer, and that simulations in the style of videogames may be useful training tools for novices.

Applications for these experiments may include the development of a valid and reliable fidelity scale on par with existing presence questionnaires, so that subjective evaluation of fidelity and presence could be more evenly compared. These results may also encourage the use of fidelity, presence and performance measures in a variety of simulator-based tasks. Rather than relying upon a single metric of presence, fidelity or performance, these results may encourage the complementary use of subjective and objective measures to more holistically evaluate people's perceptions and behaviour in simulated environments.

Acknowledgements

This work was supported by the Economic and Social Research Council and the Engineering and Physical Sciences Research Council. The author would also like to thank her supervisors, Dr. Georg Meyer and Dr. Mark White, for their support and guidance.

References

Aggarwal, R., Black. S.A., Hance, J.R., Darzi, A., & Cheshire, N.J.W. (2006). Virtual reality simulation training can improve inexperienced surgeons' endovascular skills. *European Journal of Vascular and Endovascular Surgery, 31*(6), 588-593. doi: 10.1016/j.ejvs.2005.11.009.

Alexander, A.L., Brunyé, T., Sidman, J., & Weil, S.A. (2005). *From Gaming to Training: A Review of Studies on Fidelity, Immersion, Presence, and Buy-in and Their Effects on Transfer in PC-based Simulations and Games.* Retrieved from http://www.aptima.com/publications/2005_Alexander_Brunye_Sidman_Weil.pdf

Baños, R.M., Botella, C., Alcañiz, M., Liaño, V., Guerrero, B., & Rey, B. (2004). Immersion and Emotion: Their Impact on the Sense of Presence. *Cyberpsychology and Behavior, 7*(6), 734-741. doi: 10.1089/cpb.2004.7.734.

Barnett-Cowan, M., Meilinger, T., Vidal, M., Teufel, H., & Bülthoff, H.H. (2012). MPI CyberMotion Simulator: Implementation of a Novel Motion Simulator to Investigate Multisensory Path Integration in Three Dimensions. *Journal of Visualized Experiments, 63,* e3436. doi: 10.3791/3436.

Berger, D.R., Schulte-Pelkum, J., & Bülthoff, H.H. (2010). Simulating believable forward accelerations on a stewart motion platform. *ACM Transactions on Applied Perception, 7*(1), article no. 5. doi: 10.1145/1658349.1658354.

Bormann, K. (2008). Visuals are not how they look. *Virtual Reality, 12*(2), 115-123. doi: 10.1007/s10055-007-0068-4.

Bowman, D.A. & McMahan, R.P. (2007). Virtual Reality: How Much Immersion Is Enough? *IEEE Computer, 40*(7), 36-43. doi:10.1109/MC.2007.257.

Bürki-Cohen, J. & Go, T.H. (2005). The Effect of Simulator Motion Cues on Initial Training of Airline Pilots. *Proceedings of the AIAA Modeling and Simulation Technologies Conference and Exhibit, San Francisco, CA.*

Bürki-Cohen, J., Sparko, A.L., & Go, T.H. (2007). Training Value of a Fixed-Base Flight Simulator with a Dynamic Seat. *Proceedings of the AIAA Modeling and Simulation Technologies Conference, Hilton Head, SC.*

Butler, J.S., Campos, J.L., & Bülthoff, H.H. (2015). Optimal visual–vestibular integration under conditions of conflicting intersensory motion profiles. *Experimental Brain Research*, *233*(2), 587-597.

Casner, S.M., Geven, R.W., & Williams, K.T. (2013). The effectiveness of airline pilot training for abnormal events. *Human Factors: The Journal of Human Factors and Ergonomics*, *55*(3), 477-485.

Chen, A., DeAngelis, G.C., & Angelaki, D.E. (2011). Representation of Vestibular and Visual Cues to Self-Motion in Ventral Intraparietal Cortex. *The Journal of Neuroscience*, *31*(33), 12036-12052. doi: 10.1523/JNEUROSCI.0395-11.2011.

Cho, S.J., Ovcharenko, A., & Chong, U.P. (2006). Front-Back Confusion Resolution in 3D Sound Localization with HRTF Databases. *The 1st International Forum on Strategic Technology*, 239-243. doi: 10.1109/IFOST.2006.312297.

Cooper, G. & Harper, R. (1969). *The use of pilot rating in the evaluation of aircraft handling qualities.* Technical Report TN D-5153, NASA.

Dahlstrom, N., Dekker, S., van Winsen, R., & Nyce, J. (2009). Fidelity and validity of simulator training. *Theoretical Issues in Ergonomics Science*, *10*(4), 305-314. doi: 10.1080/14639220802368864

Dinh, H.Q., Walker, N., Song, C., Kobayashi, A., & Hodges, L.F. (1999). Evaluating the importance of multi-sensory input on memory and the sense of presence in virtual environments. *Proceedings of the IEEE Virtual Reality Conference, Houston*, *TX*, 222-228. doi: 10.1109/VR.1999.756955.

Duncan, J. (2006). Fidelity versus cost and its effect on modeling and simulation. *12th International Command and Control Research and Technology Symposium*, *Newport, RI*.

Durlach, P.J., Fowlkes, J., & Metevier, C.J. (2005). Effect of Variations in Sensory Feedback on Performance in a Virtual Reaching Task. *Presence: Teleoperators and Virtual Environments, 14*(4), 450-462. doi: 10.1162/105474605774785299.

Ferwerda, J.A. (2003). Three Varieties of Realism in Computer Graphics. *Human Vision and Electronic Imaging VIII, Proceedings of SPIE 5007,* 290-297. doi: 10.1117/12.473899

Gerathewohl, S.J. (1969). Fidelity of Simulation and Transfer of Training: A Review of the Problem. *Federal Aviation Administration Office of Aviation Medicine Report*.

Grabherr, L., Nicoucar, K., Mast, F.W., & Merfeld, D.M. (2008). Vestibular thresholds for yaw rotation about an earth-vertical axis as a function of frequency. *Experimental Brain Research, 186*, 677-681. doi:10.1007/s00221-008-1350-8.

Gurusamy, K., Aggarwal, R., Palanivelu, L., & Davidson, B.R. (2008). Systematic review of randomized controlled trials on the effectiveness of virtual reality training for laparoscopic surgery. *The British Journal of Surgery, 95*(9), 1088-1097. doi: 10.1002/bjs.6344.

Hale, K.S., Stanney, K.M., & Malone, L. (2009). Enhancing virtual environment spatial awareness training and transfer through tactile and vestibular cues. *Ergonomics, 52*(2), 187-203. doi: 10.1080/00140130802376000.

Hancock, P.A., Mercado, J.E., Merlo, J., & Van Erp, J.B.F. (2013). Improving target detection in visual search through the augmenting multi-sensory cues. *Ergonomics, 56*(5), 729-738. doi: 10.1080/00140139.2013.771219.

Harris, D. (2011). *Human Performance on the Flight Deck* (pp. 214). UK: Ashgate Publishing.

Harrison, N.R., Wuerger, S.M., & Meyer, G.F. (2010). Reaction time facilitation for horizontally moving auditory-visual stimuli. *Journal of Vision, 10*(14):16, 1-21. doi: 10.1167/10.14.16.

Hendrix, C. & Barfield, W. (1996). Presence in virtual environments as a function of visual and auditory cues. *Proceedings of the Virtual Reality Annual International Symposium, Research Triangle Park, NC*, 74-82. doi: 10.1109/VRAIS.1995.512482.

Hyltander, A., Liljegren, E., Rhodin, P.H., & Lönroth, H. (2002). The transfer of basic skills learned in a laparoscopic simulator to the operating room. *Surgical Endoscopy and Other Interventional Techniques, 16*(9), 1324-1328.

Jeon, J.Y., Kim, Y.H., Cabrera, D., & Bassett, J. (2008). The effect of visual and auditory cues on seat preference in an opera theater. *Journal of the Acoustical Society of America, 123*(6), 4272-4282. doi: 10.1121/1.2912435.

Judkins, T.N., Oleynikov, D., & Stergiou, N. (2009). Electromyographic response is altered during robotic surgical training with augmented feedback. *Journal of Biomechanics, 5*(42), 71-76. doi: 10.1016/j.jbiomech.2008.09.039.

Kasarskis, P., Stehwein, J., Hickox, J., Aretz, A., & Wickens, C. (2001). Comparison of expert and novice scan behaviors during VFR flight. *Proceedings of the 11th International Symposium on Aviation Psychology, Columbus, OH.*

Lackey, S.J., Maraj, C.S., & Barber, D.J. (2014). Immersion, Presence, and Flow in Robot-Aided ISR Simulation-Based Training. *Proceedings of the 2014 Winter Simulation Conference, Savanah, GA.*

Lee, D., Sezer-Uzol, N., Horn, J.F., & Long, L.N. (2005). Simulation of Helicopter Shipboard Launch and Recovery with Time-Accurate Airwakes. *Journal of Aircraft, 42*(2), 448-461.

Lele, A. (2013). Virtual reality and its military utility. *Journal of Ambient Intelligence and Humanized Computing, 4*(1), 17-26. doi: 10.1007/s12652-011-0052-4.

Gallace, A., Ngo, M.K., Sulaitis, J., & Spence, C. (2011). Multisensory presence in virtual reality: possibilities and limitations. In G. Ghinea, F. Andres, & S.

Gulliver (Eds.), *Multiple New Developments in MulSeMedia Sensorial Media Advances and Applications* (pp.1-40). Hershey, PA: IGI Global.

Liu, D., Blickensderfer, E.L., Macchiarella, N.D., & Vincenzi, D.A. (2009). Transfer of Training. In D.A. Vincenzi, J.A. Wise, M. Mouloua, & P.A. Hancock (Eds.), *Human Factors in Simulation and Training*. FL: Boca Raton.

Longridge, T., Bürki-Cohen, J., Go, T.H., & Kendra, A.J. (2001). Simulator fidelity considerations for training and evaluation of today's airline pilots. *Proceedings of the 11th International Symposium on Aviation Psychology, Columbus, OH.*

Mania, K. & Chalmers, A. (2001). The effects of levels of immersion on memory and presence in virtual environments: a reality centered approach. *Cyberpsychology and Behavior: The impact of the Internet, multimedia and virtual reality on behavior and society, 4*(2), 247-264.

Mania, K. & Robinson, A. (2004). The effect of quality of rendering on user lighting impressions and presence in virtual environments. *Proceedings of the 2004 ACM SIGGRAPH international conference on Virtual Reality continuum and its applications in industry, New York, NY*, 200-205. doi: 10.1145/1044588.1044629.

McCauley, M.E. (2006). *Do Army Helicopter Training Simulators Need Motion Bases?* Technical Report from the United States Army Research Institute for the Behavioral and Social Sciences. Retrieved from http://www.researchgate.net/publication/235134325_Do_Army_Helicopter_Training_Simulators_Need_Motion_Bases.

McMahan, R.P., Bowman, D.A., Zielinski, D.J., & Brady, R.B. (2012). Evaluating Display Fidelity and Interaction Fidelity in a Virtual Reality Game. *IEEE Transactions on Visualization and Computer Graphics, 18*(4), 626-633. doi: 10.1109/TVCG.2012.43.

Meehan, M., Insko, B., Whitton, M., & Brooks Jr., F.P. (2002). Physiological Measures of Presence in Stressful Virtual Environments. *Proceedings of the*

29th Annual Conference on Computer Graphics and Interactive Techniques, New York, NY, 645-652. doi: 10.1145/566570.566630.

Meyer, G.F., Wong, L.T., Timson, E., Perfect, P., & White, M.D. (2012). Objective Fidelity Evaluation in Multisensory Virtual Environments: Auditory Cue Fidelity in Flight Simulation. *PLoS ONE*, *7*(9), e44381. doi:10.1371/journal.pone.0044381.

Miller, J. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology, 14,* 247-279.

Morrison, J.E. & Hammon, C. (2000). *On measuring the effectiveness of large-scale training simulations.* Retrieved from the Institute for Defense Analyses, http://www.dtic.mil/dtic/tr/fulltext/u2/a394491.pdf.

Nykänen, A., Zedigh, A., & Mohlin, P. (2013). *Effects on localization performance from moving the sources in binaural reproductions*. Paper presented at the 42nd International Congress and Exposition on Noise Control Engineering (INTER-NOISE 2013), Innsbruck.

Ooms, M. (2004). *Measuring presence and awareness of external events*. Paper presented at Le Twente Student Conference on IT, Enschede. doi: 10.1.1.59.9238.

Patrick, J. (1992). *Training: Research and Practice*. London: Academic Press.

Pausch, R., Proffitt, D., & Williams, G. (1997). Quantifying immersion in virtual reality. *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques,* 13-18. doi: 10.1145/258734.258744.

Perfect, P., Timson, E., White, M.D., & Padfield, G.D. (2014). A Rating Scale for Subjective Assessment of Simulator Fidelity. *The Aeronautical Journal, 118*(1206), 953-974.

Riva, G., Davide, F., & IJsselsteijn, W.A. (Eds.). (2003). *Being There - Concepts, Effects and Measurements of User Presence in Synthetic Environments.* Amsterdam: IOS Press.

Sallnäs, E., Rassmus-Gröhn, K., & Sjöström, C. (2000). Supporting presence in collaborative environments by haptic force feedback. *ACM Transactions on Computer-Human Interaction (TOCHI) - Special issue on human-computer interaction and collaborative virtual environments, 7*(4), 461-476. doi: 10.1145/365058.365086.

Sanchez-Vives, M.V. & Slater, M. (2005). From presence to consciousness through virtual reality. *Nature Reviews: Neuroscience, 6,* 332–339. doi: 10.1038/nrn1651.

Schricker, B.C., Franceschini, R.W., & Johnson, T.C. (2001). Fidelity Evaluation Framework. *Proceedings of the 34th Annual Simulation Symposium*, 109-118.

Seeber, B.U. & Fastl, H. (2003). Subjective selection of non-individual head-related transfer functions. *Proceedings of the International Conference on Auditory Display, Boston, MA.*

Seitz, A.R., Kim, R., van Wassenhove, V., & Shams, L. (2007). Simultaneous and independent acquisition of multisensory and unisensory associations. *Perception, 36*, 1445-1453. doi:10.1068/p5843.

Seymour, N.E., Gallagher, A.G., Roman, S.A., O'Brien, M.K., Bansal, V.K., Andersen, D.K., & Satava, R.M. (2002). Virtual Reality Training Improves Operating Room Performance: Results of a randomized, double-blinded study. *Annals of Surgery, 236*(4), 458-464.

Sherman, W.R. & Craig, A.B. (2003). *Understanding Virtual Reality*. San Francisco, CA: Morgan Kaufmann.

Slater, M. & Wilbur, S. (1997). A Framework for Immersive Virtual Environments (FIVE): Speculations on the Role of Presence in Virtual Environments, *Presence, 6*, 603-616.

Slater, M., Khanna, P., Mortensen, J., & Insu, Y. (2009). Visual Realism Enhances Realistic Response in an Immersive Virtual Environment. *IEEE: Computer Graphics and Applications, 29*(3), 76-84. doi: 10.1109/MCG.2009.55.

Soto-Faraco, S., Kingstone, A., & Spence, C. (2003). Multisensory contributions to the perception of motion. *Neuropsychologia, 41*(13), 1847-1862. doi: 10.1016/S0028-3932(03)00185-4.

Taylor, H.L., Lintern, G., Hulin, C.L., Talleur, D.A., Emanuel Jr., T., & Phillips, S.I. (1999). Transfer of Training Effectiveness of a Personal Computer Aviation Training Device. *The International Journal of Aviation Psychology, 9*(4), 319-355.

Tracey, M.R. & Lathan, C.E. (2001). The interaction of spatial ability and motor learning in the transfer of training from virtual to a real task. In J.D. Westwood, H.M. Hoffman, G.T. Mogel, D. Stredney, & R.A. Robb (Eds.), *Medicine Meets Virtual Reality* (pp. 521-527). Amsterdam: IOS Press.

Usoh, M., Catena, E., Arman, S. & Slater, M (2000). Using Presence Questionnaires in Reality. *Presence: Teleoperators and Virtual Environments, 9*(5), 497–503. doi: 10.1162/105474600566989.

Väljamäe, A. (2007). *Sound for multisensory motion simulators* (Doctoral dissertation). Retrieved from Chalmers University of Technology, http://publications.lib.chalmers.se/publication/44880-sound-for-multisensory-motion-simulators.

Väljamäe, A. (2009). Auditorily-induced illusory self-motion: a review. *Brain Research Reviews, 61*(2), 240-255. doi: 10.1016/j.brainresrev.2009.07.001.

Väljamäe, A., Larsson, P., Västfjäll, D., & Kleiner, M. (2008). Sound representing self-motion in virtual environments enhances linear vection. *Presence, 17*(1), 43-56. doi: 10.1162/pres.17.1.43).

Watterson, J.D., Beiko, D.T., Kuan, J.K., & Denstedt, J.D. (2002). A Randomized Prospective Blinded Study Validating Acquisition of Ureteroscopy Skills Using A Computer Based Virtual reality Endourological Simulator. *Journal of Urology, 168*(5), 1928-1932. doi: 10.1016/S0022-5347(05)64265-6.

Wiggins, M.W. & O'Hare, D. (2003). Expert and Novice Pilot Perceptions of Static In-Flight Images of Weather. *International Journal of Aviation Psychology, 13*(2), 173-187. doi: 10.1207/S15327108IJAP1302_05.

Wightman, F.L. & Kistler, D.J. (1999). Resolution of front-back ambiguity in spatial hearing by listener and source movement. *The Journal of the Acoustical Society of America,105*(5), 2841-2853.

Wignall, G.R., Denstedt, J.D., Preminger, G.M., Cadeddu, J.A., Pearle, M.S., Sweet, R.M., & McDougall, E.M. (2008). Surgical Simulation: A Urological perspective. *The Journal of Urology, 179*(5), 1690-1699. doi: 10.1016/j.juro.2008.01.014.

Witmer, B.G. & Singer, M.J. (1998). Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence: Teleoperators and Virtual Environments, 7*(3), 225-240. doi: 10.1162/105474698565686.

Zimmons, M.S. & Panter, A. (2003). The Influence of Rendering Quality on Presence and Task Performance in a Virtual Environment. *Proceedings of the IEEE Virtual Reality*, 293-294. doi: 10.1109/VR.2003.1191170.

APPENDIX – About HELIFLIGHT

HELIFLIGHT is a full-motion helicopter simulator which has the ability to simulate both rotary and fixed-wing aircraft. It was first installed in the University of Liverpool's School of Engineering in 2001. As well as containing a realistic cockpit and flight controls and collimated visual displays, it also possesses a Maxcue 6 degree freedom-of-motion platform, which affords the advantage of providing simulated motion cues. These together enable the pilot to be immersed in the simulated environment and hence engage in the virtual flying experience. HELIFLIGHT runs under the Linux Operating System on a PC-based architecture (Padfield & White, 2001).

Fig. 1: HELIFLIGHT is pictured during its first year in operation at the University of Liverpool.

Software

Through the use of HELIFLIGHT's specialist software, FLIGHTLAB, one is able to configure the simulator to selectively control features concerning simulation fidelity and flight dynamics. Two graphical user interfaces (GUI) were used to create the flight models used in the two experiments: the Control System Graphical Editor (CSGE) and X-analysis (Padfield & White, 2001).

The flight dynamics model used in Experiments One and Two is shown below, which was created in CSGE by Dr. Philip Perfect. The CSGE allows the user to specify the architecture of a flight dynamics model by using icons, representing control elements, and connections to form a schematic diagram. The model was restricted to up/down movements for Experiment One, and restricted to movement along the azimuth in Experiment Two. The control input (Xc) was controlled via the collective lever (Experiment One) or the rudder pedals (Experiment Two) by the pilot. Two parameters influenced the flight dynamics; whilst the input gain was constant in both experiments (Zo = 4.8), the damping coefficient (Zw) was manipulated in Experiment One to create difficult to control flight dynamics (Zw= -0.1) or to afford easier control (Zw= -0.5); more damping made the simulator easier to control (Fig. 2).

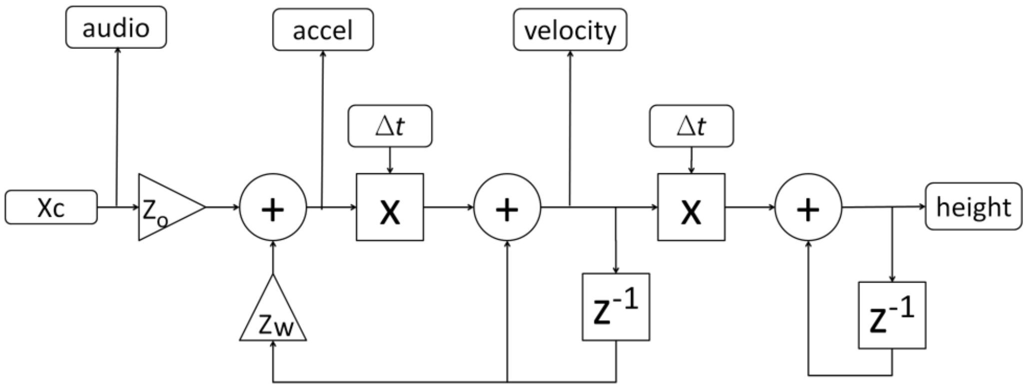

Fig. 2: Schematic showing the flight dynamics model used in Experiments 1 and 2. Control input (XC) controlled the altitude of the aircraft via two feedback loops.

X-analysis is a GUI which was used for the analysis of the flight dynamics model used in the experiments. This GUI allowed the user to select parameters for the measurement of task performance. In the case of Experiment One, control input, basket height, participant height and the error measure between basket and participant heights were recorded. In Experiment Two, control input and the angular distance in between the participant and the 0˚ were recorded.

These GUIs were represented to the participant in the form of a pilot interface called PilotStation. This formed a link between the Model Editor and X-analysis during flight. PilotStation enabled the rendering of the images used for the out-of-window displays: in Experiment One, participants saw a re-fuelling basket attached to the back of a tanker plane, whilst in Experiment Two they saw a green field with radial white lines separated every fifteen degrees.