On Approach to Reality: The Impact of a Simulated Air Traffic Control Environment (SATCE) on Workload and Situational Awareness in Military Aviators

Jonathan Allsop, Richard Keeling
RAF Central Flying School
Sleaford, Lincolnshire, UK
Jonathan.Allsop102@mod.gov.uk, Richard.Keeling926@mod.gov.uk

ABSTRACT

The use of flight simulators in flying training has become increasingly prevalent due to advances in technology and the benefits that they offer. Advances have led to high levels of fidelity in terms of flight dynamics and systems representation. In relative contrast, little attention has been directed towards the psychological fidelity of simulators, which often means that core aspects of airmanship (e.g., communication, decision-making, situational awareness) are not currently adequately fostered by their use. Simulated Air Traffic Control Environment (SATCE) solutions offer an under-examined method to partially address these shortfalls.

To examine this, a virtual reality based commercial-off-the-shelf flight simulator was coupled with SATCE software. The software injected representative airborne entities into the simulator and provided interactive Air Traffic Control (ATC) via natural language processing. Eighteen participants completed two flight tasks: a normal circuits task, and a general handling task. These tasks were completed in several conditions which were randomised across participants: (1) sterile control conditions, where the SATCE was not enabled, (2) ATC alone enabled, (3) both ATC and entities enabled. Workload, situational awareness, presence, instructor ratings of flight performance, and system voice recognition accuracy were measured.

Results showed that overall workload only increased when both ATC and airborne entities were injected into the simulator. The results from the situational awareness measure, suggest that this was due to the increased attentional demands inherent within this condition. Certain ratings of presence increased in the ATC alone condition. Voice recognition accuracy was acceptable but could have been improved through more regional-specific amendments to the speech recognition engine. In conclusion, a SATCE offers a promising method to increase the workload and attentional demands placed on aviators in simulators. Further work is required to test SATCE solutions, in an experimentally controlled manner, during actual flying training pipelines.

ABOUT THE AUTHORS

Jonathan Allsop, PhD is a chartered psychologist who works within the Royal Air Force's (RAF) Central Flying School (CFS) as a Research & Development Psychologist. Jon has a range of research interests, but with a background predominantly examining the effects of acute psychological stressors on performance, and skill acquisition. Over the past 6 years in his role, Jon has conducted research into various novel technologies and techniques that have the potential to enhance aircrew instruction. This includes multiple studies investigating the effectiveness of XR technologies for aircrew training. Jon also supports several human performance initiatives that aim to enhance trainee and instructor performance.

Richard Keeling, BA is an A2 Qualified Flying Instructor within the RAF CFS. Flt Lt Keeling has amassed over 4500 hours on the C130, Hawk T1 and Tucano T1 which includes 2500 instructional hours. In his current capacity, he delivers courses to new Flying Instructors, provides Performance Coaching to instructors and trainees and works as part of the Research and Development section of the Development and Delivery wing, CFS.

On Approach to Reality: The Impact of a Simulated Air Traffic Control Environment (SATCE) on Workload and Situational Awareness in Military Aviators

Jonathan Allsop, Richard Keeling
RAF Central Flying School
Sleaford, Lincolnshire, UK
Jonathan.Allsop102@mod.gov.uk, Richard.Keeling926@mod.gov.uk

INTRODUCTION

Flight simulators are having an increasingly prominent role in the training of aircrew within military flying training contexts, a trajectory that is likely to continue given strategic and environmental imperatives (e.g., Royal Air Force Strategy, 2022). The potential advantages that flight simulators offer have been recognised for decades (e.g., Orlansky & String, 1977), with simulators having the potential to offer a safe, effective, and resource efficient setting to train a variety of skills, ranging from fundamental procedural skills (Jacobs, Prince, Hays & Salas, 1990) to complex combat skills (e.g., Schreiber, Schroeder & Bennett, 2011). This increased utilisation has been particularly enabled through the wide variety of technological advances that have taken place over a similar time frame, including for example, large advances in computer processing power and simulation visual systems. Indeed, there are still clear opportunities to further leverage and utilise flight simulators within military flying training, by embracing even the most recent advances in simulation technologies. For example, recent rapid advancements in virtual reality (VR) and mixed reality (MR) technologies have been stated to offer several potential advantages (e.g., smaller simulator footprint, reduced cost, ease of reconfiguration) when compared to established flight simulation technologies, while maintaining training effectiveness (McCoy-Fisher et al., 2019), with recent evaluations providing initial support for these assertions (e.g., Mishler et al., 2022). While flight simulators offer a number of benefits, and there is the potential for increased utilisation, there are also a number of barriers to increased adoption, with one such barrier relating to simulation fidelity.

Simulation fidelity can be defined as the extent to which a simulation replicates the actual environment (Alessi, 1988; Gross et al., 1999). It has been suggested to be a composite term that consists of various forms of fidelity, however, these mainly fall within two sub-categories: physical fidelity and psychological-cognitive fidelity (Liu, Macchiarella & Vincenzi, 2008). Liu et al., (2008) describe the physical fidelity category as being concerned with the extent to which the simulation looks, feels, and sounds like the real-world equivalent. This category includes: visual-audio fidelity, the extent to which visual-audio stimuli are replicated; equipment fidelity, the extent to which hardware and software capabilities are replicated; and motion fidelity, the extent to which motion cues are replicated. The second sub-category, psychological-cognitive fidelity (termed psychological fidelity from this point), refers to the extent to which the simulation fosters the same psychological and cognitive processes that the actual equipment and environment would engender. For example, do both the simulation and the real-world invoke similar levels of anxiety, or similar decision-making processes.

Within military flying training, aspects related to physical fidelity are often the focus of much attention, both from aircrew and industry. Over time this has contributed to the increased availability and adoption of flight simulators with high levels of physical fidelity, where for instance, the flight and systems modelling can be almost indistinguishable from the real aircraft. However, psychological fidelity has received relatively less attention, despite its significant contribution to achieving successful training transfer (e.g., Kozlowski & DeShon, 2004). The relative lack of focus on psychological fidelity could, arguably, be entirely appropriate for the type of training that is currently being conducted, or the extent to which flight simulators are being employed at present. However, if simulation's role within military flying training is to be expanded, a greater focus is required on the psychological aspects of the simulated experience, in order to help ensure that airmanship is adequately fostered during training.

Proficient aircrew flying performance is acknowledged to rely on a combination of both airmanship and motor skills, to operate and fly the aircraft (Kern, 2010). A pilot with strong airmanship, for example, would adeptly manage unforeseen events, complex weather conditions, and external communication while competently handling the aircraft. Airmanship has however been formally defined in a variety of ways (e.g., Federal Aviation Administration [FAA], 2021; Heycock & Brown, 2000; Kern, 2010; Mane, 1981; Negård, 2014), with a commonality being the implicit agreement that it is a multidimensional, higher-order construct, that consists of at least several cognitive skills. Following definition and measurement development work led by aircrew subject matter experts (SMEs) within the Royal Air Force (RAF), Heycock & Brown (2000) proposed that airmanship comprises the following constructs: situational awareness, mental capacity, communication, decision making and resource management. The extent to which each of these airmanship constructs is fostered during simulated training is determined by multiple factors related to a simulator's psychological fidelity. In relation to the present study, situational awareness, mental capacity, and communication are of most importance, therefore, the relationship between each of these three airmanship constructs and psychological fidelity will be explored below.

Situational awareness (SA) is one of the most important constructs related to airmanship, with the generation and maintenance of SA being acknowledged as a critical requirement for successful performance in aviation settings (e.g., Endsley, 1995). At a high-level, SA is defined as a state of knowledge about a dynamic environment, and it incorporates "the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (Endsley, 1995, p. 36). Aircrew with good SA use this knowledge to make appropriate decisions, and support optimal performance (Endsley, 1995). Multiple cognitive processes interact to generate and maintain SA (Endsley, 2015), including: perception, attention, memory, and mental models. In relation to psychological fidelity, it has been previously noted, "to give this phenomenon [SA] a chance to occur, it is necessary to stage complex dynamic situations that require resources comparable to high-fidelity full-mission simulations" (Sarter & Woods, 1991, p. 53), this staging would concomitantly assist in fostering the underlying cognitive processes. However, currently within military flying training, aircrew often perform simulator sorties in sterile simulated environments, where no other air traffic is present. On the occasions where (often minimal) traffic is present, it commonly exhibits simplistic behaviours that have little impact on the aircrew's goals for the sortie, thus requiring minimal attentional resources. Similarly, weather is often pre-set and fixed, again requiring minimal attention resources and decision making throughout the sortie. These specific examples of relatively low psychological fidelity give an indication of the types of factors that must be improved if SA is to be better fostered within simulated training. Mental capacity is a second airmanship construct that could be better fostered within military flying training via a greater emphasis on psychological fidelity in simulator sorties.

Mental capacity is an airmanship construct operationalised for use within military aircrew training and assessment, in the psychological literature it is most closely linked to working memory and mental workload (RAF CFS, 2014). Workload consists of multiple dimensions, with mental workload referring to the cognitive and perceptual processing costs associated with completing a task (Estes, 2015, Hart & Staveland, 1988). This processing involves the attentional system supporting the storage, maintenance, manipulation and retrieval of information from within limited working memory resources and long-term memory (Estes, 2015). The magnitude of workload emerges from an interaction between task requirements, the circumstances under which the task is performed, and the skills, behaviours and perceptions of the operator (Hart & Staveland, 1988). In relation to psychological fidelity, the task requirements and performance circumstances often differ significantly between simulated training and real-world training. While the primary sensorimotor components of the task can now be very similar, as noted previously, the complexity and variability of the environment (e.g., ATC, weather, other aircraft) that the task is performed in is often currently dissimilar. This can lead to a relative lack of multitasking demands, a key influence on perceived workload (e.g., Yeh & Wickens, 1988). As a consequence, the simulation may not engender the same attentional processes, and workload-management strategies (e.g., task prioritisation, task shedding, communication, resource usage etc.) that are required in the airborne environment. In relation to the present study, communication is the remaining airmanship construct that could be better fostered within military flying training via a greater emphasis on psychological fidelity.

Proficient airmanship also requires effective communication skills, with their development and application being important for effective performance (e.g., Foushee & Manos, 1981; Helmreich, Merritt & Wilhelm, 1999; O'Neil & Andrews, 2000). Aircrew are required to communicate both within aircraft or formations, and with external agencies (e.g., ATC), they are also required to communicate in at least two distinct language types (Kim, 2018): plain language and phraseology (e.g., Civil Aviation Authority, 2020). The utilisation of correct phraseology is critical for flight safety and requires continual training and assessment. In training conducted within military flight simulators, communication with external agencies is conducted, although not always. When it is conducted, the instructor or evaluator will often role-play as the ATC agencies or other entities (Bell, Billington, Bennett, Billington & Ryder, 2010). However, this practice has been shown to unrealistically lower pilot communication load compared to actual airborne training, undermining training effectiveness (Bürki-Cohen et al., 2000). Aircrew also anecdotally describe the practice of playing recordings of real-world ATC communications during simulated training. However, these recordings provide little, if any, relevant contextual information to the aircrew, and are obviously not matched to any visible entities in the simulator. Therefore, this practice almost certainly does not foster the same cognitive processes that are required in the real-world environment. These examples are illustrative of psychological fidelity challenges related to the development of communication skills in simulated training. While there are multiple possible ways to enhance the psychological fidelity in simulated training (Kozlowski & DeShon, 2004), this study focuses on investigating the initial potential of a simulated air traffic control environment (SATCE).

A SATCE system consists of two components: an air traffic management or ATC component, and a traffic generator component (Papadopoli, 2017). The ATC component is made up of speech recognition, speech synthesis and ATC behaviour systems. These work together to enable a pilot to interact with the system over voice, in a relatively natural manner. The traffic generator component injects other representative aircraft visuals, radio calls and behaviours into the simulator environment. These components work together to produce a SATCE that can result in a more dynamic and representative simulator environment. However, despite their potential, there is a dearth of studies that have experimentally examined SATCE systems. Therefore, this initial study aims to address this research gap by investigating the effects of a SATCE on key aspects of military aviators' airmanship - workload and situational awareness. The study also conducted an initial examination of the recognition accuracy of a SATCE system to provide an impression of a SATCE system's adequacy for fostering communication skills, and to inform potential follow-on work.

METHODS

Equipment

A commercial-off-the-shelf flight simulator coupled with a virtual reality headset was used as the testbed for the study. Specifically, Prepar3D Version 4.5 (Lockheed Martin) was used in conjunction with a Reverb G2 (Hewlett-Packard) Virtual Reality (VR) Head Mounted Display, and a Thrustmaster Warthog Stick and Throttle (Guillemot, Montreal, Canada). A geo-specific UK terrain database was used (OrbX) along with a representative version of RAF Cranwell airfield (UK2000 Scenery). All participants flew a Grob Tutor T.1 aircraft model (Iris Simulations) which was chosen due to its familiarity and simplicity of use for participants, and its active service within the RAF. For ease of interaction while in VR, radio frequencies were pre-set and changed by the experimenter. All other push-to-talk, aircraft controls, and services were operated by the participant from the Hands-on Throttle and Stick (HOTAS).

The SATCE system employed was the Simulated Environment for Realistic Air Traffic Control (SERA; Advanced Simulation Technologies inc.) software package. This was run locally on a separate machine and connected to Prepar3D over a LAN. The SATCE software injected other aircraft entities into Prepar3D via a Distributed Interactive Simulation (DIS) interface and provided interactive ATC through voice recognition and natural language processing. It is worth noting that the SATCE system had not been extensively tailored to UK military phraseology or procedures for this initial study.

Participants

Eighteen participants (17 Male, 1 Female; Average age = 29.8, SD = 7.8) voluntarily took part in the study and provided informed consent. All participants were military aviators who had completed at least UK Military Elementary Flying Training (self-reported flying hours average = 776.4 hours, SD = 1536.1). Thirteen participants had experience of VR in an aviation context (self-reported VR flying hours average = 44.9, SD = 118.3).

Tasks

Circuits Task: This task involved a representative version of the RAF Cranwell visual circuit using runway 26L and a circuit height of 800ft above ground level (AGL). Conditions were Ceiling and Visibility OK (CAVOK) and the surface wind was 260° at 8 kts. The task required participants to complete a short taxy, take off and complete 2 visual circuits - the first one being either a successful touch and go or go around, then followed by a landing (no taxy).

Practice Area Task: This task was again performed using the same runway and environmental conditions as the circuits task, except for scattered cloud at 2000ft. It involved a short taxy, take off, Visual Flight Rules (VFR) departure and turn to a height of 3000ft. At 3000ft, participants entered a general handling sector and were then required to perform a level turn through 360° using 30° angle of bank. When the turn was complete, a VFR recovery via the initial point for runway 26L.

Research Design

A within-subjects design was employed, with the ordering of conditions and tasks being randomised across participants. The circuits task was completed in 3 experimental conditions (Sterile, ATC, and ATC + entities), whereas the practice area task was completed in 2 experimental conditions (Sterile, and ATC + entities) due to time constraints related to the availability of participants. For the sterile condition, the SATCE system was not activated at all; for the ATC condition, only the air traffic agencies were activated; for the ATC + entities condition, ATC was again activated but with other aircraft entities also being injected into the simulator. The maximum density of other aircraft entities was such that there could be a maximum of four aircraft in the circuit, with the exact number and their behaviours being randomly determined by the SATCE system for each task attempt.

Measures

Workload: Workload was measured using the NASA Task Load Index (NASA TLX; Hart & Staveland, 1988), which assesses participants' perceptions (possible scores ranging from 0 to 100) of various workload domains, including: mental, physical, temporal, effort and frustration. An unweighted administration procedure was followed due to its relative ease of employment and high correlation with the original weighted procedure (see Moroney et al., 1992).

Situational Awareness: The Situational Awareness Rating Technique (SART; Taylor, 1990) was used to quantify participants' perceptions of SA. This scale consists of ten questions (possible range from 1 to 7) grouped into three dimensions: attentional demands, attentional supply, and understanding. Only the demand and resources subscales were employed in the present study. Mean ratings were calculated for the demand and resources dimensions (Endsley, Sollenberger, Nakata, & Stein, 2000) and are presented independently (e.g., Nuamah et al., 2020).

Presence: Two items from Witmer & Singer's (1998) presence questionnaire were used to assess participants' perceptions of: the degree of similarity between the simulated experience and the real world (possible range from 1 = not at all, to 7 = very consistent with the real word), and their degree of involvement within the simulated experience (possible range from 1 = not involved, to 7 = completely engrossed).

Flight Task Performance: A Qualified Flying Instructor (QFI) with significant operational and instructional experience (self-reported approximate flying hours = 4500, instructional hours = 2500), evaluated the performance of each flight task using the standard assessment criteria and instructor mark sheet used in normal flying training. Performance is rated on a scale of 1 to 5, with the mark sheet consisting of items related to both the behavioural aspects of task performance (e.g., the mechanics of competing a finals turn), and airmanship (e.g., decision making, lookout). An overall flight task performance score was obtained for each task iteration by averaging across all assessment criteria.

SATCE Recognition Accuracy: Audio files for all radio calls (n = 1128) were manually transcribed by a military aviation SME. Relatedly, text files for all radio calls were extracted from the SATCE system, these contained the SATCE system's recognised version of the radio call. Each radio call was then rated by the SME as either:

- Correctly Accepted (CA) – an aviator's correct radio call was correctly recognised and actioned by the system.
- Correctly Rejected (CR) – an aviator's incorrect radio call (e.g., wrong frequency read-back, wrong callsign etc.) was correctly rejected by the system.
- Incorrectly Accepted (IA) – an aviator's incorrect radio call was incorrectly accepted by the system (e.g., aviator vocalises cleared for take-off instead of line-up and wait, and it is accepted by the system).

- Incorrectly Rejected (IR) – an aviator's correct radio call was incorrectly rejected by the system (e.g., correct frequency vocalised, but the system recognised it incorrectly).

The examples provided are illustrative for explanation purposes and are not necessarily representative of common errors made by the SATCE system. A second SME independently rated a proportion (25%) of the radio calls, with interrater reliability then being calculated using the interobserver agreement method (Thomas & Nelson, 2001), this analysis revealed an acceptable level of agreement (92.9 %). Overall recognition accuracy was then calculated for all applicable experimental conditions using equation one below (Cucchiarini, Neri & Strik, 2009):

$$\text{Recognition Accuracy} = \left(\frac{\text{CA} + \text{CR}}{\text{CA} + \text{CR} + \text{IA} + \text{IR}}\right) \text{x } 100 \qquad (1)$$

Procedure

Participants first read an information pack containing details about the study, the flight tasks, and phraseology differences related to the SATCE system. This information was supported by an in-person follow-up brief which also covered information on the HOTAS controls, VR headset, and simulated aircraft. Participants then provided informed consent and completed a demographics questionnaire. To minimise the potential for cybersickness, and to ensure familiarity with the system, participants were provided with a familiarisation flight for approximately 10 minutes. This involved taking off, climbing to 2000ft and conducting general handling exercises. Upon completion, participants were familiarised with the psychological measures. Participants were then asked to complete the flight tasks in the experimental conditions detailed above, in a random order. Following each iteration of the task, participants removed the VR headset and completed the psychological measures. Upon completion of the study, participants were debriefed and thanked for their participation.

RESULTS

Workload

Circuits task

Workload data for the circuits task is shown in Figure 1, for all figures throughout this paper error bars represent standard error of the mean. Repeated measures ANOVAs (RM-ANOVA) revealed a significant effect of experimental condition on all components of workload; Mental ($F(2,34) = 29.6$, $p < .001$), Physical ($F(2,34) = 3.9$, $p = .029$), Temporal ($F(2,34) = 19.0$, $p < .001$), Performance (F(2,34) = 5.2, $p = .01$), Effort ($F(2,34) = 10.1$, $p < .001$), Frustration ($F(2,34) = 9.4$, $p < .001$) and Average workload scores ($F(2,34) = 23.1$, $p < .001$). Follow-up tests revealed that workload scores incrementally increased between all three conditions for mental workload only (p's < .02). For all other components of workload, except physical and performance, scores were significantly higher (p's < .05) between the ATC & Entities conditions when compared to the sterile and ATC alone conditions. Follow-up tests did not indicate where differences in physical workload occurred.

Practice area task

Workload data for the practice area task is shown in Figure 2. Paired-samples t-tests revealed that workload scores were significantly higher in the ATC & Entities condition than the Sterile condition, for all components of workload and overall average workload; Mental ($t(17) = 6.9$, $p < .001$), Physical ($t(17) = 3.7$, $p < .01$), Temporal ($t(17) = 5.4$, $p < .001$), Performance ($t(17) = 2.3$, $p < .037$), Effort ($t[17] = 3.9$, $p < .001$), Frustration ($t[17] = 3.4$, $p < .01$) and overall average workload ($t[17] = 5.51$, $p < .001$).

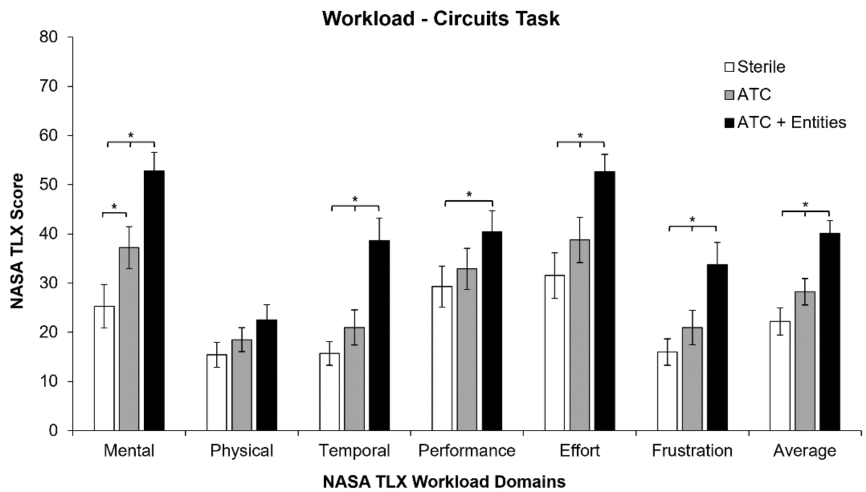

Figure 1. Average NASA TLX workload scores for the Circuits task.

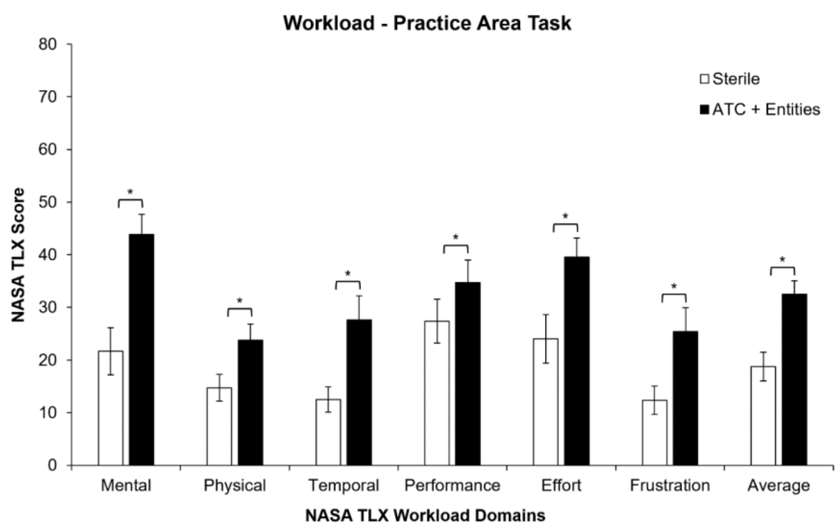

Figure 2. Average NASA TLX workload scores for the Practice Area task.

SART

Circuits task

SART data for the circuits task is shown in Figure 3. RM-ANOVAs on these data revealed a significant effect of experimental condition on both attentional supply ($F(2,34) = 11.7$, $p < .001$) and attentional demands ($F(2,34) = 98.1$, $p < .001$). For attentional supply, follow-up tests revealed a significant increase between the Sterile condition and both the ATC, and ATC & entities conditions. For attentional demands, follow-up tests revealed no difference between the Sterile and ATC conditions, but a significant difference between the Sterile and ATC & Entities conditions ($p < .001$).

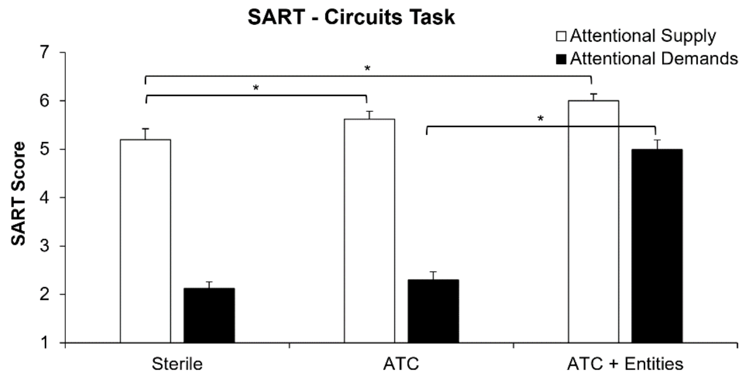

Figure 3. Average attentional supply and attentional demand from the SART in the Circuits Task.

Practice area task

The SART data for the practice area task is shown in Figure 4. Paired-samples t-tests revealed that both attentional supply ($t(17) = 5.3$, $p < .001$) and attentional demand ($t(17) = 8.7$, $p < .001$) were significantly higher in the ATC & Entities condition than the Sterile condition.

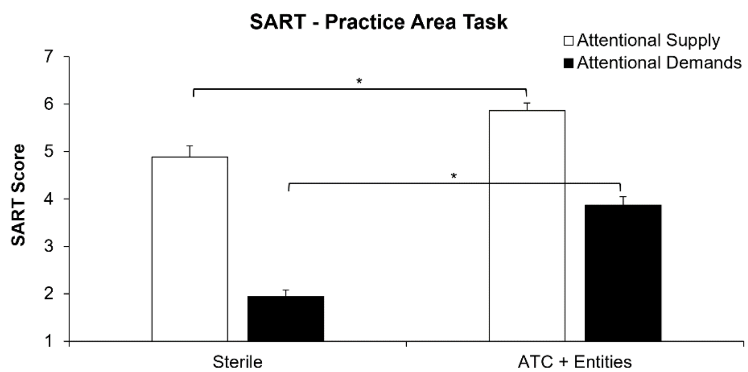

Figure 4. Average attentional supply and attentional demand from the SART in the Practice Area Task.

Flight Task Performance

Instructor ratings of flight task performance results are shown in Table 1. Experimental condition had a significant effect on instructor ratings of flight task performance for the circuits task ($F(2, 34) = 4.2$, $p = .02$). Follow up tests suggested that this was most likely due to a decrease in flight task performance for the ATC & Entities condition when compared to the Sterile condition. For the practice area task, there was no significant difference in flight task performance between the sterile condition and the ATC + entities condition ($t(17) = 1.3$, $p < .22$).

Presence

Results for the presence questions are shown in Table 1. For the circuits task, RM-ANOVAs showed a significant effect of condition on perceptions of real world similarity ($F(2, 34) = 4.1$, $p = .03$) and involvement ($F(2, 34) = 4.95$, $p = .01$). For the circuits task, post-hoc tests revealed a significant difference for perceptions of real-world similarity between the sterile and ATC conditions only. Similarly, for perceptions of involvement, differences approached significance between the sterile and ATC conditions only. For the practice area task, dependent samples t-tests revealed that perceptions of real world similarity ($t(17) = 2.9$, $p < .05$) and involvement ($t(17) = 3.6$, $p < .05$) were significantly higher in the ATC & Entities condition, than the Sterile condition.

Recognition Accuracy

Results for the overall recognition accuracy of radio calls by the SATCE system are shown in Table 2. For the circuits task, a dependent samples t-test revealed no significant difference in recognition accuracy between the ATC and ATC & entities conditions ($t(17) = -0.19$, $p = .84$).

Table 1. Average flight task performance and presence, in the circuits and practice area tasks

	Experimental Condition		
	Sterile Mean (*Std Dev*)	ATC Mean (*Std Dev*)	ATC + Entities Mean (*Std Dev*)
Circuits Task			
Flight Task Performance	4.80 (*0.20*)	4.77 (*0.29*)	4.55 (*0.46*)
Presence: Real world similarity	4.14 (*1.41*)	5.08 (*1.31*)	4.89 (*1.36*)
Presence: Involvement	5.03 (*1.36*)	5.67 (*0.79*)	5.78 (0.60)
Practice Area Task			
Flight Task Performance	4.78 (*0.32)*	-	4.67 (0.36)
Presence: Real world similarity	4.14 (*1.43*)	-	5.19 (*1.21*)
Presence: Involvement	4.83 (*1.32*)	-	5.75 (*0.72*)

Table 2. Average recognition accuracy of radio calls by the SATCE system in the circuits and practice area tasks

	Experimental Condition	
	ATC Mean (*Std Dev*)	ATC + Entities Mean (*Std Dev*)
Circuits Task	81.9% (*7.8*)	81.5% (*6.7*)
Practice Area Task	-	80% (*9.5*)

DISCUSSION

The aim of this study was to investigate the effects of a SATCE system on military aviators' workload and situational awareness. The study also aimed to conduct an initial examination of the recognition accuracy of the system in order to provide an impression of a SATCE system's utility in fostering communication skills. Across two flight tasks, the results converged on the conclusion that a SATCE system has the potential to impact military aviator's workload. The results from the circuits task illustrate this point. When compared to sterile simulator conditions, the combined integration of both air traffic control and other representative aircraft entities resulted in mental workload scores, and overall average workload scores, being 109% and 81% higher, respectively. The inclusion of both these components of the SATCE system seems to be crucial in producing workload increases across a variety of workload dimensions. Specifically, an incremental increase between the conditions was only found for mental workload. Almost all other workload domains, and overall average workload, required the addition of both components, not just ATC on its own, for workload to statistically increase. The instructor-rated task performance also lends support for the importance of both SATCE components, with performance only declining in these conditions. This suggests that workload had increased to the point that task execution was negatively impacted. These findings are in-line with previous workload studies of a similar nature (e.g., Morris & Leung, 2006; Raby & Wickens, 1994). For instance, in a simplified aviation task with trainee pilots, Morris and Leung (2006) found that across three levels of workload, increases in workload and auditory inputs led to both declines in performance, and prioritisation errors, where handling performance was deprioritised instead of more suitable aspects of the task (e.g., radio, mental arithmetic). Interestingly however, for the practice area task, workload increased but instructor-rated performance did not deteriorate. This was most likely due to workload only being high for specific phases of the practice area task (i.e., departure and arrival), whereas workload would have been comparatively high for the totality of the circuits task. Therefore, differences in performance may have been difficult to detect when averaging across all of the task. The findings related to situational awareness offer additional insights into these workload and performance findings.

The SART results indicated an incremental increase in attentional supply across the experimental conditions in the circuits task, suggesting that aircrew devoted more attention to the task as more SATCE system components were activated. However, it was only when both ATC and other aircraft were activated that attentional demands saw an increase. This was most likely due to this experimental condition requiring aircrew to continuously orient their attention to the location of other aircraft in the environment, both overtly through gaze behavior, and covertly through the maintenance of an internal representation of other aircrafts' current and future positions. The attentional demands dimension of the SART includes elements such as situation instability, variability, and complexity. This highlights a key aspect of the employed SATCE system. Specifically, the behaviors of the other aircraft injected by the SATCE system were not scripted. Instead, the aircraft were initiated into appropriate locations (e.g., holding short of the runway, about to join the circuit etc.) in a random manner upon activation, with appropriate aviation procedures, rules and guidance being followed to make their behaviors representative. More complex, variable, multi-ship interactions and situations then emerged naturally – a key hallmark of psychological fidelity (Kozlowski & DeShon, 2004), and a key factor in fostering situational awareness (Sarter & Woods, 1991). In contrast, it is likely that more scripted entity behaviour, or having more foreknowledge of their likely behaviour, would have resulted in less attentional demands. It is worth noting that the changes in attentional demands across experimental conditions were comparable to the overall workload results, therefore future research could statistically investigate whether changes to attentional processes served as a mediator behind the workload effects. An individual's sense of presence can also impact how they react psychologically and behaviourally to a simulated environment, therefore attention is now turned to these results.

Perceptions of presence increased between the sterile and ATC conditions for the circuits task, with no statistically significant change in presence being found when both ATC and other aircraft entities were activated. However, for the practice area task, increases in presence were found when both ATC and other entities were activated. The perceived realism of other aircraft entities' behaviour may offer an explanation for this pattern of results. Specifically, the rules and maneuver patterns of the injected aircraft entities were not fully adapted to the military and regional-specific procedures that the participants were accustomed to. This could have led to a discrepancy between actual and expected entity behaviour, and therefore led to reduced perceptions of presence. This may have been most noticeable during the circuits task where it is relatively more important to have continuous awareness of other aircrafts' behaviours. As an alternative explanation, it has previously been shown that perceptions of presence can plateau when workload increases past a certain magnitude (Sepich, Jasper, Fieffer, Gilbert, Dorneich & Kelly, 2022). This may have been the case in the present study, with the observed increases in workload in the ATC and other entities condition being at least partially reflective of a re-direction of cognitive resources away from processing environmental cues that contribute to a sense of presence (e.g., visual features of the terrain, aircraft cockpit details, weather etc.), and instead allocating them towards successful task completion. Voice recognition accuracy of the SATCE system is another factor that could have influenced presence. However, in the present study, recognition accuracy was approximately 81% in all conditions, with no differences between conditions. Therefore, it is argued that recognition accuracy was not responsible for the pattern of presence results. While investigating voice recognition accuracy was not the primary objective for the current study, it is still a key factor that is worthy of note and future investigation, especially when considering the likely importance of recognition accuracy in ensuring user acceptance in military aviation contexts. In the present study, the voice recognition aspects of the SATCE system were not fully trained or adapted to regional accents, or military-specific procedures or phraseology, this led to specific phrases or "problem words" (c.f., Barbato, 1998) being most responsible for recognition inaccuracies. Therefore, while overall voice recognition accuracy was likely below desired levels for actual system deployment, future work should ensure these adaptions are conducted in order for recognition accuracy to be evaluated most appropriately.

In conclusion, the present study showed that a SATCE system can both increase workload, and influence aspects of situational awareness, in military aviators. In particular, increased attentional demands seemed critical in achieving these effects. From a wider point of view, if the use of simulation continues to expand within military flying training, focusing on creating sufficiently unstable, variable, and complex simulated scenarios may support the development of such key airmanship constructs, regardless of the method employed to achieve these scenario properties. For example, Mishler et al., (2022) investigated the use of networked simulators, coupled with human air traffic controllers, to achieve similar aims. Future research should investigate the use of fully regionally adapted SATCE systems, in actual flying training pipelines, across a wider variety of tasks. Research should also aim to determine the optimal employment of a SATCE system within simulated flying training syllabi and sorties, to best support various stages of skill acquisition.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the RAF leaders, instructors and students for their support, input, and participation in the study reported here. We are also grateful to Advanced Science and Technologies inc (ASTi) for engaging in a Memorandum of Understanding which enabled us to use their SERA software to conduct this independent study of the SATCE concept. Finally, the authors would particularly like to thank Paul Dickson for his assistance in data collation, and Andrew Schneider for his insightful discussions related to analysing speech data in aviation contexts.

REFERENCES

Alessi, S. M. (1988). Fidelity in the design of instructional simulations. *Journal of Computer Based Instruction, 15*(2), 40–47.

Barbato, G. J. (1998). Integrating voice recognition and automatic target cueing to improve aircrew-system collaboration for air-to-ground attack. *Proceedings of the Research and Technology Organization Panel: Sensor Data Fusion and Integration of the Human Element*; RTO-MP-12. *System Concepts and Integration (SCI) Symposium*. Ottawa, Canada, pp. 24-1, 24-11. https://apps.dtic.mil/sti/pdfs/ADA360838.pdf#page=216

Bell, B., Billington, S., Bennett, W., Billington, I., & Ryder, J. (2010). Performance gains from speech-enhanced simulation in military flying training. *Journal of Defense Modelling and Simulation, 7*(2), 67–87. https://doi.org/10.1177/1548512910365698

Bürki-Cohen, J., Kendra, A., Kanki, B., & Lee, A. (2000). *Realistic Radio Communication in Pilot Simulator Training*. Final Report No. DOT/FAA/AR- 00/62, Washington, DC.

Civil Aviation Authority (2020). *CAP 413: Radiotelephony Manual* (23rd Ed). Gatwick, UK: Safety Regulation Group, Civil Aviation Authority. https://publicapps.caa.co.uk/docs/33/CAP413%20E23%20corr%208%20June%202020.pdf

Cucchiarini, C., Neri, A., & Strik, H. (2009). Oral proficiency training in Dutch L2: The contribution of ASR-based corrective feedback. *Speech Communication, 51*(10), 853–863. https://doi.org/10.1016/j.specom.2009.03.003

Endsley, M. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37*(1), 32–64. https://doi.org/10.1518/001872095779049543

Endsley, M. (2015). Situation Awareness Misconceptions and Misunderstandings. *Journal of Cognitive Engineering and Decision Making, 9*(1), 4–32. https://doi.org/10.1177/1555343415572631

Endsley, M., Sollenberger, R. L., Nakata, A., & Stein, E. S. (2000). *Situation awareness in air traffic control: Enhanced displays for advanced operations*. Atlantic City, NJ: U.S. Department of Transportation. https://hf.tc.faa.gov/publications/2000-situation-awareness-in-air-traffic-control/full_text.pdf

Estes, S. (2015). The Workload Curve: Subjective Mental Workload. *Human Factors, 57*(7), 1174-1187. https://doi.org/10.1177/0018720815592752

Federal Aviation Administration. (2021). Airplane flying handbook (FAA-H-8083-3C). Washington, DC: U.S. Department of Transportation. https://www.faa.gov/regulations_policies/handbooks_manuals/aviation/airplane_handbook

Foushee, H., & Manos K. (1981). Information transfer within the cockpit: problems in intracockpit communications. In: Billings CE, Cheaney ES (eds) *Information transfer problem in the aviation system*. NASA technical paper no. 1875. Moffett Field, CA: NASA/Ames Research Center, 1981; pp. 63–71.

Gross, D. C., Tucker, W. V., Harmon, S. Y., & Youngblood, S. M. (1999). Fidelity Theory Advances in Distributed Simulation. *Summer Computer Simulation Conference*, Society for Computer Simulation International, 415–420.

Hart, S., & Staveland, L. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. Hancock, & N. Meshkati (Eds), *Human mental workload* (pp.139–83). Amsterdam: North Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

Helmreich, R.L., Merritt, A.C., & Wilhelm, J.A. (1999). The evolution of crew resource management training in commercial aviation. *International Journal of Aviation Psychology, 9*, 19–32. https://doi.org/10.1207/s15327108ijap0901_2

Heycock, S. A., & Brown, M. F. (2000). The Subjective Objective Assessment of Airmanship. *Proceedings of the 2000 Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*, Orlando, Fl.,

Jacobs, J. W., Prince, C., Hays, R. T. & Salas, E. (1990). *A meta-analysis of the flight simulator training research. Orlando*, FL: Naval Training System Center. https://doi.org/10.21236/ada228733

Kern, T. (2010). *Foundations of Professional Airmanship and Flight Discipline*. Convergent Performance. Colorado Springs, Colorado. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=b618b39504ad90612bdb1f70e661b53159c2bfb1

Kim, H. (2018). What constitutes professional communication in aviation: Is language proficiency enough for testing purposes? *Language Testing, 35*(3), 403-426. https://doi.org/10.1177/0265532218758127

Kozlowski, S. W. J. & DeShon, R. P. (2004). A psychological fidelity approach to simulation-based training: Theory, research, and principles. In E. Salas, L. R. Elliott, S. G. Schflett, & M. D. Coovert (Eds.), *Scaled Worlds: Development, validation, and applications* (pp. 75-99). Burlington, VT: Ashgate Publishing

Liu, D., Macchiarella, N. D., & Vincenzi, D. A. (2009). Simulation fidelity. In D. A. Vincenzi, J. A. Wise, M. Mouloua, & P. A. Hancock (Eds.), *Human factors in simulation and training* (pp.61–73). Boca Raton, FL: CRC Press. http://dx.doi.org/10.1201/9781420072846.ch4

Mane, A. (1981). Airmanship: An introduction. *First symposium on aviation psychology, 1981. Proceedings of the symposium on aviation psychology*, 161-166. Columbus, Ohio. https://files.eric.ed.gov/fulltext/ED219236.pdf

McCoy-Fisher, C., Mishler, A., Bush, D., Severe-Valsaint, G., Natali, M., & Riner, B. (2019). Student Naval Aviation Extended Reality Device Capability Evaluation. Technical Report: Naval Air Warfare Center Training *Systems Division - NAWCTSD-TR-2019-001.* https://apps.dtic.mil/sti/pdfs/AD1103227.pdf

Mishler, A., Severe-Valsaint, G., Natali, M., Seech, T., McCoy-Fisher, C., Cooper, T. & Astwood, R. (2019). Project Avenger Training Effectiveness Evaluation. *Technical Report: Naval Air Warfare Center Training Systems Division - NAWCTSD Public Release 22-ORL006.* https://apps.dtic.mil/sti/pdfs/AD1162306.pdf

Moroney, W. F., Biers, D. W., Eggemeier, F. T., & Mitchell, J. A. (1992). A comparison of two scoring procedures with the NASA Task Load Index in a simulated flight task. *Aerospace and Electronics Conference, 1992. NAECON 1992, Proceedings of the IEEE 1992 National*, Vol. 2, 734-740. Dayton, OH: NAECO. https://doi.org/10.1109/NAECON.1992.220513

Morris, C. H., & Leung, Y. K. (2006). Pilot Mental Workload: How Well do Pilots really Perform? *Ergonomics, 49* (15), 1581-1596. https://doi.org/10.1016/j.adro.2020.09.012

Nergård, V. (2014). Airmanship - a qualitative approach. *Aviation, 18*(3), 147-156. http://dx.doi.org/10.3846/16487788.2014.969882

Nuamah, J.K., Mosaly, P.R., Adams, R., Adapa, K., Chera, B.S., Marks, L.M., Mazur, L.M. (2020). Assessment of Radiation Therapy Technologist's Workload and Situation Awareness: Monitoring Two vs. Three Collocated Display Monitors. *Advances in Radiation Oncology, 6*(1), 100572. https://doi.org/10.1016%2Fj.adro.2020.09.012

Orlansky, J., & String, J. (1977). *Cost-effectiveness of flight simulators for military training: Vol. 1. Use and effectiveness of flight simulators*. Alexandria, VA: Institute for Defense Analyses. https://apps.dtic.mil/sti/pdfs/ADA052801.pdf

O'Neil, H. F., Jr., & Andrews, D. H. (2000). *Aircrew training and assessment*. Mahwah, NJ: Erlbaum. https://doi.org/10.1177/106480460100900310

Papadopoli, N. (2017). ATC simulation for flight training: The missing link. *International Journal of Modeling, Simulation, and Scientific Computing, 8*(4), 1-9. http://dx.doi.org/10.1142/S1793962317430012

Raby, M., & Wickens, C. D. (1994). Strategic workload management and decision biases in aviation. *The International Journal of Aviation Psychology, 4*(3), 211-240. https://doi.org/10.1207/s15327108ijap0403_2

Royal Air Force Central Flying School (2014). An explanation of the CFS airmanship Model. Internal RAF report: Unpublished

Royal Air Force (2022). Royal Air Force Strategy. *Ministry of Defence*. https://www.raf.mod.uk/documents/word/raf-strategy/

Sarter, N. B., & Woods, D. D. (1991). Situation awareness: A critical but ill-defined phenomenon. *International Journal of Aviation Psychology, 1*, 45-57. https://doi.org/10.1207/s15327108ijap0101_4

Schreiber, B.T., Schroeder, M., & Bennett, W. (2011) Distributed Mission Operations Within-Simulator Training Effectiveness. *International Journal of Aviation Psychology. 21*, 254–268. https://doi.org/10.1080/10508414.2011.582448

Sepich, N. C., Jasper, A., Fieffer, S., Gilbert, S. B., Dorneich, M. C., & Kelly, J. W. (2022). The impact of task workload on cybersickness. *Frontiers in Virtual Reality, 3,* 943409. doi: 10.3389/frvir.2022.943409

Taylor, R. M. (1990). Situation awareness rating technique (SART): the development of a tool for aircrew systems design. In *Situational Awareness in Aerospace Operations* (Chapter 3). France: Neuilly sur-Seine, NATO-AGARD-CP-478

Thomas, J. R., & Nelson, J. K. (2001). *Research methods in physical activity*. Champaign: Human Kinetics.

Witmer, B. G., & Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence, 7*(3), 225–240. https://psycnet.apa.org/doi/10.1162/105474698565686

Yeh, Y.Y. & Wickens, C. (1988). Disassociation of performance and subjective measures of workload, *Human Factors*, *30* (1), 111-120. https://doi.org/10.1177/001872088803000110