Evaluating System Usability of Augmented Reality in Flight Operations

Wen-Chin Li[1], Tim Bord[2], Jingyi Zhang[3], Graham Braithwaite[1] and Mudassir Lone[1]

[1] Safety and Accident Investigation Centre, Cranfield University, UK, [2] l'Ecole de l'Air – 13661 Salon Air, France, [3] Flight Technology College, Civil Aviation University of China, China

ABSTRACT

The human-centred design of augmented visualisation aids can have significant effect on human performance and cognitive processes by increasing an operator's capability to manage complex checklists. This study investigated the use of an Augmented Reality (AR) device as a cockpit integration tool and the possible new challenges relating to Human-Computer interactions it induces. Seventeen aviation professionals (pilots, engineers, and training pilots) aged from 23 to 53 (M=29.82, SD=8.93) participated in this experiment. Their flight experience ranged from zero flight hours to 3000 flight hours (M=605.00, SD=1051.04). Two types of interaction - by gesture and voice control, have been compared with traditional paper checklists. The results show that gesture control AR gives rise to unnecessary complexity and tends to be cumbersome to use. On the other hand, voice control AR checklists could constitute an improvement in terms of usability of checklists completion in flight operations. Paper checklists tend to score higher in terms of 'learnability' as it is the simplest way to use a traditional checklist. It is also interesting to find that voice control AR checklists tend to be rated as the highest on both the total score of System Usability Scale (SUS) and in terms of 'usability'. These phenomena are consistent with the comments of participants that they would prefer to apply a voice control AR checklist over a paper checklist, if they were more familiar with it in the future. The improvement in modes of interaction and the presentation of information could lead to changes in usability and operational procedures. There is a need for further exploration of the implications of AR technology on the flight deck before implementation.

KEYWORDS

Augmented Reality, Aviation Safety, Flight Deck Design, System Usability Scale

Introduction

Augmented Reality (AR) is a tool which can be used to improve human-computer interaction in aviation (Luzik & Akmaldinova, 2006). The application of AR can facilitate pilots interacting with the interfaces in the flight deck to analyse various sources of information simultaneously. System developers take the importance of human-computer interaction into account when design new operational systems to optimise pilots' situational awareness and minimise workload (Dorneich, Rogers, Whitlow, & DeMers, 2016). Augmented Reality differs from Virtual Reality (VR) as AR uses overlaid images in the real-world environment, whereas VR is based on a digital environment where the user cannot see or interact with the real-world. AR in aviation has existed since WWII with the first Head-Up Display (HUD) displaying an aiming sight in fighter aircraft cockpits. The extensive use of HUDs in both civilian and military aviation can explain the benefits of allow pilots to access primary flight parameters while searching exterior dynamic targets, therefore facilitating collision avoidance and enhancing flight safety. Nevertheless, pilots still have to switch their

attention between the far domain of operational environments and the near domain of HUD (Prinzel III & Risser, 2004). AR headsets are the logical evolution of HUDs because they represent a physical manifestation of human-centred design of interface displays. AR is gaining momentum with the release of commercial grade AR headset and pilots can now interact with augmented visualisation cues to see information superimposed on the operational environment (Li, Zhang, Le Minh, Cao, & Wang, 2019). However, such innovative AR devices need to be validated before implementation in aviation.

The usability of advanced systems is an essential part of innovative technology. The acceptance of one system depends on whether the system can fulfil the requirement of usability in real-world operations. The application of human factors engineering to the design of the human-computer interaction has grown, focused mainly on objective usability, effectiveness and efficiency. The System Usability Scale (SUS) was developed for measuring the usability of a new system (Brooke, 1996 Lewis, 2018) and has been used for evaluation of mobile devices and their applications, 3D exploration games for older adults (Money et al., 2019); surface projection for military applications. Moreover, the different types of systems have proved that the SUS can be used very well in assessing the usability of prototype and providing a valuable reference for production systems (Baumgartner, Sonderegger, & Sauer, 2019; Boyce et al., 2019; Kortum & Sorber, 2015). SUS is not a questionnaire on usability *per se*, but rather on users' perception of usability. Participants who believe they are successfully operating a system tend to give higher SUS ratings, therefore perceived success strongly correlates with SUS ratings (Drew, Falcone, & Baccus, 2018). The SUS has ten questions, and the individual SUS scores range from zero to 100. Even if many studies have used an unidimensional SUS, Lewis and Sauro (2009) demonstrated that SUS could be divided as two dimensions, "Learnable" made of both items four and ten, and "Usable" made of the other eight items. These two dimensions provide additional data for the practitioner to analyse in addition to the overall SUS score. The aim of this study is to evaluate the usability of an AR device in flight operations for pre-landing checks. The usability of the two operational approaches of AR checklists (voice control vs gesture control) needs to be investigated and compared with the traditional paper checklist.

Method

Participants

Seventeen aviation professionals (pilots, engineers, and training pilots) aged from 23 to 53 (M=29.82, SD=8.93) participated in this experiment. Their flight experience ranged from zero flight hours to 3000 flight hours (M=605.00, SD=1051.04). The collected data are gathered from human subjects; therefore, the research proposal was submitted to the Cranfield University Research Ethics System for ethical approval (CURES/8477/20198). As stated in the consent form, participants have the right to terminate the experiment at any time and to withdraw their provided data at any moment even after the data collection.

Apparatus

The experiment was run on the Cranfield University Large Aircraft Flight Simulator with a representative model of the Boeing 747 simulator (Hanke, 1971). The simulator is used predominantly for human factor studies and other research programmes focusing on flight dynamics and handling qualities of current and future aircraft. It is comprised of a realistic mock-up of a cockpit of Boeing commercial aircraft with functioning flight controls, stick-shaker stall warning, over-speed alerts, primary flight and navigation displays, and landing gear lever to name a few. The simplified overhead panel is composed of light switches, engine fire emergency levers and engine ignition switches. Three projectors provide the collimated 180-degree horizontal and 40-degree

vertical field of view, which together with the audio cues and multifunctional displays provide subjects with an immersive and realistic simulation environment (Figure 1).

Figure 1: Flight simulator for developing AR apps

Augmented Reality Device: The AR device used in the experiment is a Microsoft HoloLens headset (Figure 2). These glasses comprise see-through holographic waveguides, two HD 16:9 light engines and built-in processors that can display holograms with a resolution of 1280 x 720 px per eye, a field of view of 30° x 17.5° and a refresh rate of 60 Hz. Brightness and audio volume can be adjusted by 4 buttons located on top of the headset. The HoloLens comes with built-in sensors: an Inertial Measurement Unit (IMU), four environment understanding cameras, one depth camera, one 2MP photo/HD video camera, four microphones and one ambient light sensor. Its audio output consists in two speakers located near the user's ears that can emit spatial sound. The depth camera is used to carry out user's hand gesture recognition and spatial mapping of the surrounding environment. The user can extend or retract the headband and can slide the visor forward or backward in order to wear the headset more comfortably.

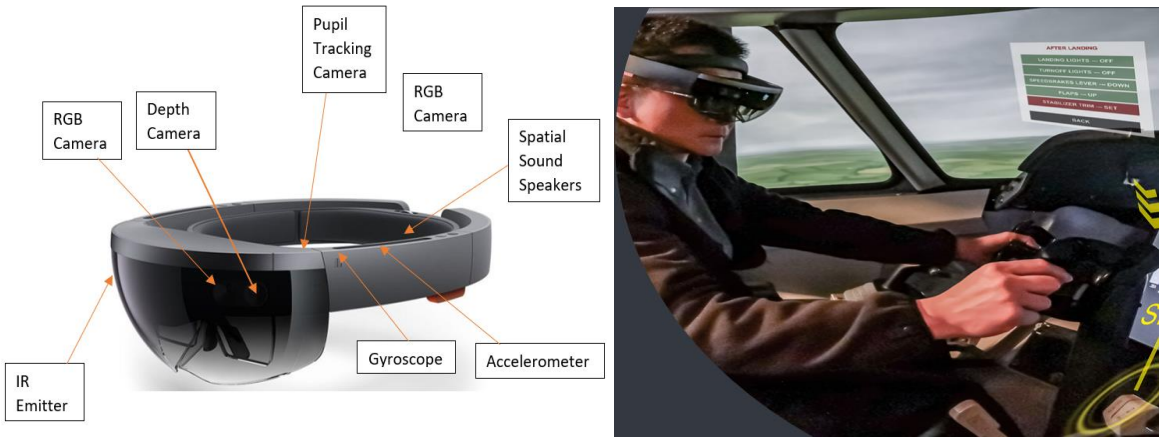

Figure 2: The AR device of Hololens (left) and participant wearing Hololens in the flight deck (right) performing pre-landing checklist

Development of AR Applications in the Flight Deck

There are certain augmented visual cues located on the flight deck using the spatial mapping technology of HoloLens. The B747 flight deck has first been mapped via the depth camera to obtain a 3D model of the scene (Figure 3). The AR checklist needs to be calibrated at launch in order to correctly position the highlights according to the cockpit. Calibration is achieved by the Vuforia Engine: the user scans a QR code located near the throttle levers to begin using the app.

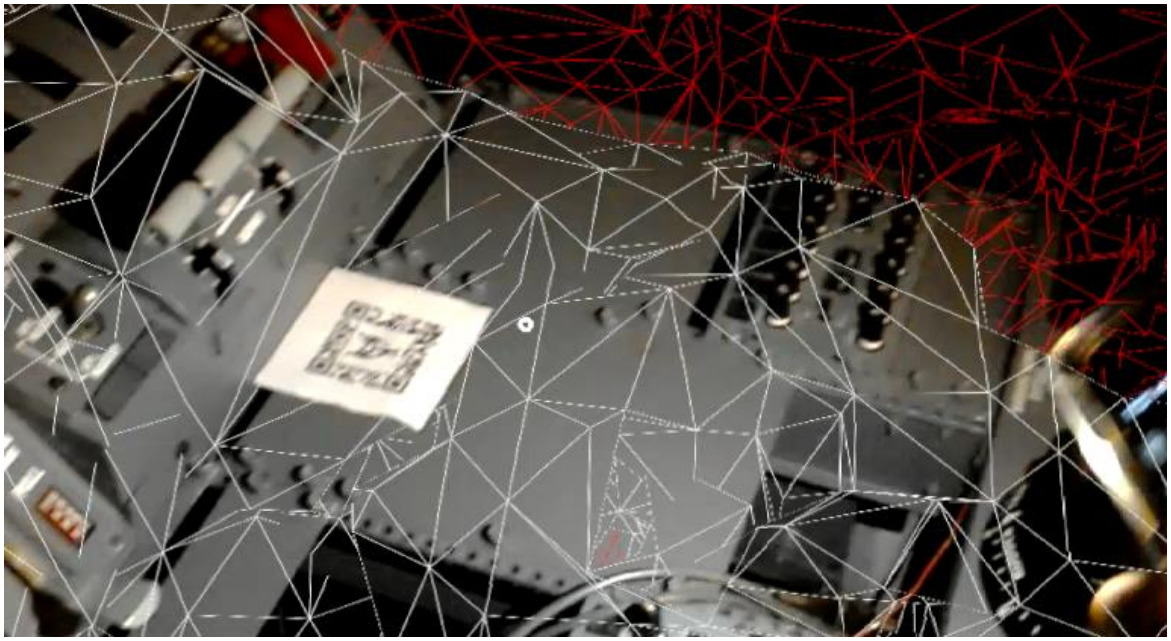

Figure 3: Flight deck mapping via the AR depth camera to create 3D model with a positioned marker (in the centre) highlight in the Unity editor

Scenario

The scenario is based on an Instrumented Landing System (ILS) in the final approach. The aircraft is set at 2000 ft and eight nautical miles (NM) from the airfield. As soon as the simulation starts, participants must execute a pre-landing checklist by interacting with the AR device and flying the aircraft for landing. The AR checklist application has been developed with Unity game engine and Microsoft Mixed Reality Toolkit (MRTK). All of its custom scripts have been written in C# language. It has been especially designed to be used with Cranfield University's B747 simulator.

Research Design

All participants undertook the following: (1) completed the demographical data including age, gender, qualifications, type hours and total flight hours (five minutes); (2) completed a briefing regarding the purpose of the study and how to use Hololens AR device (15 minutes); (3) sat in B747 simulator to practice how to use flight control to land the aircraft using a checklist (ten minutes); (4) completed a briefing on the AR checklist app, with a detailed explanation of the item highlights by voice control and gesture control (ten minutes); (5) performed a landing by using Hololens AR device on both voice control (five minutes) and gesture control (five minutes); (6) completed an evaluation of system usability of AR application on both voice control and gesture control compared to traditional paper checklist by SUS (ten minutes). It took around 60 minutes for each participant to complete the experiment.

Results and Discussions

There are 17 participants conducting three modes of flight operations, traditional paper checklist, gesture control AR checklist, and voice control AR checklist. One-way ANOVA was applied for data analysis. Bonferroni tests were performed to identify pairwise differences for factors with more than two levels. Partial eta-square (ηp2) is a measure of effect size for ANOVA. The descriptive results of SUS, Learnable and Usable scores on three checklist modes are shown as Table 1.

Table 1: The means and standard deviations of SUS scores on three modes of checklist

SUS dimension	Checklist mode	N	M	SD
Total	Traditional paper checklist	17	67.50	20.63
	Gesture controlled AR	17	41.47	16.06
	Voice controlled AR	17	72.65	12.97
Learnable	Traditional paper checklist	17	84.56	25.97
	Gesture controlled AR	17	52.21	25.09
	Voice controlled AR	17	61.77	22.74
Usable	Traditional paper checklist	17	53.68	8.43
	Gesture controlled AR	17	43.01	14.85
	Voice controlled AR	17	68.93	10.68

There is a significant difference of SUS scores on three modes of flight operations, $F(2, 48) = 16.72$, $p < .001$, $\eta p2 = 0.41$. Post-hoc comparison indicates that SUS total score on AR gesture checklist ($M = 41.47$, $SD = 16.06$) is lower than on traditional paper checklist ($M = 67.50$, $SD = 20.63$) and AR voice checklist ($M = 72.65$, $SD = 12.97$). Furthermore, there is a significant difference of Learnability scores on three modes of flight operations, $F(2, 48) = 7.74$, $p < .01$, $\eta p2 = 0.24$. Post-hoc comparison shows that Learnability score of traditional paper checklist ($M = 84.56$, $SD = 25.97$) is significantly higher than AR gesture checklist ($M = 52.21$, $SD = 25.09$) and AR voice checklist ($M = 61.77$, $SD = 22.74$). There is a significant difference of Usability scores on three modes of flight operations, $F(2, 48) = 21.34$ $p < .001$, $\eta p2 = 0.47$. Post-hoc comparison shows that Usability score AR on gesture checklist ($M = 43.01$, $SD = 14.85$) is significant lower than traditional paper checklist ($M = 53.68$, $SD = 8.43$) and AR voice checklist ($M = 68.93$, $SD = 10.68$); traditional paper checklist also significant lower than AR voice checklist. Generally, AR gesture checklist demonstrated the poorest usability among three modes and AR voice checklist represented the best usability among three modes of flight operations (Figure 4).

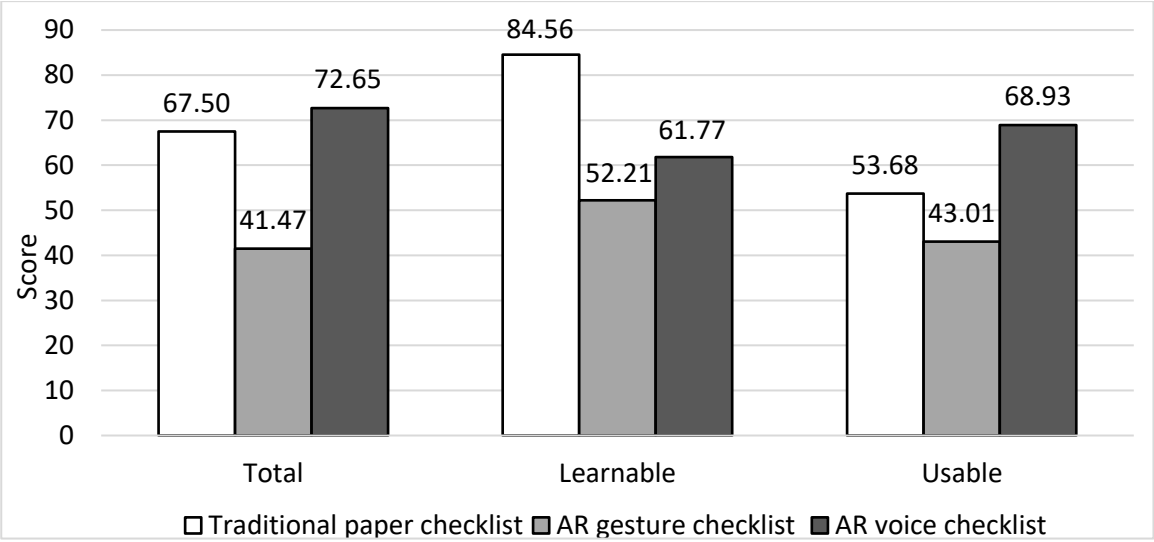

Figure 4: SUS scores of three modes of checklist

Participants interacted with AR technology in the flight deck experiencing certain levels of HCI challenges. The perceived usability of the AR gesture checklist is significantly lower than AR voice checklist on both total SUS score and Usable score. The lowest "Learnable" score of the AR gesture checklist can be explained by the fact that the understanding and the assimilation of the Air Tap gesture was very capricious among participants and limited the learnability of this approach. Participants were struggling to press fingers down to tap or click by gesture to activate AR checklist navigation (Figure 5). On the other hand, human operators tend to get higher marks if he/she is more familiar with the objects. This is proved by current research which demonstrated the traditional paper checklists score the highest mark on the Learnability score among three modes. This was particularly the case for the senior pilots who revealed that they prefer traditional checklist and disliked the AR gesture checklist. Furthermore, the low "Usability" score can be explained by the high physical demand related to moving fingers in front of camera on the Hololens – something participants found challenging and frustrating. The difficulties of HCI in the flight deck resulted in both low performance and lower perceived usability (Kortum & Peres, 2014).

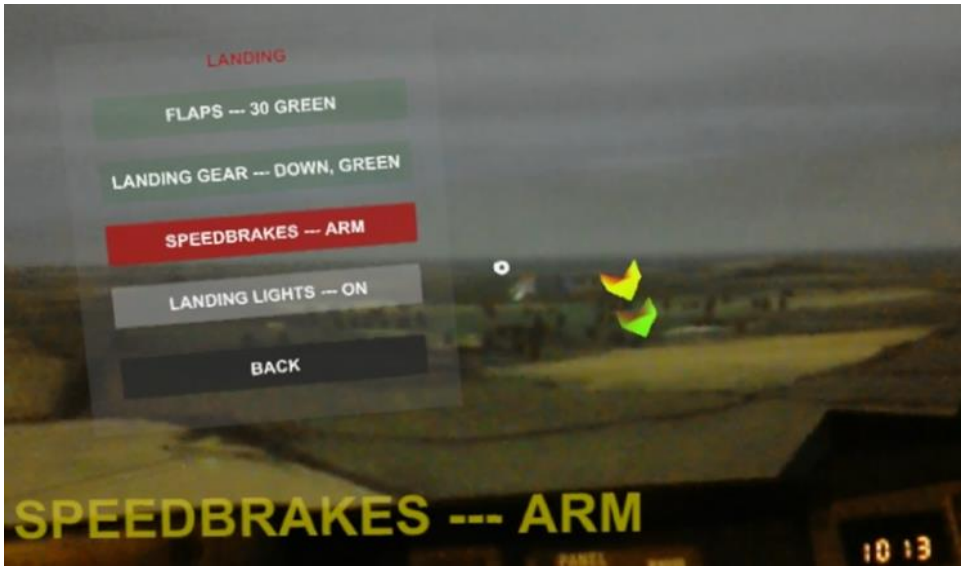

Figure 5. The pre-landing checklist embedded in AR device controlled by both gesture and voice in the flight deck

Paper checklist tend to score higher in "Learnability" as it uses a simple medium that participants do not need to be briefed on. On the contrary, voice control checklists tend to be rated as the highest in terms of "Usability". This phenomenon aligns with the comments of participants who would like to apply voice control AR checklist over paper checklist if they could have more practice and get familiar with it. The findings of this research are consistent with McLellan, Muddimer, and Peres (2012) who highlighted that experienced consumers tend to grant higher SUS scores than new consumers (up to 15% difference). Admittedly, the effect of practise for the voice control approach should be investigated. Besides we informally observed that experienced pilots perform a checklist significantly quicker, thus with higher perceived usability. During the experiment a potential lack of compliance was witnessed as some participants appears to become frustrated with the physical effort and mental demand required to operate both AR device by Air Tap and the landing gear in particular. Although these instances did not result in non-compliance during the study it is likely that some may become non-compliant in a single pilot operation scenario (Stanton, Plant, Roberts, & Allison, 2019).

Conclusion

The compliance of checklist and procedures are of great interest for Human Factors research as they may be accountable for a significant number of aviation accidents. The human-centred design of augmented visualisation aids have significant effects on human performance and cognitive processes by increased operator's capability to manage complex checklists. This study was aimed at investigating the use of AR device as a cockpit integration tool and the possible new challenges related to human-computer interactions that it induces. A checklist application has been developed on a Microsoft HoloLens headset in flight operations. There are two types of interaction by gesture and voice controlled have been compare to traditional paper checklists. The results show that gesture control AR gives rise to unnecessary complexity tends to be cumbersome to use. On the other hand, voice control AR checklists could constitute a real improvement in terms of usability of checklists completion in flight operations. Some considerations on the hardware used for this study need to be highlighted. The AR checklist application has been relying on the use of the default HoloLens interactions (i.e. cursor movement with head movements, Air Tap gesture, Microsoft voice recognition system). The current technological features embedded in the headset do not allow a reasonable and safe use in the cockpit. However, improvement in the types of interaction and displays could lead to changes in usability and operational procedures. There is a need for further exploration the applications of AR technology in the flight deck before implementation.

References

Baumgartner, J., Sonderegger, A., & Sauer, J. (2019). No need to read: Developing a pictorial single-item scale for measuring perceived usability. *International Journal of Human-Computer Studies, 122*, 78-89. doi:10.1016/j.ijhcs.2018.08.008

Boyce, M. W., Rowan, C. P., Shorter, P. L., Moss, J. D., Amburn, C. R., Garneau, C. J., & Sottilare, R. A. (2019). The impact of surface projection on military tactics comprehension. *Military Psychology, 31*(1), 45-59. doi:10.1080/08995605.2018.1529487

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability Evaluation in Industry, 189*(194), 4-7.

Dorneich, M. C., Rogers, W., Whitlow, S. D., & DeMers, R. (2016). Human performance risks and benefits of adaptive systems on the flight deck. *The International Journal of Aviation Psychology, 26*(1-2), 15-35. doi:10.1080/10508414.2016.1226834

Drew, M. R., Falcone, B., & Baccus, W. L. (2018). What does the system usability scale (SUS) measure? In *International Conference of Design, User Experience, and Usability* (pp. 356-366). Springer, Cham. doi: 10.1007/978-3-319-91797-9_25

Hanke, C. R. (1971). The Simulation of a Large Jet Transport Aircraft. Volume 1: Mathematical Model (NASA CR-1756). National Aeronautics and Space Administration, Washington, DC. https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19710009764.pdf

Kortum, P., & Peres, S. C. (2014). The relationship between system effectiveness and subjective usability scores using the System Usability Scale. *International Journal of Human-Computer Interaction, 30*(7), 575-584. doi:10.1080/10447318.2014.904177

Kortum, P., & Sorber, M. (2015). Measuring the usability of mobile applications for phones and tablets. *International Journal of Human-Computer Interaction, 31*(8), 518-529. doi:10.1080/10447318.2015.1064658

Lewis, J. R. (2018). The system usability scale: past, present, and future. *International Journal of Human–Computer Interaction, 34*(7), 577-590. doi:10.1080/10447318.2018.1455307

Lewis, J. R., & Sauro, J. (2009). The factor structure of the system usability scale. In *International conference on human centered design* (pp. 94-103). Springer, Berlin, Heidelberg. doi: 10.1007/978-3-642-02806-9_12

Li, W.-C., Zhang, J., Le Minh, T., Cao, J., & Wang, L. (2019). Visual scan patterns reflect to human-computer interactions on processing different types of messages in the flight deck. *International Journal of Industrial Ergonomics, 72*, 54-60. doi:10.1016/j.ergon.2019.04.003

Luzik, E., & Akmaldinova, A. (2006). Psychological aspects of ensuring flight safety in civil aviation. *Aviation, 10*(1), 25-35. doi:10.1080/16487788.2006.9635924

McLellan, S., Muddimer, A., & Peres, S. C. (2012). The effect of experience on System Usability Scale ratings. *Journal of Usability Studies, 7*(2), 56-67.

Money, A. G., Atwal, A., Boyce, E., Gaber, S., Windeatt, S., & Alexandrou, K. (2019). Falls Sensei: a serious 3D exploration game to enable the detection of extrinsic home fall hazards for older adults. *BMC Medical Informatics and Decision Making, 19*(1), 85. doi:10.1186/s12911-019-0808-x

Prinzel III, L., & Risser, M. (2004). *Head-up displays and attention capture (NASA/TM-2004-213000).* Retrieved from Langley Research Center, Hampton, VA: https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20040065771.pdf

Stanton, N. A., Plant, K. L., Roberts, A. P., & Allison, C. K. (2019). Use of Highways in the Sky and a virtual pad for landing Head Up Display symbology to enable improved helicopter pilots situation awareness and workload in degraded visual conditions. *Ergonomics, 62*(2), 255-267. doi:10.1080/00140139.2017.1414301