Augmented Reality as an Aid and a Distractor:
Pilot Behavior in Normal and Abnormal Flight Scenarios

by

Weiss OConnor

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Auburn, Alabama
August 08, 2026

Approved by

Dr. Missie Smith, Chair, Assistant Professor of Industrial and Systems Engineering
Dr. Nicoletta Fala, Assistant Professor of Aerospace Engineering
Dr. Mark Schall, Professor of Industrial and Systems Engineering
Dr. Ike Stutts, Department Chair of Cyber Science, United States Naval Academy

Abstract

In the information-dense and time-compressed cockpit environment, augmented reality (AR) displays are a promising approach to provide pilots with supplementary information. This research investigated head-worn, optical see-through AR systems, which project digital images while providing a clear view of the real world. While the benefits of having helpful information readily accessible are clear, AR displays bring several unique concerns from a human factors perspective. This research pursues a more complete understanding of both favorable and adverse impacts of AR displays on pilot performance through two user studies. The first study investigated the effects of AR display occlusion of the real world on pilot performance, assessing the impact of AR graphic presence and location. The second study investigated the influence of AR on pilot objective flight performance, workload, situational awareness, and usage when serving as either an aid or a distractor for an aircraft landing task.

The findings from both studies reflect the feasibility for AR display usage in the cockpit, as well as the importance of understanding the potential for adverse performance, trust, or cognitive impacts from its use. This work contributes a novel investigation into the drawbacks of AR displays in the aviation domain, utilizing a multi-modal approach to understand both flight performance and cognitive impacts of AR systems. Further research is needed to understand the complete picture of AR use in the cockpit and ensure that AR displays support pilot safety, even in the case of malfunction or failure.

Artificial Intelligence (AI) Use Disclosure Statement

In the preparation of this thesis, the following Artificial Intelligence (AI) tools were used: Claude (Sonnet 4.5/4.6/4.7, Opus 4.6/4.7/4.8). These tools were used exclusively to develop code for data cleaning, processing, and visualization in R and Python. No AI tools were used for any written component of the work. The author acknowledges full responsibility for the intellectual content of this work and has ensured that all AI-assisted material has been reviewed and revised for accuracy and appropriate academic style. All AI-generated content was reviewed and validated for relevance, appropriateness, and accuracy before incorporation into the final document to maintain scholarly integrity of this research.

Digital Accessibility Disclosure Statement

In the preparation of this thesis, the following digital accessibility tools were used to ensure this document complies with federal requirements: Microsoft Word Accessibility Assistant, Adobe Acrobat Accessibility Checker. The author acknowledges full responsibility for the intellectual content of this work and has made a good faith effort to comply with digital accessibility requirements in publishing, wherein the nature of the content does not significantly change in order to do so. Furthermore, all content has been reviewed and revised to meet these requirements prior to final publication.

Table of Contents

List of Tables

List of Figures

1 Introduction

To fly an aircraft safely, pilots must process continuous streams of visual information to accomplish tasks, make decisions, and execute timely and correct manual inputs. In the modern cockpit, pilots have access to more information and assistive technology than ever before, allowing for improved awareness about the state of the aircraft and the environment. However, the cockpit system is limited by the physical space available, the large quantity of information sources and supplemental materials, and human cognitive limitations. To this end, augmented reality (AR) displays represent a promising format for consolidating and delivering additional relevant information to pilots in the time-compressed, safety-critical domain of flying. AR displays enhance ease of access and awareness of the real world as information is projected directly over a pilot's existing field of view.

As AR displays become increasingly capable and affordable, they can supplement operators in a wide range of applications. In aviation, parallel advances in aircraft automation and data synthesis have made integrated AR displays feasible. AR displays project information directly over a user's visual field, while still allowing a simultaneous view of the real world (Grabowski et al., 2018). As a result, users may expend less cognitive effort to access information from an AR device, as the display can always be available in the visual field. The utility of an AR display results from the synergistic effect of providing operators with relevant information and low barriers in accessing that information, while still supporting awareness of the world around them.

1.1 Contributions

While AR displays may have potential to help pilots, gaps exist in understanding how they fit within the demands of abnormal in-flight situations. An *abnormal* or *emergency* situation can occur from mechanical factors or environmental factors, requiring deviation from planned routes or procedures (*Federal Aviation Administration*, 2021; *SKYbrary*, 2026). While automation may improve precision for routine procedures (Causse et al., 2025), human pilots are expected to adapt to abnormal or emergency scenarios to maintain safety. Pilots can adaptively and creatively problem-solve in unanticipated situations, beyond the pre-programmed limits of automation (Anand, 2025). It is vital to understand how AR technologies might affect this critical competency of human pilots before they are implemented in the cockpit at scale.

In this thesis, we describe two studies which investigate the influence of AR displays in the cockpit environment. Study 1 evaluates the impact of AR display location on pilot performance and informs the design of the AR display implementation used in subsequent research. Study 2 examines how inaccurate information presented by an AR display affects pilot workload, situational awareness, and performance. We implemented a notional assistive AR display interface that presents simple cues, in sequence, to assist with checklists, navigation, and task management. Specifically, we explore the potential for detrimental effects on pilot capability due to incorrect cues from an AR display during a simulated in-flight emergency. In both cases, we place emphasis on unanticipated effects of AR technology within the task-specific context of flying. A more complete understanding of AR displays may guide future employment of the technology in the safety-critical aviation domain.

2 Problem Statement

Although AR displays have potential for improving workflow and information access, responsible implementation in the cockpit also requires an understanding of potential disadvantages. First, it is helpful to define a possible AR assistance system and environment. For our research goals, a head-worn AR display in the cockpit may 1) consolidate routine supplementary information, 2) provide cues for task accomplishment, and 3) adapt information to the phase of flight. These elements are derived from characteristics of current head-up display implementations, while reflecting the core functions of aviation displays and strengths of AR displays (Biberman & Alluisi, 1992; Wickens, 2002; Willett et al., 2017). An AR display system could display prompts for checklists, tasks, or flight maneuvers based on established procedures for a standard flight. By integrating tasks and information, such a display may enhance operator performance and improve situational awareness. This notional display implementation reflects other AR display systems in research from the nautical domain (Rowen et al., 2019) and at higher levels of integration with cockpit systems (Li et al., 2022) in terms of core functions.

A preliminary concern for a cockpit-based AR display is that its presence might adversely affect pilots in information acquisition and processing. A key advantage of AR, as we will define it, is that users can access digital information while maintaining awareness of the real world (Skarbez et al., 2021). Nevertheless, in the information-dense cockpit environment (Wickens, 2007), new AR graphical elements may add visual complexity. Projections from an AR display might block, or *occlude*, real-world elements behind it, depending on the placement within the visual field. In the upper half of the visual field, an AR display might disrupt access to the horizon or navigational landmarks. In the lower half, the AR display is likely to occlude flight instrumentation or controls. To this end, we seek to understand the effect of AR display placement on pilot performance. This line of effort will additionally assess whether the mere presence of an AR display might be a detrimental influence, a relevant check considering the pace of AR display research in the aviation domain.

Beyond the concern of visual occlusion, the information provided through an AR display might influence pilot behavior and performance. If a flight situation becomes abnormal, assistive technology may lack the context available to the pilot to evaluate the safest course of action. While humans can quickly perceive and assess multisensory cues (Shams & Kim, 2010), even intelligent automation systems are limited by sensor capability and data availability (Skraaning & Jamieson, 2024). If the capabilities of the automated system are insufficient for the abnormal flight situation, then the assistive cues may no longer align with the pilot's goals and experience. This misalignment has the potential to introduce *cognitive friction*, or additional effort required from users to rectify a difference between their expectations and the system's capabilities, causing the automation to transition quickly from reducing workload to increasing it (Cooper, 1999). After a failure or error that constitutes an in-flight emergency, even a small mismatch in assessment between pilot and assistive system introduces resistance into the pilot's workflow. Pilots must transition from *following* guidance provided by the system to *weighing the judgement* of an autonomous agent against their own, requiring additional effort in evaluating multiple information sources (Walliser et al., 2019). We are interested in the demand on pilots due to mismatched cognitive efforts with automation.

While AR displays represent a promising medium for providing supplementary information to pilots, the potential for negative externalities with their use is not yet understood. Existing research emphasizes the potential for AR displays to assist operators with highly salient cues, as well as the prevalence of automation failure in complex systems. This research aims to expand our understanding at the intersection of these research areas. Though AR-based cues may be helpful in routine flight operations, it is vital to understand the degree to which a failure of this system in an abnormal flight could impact pilot performance and safety. An assessment of the capacity of pilots to maintain a high level of performance amidst unreliable AR cues can help to inform procedures, training, or regulations surrounding the technology. The fundamental research objective is to ***evaluate the influence of augmented reality displays on pilot behavior and performance in dynamic flight environments.***

To address this objective and guide research efforts, we identify several research questions.

- RQ1: How does the presence of an AR graphic in the visual field affect pilot performance? *(Study 1)*
- RQ2: What are the impacts on performance, workload, and situational awareness from a functional AR display system? *(Study 2)*
- RQ3: Conversely, what are the impacts on performance, workload, and situational awareness from an AR display system malfunctioning in an emergency situation? *(Study 2)*

3 Background

3.1 Cockpit Environment

The modern aircraft cockpit represents a complex, dynamic human-machine system. To fly an aircraft, the pilot relies on flight instruments to assess the state of the aircraft and the flight controls to apply inputs (Federal Aviation Administration, 2023). Other equipment, like the radio, navigation aids, or engine controls, allow the pilot to manage important subsystems required for flight. An aircraft may be flown to accomplish any range of tasks or missions, but core functions remain the same. Overall, the cockpit system must allow for controlled flight in the daytime or nighttime, as well as during favorable or unfavorable weather (Federal Aviation Administration, 2023).

Both the intricacy of the system and consequences of failure distinguish the cockpit system from a typical operating domain. In the aviation domain, failures are generally rare, but costly. The cockpit may be considered a safety-critical system, as mishaps may risk human lives and have significant environmental impacts (Grabowski et al., 2018). Pilots are responsible for maintaining safety and must prevent abnormal situations from progressing into true emergencies (Federal Aviation Administration, 2021). Flying carries inherent danger with very real possibilities of equipment failure or environmental hazards. Although the accident rate continues to decline, there remain more than 1,000 accidents per year with general aviation (GA) aircraft (General Aviation Accident Dashboard, 2021). GA, defined as all civilian flying except for airline service, accounts for more than 80% of pilots in the United States (Davis, 2019). Of the GA accidents, approximately 200 per year are fatal, with the majority of mishaps attributed to human task performance failures (General Aviation Accident Dashboard, 2021). The integral components, including the core pilot tasks, displays, and checklists, must integrate harmoniously due to the high stakes of failure.

3.1.1 Pilot Tasks

In the cockpit, the pilot is ultimately responsible for the safe and effective operation of the aircraft. At all times, the pilot must aviate (manually control the plane), navigate (monitor position and course), communicate (coordinate with ATC), and manage (monitor aircraft systems) (Liu et al., 2016). Each of the four core tasks require constant feedback and adjustment, and must be performed concurrently to maintain flight safety (Loukopoulos et al., 2003). This complexity characterizes the role of the pilot in the cockpit system.

Manual control of an aircraft is a complex and dynamic task. Pilots must control the high-level parameters of flight, like altitude and flight path, through inputs of the controls that affect lower-order elements, like turn rate or climb rate (Ebbatson et al., 2011). The task of aviating can be understood as a four-dimensional (three physical dimensions plus time) motor coordination task (Landry, 2012). The control inputs are continuous and interconnected, requiring pilots to respond to changes in the environment and aircraft state to maintain stable flight. The temporal element of manual control requires inputs not only be correct, but appropriately timed. Beyond basic manual control, the task of flying an aircraft is characterized by the density of tasks and information that must be managed in an ever-changing environment (Wickens, 2007). Pilots are expected to maintain primary control of the aircraft while managing interconnected systems, communication, and tasks on the various displays within the aircraft (Schutte & Trujillo, 1996). The activity of flying is primarily visual, as pilots must perceive and process information from instruments and the outside world to maintain aircraft control and navigate (Wickens et al., 2003). In many cases, simply looking outside the window can provide spatial cues sufficient for orientation.

3.1.2 Aircraft Displays

To fly an aircraft, pilots rely on several sources of information beyond what they can see outside the window. Displays within the aircraft integrate flight data, navigational aids, and aircraft status information, characterizing the complexity of the cockpit. Increases in display quantity over time have further contributed to concerns about the demands on pilot processing (Carroll et al., 2019). The form of display is a vital consideration, as appropriate presentation of vital information can prevent overload for pilots (Devean & Kewley, 2009). Research into aviation displays intends to ensure that the presentation of flight information supports pilot effectiveness and safety (Wickens, 2002). This involves supporting pilots in information acquisition and attention allocation.

Typical head-down displays (HDDs), characteristic of traditional “legacy” cockpit designs, are limited in several respects. HDDs take the form of several large, color screens, showing flight data, radio frequency selection, and navigation cues, which replace a multitude of small gauges in older models. The space allocated for displays in the cockpit is limited at the time of the aircraft’s physical design and manufacture. In addition, HDDs are constrained in their design by requirements for consistency and legibility, among other regulatory standards (*AC 25-11B - Electronic Flight Displays*, 2014), limiting their ability to adapt to novel assistive technologies. Head-up displays (HUDs), projected displays on clear vehicle-mounted screens, allow for information to be graphically overlaid on the real world, but remain limited by space and their mounting location within the cockpit.

3.1.3 Checklists

With the demands inherent in safely operating an aircraft, reliance on checklists (Ross, 2004) and standard procedures (Degani & Wiener, 1997) are essential to flight safety. This is further compounded by the range of actions available to pilots, given the increasing complexity of the cockpit environment. Checklists codify the exact sequence of tasks that must be completed at various points of a flight, while standard procedures ensure consistency and predictability when operating in the vicinity of other aircraft. Checklists are commonly printed and bound, or uploaded to tablet devices as digital documents. Past failures to use a checklist have contributed to aircraft accidents (Degani & Wiener, 1991). Specifically, checklist-related errors have contributed substantially to the commission of human error in aircraft accidents (Dye, 1995). The top three errors, “(a) Crew failed to use the checklist. (b) Crew overlooked item(s) on the checklist. (c) Crew failed to verify settings visually,” all reflect neglectful use of the checklist (Dye, 1995). Unfortunately, many characteristics of the modern cockpit environment may exacerbate reasons for non-compliance with checklists, to include the limitations of paper checklists, space/time saving measures (Dismukes & Berman, 2010), distraction, and frustration (Degani & Wiener, 1991).

3.2 AR Displays

AR is an emerging technology across many domains, with no universal criteria to define it. Milgram and Kashino's seminal work defined AR as the merging of real and virtual environments (Milgram & Kishino, 1994). AR can be understood as a subset of *mixed reality*, an environment in which "real world and virtual world objects and stimuli are presented together within a single percept" (Skarbez et al., 2021). In other words, AR displays enable users to simultaneously access elements from the display and the real world. AR environments are characterized by virtual elements, like text, 3D objects, or navigational cues, superimposed on the real world (Azuma, 1995). To understand how AR can be applied to the cockpit environment, we consider historical applications and factors concerning the design of AR systems.

3.2.1 Use and Development

As technology has progressed, AR technology has become increasingly capable. By retaining a view of the real world, AR displays can provide visual, spatial, or written information without fully separating the user from their surroundings. Though this research considers only head-worn AR displays, current devices of this type have relied on the innovation of two previous technologies. The first applications of AR in aviation came with the rise of modern HUDs in the late 1960s and helmet-mounted displays (HMDs) in the 1980s (Biberman & Alluisi, 1992). HUDs are a well-established example of AR in both aviation and automotive contexts (Crawford & Neal, 2006). HMDs allow for improved situational awareness, as pilots retain access to relevant flight or weapons information no matter the direction they are looking (Knabl et al., 2014). These devices, primarily utilized in military aviation, are distinct in the specificity of their intended use, as they primarily support weapon employment for a given aircraft platform.

In contrast, head-worn AR displays refer to a broader category of lightweight, general, and accessible wearable systems. These devices may take on many forms, but are often enabled by a micro-display projection and several refractive optical lenses that align the digital image to the user's eyes (J. Rolland & Hua, 2005). Head-worn AR displays have supported personal navigation, assembly and maintenance visualization, and medical procedures, providing additional context to operators (Van Krevelen & Poelman, 2010). Current research reflects the potential benefits of AR displays through a diverse range of designs, contexts, and implementations. *In this work, we will specifically consider lightweight head-worn, see-through AR displays that take on the form of glasses, commonly referred to as "smart glasses" in commercial and industry applications.*

3.2.2 AR Design Factors

Hardware Implementation. The experience using a head-worn AR display depends on several hardware-related factors. There are generally two approaches to AR devices: optical see-through and video see-through. With the former, users look through optical elements to see the real world directly, while the latter provides a real-time video feed from cameras mounted on the device (Aukstakalnis, 2016). Lee et al. (2020) identified several practical issues with AR displays which impact their use and the resulting user experience. First, there is a tradeoff between resolution and field-of-view, or the angular region in which the device projects an image. If a virtual image is made wider to fill the visual field, its sharpness, or angular resolution, will decrease (Xiong et al., 2021). Second, field-of-view has a tradeoff relationship with eyebox, or the region in which the user can be positioned to see the image (Lee et al., 2020). A larger eyebox can make the system more tolerant to user or vehicle motion, avoiding distortion of the virtual image (C. Zhou et al., 2024). However, the size and bulk of the device must increase to support a larger eyebox. Third, the AR device has a physical form factor in terms of both size and weight that may impact its practical use (Lee et al., 2020). A smaller, lighter device may be less capable in terms of the other hardware factors, yet more suited to sustained use. These design factors of AR displays are important to consider in evaluating both the strengths and limitations of specific implementations.

Compared to other mixed reality display technologies, head-worn, optical see-through AR displays offer an advantage in real-world awareness. As described in Section 3.2, mixed reality refers to the spectrum of environments in which real and virtual elements are merged, and includes technologies like virtual reality (VR), AR, or hologram projections. As a subset of mixed reality, optical see-through AR displays enable the user to see a view of the real world no more obstructed than wearing eyeglasses, with virtual elements superimposed. Many virtual reality (VR) headsets, like the Meta Quest line, offer a video see-through function to achieve a mixed reality experience. Cameras on the outside of the headset stream directly to the display, allowing users a live-feed video view of their surroundings.

From a visual perception standpoint, there are several downsides to the video see-through approach. First, video resolution may be limited by the camera, digital display, or viewing optics, resulting in a view less clear than the naked eye (J. P. Rolland et al., 1995). Second, processing the image from the cameras to the display comes at a cost of latency. In the case of typical VR systems, this is approximately 35-40ms (Bezmalinovic, 2024). Research suggests that even latencies below 100ms affect human motor performance, and may be perceptible as low as 20ms (Hadjiosif et al., 2025). Perceived visual latency diminishes user satisfaction, reduces presence, and may exacerbate simulator sickness (Gruen et al., 2020). In dynamic, high-stakes environments like flying, visual latency is unacceptable. Pilot-induced oscillations, a dangerous feedback loop resulting from pilot control inputs, are worsened by excessive lags in the closed-loop control process with the aircraft (McRuer, 1995). If pilots cannot see the effect of their control input quickly, they will tend to enter opposing inputs of increasing or consistent magnitude, causing the aircraft to oscillate. This behavior is dangerous in critical phases of flight, supporting the choice of optical see-through implementations of AR for aviation applications.

AR Graphic Design. There are several approaches to the design of digital elements, or graphics, that affect how an AR display is implemented and used. In this work, we define AR through the lens of AR integration, or the extent to which the display conforms to the real-world environment. To understand the level of AR integration, we can consider design decisions with respect to spatial location and information representation (Smith et al., 2025). A fully integrated display (or, *world-referenced*) will display items that appear to be part of the real world, both spatially and perceptually. While still spatially integrated into the world, graphics from a localized AR display will not appear to be part of it. The dynamic projection of objects in the world may reduce clutter and provide location context, but will increase complexity (Smith et al., 2025). With no complex spatial synthesis, an *AR overlay* display is the simplest implementation. Without outside-world integration, displays may simply overlay information statically over the existing field of view (Rowen et al., 2019). This would be considered a *device-referenced* placement, as the projection remains in the same location within the display, even as the user moves the display (Kruijff et al., 2011). Despite the simplicity of the graphics, the proximity of graphics makes pertinent information more accessible to users (Willett et al., 2017). The improved access to information can, in turn, improve operator performance (Kim et al., 2019; Smith et al., 2025). In this work, we focus specifically on *device-referenced AR overlays* in the context of the cockpit environment.

In the design of AR graphics, several characteristics may affect how users perceive and process information from an AR display. With the real-world environment as the background, since the display itself is transparent, graphics are perceived based on their relative brightness in the scene (Eisenberg & Jensen, 2020). Darker shades may appear more transparent, making details more difficult to resolve (Herbeck et al., 2024). For this reason, AR displays will be easiest to differentiate when they are bright, opaque, and contrast significantly from the background visual scene. In terms of visual occlusion, the opacity of an AR graphic represents a tradeoff; improved readability of AR graphics reduces the user's ability to perceive real-world elements beyond it.

3.3 Concerns with AR Displays

Though head-worn, optical see-through AR systems have several advantages, a thorough evaluation of these devices must consider the potential disadvantages.

Specifically, we note established concerns with visual processing and information processing. These concepts will help us to understand and explain any observed adverse effects of AR displays on pilot performance in the research.

3.3.1 Visual Processing Concerns

In this section, we identify several areas in which AR displays may influence human perception and processing. Although far from exhaustive, three of these concerns are especially relevant to the implementation of AR in the cockpit. First, there is a disconnect between the distance of the apparent object and the focal distance of the display. This *vergence-accommodation conflict* can result in user discomfort or nausea, as the projected display is unnatural to the visual processing system (Baraas et al., 2021). Even though the screen (or lens) is no more than an inch away, the projection is only in focus when the eyes are accommodated (physically adjusted to be in focus) several feet away. Fixed-mounted aircraft HUDs are designed to collimate at optical infinity, meaning that the projected image is perceived to be far away, allowing pilots to perceive distant objects without refocusing. Specifically, optical infinity refers to the distance at which rays of light are parallel and the image is in focus when the eye is relaxed (i.e. no accommodation is necessary) (Gabbard et al., 2019). The HUD therefore intends to prevent issues with accommodation when looking far away, as such systems primarily overlay the horizon (Crawford & Neal, 2006).

Over the course of a flight, pilots must perceive visual objects both inside and outside of the threshold of optical infinity accommodation (Biberman & Alluisi, 1992). While the HUD is positioned to overlay the horizon, a head-worn AR display might overlay the horizon and the inside of the cockpit at different points, or simultaneously throughout a given flight. Especially with a device-referenced display, where projected elements move with the user's head, the optical depth of the real-world background is not consistent. As a result, the apparent depth of the AR display, which remains constant, may frequently be mismatched to real-world depth cues, further exacerbating this issue (Xiong et al., 2021). This is known as *focal rivalry,* reflecting the conflict between real and virtual elements at different distances that cannot both be in focus simultaneously (Ganeswaran & Stevens, 2024).

Eyes must physically refocus to perceive objects at different depths in the visual field. Since the focal distance of the AR display may not match other objects in the visual field, users must switch between accommodating for the virtual display and the real world, known as *focal distance switching* (Gabbard et al., 2019). Despite the perceived benefit of AR displays in allowing users to simultaneously see the real world and the virtual projection, the AR display adds an additional layer to the visual field, so both cannot fully be in focus simultaneously. The physical act of refocusing the eyes puts a burden on users, contributing to physical fatigue (Lambooij et al., 2009). In an environment that relies critically on visual information, this strain on the optical system is a concern. Not only does the task of switching between viewing different objects contribute to visual fatigue, but the cognitive and visual attention shift between the AR display and the real world, known as *context switching*, requires effort (Arefin et al., 2020; Gabbard et al., 2019). The burden of this visual switching has been shown to increase eye fatigue and decrease task performance (Arefin et al., 2022), as well as increase error rate and task completion time (Eiberger et al., 2019).

3.3.2 Information Processing Concerns

Users must switch not only between focal depths, but also between contexts, creating a concern in terms of information processing. Specific information is needed for decision-making over the course of a flight, requiring pilots to direct attention towards the correct source at the correct time (Gray et al., 2016). Selective attention, in this case, helps to guide and filter the information that pilots will process. Safety of flight depends on pilots noticing and processing relevant information regarding the aircraft or environment (Wickens, 2007). Pilots must divide their visual attention within the cockpit to elements of different optical depths, from the horizon to the flight displays (Ziv, 2016). An AR display adds an additional information context that users must switch between. The cognitive burden of context switching is notable in itself, especially for potential periods of extended use (Gabbard et al., 2019). Refocusing cognitive and visual attention from real-world elements to an AR display is demanding, resulting in reduced user performance (Arefin et al., 2020). The potential for an AR display to impose additional demands on attention, through context switching, is a concern for its implementation.

A significant concern with a supplementary AR display may be that its presence negatively influences pilot scan. Inherently, the display will *occlude*, or visually block, the world behind it, to some extent, requiring pilots to adjust their scan. This compensatory gaze pattern may use both directed behaviors (top-down) or salient cues (bottom-up) (Li et al., 2014), suggesting that a change in information availability will change typical processing patterns. In a study on pilot situational awareness and gaze behaviors with a projected HUD, pilots focused their attention on the HUD region, leading to fewer glances to primary flight instruments (Yuan, Ng, Lau, et al., 2024). This result may suggest that pilots were able to look outside the window for longer durations, but may also carry negative implications if pilots fail to detect information that is only conveyed in primary cockpit displays.

An additional concern in the aviation domain lies in the capability of AR displays to project information anywhere in the central visual field. This projection over existing information sources may lead users to attend, or direct their attention towards, AR with higher priority. Display elements that are rendered atop the real world may demand a disproportionate amount of attentional resources (Crawford & Neal, 2006). This effect, *cognitive tunneling*, may impede a pilot's ability to detect changes in the external environment or in other aircraft displays (Crawford & Neal, 2006). The display itself may contribute to fixations and, consequently, less attentional resources dedicated to other areas (Biberman & Alluisi, 1992). Information that distracts will detract from the user's focus on the task at hand, reducing the level of performance (Bonanni et al., 2005). Furthermore, users may switch more quickly to information presented on AR displays, which can be disruptive if the information is irrelevant or distracting to the task at hand (He et al., 2018).

3.4 Flying Applications of AR

The advantages of AR technology are especially relevant in the aviation domain. According to Wickens et al. (2004), HUDs have three primary goals: reducing visual scan demands, reducing visual re-accommodation, and directly overlaying imagery. These principles may also be applied to AR displays, given the similar intent of head-worn AR to HUDs in cockpits. As described in Section 3.2.1, HUDs and AR displays both rely on projected information, but HUDs are statically mounted within the cockpit. Research has explored the potential for AR displays to aid in navigation tasks, especially in low-visibility situations (Pradhan et al., 2023; Sarbach & Weber, 2025). In adjacent areas, AR displays have been employed for aircrew training, as well as aircraft manufacturing and maintenance (Koc, 2024).

Recent studies have shown the potential for pilots to leverage the advantages of AR displays. In an early implementation of AR-based checklists, a dynamic overlaid display that guided airline pilots to the next step in an emergency procedure reduced both reaction time and workload (Tran et al., 2018). The researchers attributed the improvements to a reduction in scanning time and decision-making, allowing for higher situational awareness overall. In a study comparing AR and paper checklists, AR checklists controlled by voice commands were assessed to be the most usable (Li et al., 2020). These findings suggest that the augmented reality medium improved the pilots' perception of their own ability to manage the checklists for a simulated aircraft landing. A further study by the team revealed that the voice-command AR display had a promising, but not significant, improvement over traditional checklists with respect to perceived workload and performance (Li et al., 2022). The study also found no significant effect on situational awareness from the augmented reality display.

3.4.1 Regulatory Context for AR in the Cockpit

In exploring aviation applications for AR, the use of commercially available devices may support users in accessing the technology (Sarbach & Weber, 2025). Integrating existing devices may reduce the cost of entry and permit a faster pace of innovation. Currently, there is no outright ban on AR technology in the cockpit. While advanced helmet-mounted displays have been used in the military context for many years, the first head-worn AR device, the *AerAware*, was only recently certified for use in the United States (Hughes, 2024). This display, intended for use by pilots of the Boeing 737, extends the capabilities of a typical HUD by providing spatial information about the surrounding environment with a full 180-degree field of view. While this display represents a real-world deployment of AR technology in the cockpit, it is far from accessible in the general aviation context. Outside of limited commercial airline operations, pilots do not currently have access to certified AR devices. The FAA certification is the central mechanism preventing premature or unsafe integration of AR displays into the cockpit. To meet FAA requirements for certification, AR displays must conform to two relevant requirements, included in Table 1 (*emphasis added*) (Federal Aviation Administration, 2024a).

Table 1: Relevant FAA Equipment Regulations

Subsection	**Regulation**
§ 25.1309 Equipment, systems, and installations.	(a) The airplane's equipment and systems must be designed and installed so that: (1) The equipment and systems required for type certification or by operating rules, or *whose improper functioning would reduce safety, perform as intended under the airplane operating and environmental conditions*; and (2) Other equipment and systems*, functioning normally or abnormally, do not adversely affect the safety of the airplane or its occupants* or the proper functioning of the equipment and systems addressed by paragraph (a)(1) of this section.
§ 25.1302 Installed systems and equipment for use by the flightcrew.	(c) Operationally-relevant behavior of the installed equipment must be: (1) *Predictable and unambiguous;* and (2) Designed to *enable the flightcrew to intervene in a manner appropriate to the task*.

While research has supported the potential for head-worn AR to assist pilots, the implications of AR use on performance and human behavior require further investigation. The FAA standards set a clear threshold for inclusion: AR display systems must “not adversely affect the safety of the airplane or its occupants” (Federal Aviation Administration, 2024a). Given that current research into AR displays for flying applications has yielded limited and mixed results, this work intends to investigate pilot performance and behavior further. More specifically, the FAA imperative that equipment must not negatively impact pilots "functioning normally or abnormally" motivates the subsequent research in this thesis (Federal Aviation Administration, 2024a).

3.5 Automation in Flying

Among the most significant innovations in the cockpit is the rise of advanced automation technology. The most well-known form of automation is the autopilot system, which is often installed even in small general aviation aircraft. Through an onboard computer, the autopilot can manage pitch, roll, heading, altitude, and radio navigation, taking manual control tasks over from the pilot (Federal Aviation Administration, 2021). Automation in the cockpit is characterized by triggers and adaptations, adjusting to both control inputs and aircraft state outputs (Dorneich et al., 2016). The use of autopilot intends to provide more precise flight path control than a human pilot (Causse et al., 2025) while also reducing workload demands (Wickens et al., 2023). The use of autopilot systems shifts the responsibilities of human pilots from control to supervision, with the goal of improving overall safety and performance. This section explores the relevant context for automation that would enable an AR display system to assist pilots over the course of a flight.

3.5.1 Information Automation

Aircraft automation, although normally considered in terms of manual control, can also connect to the information domain. Although future systems may shift the pilot's role from active control to passive monitoring, pilots will still require real-time, effectively conveyed knowledge about the state of the aircraft (Rohacs et al., 2019). Current research in the space has explored the use of automation technology to support pilots in perceiving and processing information. Cahill et al. (2017) proposed an adaptive automated system that works collaboratively with pilots, as if it was an additional pilot itself. Integrating AR, Yuan, Ng, Li, et al. (2024) evaluated a HUD-based system that provided information and guided pilots' attention to reduce the burden of flying with only one pilot. Automation technology can be employed not only to supplant, but to assist human pilots in accomplishing complex tasks.

Adaptive systems that provide decision support to pilots are both valuable and feasible in the short term. We will consider *adaptive automation* as an automated output that is implemented dynamically based on the current task demands on the pilot (Parasuraman et al., 1992). For instance, Hilton (2012) implemented an intelligent auditory checklist system, which read checklist items aloud to pilots based on when the items were required to be accomplished in flight. In the user study, pilots showed an increase in compliance from 22.7% to 97%, on average (Hilton, 2012). Compliance with the prescribed actions on the checklist is critical, given that a substantial proportion of cockpit errors that have caused accidents have been attributed to neglect of checklist items (Dye, 1995). The auditory support system is emblematic of the advantages of automation in aircraft displays, even in unique modalities.

In the context of automation, AR can provide timely cues, checklists, and applicable information directly to the pilot. The information needed will depend on the phase of flight, flying environment, and aircraft status. According to the framework presented in Dorneich et al. (2016), a notional AR-based automated assistance technology would rely on “mission event,” “location,” and “time” triggers to produce “interface feature” responses, changing the presentation or salience of information. In future applications, augmented reality could be integral in enabling a pilot to monitor a highly automated system. For now, we focus on the potential for AR to present cues and information derived from adaptive automated systems.

3.5.2 Automation Failures

Despite the benefits of automated systems, failures can still occur. Automation failures might include several cases, including issues with software or hardware, use of the automated system in situations for which it was not intended, or performing as designed, but not in a way that the user anticipates (Sebok & Wickens, 2017). With this definition, automation failure may result from the system itself or the interaction between humans and the system. At the system level, Skraaning and Jamieson (2024) further differentiate between elementary failures, involving isolated components or functions, and systemic failures, in which interactions between systems lead to breakdowns. A systemic failure could result from sensor limitations, operating outside of the design envelope, or misconfigured inputs to the system from operators (Skraaning & Jamieson, 2024). With respect to an adaptive AR system, automation failure and its downstream may be impossible to fully eliminate.

Though an outright mistake from an automation system might be most alarming to a crew, failures may occur even without errors from the autopilot. Even if the automation performs well, the potential for a decline in pilot vigilance due to complacency or overreliance on automation are both concerns (Mitchell et al., 2009; Rudisill, 2000). Automation mode error, in which pilots lose awareness of the status or current task of an automated system, may lead to unsafe flight situations (German & Rhodes, 2016). This breakdown in human-machine interaction may lead to conflicting inputs between the pilots and autopilot, especially when the automation mode shifts without human input (Rudisill, 2000). Both working memory and sustained attention seem to be important in interacting with automation and recovering from errors (Jipp, 2016; Rovira et al., 2017). Each of these situations, resulting from pilot unpreparedness to effectively interact with the new technology, can challenge pilots in maintaining safety.

With automation capabilities, pilots may be prone to errors related to their perception of or interaction with the autopilot system. In research of the influence of automation on safety, pilots may not commit fewer errors, but will commit different errors (Amalberti & Wilbaux, 1994). In other words, since the required tasks and information processing demands are shifting, new human failure modes may arise. Pilots might demonstrate *automation bias*, the tendency to forgo their own information seeking and blindly follow decisions made by the automated system (German & Rhodes, 2016). Pilots should maintain responsibility for flying safely, but may become more passive, deferring to the automated system (Rudisill, 2000). In the context of an adaptive information system, this deference to automation could manifest in no longer seeking out necessary information in the cockpit, but simply waiting for the system to provide it. Sustained use may further develop this tendency, known as *complacency*, contributing to the decay of information acquisition skills (Klostermann et al., 2022).

While automation can provide cognitive support to pilots, it may lead to an increasing disconnect in the pilot's awareness of the system. This phenomenon, referred to as the *automation conundrum*, explains why automated assistance can improve overall reliability while reducing human awareness and ability to intervene manually (Endsley, 2017). Reliance on the automated information system may contribute to humans being out-of-the loop, unaware, and late to respond when a change in action is required (Sarter, 1994). Pilots that rely on a notional automated information system to prompt their processing of flight data or upcoming actions may find themselves behind if the system fails in some way. As previously mentioned, this failure may occur at any level, including a simple misalignment of current objectives. The potential for decline in operator skill when interacting with an automated system has been noted (Volz et al., 2016), as such a shortcoming is unacceptable when pilots must revert to fundamental skills for safety. When considering a notional automated AR system, the secondary effects of automation on pilot safety are a key consideration.

3.6 Flying Performance Assessment

Given that the highest priority of the pilot is to aviate, or safely control the aircraft, several means of quantifying pilot performance have been developed. Basic "stick-and-rudder," or basic manual input, performance is a core skill for pilots, as improper manual flying can contribute directly to mishaps (Abbott, 2023). Assessment of pilot performance is important for predicting safety, evaluating novel technologies in the cockpit, or selecting pilots that will be successful in future flying (Bates et al., 1997). However, many factors contribute to individual performance in manual control. In understanding human behavior amidst technology or automation failures, the quantification of pilot performance, in conjunction with assessments of cognitive factors, is vital (Kyle et al., 2025). Changes in manual control performance can indicate when pilots are flying less safely.

Through the parameters generated by flight simulation software, objective measures of pilot performance can be compiled and compared. The issue of assessing flight performance is complex, as pilots can only control high-level performance (like flight path) through lower-level inputs (like pulling the yoke back, thereby increasing the pitch to change the climb rate) (Ebbatson et al., 2011). Several methods for summarizing pilot performance have been developed, often varying in the tasks that are included. The earliest implementations of objective measurement considered performance relative to defined tolerances, recording the quantity and magnitude of deviations (Dickman, 1982). This approach may suffer if standards are defined arbitrarily, pilots use different strategies, or external conditions are inconsistent. In another approach, performance is recorded continuously, so standard deviation of both aircraft (pitch, roll, yaw) and control (rudder, aileron, elevator deflection) parameters are commonly collected (Mixon & Moroney, 1982) (Yesavage et al., 1999). Root mean square error (RMSE) can also be used, especially to assess the cumulative deviation in tracking performance (Kyle et al., 2025; Rantanen et al., 2001). More sensitive measures exist, like transforms of control inputs to the frequency domain, but these methods may be more reflective of pilot experience and general skill than real-time task performance (Rantanen et al., 2001).

For this work, we will use measures of variability to assess objective flying performance. This selection reflects the experimental design choices in the two user studies. We do not need to definitively assess the absolute level of performance, but rather the relative changes in performance based on the presence of an AR system. The flying task changes substantially between the two studies, so no measure of tolerance or unacceptable deviations could allow for a generalized understanding. The use of standard deviations of flight parameters enables comparison that is agnostic of the flying task and less dependent on uniform pilot skill. The simplicity of standard deviations helps to computationally mitigate the large volume of data from continuous streams, especially in Study 2 analysis.

3.7 Cognitive Assessment

3.7.1 Workload

Beyond objective performance, workload and situational awareness are essential real-time indicators for pilot performance and well-being. Both workload (Kantowitz & Campbell, 1996) and situational awareness (Hicks et al., 2014) have been connected to pilot error and aircraft mishaps during complex task performance. Workload can be understood as an intervening variable that is impossible to directly measure, yet affects the operator's capacity to handle environmental demands (Kantowitz & Campbell, 1996). When task demands are too high, an individual may become overloaded, negatively affecting their efficiency, safety, and comfort (Rubio et al., 2004). The conflict between cognitive resources contributes to the potential for unsafe situations in the air; pilots rely on limited cognitive resources to synthesize, process, and respond to information (Devean & Kewley, 2009). Competition for cognitive resources could disrupt human performance at every stage of information processing (O'Hare et al., 1994). In practice, this could manifest as a delay in processing information or acting on it, both of which may be critically dangerous in the air.

A deficit of cognitive resources relative to the task demands of the situation may lead to degraded or unsafe performance. The NASA TLX is a widely-used subjective assessment of perceived workload (Hart & Staveland, 1988). This questionnaire instrument captures several factors that affect workload, including mental and physical effort, time-based demands, performance, effort, and frustration when executing a given task (Hart & Staveland, 1988). This tool is especially useful for comparing between conditions; while two individuals may report their workload differently for the same task, the change in workload between tasks will be less affected by their subjective scoring. The interpretation of composite or weighted values, as well as interpreting a "baseline" workload, have been concerns related to the TLX (Hertzum, 2021). The TLX has also been criticized for shortcomings in the discriminability between the various categories; each workload factor seems to influence the others (McKendrick & Cherry, 2018). In another study, researchers found that the introspective measures of perceived workload within the TLX are stable and sensitive, though perception of external events can vary between individuals (Braarud, 2021). In studies assessing the effects of AR on workload, the TLX has been capable of showing significant differences between conditions (Li et al., 2022). The convenience of a self-reported questionnaire is high, a further benefit to the NASA TLX approach. Therefore, we will use this tool in our work as a means of assessing participant workload between conditions.

3.7.2 Situational Awareness

In the aviation domain, pilots' awareness within the ever-changing flight environment is essential for performing the correct actions at the correct time. Situational awareness (SA) refers to an operator's ability to perceive, comprehend, and project information about their environment to further the mission (Endsley, 1995). Pilots may make incorrect decisions without an understanding of the situation at hand (Stanton et al., 2001). As flight safety depends on continuous pilot input and accurate judgement (Loukopoulos et al., 2003), a clear understanding of the situation is clearly valuable. A loss of situational awareness occurs when the pilot's mental model begins to shift from reality, leading to a cognitively-demanding collapse and rapid rebuilding of the model (Nguyen et al., 2019). If the situation does not allow the time to regain awareness or the pilot is critically overloaded, a mishap can result.

Given the importance of situational awareness to pilot effectiveness and flight safety, understanding pilot awareness in a given span is desirable. To measure SA indirectly, the Situational Awareness Rating Technique (SART) is widely used as a self-reported assessment (Taylor, 1989). The validity of the SART for assessing situational awareness remains a topic of contention, however. Research suggests that the assessment of perceived situational awareness may be distinct from actual situational awareness (Endsley et al., 1998) and that the measure may be confounded with the mental processes assessed by the NASA-TLX (Braarud, 2021). The SART may assess perceived SA and confidence more than actual awareness when compared to objective measures (Moens et al., 2026). Individual elements of the assessment seem to be valid measures, but the mathematical combination of these elements to create a single value for SA is questionable (Bolton et al., 2022).

SA may also be assessed using objective measurement tools, like probes or external expert evaluation. With a probe, the scenario is paused and participants are asked questions about the situation to assess their level of understanding (Yuan, Ng, Lau, et al., 2024). This technique requires a disruption of the task at hand, potentially changing participant behavior. Expert evaluation, though minimally disruptive and independent of participant self-assessment, is heavily reliant on subjective perceptions of outside agents (Moens et al., 2026). Given these limitations of objective measurement techniques, the SART remains a common tool for comparative purposes (Hicks et al., 2014) and expanding understanding of performance (Bolton et al., 2022). For its advantages in simplicity, feasibility, and task continuity, this paper will rely on the SART as a basic assessment of situational awareness. To mitigate concerns with the validity of a single composite rating, we will consider the component subscales in the analysis.

3.8 AR System Evaluation

The evaluation of AR systems in applied environments remains a developing area of study. Considerations of user interactions and practical applications of AR systems were largely neglected throughout the first four decades of its existence (Dünser et al., 2008; Swan & Gabbard, 2005). As AR device technology has become more available, the importance of user-based evaluations has become more clear. Grabowski et al. (2018) present a review and comprehensive framework for evaluating wearable, immersive augmented reality (WIAR) technology in safety-critical systems. The researchers noted a gap in understanding how to assess AR as a tool to improve safety and operator performance in high-consequence environments. Recognizing the accessibility of digital and real-world information in operational environments with an optical see-through AR display, the team notes the potential for situational awareness and performance improvements. The specific shortcoming that the researchers identify is in the methods used for evaluating AR technologies and their impact within larger systems. To this end, the WIAR Technology Evaluation Research model provides a framework for assessing these technologies in applied environments. The model includes technology characteristics, operator characteristics, operator performance and processes, and technology use and trust. Though not all of the factors are relevant to this work, the approach of assessing multiple factors, including the environment, operators, and technology, has informed our evaluation methods.

3.8.1 Usage and Acceptance Assessment

Beyond the critical areas of performance, there may be other informative cognitive constructs to consider. Pilot preferences and perceived usability are relevant, as pilots may have a choice in whether or not to use a new display type (Brzeska et al., 2024; Wright & O'Hare, 2015). Usability refers to the ease of use and acceptability of a product, reflecting the interaction between a user and that technology (Lewis, 2012). In the complex and demanding environment of aviation, pilots should be in harmony with the interfaces they rely on (Clamann & Kaber, 2004). Real-world use of novel interfaces in the cockpit depend on usability due to both acceptance of the technology (Li et al., 2020) and the criticality of human performance (Clamann & Kaber, 2004). Safe and effective flying depends on human performance, elevating the importance of usability beyond convenience or enjoyment. In aviation applications, usability should reflect the range of environments in which the technology may be used, to include abnormal flight situations (Burke et al., 2016; Maybury, 2012). Though satisfaction may be assessed as less critical than safety, pilot perceptions of AR are important in understanding the feasibility of voluntary use of such devices.

The assessment of AR acceptance has been a notable area of research. The Technology Assessment Model (TAM) considers system and usage characteristics alongside the behavioral factors of *perceived ease of use* and *perceived usefulness* (Marangunić & Granić, 2015). Though this research will not assess the feasibility of a large-scale rollout of AR technology into the cockpit, several dimensions of acceptance are useful for understanding the potential for real-world use. Performance expectancy refers to the degree to which a user believes that technology will improve their performance, while use behavior describes their response to the technology in context (Guest et al., 2018). Pilots would logically reject the use of a supplementary AR system if they believed it would decrease their performance. This research specifically intends to examine the use behavior, or the observed effects of using an AR device while completing flying tasks. In addition, factors like output quality and result demonstrability refer to the expected precision and consistency of performance improvements from using the technology (Wild et al., 2017). The factors of pilot precision and consistency are useful to understand, especially in evaluating objective flight performance measures.

3.8.2 System Trust

In understanding human interactions with an AR system, trust is an impactful mediating variable. Specifically, we are interested in user trust given that automated systems are likely to fail, as described in Section 3.5.2. Trust, typically considered with respect to automation, refers to the human operator's belief that the system will help to achieve their goals in an uncertain situation (J. D. Lee & See, 2004). The willingness of operators to rely on an automated system may be calibrated by their disposition, the situation, and their experiences (Hoff & Bashir, 2015). Both a lack of trust and an excess of trust in a system can be problematic. The latter, which characterizes *misuse*, can lead to human operators accepting mistakes from an automated system over their own correct judgement (J. D. Lee & See, 2004). Intuitively, trust is difficult to regain once lost. Even after initially trusting an automated system, participants that observed mistakes lost confidence in both reliable and unreliable versions of the system (Dzindolet et al., 2003). For this reason, it is important to understand how pilots respond to unreliable behavior from an automated AR system.

Given that trust can affect user behavior, there are real consequences for loss of trust. The notional improvements to pilot performance when using the system may be lost due to subsequent *disuse*, or failure to sufficiently utilize the system (Parasuraman & Riley, 1997). In other words, pilots that no longer trust an automated AR system will not elect to use it, even if use typically enhances performance. To this end, we must understand the extent to which users trust a cockpit AR display. Though several questionnaires for system trust exist (Scharowski et al., 2024), we intend to ask participants directly the extent to which they trust the system. This choice intends to simplify and expedite the participants' experience during data collection, given the extensive survey demands previously described for workload and SA assessment, while still gathering the critical information on user trust.

3.9 AR and Human Information Processing

At the foundational level, this research extends our understanding of the effects of augmented reality on human information processing. The conceptual framework we will consider for this research is Wickens' Human Information Processing (HIP) Model (Wickens, 2013). Figure 1 is adapted from the Wickens HIP model. As a novel and unique medium, AR displays may be better understood through the principles of information processing. Specifically, AR displays may change the availability and presentation of information in a system, influencing how users gather and utilize information. Although significant applied research has been conducted with AR, understanding the ways in which the technology affects the stages of information processing can guide future research and technological implementation. By connecting specific *in situ* findings of AR research to information processing principles, we can later synthesize across domains, supporting research through generalized, actionable conclusions.

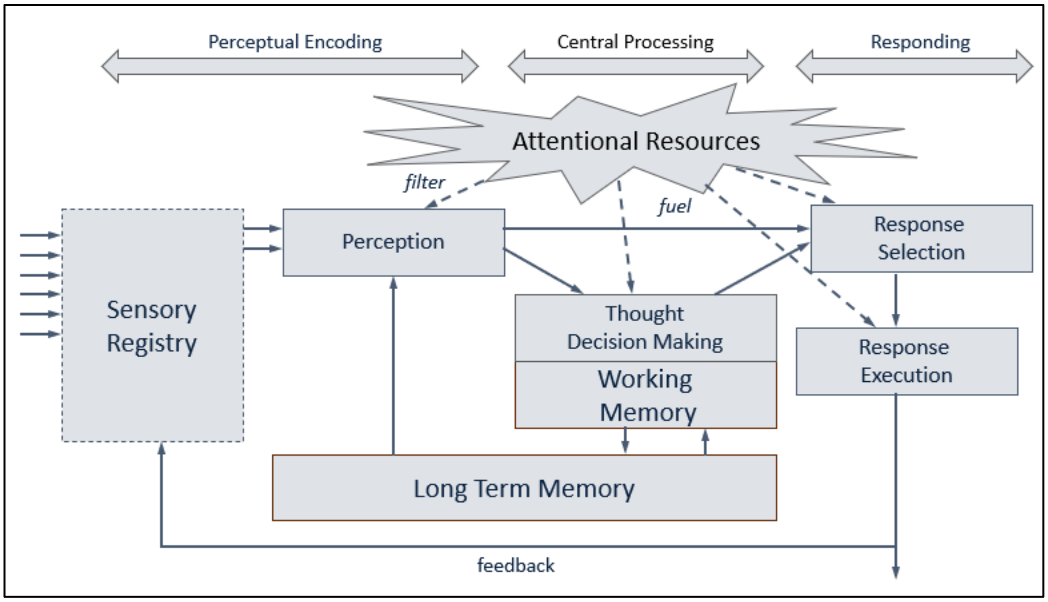


Figure 1: Human Information Processing Model

In context of the cockpit environment, AR technology will most critically affect sensory processing, perception, and response selection through the allocation of attentional resources. First, *sensation* refers to the reception of information from the physical environment through the sensory registry. In this case, light from an AR display must be received by the visual system and converted into neural signals to the brain for optical processing (Proctor & Proctor, 2012). The design of the AR display can adapt to human sensory capabilities through elements like resolution, contrast, and brightness (Baraas et al., 2021). Since AR displays can only add light, the ability to discern information on the display depends on the baseline brightness and colors of the environment (Erickson et al., 2020). A failure at the sensory level might be, for instance, a user's failure to visually discriminate an element on the AR display from the background. If a line of text or icon is too small or dim, it may not be detectable to a user. In addition, a digital element may occlude some elements of the real world, even after conscious focal switching.

When information is successfully sensed, the next stage in information processing is *perception*. Perception refers to the interpretation of sensory information in context (Wickens & Carswell, 2012). Perception of visual data allows the user to obtain meaningful information that can be used or acted upon. AR may be especially helpful in promoting bottom-up processing, in which perception begins with the detection of external sensory stimuli (Wang et al., 2013). A principal theoretical advantage of AR lies in the ability to simultaneously perceive the real world and display elements. Nevertheless, perception may be negatively affected by optical clutter from digital display elements, increasing the difficulty for users to locate and process relevant information (Smith et al., 2025). The information processing model framework can help in understanding the extent to which AR displays help or hinder human perception.

An evaluation of human interactions with augmented reality would be incomplete without considering attention. Perception, along with the cognitive tasks of working memory, decision and response selection, and response execution, is mediated by attentional resources (Wickens & Carswell, 2012). It is important to consider the demands of an AR display on attention, which is limited in capacity. Working memory, which can support the temporary storage of information for further processing (Cowan, 2010), may be supported by the accessibility of information through AR displays. If important information can be readily available to the operator, demands on working memory may be offloaded to the display. Through the theoretical framework of the human information processing model, we can understand the cognitive elements that are most affected by augmented reality and understand the importance of critically evaluating the technology.

Although the human information processing model is useful for considering many cognitive tasks within the cockpit, it may not explain the range of pilot behavior. To expand our understanding of the system and information flows, we may also consider naturalistic decision making (Wickens & Carswell, 2021). Naturalistic decision making (NDM) can be defined as “the way people use their experience to make decisions in field settings” (Zsambok & Klein, 2014). NDM is useful for explaining human processing in environments that are uncertain and dynamic, with continuous action/feedback loops and time constraints (Hare, 2017). This type of information processing and response execution relies more on intuition than analytical strategies (Simpson, 2001). Operators gain intuition about the proper action to take through extensive experience in the cockpit environment. Specifically, pilots perceive the situation, match patterns of information, and then assess goals, cues, expectancies, and potential courses of action (Simpson, 2001). Expert pilots seem to “circumvent” typical processing limits through their organization of knowledge, allowing rapid assessment of the current situation and anticipation of future events (Vidulich et al., 2010). This model can help us to assess the emergent cognitive effects of expertise in the cockpit. By considering flying as a holistic, unified task, we can avoid reducing complex behaviors into just one stage of information processing. Both models are useful in evaluating and explaining the behavior and performance of pilots for this study.

3.10 Related Work

3.10.1 Impacts of Wearable Augmented Reality Displays on operator performance, Situation Awareness, and communication in safety-critical systems

In the space of AR-based assistance for operators in high-stakes environments, several studies have laid the foundation for this research. In 2019, Rowen et al. investigated the effects of a wearable AR display on operators performing a simulated maritime navigation task. The researchers assessed communication, objective performance, and situational awareness between subjects, with and without an AR assistive display. The AR display provided an integrated view of information available from other displays, but was purely device-referenced (not conformal to the environment). Their design of the AR system has several similarities with our intended implementation of the AR display, despite the difference in domain. The use of the AR display significantly improved operator performance and situational awareness in the study. With 211 participants and a between-subjects design, this study was able to eliminate effects of learning while maintaining statistical power, especially for measures of objective performance. The scale is a notable strength of this study. This research extends understanding in the space of AR assistive displays with a practical display implementation and relevant measures of task performance for the domain.

Unlike many studies of AR technology in safety-critical systems, the study did assess how operators performed under abnormal conditions. In one leg of the navigation task, participants had to contend with “reduced visibility, high currents, bridge equipment failures, and unpredictable behavior from other vessels” (Rowen et al., 2019). This reflects our abnormal/emergency condition, given the presence of equipment failures. However, the authors do not differentiate the effects of AR on performance by condition (normal vs. abnormal). The introduction of nonstandard conditions allowed the researchers to assess additional measures of performance (like threat avoidance) for both display conditions. Our work extends the Rowen et al. study by determining if there is a change in performance during abnormal conditions.

Additionally, the expected behavioral impacts from an emergency equipment failure between the maritime and aviation domains are vastly different. Engine failure in flight necessitates an urgent and accurate response, characteristics that the nonstandard conditions in the maritime study do not embody. Put simply, a failure of ship equipment leaves the ship in a stable position; even if immobile, the ship keeps floating. While the vessels were able to complete their course and faced no timeline compression, pilots must respond quickly since sustained flight depends on engine power. Thus, the requirement to quickly and substantially alter the planned course of action in our study supported finding a difference between normal and emergency conditions.

3.10.2 The influence of augmented reality interaction design on Pilot's perceived workload and situation awareness

Within the aviation domain, Li et al. (2022) studied the effects of an AR checklist and associated interaction methods on pilot workload and situational awareness. The study assessed commercial pilots performing an instrument landing in a Boeing 747 flight deck simulator. Participants performed three landings under each of the three conditions: AR with voice control, AR with gesture control, and traditional checklist. After completing the landings, participants completed the NASA-TLX and SART for workload and situational awareness assessment, respectively (Endsley, 1995; Hart & Staveland, 1988). The information on the AR display was fully integrated, meaning cues were spatially overlaid onto real-world objects. By mapping the cockpit environment and integrating the display items, the display guided participants to the specific control they had to manipulate next to accomplish the checklist.

The study found a significant effect on situational awareness by interaction mode, as the gesture-control AR checklist decreased SA compared to voice-controlled AR and traditional checklist. In terms of workload, the AR voice-command mode reduced perceived workload compared to traditional and gesture-controlled modes. The authors also note no significant difference with workload and SA between the traditional and AR voice-command checklists. The researchers view semantic, written feedback and integrated display elements as key advantages of a cockpit-based AR display. This supports the idea of integrating AR with automated systems to provide targeted information. In their implementation, specific cues guided pilots to the location of a control and gave written instructions for how to manipulate it. While our implementation will not use spatial cues or integrated indicators to guide attention, we are nevertheless interested in the same measures of performance in a very similar task. If we can see an effect of workload on a non-integrated display, integration may be less essential for future implementation than the researchers indicate.

Our work extends these research efforts in several ways. First, the Li et al. (2022) work assessed only the potential for workload and performance improvements in standard flight conditions. This limited scope presents a clear area for further exploration in abnormal conditions with nonstandard pilotage. If there is not a significant difference between AR and traditional checklists in normal conditions, might there be a difference in abnormal conditions? The lack of objective flight performance data also represents a gap, especially if AR has a substantial, yet unperceived, effect on manual control. This could come as a result of incorrect control response selection or difficulty in executing the selected response (Wickens, 2013). In either case, the unique attentional demands of the AR display could manifest in the task of piloting the aircraft.

3.10.3 *Fast, Accurate, But Sometimes Too-Compelling Support: The Impact of Imperfectly Automated Cues in an Augmented-Reality Head-Mounted Display on Visual Search Performance*

Using an abstract visual search task, this study represents an investigation of the performance impacts of an AR system that is not perfectly reliable (Warden et al., 2023). The study investigated the response time and accuracy of participants in finding a unique object with and without the assistance of an AR system providing spatial or visual cues. The researchers used three distinct types of cues, including an arrow that pointed towards the object's location, an icon showing the appearance of the object, and a minimap indicating the location of the object in the space. For each of these cue types, half of participants experienced trials in which the AR system provided unreliable cues. While the Warden et al. work was completely independent of the aviation domain and used an abstract task, it reflects a similar line of effort as our study. It represents useful context regarding our investigation into incorrect AR cues for pilots in an emergency flight condition.

The study found that spatial cues, in both reliable and imperfectly reliable conditions, provided a benefit in response time (Warden et al., 2023). The arrow and minimap, which indicated the object's spatial location, led to a significantly faster response than the icon, which showed only the appearance of the object. Accuracy was highest for the arrow cue, followed by the icon, though all AR cues reflected improved performance over the unassisted condition. Incorrect AR cues decreased accuracy substantially, while only slightly increasing response time. This finding may indicate that participants were over-reliant on the AR system for accomplishing the task.

A notable difference between this work and our own, beyond the task itself, is the design of the unreliable AR cue condition. Warden et al. specifically utilize an AR system that is 83% accurate, only presenting an incorrect cue 17% of the time. Even for the incorrect cues, the AR system indicated an object very similar to the target one, increasing the challenge of recognizing a mistake from the AR system. In contrast, our experimental design used an entirely incorrect cue sequence. We intended this failure to be recognizable, potentially allowing participants to ignore it entirely. We selected this approach in order to assess the extent to which AR displays, even when clearly incorrect, have the potential to distract and disrupt pilots in the execution of their tasks. Nevertheless, the Warden et al. study offers great insight into the implications of incorrect AR cues on operator performance. Our work relies on very similar principles to understand the impact of AR failure more clearly in the aviation domain.

3.11 Gap Identification and Rationale

Although recent research has explored the potential for AR displays to mitigate workload and enhance situational awareness, these studies have typically focused on basic tasks in standard (non-emergency) flights, such as normal landing approaches. Despite this direction of research, well-trained pilots need the least assistance during a standard, routine flight. In familiar situations, pilots experience less stress and perform better, even when there is a high time pressure (Peng et al., 2025). Therefore, supplemental displays and automation systems may provide less value for standard, familiar operations. Just as human pilots may not need support from automated systems in standard conditions, advanced automated systems may not need a human pilot. While an automated system can handle pilot tasks with higher accuracy and consistency, the value of human pilots is in their ability to “adapt to ever-present and ever-changing operations and system/equipment failures” (Bailey et al., 2017). Therefore, as a pilot assistance tool, AR displays should also aim to assist in abnormal or emergency conditions. At the very least, they should not impede human pilots in their most critical task – adapting to dynamic conditions and making safe decisions.

To our knowledge, no research has explicitly tested the effects of incorrect or irrelevant cues through an AR system on pilot behavior. Current research has examined, at most, two of three components of the current research: pilot performance, augmented reality, and incorrect cues. We expect that the intersection of these topics will yield insights about the impact of integrating AR displays into legacy systems. If there is, in fact, a performance decrement in an emergency, measures could be implemented early in system development to mitigate the unintended negative effects.

Knowledge of the specific ways in which the AR display affects pilots would guide the design and implementation of such supplementary displays into the cockpit.

3.12 Hypotheses

Hypothesis 1: AR graphics that occlude real-world information sources in the visual field will slightly degrade pilot performance.

By definition, virtual elements from AR displays will be projected over the existing visual field, potentially obstructing the view of real-world information sources. This defines visual *occlusion*, or the obstruction of elements in the environment from AR display projections. Since we intend to use opaque AR graphics, light from the cockpit displays or outside world (represented by screens in the simulator) will be unable to reach the pilots (J. Rolland & Hua, 2005). As described in Section 3.1.1, pilots primarily rely on visual information to execute the various tasks of flying, including manual control. With fewer sources available due to AR visual occlusion, pilots may struggle to maintain their baseline level of performance. Nevertheless, we expect that this effect may be minimal, since pilots have access to redundant sources of critical position and orientation information (Schvaneveldt et al., 2000). Impacts to flying performance may also reflect an increase in pilots' cognitive demand from contending with the AR occlusion, as the AR may disrupt habitual information-gathering behaviors.

Hypothesis 2: Pilots presented with irrelevant AR information will experience higher workload.

Foremost, we expect that pilots presented with irrelevant information on the AR display will experience higher workload during their flying task. AR displays may be more likely to capture attention over other displays, increasing the likelihood that pilots will be distracted by AR cues (He et al., 2018). Furthermore, we expect that the gaze pattern of pilots will be affected by the AR display, which compounds the processing burden of separating incorrect cues (Yuan, Ng, Lau, et al., 2024). The combined effect of visual salience and mental processing with incorrect cues may increase workload and decrease the capacity of pilots to attend to other tasks.

With respect to workload, pilots presented with irrelevant AR cues in an emergency will experience increased mental demand from the conflicting and superfluous information to process. Given the salience of a dynamic AR display, maintaining attention away from the display may require mental effort. Pilots may also experience increased temporal demand, given that the AR display would add complexity to an already time-compressed scenario. To land as soon as conditions permit, pilots must accomplish vital tasks more quickly than a typical landing pattern. Pilots may spend time reviewing the cues presented by the AR display, wasting time through processing and through the cost of switching tasks.

Hypothesis 3a: Information from the AR display will aid situational awareness in normal conditions.

In terms of situational awareness, we expect that the presence of an AR display will increase the supply of cognitive resources, as assessed by the SART. The additional cues presented to pilots are not typically available without deliberate searching. Therefore, the consolidation of important information will improve both the quantity and quality of information available to pilots. Therefore, pilots may increase their awareness of the aircraft and environment, supporting the tasks they must complete. In normal flight conditions, we expect that pilots will assess their situational awareness to be higher with the assistance of cues from the AR display.

Hypothesis 3b: The same AR display cues, when presented as irrelevant information in the emergency condition, will reduce pilot situational awareness.

However, in the emergency condition, we anticipate that the presence of the AR display will increase the cognitive demand on pilots. The additional task of reading and ignoring the irrelevant AR cues will reduce the pilot’s awareness of other events in and outside of the cockpit. This will, therefore, reduce their perceived situational awareness. The task of flying an aircraft relies on many concurrent tasks being accomplished at once, and introducing yet another task (evaluating irrelevant cues) will make situational awareness lower (Loukopoulos et al., 2003). Cognitive tunneling on the information presented by the AR display could further reduce the pilots' awareness of their environment (Crawford & Neal, 2006).

Hypothesis 4: In emergency conditions, irrelevant AR assistive cues will decrease flying performance.

As described in Section 3.1.1, the task of flying is primarily dependent on visual information. Even when AR cues no longer align with the goals of the pilot, the information still takes up a portion of the visual field (Wickens, 2013). Once again, the salience of AR displays and the cognitive friction of processing irrelevant cues may place a burden on pilots. As a result, we expect to observe an increase in variability in their flying. Using the standard deviation of basic flight parameters (further described in Section 3.6), we hope to identify a decrement in manual control consistency. Manual control is a product of physiological, cognitive, and sensory dynamics (Lone & Cooke, 2014), suggesting that interference by an AR display to sensation and perception of visual cues could impact control. Manual control tasks can be mathematically modeled on their own, but the effects of both new displays and automation are not fully understood (Baron, 1988). The addition of a new source of information, as well as target of attention, may impact pilot control input behavior. As a result, we expect that irrelevant cues presented on an AR display will negatively impact pilot performance.

In the following sections, we describe the two user studies to evaluate these hypotheses. Section 4 describes the first study, evaluating how the presence of AR graphics affect pilot performance. Section 5 describes the second study, which implements and tests an AR system in both normal and abnormal conditions to assess impacts on pilot performance, workload, and situational awareness.

4 Study 1: Effect of AR Display Location on Flying Performance

4.1 Introduction

In the cockpit, AR displays may offer pilots information without the need to look away from the outside world or critical instruments. The upsides to a display that allows a view of the world behind it have been well-documented. Even so, the impact of AR on pilots flying aircraft is still not sufficiently understood. As discussed in Section 3.9, we suspect that AR displays change the characteristics of human information processing relative to typical display systems. Since AR is defined by the projection of virtual elements over a real-world scene, certain elements in the environment may be occluded. If an AR display occludes an important real-world information source, pilots may be less stable in their flying performance. This study aims to reduce the gap in understanding regarding potential detriments to fundamental flying performance due to the presence of AR displays, regardless of content.

4.2 Problem Statement and Motivation

Although research has explored the potential for AR displays that take over functions of traditional cockpit displays and automation (Minaskan et al., 2021), thereby replacing existing cockpit displays, early examples will likely occur with simple designs into legacy cockpits. As a result, information projected by early AR displays will likely visually obscure existing controls or displays. Although an implementation of an AR display might use semi-transparent graphics to retain some view of the world behind the display without the need to move head position, we considered the most severe scenario of complete occlusion. Even with a transparent background, an information-dense AR projection could still reduce access to the real-world area behind it, as discussed in Section 3.2.2. By understanding the full occlusion case, we can understand the worst-case effects on pilot performance for less intrusive designs.

For pilots, information both within and outside of the cockpit may be relevant in accomplishing a task or building awareness. Pilot information needs are task dependent (Yu et al., 2014), so gaze patterns to acquire information may shift depending on the task at hand. Therefore, it is unlikely that there is a single ideal location for AR display graphics for every flying task. Nevertheless, a supplementary AR display must not impede the most basic tasks of flying. This study investigated the effect of AR display *location* on basic flying performance. For basic flying maneuvers, we consider strong pilot performance to be low variability in flight parameters (like airspeed, vertical speed, and bank angle). This definition is a simple adaptation of widely-used objective measures of performance (Mixon & Moroney, 1982).

Visual references for basic maneuvers are redundant (e.g. bank angle may be determined by looking at the horizon, through vestibular senses, and through flight instruments). Information may be derived from real-world cues or traditional aircraft displays (*Federal Aviation Administration*, 2021). As such, we expect pilots to be able to compensate in their flying when a part of the visual field is occluded. However, instability through deviation from baseline parameters could be indications that this compensation is either demanding or unnatural to the pilot.

To assess the impact of AR displays on pilot flying performance, we systematically blocked out different areas in the pilot field of view with a head-worn AR display (described in more detail in Section 4.3.2). Pilots flew a basic turning maneuver, a steep turn, for each visual occlusion condition. The steep turn involves a sustained bank of at least 45 degrees, requiring precise control inputs. This study design enabled us to isolate the impact of AR display position on flying performance, since the AR graphic contained no information that could assist or hinder the pilots.

If AR display location does influence actual or perceived flying performance, it should be a key consideration for designers and regulators within the field of aviation. We are interested in understanding the impact of AR disrupting pilots' access to one or more of these sources. This experimental design answers the question of AR's impact on flight performance due to its presence alone. The study contributes to the measured integration of novel displays for pilot assistance into the legacy cockpit environment.

4.3 Experimental Design

The study design was approved by the IRB, in conjunction with parallel experiments by the Aerospace Engineering Operations aNalysis (AEON) Lab. Specifically, this experiment was the second station of three during the data collection event. The first station assessed steep turn performance in a VR flight simulator. We do not suspect that the order of experiments had an impact on participant performance in this study. The experience using a VR system is distinct from the fixed-base simulator used, meaning that participants had to approach the task anew. We also gave participants the opportunity to take a break between stations, mitigating any effects of fatigue.

4.3.1 Participants

A total of 28 participants took part in the study. All participants were student or licensed pilots currently attending (n=26, age 18-29) or instructing (n=2, age 60+) at Auburn University. Outside of the instructors (with 5200 and 19500 flight hours), participants had 177.7 hours on average (SD = 128.2). Of the participants, 25 were male and 3 were female. All had experience in the Cessna 172-type aircraft. Eligibility criteria required that participants have existing flight experience of at least 30 hours and not require glasses to fly the fixed-base flight simulator task. The flight component of the study took 15 minutes and the survey took 5 minutes. Participants were not compensated directly, but could log the total time of flight on the simulator at their discretion.

4.3.2 Apparatus

The study utilized the Precision Flight Controls DCX MAX fixed-base simulator, which was previously certified to FAA 14 CFR Part 60 standards for flight training. The simulator features a functional Garmin G1000 display, 180-degree panoramic displays, and all representative control components of a Cessna 172 cockpit. It also employs 6-degree-of-freedom motion through four actuators on its base (Marsh, 2016). Data from the flight simulation, running XPlane 9 Commercial, was streamed from a wireless modem to a data collection laptop.

For the AR display, participants wore XReal Air 2 Pro glasses. These optical see-through glasses have a 46-degree field of view (FOV) and a 1920*1080-pixel resolution per eye (*XREAL Air 2 Pro*). They are commercial off-the-shelf AR glasses available today and represent likely implementations of early AR displays. Participants' view during the occlusion condition is shown in Figure 1. We developed custom graphics in Microsoft PowerPoint, which the glasses mirrored, to project the occlusion. These graphics were divided into four quadrants, as shown in Figure 2. For a given condition, a plain, opaque white rectangle was placed into the desired quadrant to occlude it.

Figure 2: AR Occlusion Locations from Pilot View


Figure 3: AR Occlusion Overlay Cockpit View

4.3.3 Experimental Design

The study utilized a 6x6 Latin Square, within-subjects design. Each participant experienced all six possible AR conditions: *no glasses*, *nothing displayed*, and occlusion in each quadrant of the display (*top left*, *top right*, *bottom left*, *bottom right*), as shown in Figure 2. The assigned conditions were counterbalanced for every six participants, as we do not foresee any effects by participant order based on recruitment. For each condition, participants performed a set of two steep turns in the same direction (a 720-degree turn, in total).

4.3.4 Procedure

After arriving, participants sat down in the simulator and adjusted their seat position for comfort and full range of motion of the controls. We then provided participants with the following instructions:

> *In this section, you will be wearing augmented reality glasses. These glasses may project a shape intended to block part of your vision. If you are instructed to take the glasses off, just place them on your head or in your lap. You will be flying steep turns at 45 degrees of bank. For this section, you will perform two 360-degree turns, then roll out, repeating for a total of six sets (or twelve total turns)*.

We then started the simulation, giving participants an opportunity to familiarize themselves with the simulator and practice stabilizing the aircraft. After they practiced in the simulator for 2-3 minutes and confirmed that they were ready, participants were instructed to begin the set of turns. The AR occlusion condition was only presented after participants began to turn, ensuring they could not select a turn direction preferentially based on the location of the occlusion, such as turning left to avoid an AR occlusion to the right. After each set of turns, participants were asked to rate their overall performance and the smoothness of the maneuver on a scale of 1-10. This rating scale, drawn from a concurrent research project, did not have any specific benchmarks, so participants responded based on their own intuition (Alarcon-Aneiva & Fala, 2025). Participants continued to fly until they completed all 6 conditions, with no formal breaks between sets. After the final set of turns, participants exited the simulator and completed the brief survey.

4.4 Data Cleaning and Analysis

After the data was collected, we proceeded by processing and analyzing the flight simulator streams. To analyze the data, the raw continuously-recorded flight data streams were first converted to a usable form, sampled at a frequency of 4 Hz (Rahman & Schulmeyer, 2025). Then, individual turns were separated within each participant's session based on both rolling average of bank angle and total heading change. With the six turning periods identified, we calculated several metrics intended to assess the stability of the participants' flying. This meant finding the standard deviation (SD), by turn, of airspeed, bank angle, and vertical speed. These measures were selected to account for the key parameters of a steep turn, and because they yielded usable data. Other proposed metrics, like standard deviation of G-force, pitch angle, and altitude, did not have substantial relevance for assessing performance. Basic visualizations of these data were noisy, lacking consistency across participants or between conditions. Finally, the average SD for each metric within each turning condition was calculated, providing a basis of comparison between participants. Data on turn direction and the occlusion condition were also connected to the participants' self-assessed performance. This additional context helped to build a more complete understanding of how the AR display affected pilots' flying.

Upon analysis, the flight simulator data from three participants proved to be unusable, as the participants deviated too significantly from the stated procedures to properly compare their performance. There is no evidence to suggest that any external factor or demographic characteristic contributed to their irregular execution of the task. We excluded the subjective assessments of performance from these participants, since the task they executed was fundamentally different. In practice, the continuous streams of data representing their flights could not be differentiated into separate experimental conditions.

4.5 Results

4.5.1 Subjective Performance Results

As previously described, the subjective self-assessed performance data was based on a 1-10 ordinal scale, with 1 representing lowest performance and 10 representing highest. As a result, normality assumptions were not supported for data analysis using a repeated-measures ANOVA, so a non-parametric approach was selected. Specifically, the Kruskal-Wallis Test was the best fit for subsequent data analysis. Unlike the Friedman Test, it assumes each observation is independent, better reflecting the potential for a single participant to vary between turn directions for arbitrary reasons. Though participants all completed six steep turns, participants were able to complete each turn at the airspeed and in the direction of their choosing. Therefore, each set of turns for a given AR condition was best represented as an independent observation.

As a result, we used a Kruskal-Wallis Test to evaluate whether a significant difference exists between conditions. This test revealed no significant effect of condition by performance ($p = 0.489$). The self-assessed measure of "Smoothness," collected alongside performance (Alarcon-Aneiva & Fala, 2025), was not significantly different than reported measures of performance ($p = 0.595$) using a Wilcoxon Signed-Rank Test. Therefore, we considered only self-assessed performance in this analysis.

As participants were free to turn in either direction for each set of turns, the procedures inadvertently introduced the additional factor of turn direction. To investigate the effects of turn direction, we conducted a Kruskal-Wallis Test with turn direction as a blocking variable. The turn direction was not a significant factor on performance alone ($p = 0.746$). To assess the potential for interaction between turn direction and condition, we treated each combination of direction and AR occlusion as a separate condition. This subsequent Kruskal-Wallis test yielded a p-value of 0.242, again failing to reach significance. Though the choice of turn direction may have some impact on stability, it does not help to explain the variability in the data. The effect size, representing the extent to which the phenomena may be present (or, in other words, the extent to which there may be a true difference between conditions) was small (ε^2=0.030), as well (Iacobucci et al., 2023; Lakens, 2013).

Therefore, all findings on equivalent statistical tests resulted in the same conclusion – participants did not report significantly different flying performance based on AR condition. Although not significant, we identified several trends in the data. Figure 4 shows box plots for self-assessed performance by condition. Overall, performance was rated highest for the control condition, blank condition, and top left condition. The lowest rating for all conditions was with the *Bottom Right* quadrant, followed by the *Bottom Left* quadrant, occluded. In addition, the variability for the two occlusion conditions in the lower visual field appears to be higher than the *No Glasses* and *Blank* conditions. Though average performance did not seem to be affected by the AR occlusion condition, the variability in reported performance could indicate that some pilots experienced a wider range of outcomes, including some instances of lower performance.

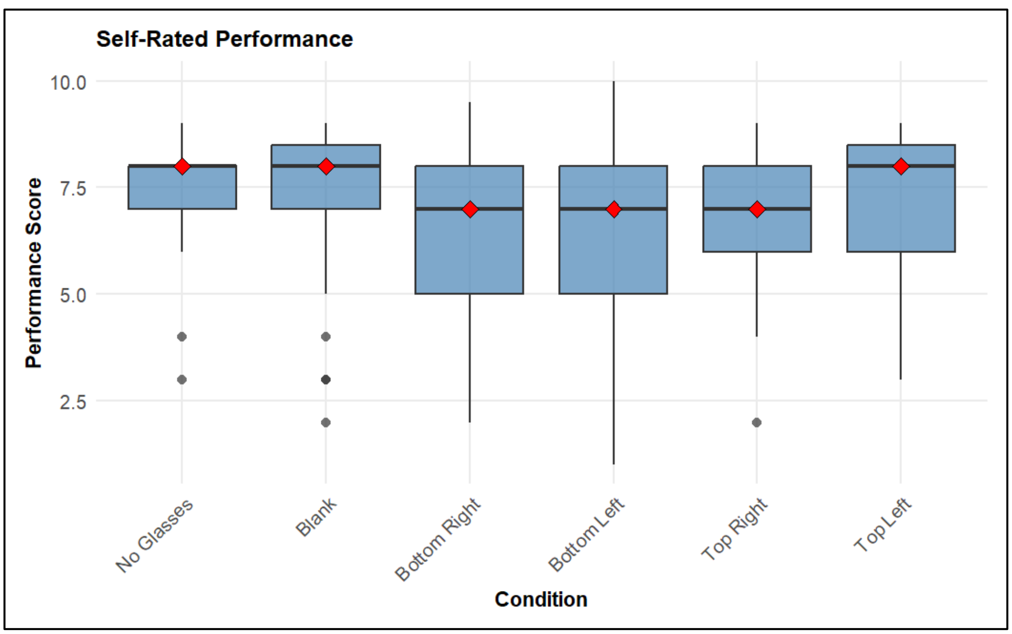


Figure 4: Box Plots for Self-Rated Performance by Condition

4.5.2 Objective Performance Results

Similarly, the objective flight performance data from the flight simulator did not yield any significant results. A Shapiro-Wilk test indicated that the distributions of standard deviation values for airspeed, vertical speed, and bank angle were non-normal ($p<0.0001$). Since the normality assumption for an ANOVA approach was violated, we again used a Kruskal-Wallis Test to assess differences. The results by condition yielded p-values of 0.333 for Airspeed standard deviation (SD), 0.911 for Bank Angle SD, and 0.582 for Vertical Velocity SD. Effect sizes were small for all variables (airspeed ε^2=0.0385; vertical velocity ε^2=0.0253; bank angle ε^2=0.0102), as well. However, these values do not account for the potential effect of turn direction on performance.

To account for the potential variability due to turn direction, each combination of turn direction and AR occlusion was again treated as a unique condition. The results by condition and turn direction yielded p-values of 0.293 for Airspeed standard deviation (SD), 0.741 for Bank Angle SD, and 0.126 for Vertical Velocity SD. Effect sizes were medium for airspeed ($\varepsilon^2 = 0.087$) and vertical velocity ($\varepsilon^2 = 0.110$), but small for bank angle ($\varepsilon^2 = 0.052$) based on common thresholds (Lakens, 2013). This suggests that there may be some difference between the sample means between conditions, but the experiment was insufficiently powerful to detect it (Iacobucci et al., 2023). Potential causes for this shortcoming are discussed in Section 4.6. The statistical test results for objective and subjective data, corrected for turn direction, are presented in Table 2. Of the objective flight performance dimensions, Bank Angle SD consistently showed the least impact from AR occlusion, with the smallest effect size. Potential explanations for this are discussed in Section 4.6, as well. Since bank angle showed no discernible statistical difference in tests, nor visual difference in plots, we elected to proceed in the analysis with only airspeed and vertical velocity.

Table 2: Study 1 Results Summary

Parameter	P-Value (Kruskal-Wallis)	Effect Size (ε^2)
Airspeed	0.293	0.087
Vertical Velocity	0.126	0.110
Bank Angle	0.741	0.052
Self-Rated Performance	0.242	0.030

Note. Epsilon squared effect size benchmarks: small = .01, medium = .06, large = .14 (Lakens, 2013)

Between airspeed and vertical velocity, we can once again examine the data for potentially relevant, but non-significant, trends. Figure 6 shows the mean variability for both airspeed and vertical velocity separated by the six conditions. Without considering turn direction, the bar chart of mean values shows modest differences in mean airspeed and vertical velocity SD between conditions. In this case, a higher value will represent a less stable flight, suggesting that the pilot was more affected by the occlusion condition. For both metrics, *the Bottom Right* occlusion condition resulted in the most variable performance. The *Bottom Left* condition produced the second highest variability in airspeed, but not in vertical speed. As previously described, participants could elect to turn in either direction, potentially affecting what was occluded by the AR display. To further understand the data, we must account for the potential influence of turn direction.

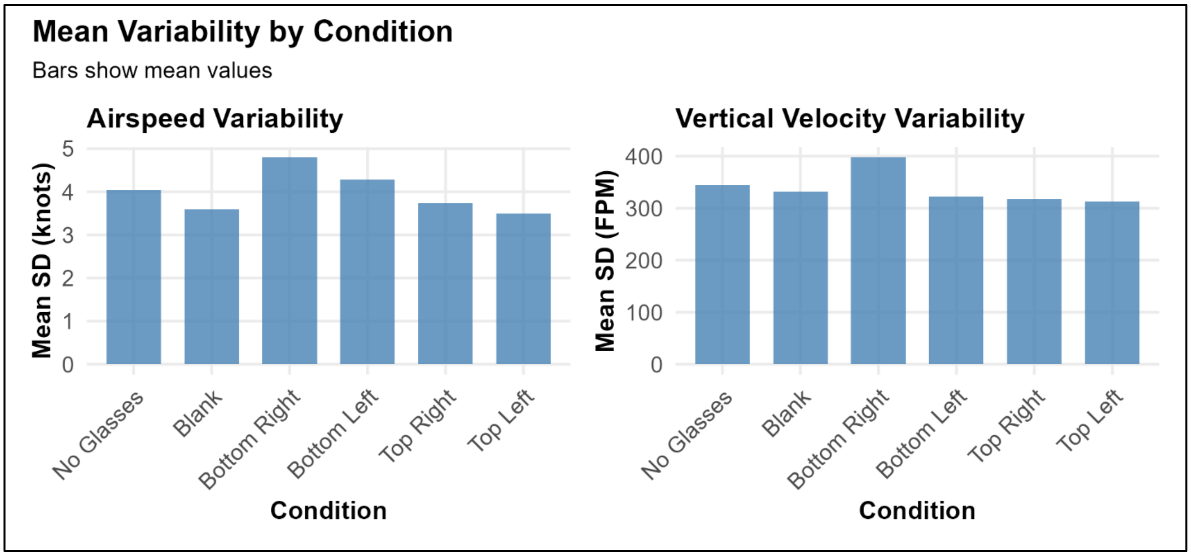


Figure 5: Mean Variability by Condition

Given that the same occlusion location will potentially affect the cockpit view differently for a right turn versus a left turn, it is important to compare performance by direction. From observing the experiment, participants tended to shift their eyes to the point where the horizon intersected the aircraft's dashboard. This meant turning their head to the right in a right turn and left in a left turn (visualized in Section 4.6). This comparison is presented in Figure 6. There is a consistent (but non-significant) trend that pervades both airspeed and vertical speed. For occlusions in the top half of the field of view, turns in the same direction as the occlusion are more variable (higher mean SD). Conversely, when the occlusion is in the bottom half of the pilots' field of view, turns in the opposite direction are more variable. This is true, to varying extent, for every turn condition for both metrics.

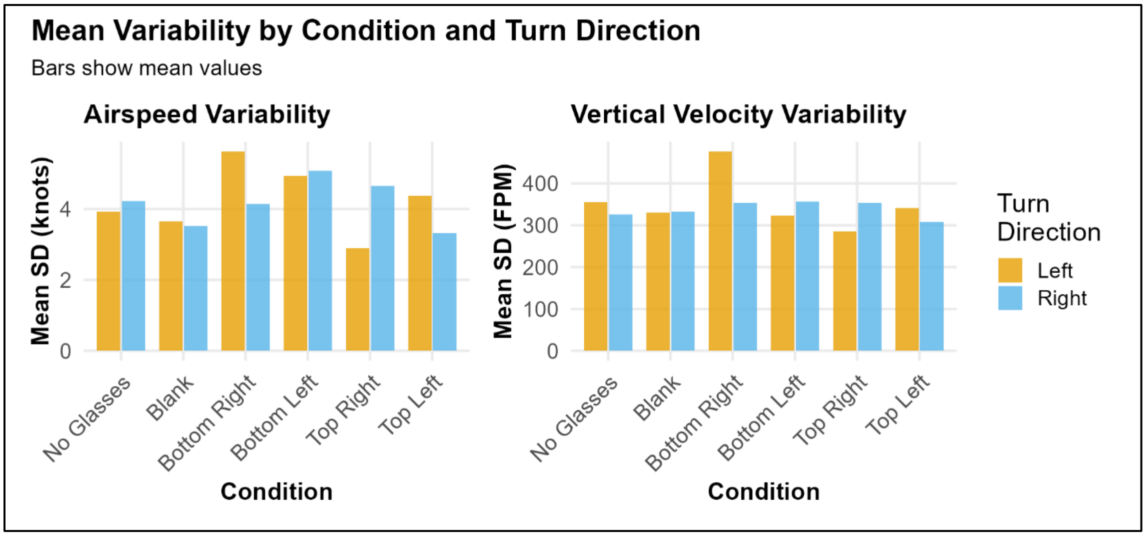


Figure 6: Variability by Condition and Turn Direction

To compare variability between the two turn conditions, we treated turn direction as a blocking variable. This way, any variability due to turn direction may be accounted for. With this source of error reduced, the effects of the AR occlusion condition can be further isolated. A visualization of this analysis is presented in Figure 7 as a series of box plots by condition. The *Bottom Right* occlusion produces the most variability for both metrics, with *Bottom Left* occlusion producing the next most for airspeed. Although the trends are more pronounced than in the baseline *No Glasses* and *Blank* conditions, these results are still not statistically significant. The significant overlap seen between box plots in for each flight parameter further reinforces the conclusion of the Kruskal-Wallis Test. No significant difference exists between the SD of airspeed and vertical velocity by occlusion condition.

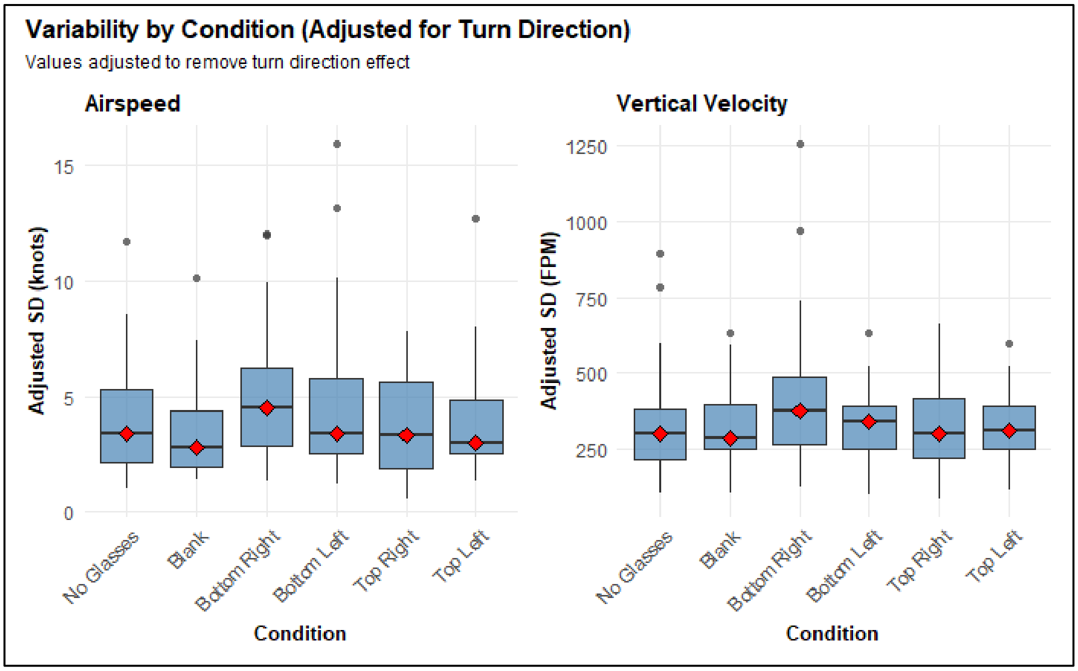


Figure 7: Variability by Condition (Adjusted for Turn Direction)

4.6 Discussion

Although the discussion will examine potential sources and implications for the trends observed, no effects were statistically significant. Thus, we can reasonably conclude that pilots can compensate for occlusion due to an AR display in basic flying performance. In other words, the placement of an AR display in the visual field did not significantly affect the pilot's performance in the fundamental task of manual control. Pilots did not perceive their performance to be significantly different for any AR occlusion location, which was validated by the objective data. This conclusion supports the investigation of the subsequent study. The effects of AR display position on performance should be negligible, allowing us to assess the influence of what is presented on the display.

As we examine trends in the data, we will consider only relative differences in average SD values for flight parameters. This approach was selected because the SD values do not carry inherent meaning in scoring or rating individual performance. In other words, we do not have a threshold value in standard deviation to classify or delineate "good" versus "bad" performance, given that performance is affected by many factors. These factors may include flying experience, simulator experience, manual control strategies, and motivation, which we cannot feasibly account for. Therefore, using relative changes in performance is far more appropriate for this analysis. Higher variability suggests less stable flying by the pilot, suggesting some influence of the display.

Overall, flight parameters with fewer redundant sources seemed to be more sensitive to the effects of occlusion from the AR display. In the steep turn, there are several perceptual inputs to determine bank angle, including vestibular sensation, bank angle indictors on the primary flight display (PFD), and the horizon. Even in the simulator, both instruments and a view of the horizon are available. Obscuring one source through the AR occlusion (the horizon for “top” conditions or the PFD for “bottom”) did not significantly affect the pilots’ consistency. On the other hand, pilots have fewer cues to assess airspeed in a relatively stable maneuver like the steep turn. Without a clear view of the PFD, pilots seemed to be less stable in this dimension. This may present a useful framework for considering both the location and content for an AR display. The AR display can be positioned in an area which primarily offers information that is available through several sources. If the AR projection must partially obscure some foundational source of information, it can improve redundancy by displaying it through the device.

Although subtle, there seemed to be a minor discrepancy between perceived performance and objective performance between the *No Glasses* and *Blank* conditions. Participants tended to rate their performance approximately half a point lower, on average, when wearing the AR glasses, even with nothing displayed on them. This is worrisome, as implementations of AR displays are at a disadvantage if the physical device itself influences performance. However, no significant trends exist in objective flight performance between the conditions. Simply wearing the glasses does not seem to impede pilots in simple manual control.

The most informative trend with respect to display location influencing flying performance came between the top and bottom of the visual field. Variability tended to be highest for occlusion lower in the visual field, especially on the right side. This was a somewhat unexpected result, as basic turning maneuvers can be accomplished primarily by looking outside the cockpit. Nevertheless, AR display occlusions within the cockpit area seemed to be more detrimental to pilots. The *Bottom Right* and *Bottom Left* conditions were likely to obscure the PFD, especially in opposite direction turn from occlusion (i.e. left turns with bottom right occlusion, or right turns with bottom left occlusion). We suspect this may be because pilots tended to shift their eyes towards the point where the horizon intersected the window/engine cowling. For a right turn, this was approximately the center of the cockpit. For a left turn, this intersection point was farther to the left. This can be attributed to parallax from the seating position, as the pilots were flying from the left seat exclusively.

The visual reference technique, using the intersection of the cowling and the horizon, provided a stable visual reference to assess the consistency of the turn. Any change in the location or angle at which the horizon intersects the cowling suggests a change in pitch or bank angle, respectively. However, by looking at this point, occlusions in the lower half and on the opposite side obscured the flight instruments. The airspeed and vertical velocity variability by turn condition reflects this, as each condition resulted in more variable flight when the turn was made in the opposite direction of the occlusion (e.g. a left turn with a *Bottom Right* occlusion). A mockup for this condition is shown in Figure 8. Note the red circle showing the visual reference point, and that the display in the bottom left, containing airspeed, altitude, and bank angle references, is inaccessible.

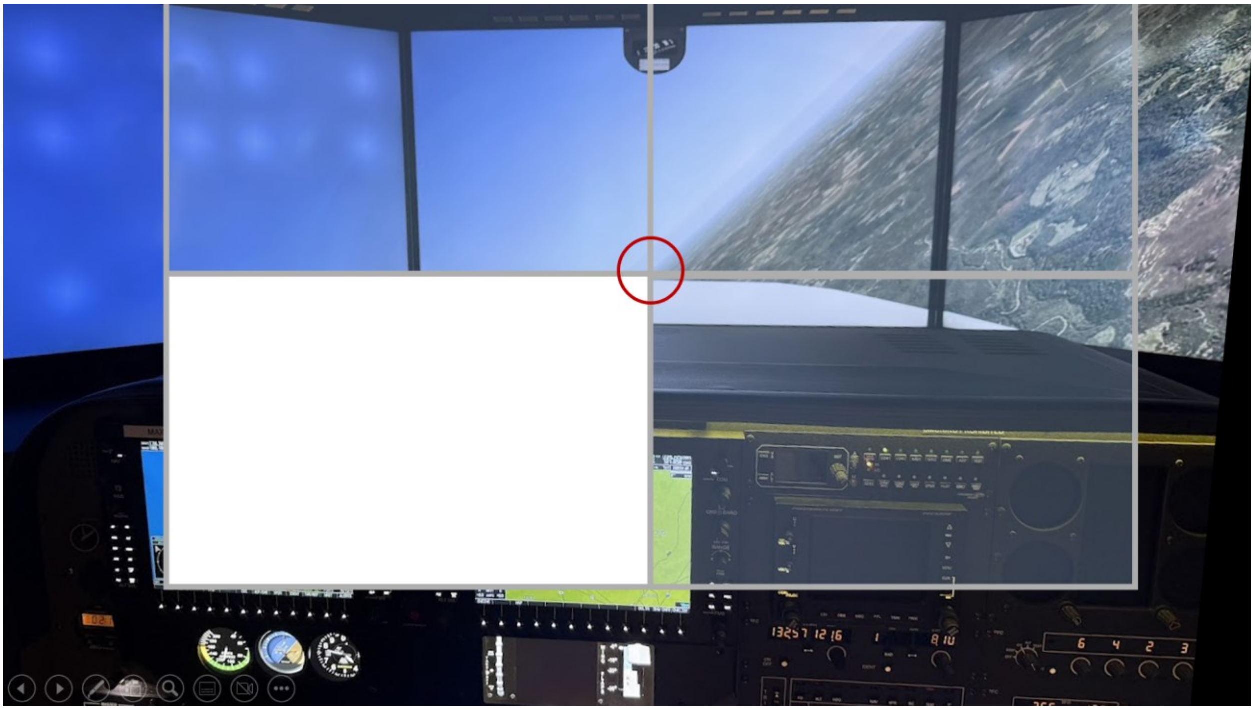


Figure 8: Bottom Left Occlusion Opposite Turn Direction Mockup

Pilots seemed to be less influenced by occlusion in the top portion of the cockpit. However, we did not examine effects of AR occlusion on visual scan or detection of other aircraft. This does represent a limitation in our basic study, as clearing for other aircraft and navigating based on ground references is an important task. Nevertheless, the research only intended to examine the effects of AR display location on basic manual control performance. A more realistic flying environment with other simulated traffic might increase the impact of occlusions in the top half of the visual field.

An additional limitation of the study relates to the tradeoffs in using a simple commercial off-the-shelf AR device. First, the use of commercial off-the-shelf AR glasses precluded the use of eye-tracking equipment. Without data on where participants were looking, we cannot be sure what strategy they used to compensate for the AR occlusion. Since the projection was device-referenced, participants could move their heads to view any occluded area. Tracking head movements may have offered additional insight into how pilots compensated for the occlusion. Nevertheless, the captured measures of performance might reflect this challenge; adjusting head position could theoretically affect the visual references for the steep turn task, adversely impacting performance. The performance of basic maneuvers represents a distinct sub-task in the complete cognitive and physical task of flying an aircraft. It is still valuable to note the trend of increased variability when obscuring references with few redundant sources and for AR display locations in the bottom of the visual field, though the effects were not significant.

4.7 Conclusion and Lessons Learned

In conclusion, our results indicate that the presence and location of an AR occlusion do not seem to significantly affect pilots' manual control performance. In other words, pilot foundational skills seemed to be robust to moderate visual occlusion from AR displays, even when access to cockpit information sources was disrupted. Participants were able to compensate for occlusion of large segments of their field of view while completing a consistent steep turn. While previous analyses have examined the costs of incorrect information presented as an AR overlay on performance (Warden et al., 2024), the present research extends this work by evaluating the cost of the overlay itself.

The methodology of this study serves to isolate the effect of an AR display's presence on flight performance. The use of a fully opaque shape to occlude the field of view represents a worst-case scenario in terms of visual access. With a smaller or semi-transparent AR display implementation, the risk of a decrement in pilot performance may be even lower (Smith et al., 2021). Because we implemented AR graphics with no value to the flying task, any potential benefit provided by AR is intentionally excluded, allowing us to isolate the negative effect of occlusion on flight performance. Therefore, this work demonstrates that the cost of occlusion may be sufficiently small that incorporating an AR display that supports pilot performance may well be worth the tradeoff.

Although no significant results were observed, the trends in self-assessed and objective performance suggest that occlusion in the lower visual field may be slightly more disruptive to pilots. These observed human-machine interaction characteristics help to inform AR graphic implementations for future research. By establishing that pilots are able to sufficiently compensate for visual occlusion due to AR displays in a naturalistic setting, future research can investigate more critical design components and interactions. For the purposes of the coming experiment, the assistive AR display implementation can theoretically be located anywhere in the visual field. Based on identified trends**,** we elect to place the projection in the top half**,** as flight performance was more variable and self-assessment was lower with occlusion in the bottom half.

Several lessons from the failure to achieve significant results in this preliminary study helped to inform design elements of the second study. In the experiment, participants were free to choose their turn direction, airspeed, and altitude. This invited several additional sources of variability between participants, making the task of standardizing results more difficult. To address this lack of experimental control, participants in the second experiment were given specific benchmark objectives. With consistent goalposts for airspeed, for instance, performance can be more precisely compared. The flying task was also more complex and representative in Study 2, supporting our efforts to identify a substantial difference between AR conditions. In terms of data collection, the task of manually separating condition from a continuous data stream led to the loss of three participants' data. By changing the procedures and simplifying the data collection methods, the second experiment was not affected by these constraints.

5 Study 2: Impact of Automated AR Technology in Normal and Emergency Flight Scenarios

5.1 Introduction

Given that the presence of an AR display did not significantly affect pilot performance, the next step in our research was to examine how the content presented on an AR display may affect pilot behavior. AR displays can provide supplementary information or automated cues to pilots. With optical see-through glasses, pilots retain their view of the outside world while gaining a highly salient viewing display. The presence of an AR display will likely change the way in which pilots process information in the cockpit (Wickens, 2002). Before AR displays can be widely implemented into commercial and general aviation cockpits, we must understand both the expected advantages and disadvantages of the technology from a cognitive perspective.

Although positive effects of AR displays in reducing workload and improving usability have been theorized and identified (Li et al., 2020; Rowen et al., 2019; Yuan, Ng, Lau, et al., 2024), we seek to address the gap in understanding how AR affects pilots when the information presented is not helpful. Specifically, we assess the effect of an AR display on pilots while handling an in-flight emergency (in the form of a loss of engine power). In this scenario, the AR display and associated automated system did not recognize the system failure, and therefore continued to provide cues in accordance with the originally planned flight. This construction serves to represent a specific, plausible instance in which the goals of an automated system may become disjointed from the user's goals and context. Under the premise of an emergency, it also intends to create a more demanding, stressful situation for the pilot. Since human pilots are uniquely capable of responding to dynamic emergencies to maintain safety (Anand, 2025; Bailey et al., 2017), it is vital to understand how the presence of an assistive AR display affects core skills. To balance this assessment, we also assess whether AR can be helpful to pilots in normal flight conditions, in which the AR cues align with their tasked course of action. In all, this work seeks to extend our understanding of AR displays and the impact of system degradation in the complex, cognitively demanding aviation environment.

5.2 Experimental Design

The study design was approved by the IRB, in conjunction with parallel experiments by the Aerospace Engineering Operations aNalysis (AEON) Lab. As with Study 1, this experiment followed other data collection events. Depending on eligibility, participants completed either one or two stations before arriving. These stations used VR systems and entirely different flying tasks. We do not suspect that the order of experiments had an impact on participant performance in this study. We again gave participants the opportunity to take a break between stations, mitigating any effects of fatigue.

The study used a two-phase mixed design. In Part 1, a within-subjects design was used, with all participants experiencing both AR and No AR conditions; presentation order was counterbalanced across participants to control for order effects. In Part 2, the design shifted to between-subjects, with participants randomly assigned to one of two conditions: AR or No AR. For each trial within the study, participants executed an appropriate traffic pattern and landing. The phase was based on aircraft status, which two levels: no issue (normal) and engine-out (emergency). In Section 5.2.1, we discuss the design of each of these factors in context of the experiment. Then, in Section 5.2.2, we describe the development and implementation of the AR cues for the experiment.

5.2.1 Emergency Condition Design

The inclusion of an emergency condition provides unique and valuable insight about the impacts of AR displays across the range of operational scenarios. Emergency situations are demanding, and may affect pilot response time and ability to correctly apply procedures (Landman et al., 2017). However, the use of an unexpected emergency in the simulation required tradeoffs with the study design. In creating the study, we expected that participant behavior would be altered after the emergency condition was presented. After a simulated emergency, we expected that pilots would be unrealistically vigilant for potential system failures, perhaps changing their typical flying or instrument scanning behavior. This could degrade their normal pattern performance in a way that is not representative of their typical flying. Physiologically, any lingering effects from the stress of the simulated emergency could affect flying stability and perceived workload for the remainder of the experiment. Pilots, especially novices, show a notable sympathetic anxiety response when faced with an unexpected emergency (Santos et al., 2022).

Conversely, the ability to anticipate an emergency could unrealistically improve pilots' flying performance and decision-making in a subsequent emergency scenario. Pilots that are familiar with an urgent emergency situation do not show the same anxiety response or performance decrement (Peng et al., 2025). These impacts also preclude two emergency conditions from being presented, as experiencing one may affect behavior on the subsequent exposure. The process of synthesizing information and making an assessment (especially against incorrect AR cues) would be compromised in successive exposures. Nevertheless, we aim to understand which anticipated impacts on workload, situational awareness, and performance can be attributed to AR in an emergency versus the emergency condition alone. The effects of this response, whether positive or negative, could linger across study conditions. For these reasons, a single emergency condition must occur after the others.

As a result of these constraints, we used a between-subjects design for the emergency landing condition in the study. The normal landing conditions were counterbalanced in order and flown by every participant to begin the experiment. In the normal condition, we assessed the effects of AR assistive cues using a within-subjects approach. The study design did not introduce any specific order effects between participants due to recruitment or scheduling, so the starting condition (AR or no AR, normal landing) alternated between every other participant. The design is outlined in Figure 9.

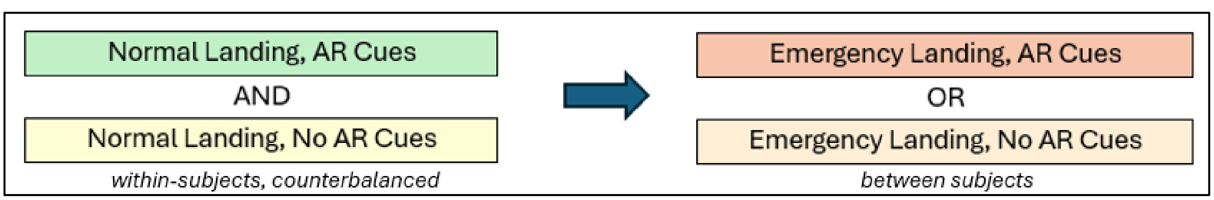


Figure 9: Study Experimental Conditions

5.2.2 AR Assistive Cues

The cues, or individual graphics presented through the AR display, present the items on the "Before Landing" checklist sequentially. This checklist was adapted from the Auburn School of Aviation Cessna 172 maneuver standards (*Cessna 172 Skyhawk*

Standardized Maneuvers Manual, 2018). The checklist is based on a standard traffic pattern. To assess the utility of an AR display implementation, we developed a notional automated interface for the AR glasses. The cues represent a potential implementation of AR in the cockpit, in which an automated system provides information based on the planned flight. For trials without the AR display, participants were provided with the paper version of the checklist and maneuver standards. A small excerpt of the maneuver standards document is provided in Figure 10, while the complete documents are presented in Appendix C.

6. On the base leg, set the flaps to 20°, and establish and maintain 70 KIAS (3° nose down). Trim as necessary.
7. Visually verify that the final approach (including the extended final and the opposite base leg) is clear, and turn final while maintaining 70 KIAS.

Figure 10: Maneuver Standards Excerpt

For the study, assistive information was created specifically for the course pilots flew in the simulator. We developed graphics through Microsoft PowerPoint, which enabled information to be presented directly through the AR glasses. This also enabled a Wizard-of-Oz approach, as fully integrating the glasses with the simulator was out of scope. The age of the simulator made such integration infeasible, as well. A representative sequence of the AR display is provided in Figure 11, corresponding to the maneuver standards shown in Figure 10. Unlike the sample image, the background was transparent for pilots.

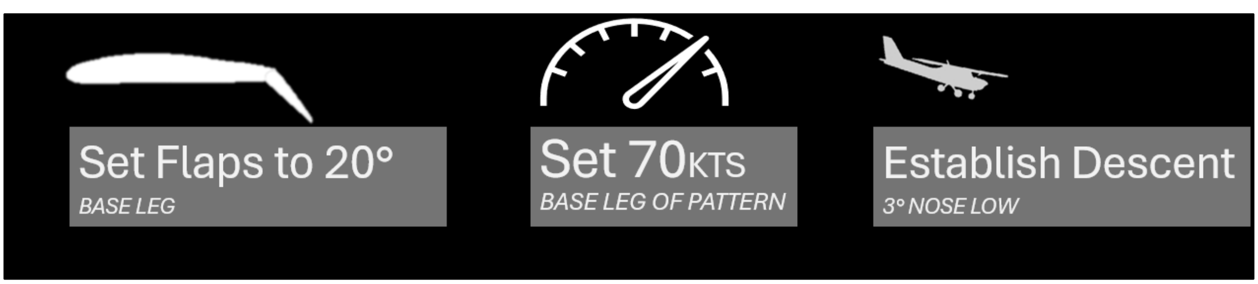


Figure 11: Sample AR Cue Sequence

The results from the first study (Section 4.5.2) reflect that the location should have no significant impact on the pilots' performance. However, avoiding the lower half reduces the risk of AR interference with cockpit displays that may have driven some of the trends observed in Study 1. Based on those results, the display was located on the top right side of the pilot’s field of view, where fewer important instruments and outside visual references are likely to be obscured.

To ensure consistent presentation of cues, each checklist item was only presented when participants reached specific locations within the simulated flight (shown in Figure 12). The use of location-dependent cues simulates a simple trigger-based automated system. This reinforces the fact that the notional automated system would not adapt to an altered flight plan, having no context of the emergency. If participants deviated from the prescribed location in a normal flight (e.g. turned earlier than expected), the sequence skipped to the next cue. Each cue was presented for approximately five seconds, then removed. No graphics were displayed for approximately two seconds, then the next cue was presented. These times were selected arbitrarily based on the total time available in the flight. The timing was adjusted for pilots flying faster (or a shorter route) than expected, with intention of keeping an AR cue in the visual field for the same proportion of their flight. The complete progression of cues is included in Appendix A, in which the slide number corresponds to the numbered boxes in Figure 12.

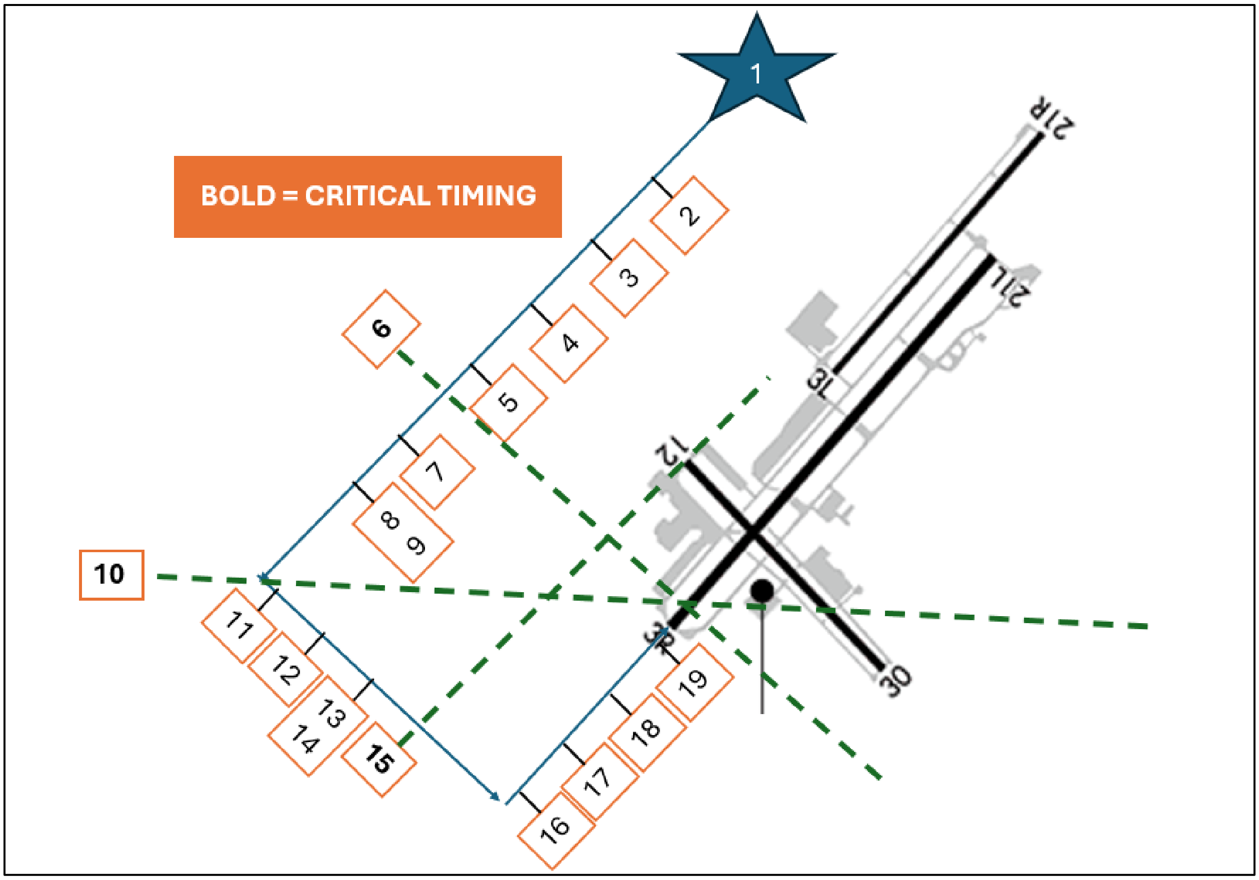


Figure 12: AR Cue Location Trigger Map

This design extends previous studies (Li et al., 2020, 2022) that adapt the checklist to an AR display. By integrating checklists into sequential graphical cues, we intend to reduce the cognitive load on pilots compared to synthesizing the information themselves. The one-by-one presentation of graphical cues reduced the amount pilots had to read, and therefore direct their attention to, at any given point. It also allowed for a high volume of novel cues to be presented over the course of the flight. In the emergency condition, this higher volume also required pilots to parse more individual cues. This increase in task demand intended to provide a higher potential to detect differences in performance.

To examine the effects of irrelevant information presented through the AR display, the same assistance cues were presented in the emergency landing condition. With less energy, participants were forced to fly towards the airport sooner, requiring adaptation and real-time judgement to fly a safe pattern. In this case, the AR cues were not necessarily applicable or even achievable. Just as pilots were unable to fly the standard pattern, the associated checklist items became irrelevant given the urgent emergency (like adjusting the aircraft lights) or impractical given the loss of engine power (like flying at 90 knots airspeed). In this way, the previously relevant, helpful AR assistance cues turned into distractors. As the pilots descended and approached the runway, they saw, in succession, all the information they would see for the normal flight condition. This required pilots to read, process, and decide to ignore the display. Pilots did not control what was currently displayed, as all cues followed the same script and sequence. Though cues were presented more quickly than the normal flight scenario (to account for the shorter flight time), we strove to present graphics the same proportion of the flight as the normal condition.

5.3 Participants

Participants were recruited from the Auburn Aerospace Engineering and School of Aviation programs. Participants were not compensated for their participation. Recruitment for the experiment took place over two rounds to accommodate IRB revisions in a separate part of the data collection event. Participants who required eyeglasses to fly were excluded, as the AR glasses were not compatible with eyeglasses. Participants (n = 15) were licensed pilots or student pilots with at least 30 flying hours logged. Of the participants, twelve were male and three were female, with mean flight hours 173.2 (SD = 101.1). All participants had experience in Cessna 172-type aircraft. Fourteen of the fifteen participants had previous experience in a flight simulator. Fourteen participants had flown within the last month, with only one participant reporting their last flight a year or more in the past. Only one individual had participated in previous data collection events with the laboratory. Since both the flying task and the AR implementation had changed since Study 1, we did not exclude this participant from subsequent analysis.

5.4 Apparatus and Materials

For practicality, experimental control, and safety, the study utilized a flight simulator, previously described in Section 4.3.2. The simulated flights were conducted in a DCX MAX simulator, representing a Cessna 172S. This fixed-base simulator was previously FAA-certified for flight training, reflecting a high level of fidelity in interface and flight characteristics. The simulator runs the commercial version of XPlane 9, allowing for simulation metadata to be streamed continuously. Within the simulator, printed documents were provided on a standard kneeboard, a small clipboard that can be fastened to the thigh for quick reference. These documents included the checklists, pattern diagram, and airport diagram, which are included in Appendix C.

For trials with AR cues, participants wore XReal Air 2 Pro glasses, an optical see-through device with a 46-degree field of view (FOV) and a 1920*1080-pixel resolution per eye (*XREAL Air 2 Pro*). They are representative of typical commercial off-the-shelf AR glasses available today. The glasses are shown in Figure 13.

Figure 13: XReal Air 2 Pro Glasses

The glasses mirrored the display of a connected laptop computer, allowing for dynamic researcher control. A six-foot cable connected the glasses to a laptop, allowing the researchers to remain out of view of the participant while still controlling the display. The graphical cues were developed through PowerPoint, allowing researchers to progress sequentially through the checklist as the flight progressed. Each checklist item was accompanied by a representative icon, representing a change in airspeed, control setting, glideslope angle, or heading. The graphical cue also provided any necessary supplementary information (like the engine revolutions per minute (RPM) to set to achieve a specific airspeed). The color, font, and size of graphics were selected to ensure visibility when projected over the screens in the simulator.

5.5 Measures and Data Collected

To assess the effect of AR on pilot behavior, data analysis included three primary lines of effort. First, objective data from the simulation showed the flight performance of each participant for their assigned conditions. In considering the variability of objective flight data, we considered higher performance to have lower variability over the course of the flying task. Second, self-assessment scales quantified the performance-mediating factors of workload and situational awareness. Third, subjective questions on usage, usability, trust, and user experience helped to provide context to the preceding data sources.

Flight data was collected through a stream from the XPlane 9 program to a separate laptop as an XDF file. Each trial, from the beginning of the pattern to landing, used a separate file. Each file was labelled using the participant number and the experimental condition. Using the pipeline developed by the AEON Lab, relevant data from the XDF files were sampled and converted into CSV files for data analysis (Rahman & Schulmeyer, 2025). The continuous data streams were sampled at a frequency of 4 Hz (or, four samples per second).

The experiment also collected data through surveys. Participants completed a survey between each trial, which included an adapted NASA TLX and the SART (Hart & Staveland, 1988) (Endsley, 1995). While the SART questionnaire used the traditional 7-point scale, the NASA TLX used a 1-100 scale. This approach matched the NASA TLX implementation used at other stations during the data collection session. Since no information was lost from smaller graduations on the scale, and we intended to mitigate confusion from participants, we maintained consistency with the other experiments.

The NASA TLX consists of six subscales: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration (NASA Human Systems Integration Division, 2020). To calculate a composite score, we rely on an overall (unweighted) mean value, also known as raw TLX (Hertzum, 2021). While the original TLX was designed for weighted subscales based on the relative importance to the task, the procedure was found to be ineffective, and is therefore not generally used today (Hertzum, 2021).

The SART consists of three primary dimensions: Demand, Supply, and Understanding (Bolton et al., 2022). Instability, complexity, and variability contribute to Demand, while arousal, concentration, division of attention, and spare capacity comprise Supply, and information quality, information quantity, and familiarity comprise Understanding (Taylor, 1989). SART can also be considered as an overall composite, calculated as (Understanding – (Demand – Supply)) (Bolton et al., 2022).

For each trial, the survey asked participants to assess their usage of the assigned checklist: paper for No AR trials or AR. The survey provided a scale from 0-100 and asked participants to estimate 1) the percentage of checklist items looked at, and 2) the percentage that were useful to them. This section aimed to assess the rate at which participants used the checklist between the No AR and AR conditions. The reporting was based on the participants' own estimations, which is discussed in Section 5.9.7, Limitations.

Following these measures, the study also used subjective inquiries. After the three trials, participants answered a series of questions focused on trust, perceived usability, and preferences (included in Appendix C). These summary questions, in addition to a free-response question, intended to capture insights not accounted for by measures of workload or situational awareness. Participants were provided with answer choices based on a Likert scale, appropriate to the given question's context (also included in Appendix C). Specifically, the questions were selected to capture preference

(Q1), ability to detect deviations (Q2), perceived utility for normal and training flying operations (Q3,4), effect of AR system on visual scan and ability to acquire information (Q5,6), and trust of the AR system (Q7).

The data contribute to a holistic assessment of the AR system. While it would be difficult to draw conclusions from the result of a single measure, the use of multiple measurements may yield clearer insights. Even subtle trends, when carried across performance, cognitive, and experiential measures, can help to indicate an influence of AR in the study. For this reason, we have collected data from the simulator, established numerical scales for workload and situational awareness, and subjective survey questions. We proceed with a description of the procedures in the study.

5.6 Tasks and Procedures

Upon arrival to the lab, participants reviewed informed consent documentation (included in Appendix B), an introduction video, and a brief demographic survey (included in Appendix C). Since recruitment was open to all individuals, we assessed eligibility for the study by inquiring about flight experience. Participants with insufficient flight experience (<30 hours) were able to take part in other portions of the data collection event, but were excused from this experiment. Before beginning the data collection, participants sat in the flight simulator and adjusted the seating position to reach all controls. Participants were given time to familiarize themselves with the specific cockpit arrangement in the simulator.

Participants then received a standardized briefing describing the experiment, including the flying task for each trial (included in Appendix C). After answering questions about the scenario or flight simulator setup, the specific flight simulator scenery was loaded. The setup for the simulator is discussed in the following paragraph. If the trial included the AR display, participants were presented a simple image of a landscape to check that the glasses were functioning. Wearing the glasses too high or low on the face could cause the image to be distorted or cut off, so participants adjusted their placement, as needed. This check ensured that the glasses were functioning properly and that the cues would be legible. For trials which did not involve the AR glasses, we instructed participants to reference the provided printed checklist, pattern diagram, and airport diagram instead.

For the emergency landing, pilots experienced a full loss of aircraft engine power, requiring them to adjust their flight path. Participants were not informed that the emergency would occur. To ensure that participants did not become preoccupied with fixing the engine failure (which was not possible in the simulation), the introductory briefing included the statement that, “you may deviate from the flight plan if safety dictates. In that case, land as soon as conditions permit.”

For each trial, the researchers first began the simulation data stream, then started the simulation at a pre-planned point. This point, and the subsequent path for the traffic pattern, is denoted by a star in Figure 14. Participants then began flying the traffic pattern, with the researchers progressing through AR graphics based on the predetermined location cues as described in Section 5.2.2. Representations of the pilots’ view of the AR cues are shown in Figure 15 (the complete cue sequence is included in Appendix A). For trials with the emergency condition, the researchers triggered an engine failure via the XPlane software after five seconds of flight time. This engine failure immediately caused the loss of engine power, and no action by the pilots could restore the engine to functioning. Pilots were not initially aware that they could not restart the engine, so some attempted to restart, as they would in a real emergency. Participants had to adjust their pattern to land safely, ignoring AR cues that were impractical or irrelevant.

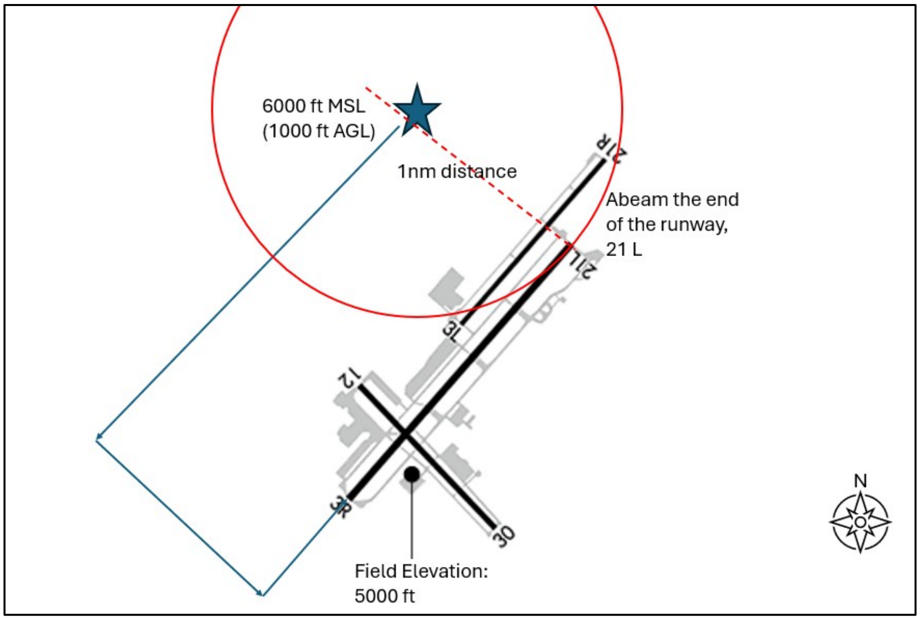


Figure 14: Pattern Setup Diagram and Starting Point

Figure 15: Pilot View of *Flaps* and *Clear* AR Cues

After participants landed or passed below the ground level away from the runway (which we described as a crash), the simulation and data stream were stopped and reset. Participants remained seated in the simulator and completed the post-flight survey. We verbally confirmed participants were willing to continue and were not experiencing symptoms of simulator sickness. Then, for willing participants, we repeated the process for subsequent trials. At the end of the three flights, participants exited the simulator and completed a post-experiment section of the survey. In addition to a set of questions about their experience with AR, participants were given the opportunity to offer any free-response feedback about the experiment. The duration of the experiment was approximately 20 minutes, on average (not including initial in-processing and demographic data collection).

5.7 Analysis

From the numerous parameters output by the XPlane stream, we selected several that capture the different dimensions of pilot performance. The selected data sources are included in Table 3. We selected these items to capture continuous variability in speed in both the forward and vertical axes (airspeed and vertical velocity) and approach stability over time (glideslope). With these dimensions captured, we did not identify any other data parameters that would be both fully independent of those selected and useful in analysis.

Airspeed refers to the forward velocity of the aircraft, while vertical speed refers to the velocity in only the z (up/down) axis. Glideslope is the angle of descent, considered as the angle between the horizontal (ground) and the aircraft's current position, with the vertex as the touchdown point. The glideslope in the final approach (when the aircraft is aligned with the runway) should typically be 3 degrees (*Federal Aviation Administration*, 2021). A steeper glideslope may be unsafe, as the aircraft is approaching the ground in less time (giving the pilot less time to make adjustments) and with more velocity (increasing the risk of a hard landing). For analysis, we specifically considered the glideslope below 500 feet, since the initial part of the pattern has a constant airspeed and no descent.

To assess performance, we use standard deviation when there is no consistent ideal value (airspeed and vertical speed) and mean value when there is (glideslope). These parameters, when considered at the summary level, are agnostic of the condition or strategy, since they represent the average or variability over the course of a flight. In other words, it does not matter whether the pilot completed the full landing pattern to obtain meaningful data from the standard deviation of these flight parameters.

Table 3: Selected X-Plane Data Streams

Data Stream	Real-Life Analog	Data Captured	Intent
sim/flightmodel/position/vh_ind_fpm	Vertical Speed Indicator (Variometer)	Variability (SD)	Consistency of vertical motion
sim/cockpit2/gauges/indicators/airspeed_kts_pilot	Airspeed (from Airspeed Indicator)	Variability (SD)	Consistency of forward motion
sim/flightmodel/position/vpath	Glideslope	Variability (SD)	Approach stability (3-degree target)

These assessments are useful for comparing outcome between conditions, but not necessarily for evaluating individual pilot performance. This suits the design of the study, but is worth noting to contextualize the data analysis approach we selected. We do not seek to classify individual trials as “good” versus “bad” performance, or compare relative to a safety threshold. One reason for this limitation is the pervasive impact of individual pilot decisions and strategy, like flying more quickly after an emergency to get on the ground sooner. A fast, yet controlled, airspeed is not indicative of lower pilot performance. While many factors could influence performance in the simulator outside of the experimental scope, like experience in similar simulator systems, comparing the variability for each participant by condition still allows us to consider the influence of AR.

To accurately capture pilot flight performance, we used several criteria to filter the data. First, we trimmed the first five seconds of the simulation from the data stream. During the first second of the simulation, it was common for the aircraft to pitch up or down unexpectedly as the simulator initialized. Pilots recovered from this initialization by holding (and therefore neutralizing) the controls, but this simulation error introduced variability that had to be removed from analysis. The flight stream was trimmed on the back end based on the altitude of touchdown on the runway. We did not include the period slowing down after touchdown, or roll-out, in the analysis. By this point, the aircraft was on the ground, meaning it had no vertical velocity or glideslope. It was slowing down, therefore affecting the integrity of the airspeed data, as well. This data trimming approach also worked for a crash-landing outcome, as well. By the point of crashing, pilots were at the ground-level attitude. We were not interested in any data that the simulation produced after the aircraft impacted the ground.

5.8 Results

To understand the data at a high level, we plotted the flight path for all participants by condition, shown in Figure 16. Through visual inspection, we verify that the starting and ending points are correct, and that no anomalous deviations are present. In the plot, each line represents an individual flight, or one trial. For the Normal scenario, there are fifteen lines for each condition. In the Emergency scenario, there are seven lines for No AR and 8 lines for AR trials, given the between-subjects design. This basic check ensures that the data is sufficiently clean for subsequent analyses.

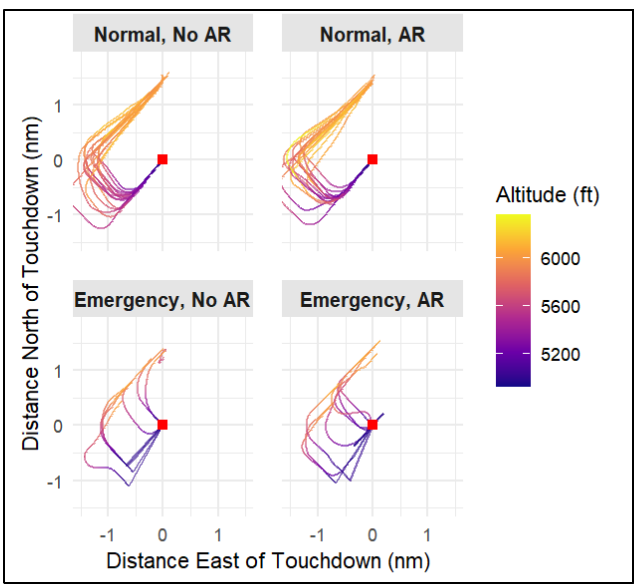


Figure 16: Flight Paths of Participants by Condition

Given the small sample size (n=15), assumptions of normality for inferential tests between conditions are not appropriate for these data. Instead of assessing the normality of each parameter, we used non-parametric Wilcoxon tests throughout. The

Wilcoxon Signed Rank Test places the absolute difference between an observed value and the suspected mean value in rank order of magnitude, then adds or subtracts these values together based on the sign of the original difference to form the test statistic *W* (Ramachandran & Tsokos, 2015). If this value exceeds a critical value (based on the sample size *n* and desired $\alpha = 0.05$), we can reject the null hypothesis of no difference between the observed and expected data. This approach supports both consistency and comparison in the analysis. In addition, Cohen's *d* is used to assess the effect size, or the extent to which a difference might exist (Goulet-Pelletier & Cousineau, 2018). Effect size can indicate whether a true influence may exist, even if the sample size was insufficient to achieve statistical significance.

5.8.1 Normal Flight Condition

To begin, we considered the results from the normal (non-emergency) flights. Wilcoxon tests for a difference between the means, shown in Table 4, yield p-values greater than 0.148 for each parameter. The effect sizes, using Cohen's *d,* suggest a small effect ($0 < d < 0.2$) favoring the no AR condition for glideslope and vertical velocity. The effect size for airspeed variability is close to zero. Therefore, objective flight performance data does not suggest a difference in pilot performance between the AR and no AR conditions. A visualization of the data, showing the change in flight parameter variability between the two flights, is presented in Figure 17. Most participants demonstrated a minor increase in variability while using the AR display. While five participants improved their glideslope and vertical velocity stability and seven participants improved their airspeed stability through use of the AR system, the majority saw a decline in performance. This change in performance was not statistically significant, but still represents a shift towards worsened performance.

Table 4: Variability of Flight Performance Parameters, Normal Flight

Flight Parameter	**No AR Median *[IQR]***	**AR Median *[IQR]***	**W**	***p***	**Cohen's *d***
SD Glideslope (deg)	2.84 [2.27, 3.51]	3.11 [2.57, 4.06]	34	0.148	-0.40
SD Airspeed (kts)	12.52 [10.61, 14.12]	12.87 [11.14, 13.52]	57	0.887	0.02
SD Vertical Velocity (fpm)	354.99 [296.18, 403.71]	379.61 [312.79, 472.11]	37	0.201	-0.40

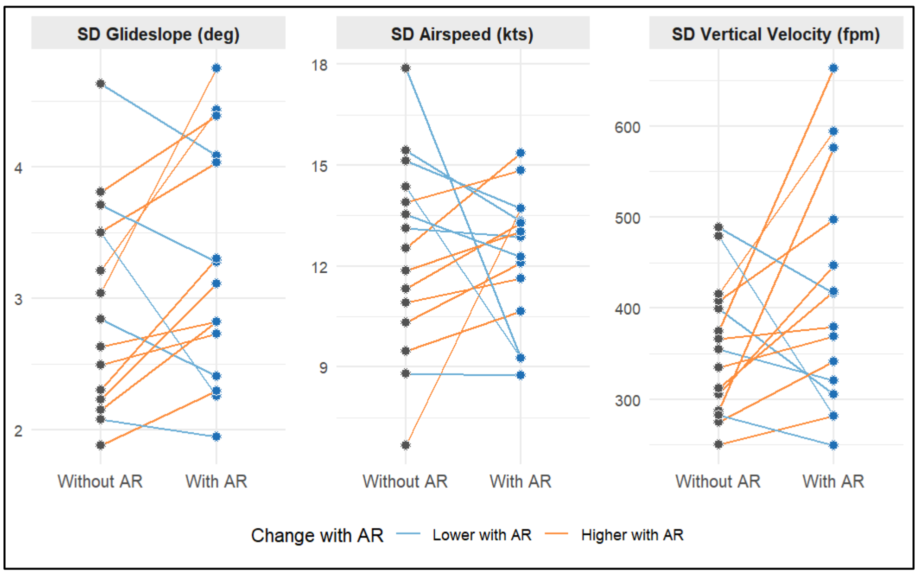


Figure 17: Variability of Flight Performance Parameters, Normal Flight

Though the mean values for glideslope, airspeed, and vertical velocity were not the focus of this investigation (as discussed in Section 5.7), results for central tendency measures are included in Appendix D, Table D1. We observe no difference in mean airspeed, glideslope, or vertical velocity, overall.

Moving to subjective measures, we consider self-assessments of workload and situational awareness for the normal flight condition. Workload, assessed through the NASA TLX, contains six subscales that were weighed equally in calculating a composite. The results show no significant difference between participant workload on the normal flight with and without AR using a Wilcoxon test (p = 0.887, Cohen's d =

0.088). The results also show no significant difference between the assessed situational awareness ($p = 0.529$, Cohen's $d = -0.184$). The summary level plot for these assessments is provided in Figure 18, showing a very slight increase in workload and decrease in situational awareness. For more granularity into the effects of the AR display, we assessed the component subscales of each measure.

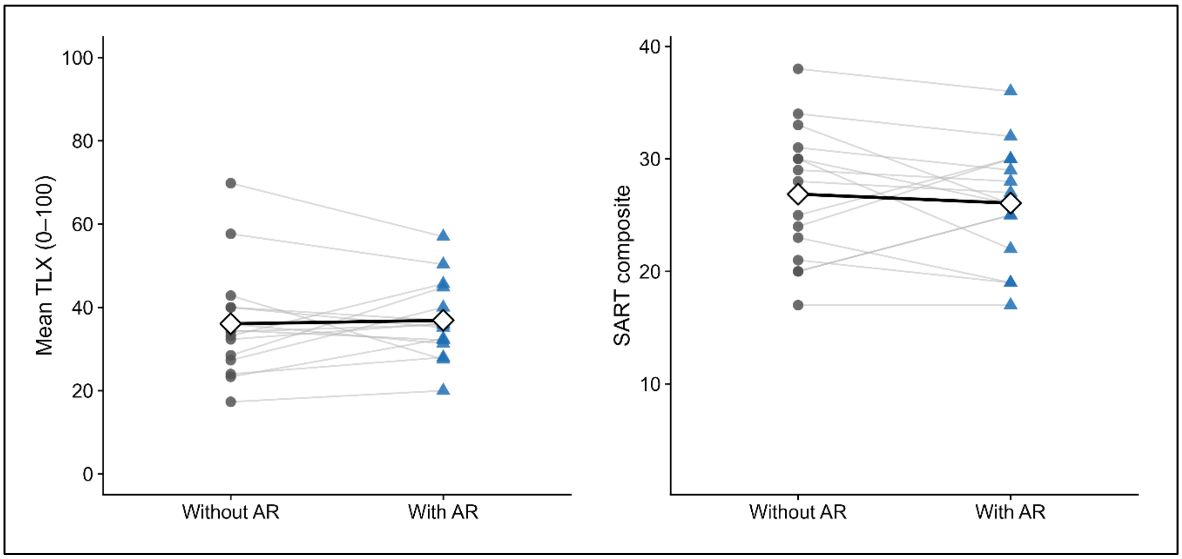


Figure 18: Normal Flight Overall TLX and SART

In general, the NASA TLX subscales show no significant difference between conditions ($p > 0.66$ for all subscales). One observed trend in the data was a slight increase in temporal demand, with a mean score of 26.9/100 without AR and 31.4/100 with AR. This may represent the moderate increase in time pressure participants felt while using the AR display. While we observed no difference between the AR and No AR conditions at the summary level, there are minor (non-significant) changes at the subscale level. The change in TLX subscale scores between the two flights, as well as the difference between the mean values, are depicted in Figure 19. Participants showed slight increases in mean mental demand (0.7), physical demand (0.4), and temporal demand (3.3) in the AR condition. Mean performance showed an increase with the AR display (2.2), while the effort required decreased (-3.4) and frustration slightly increased (1.7). While the change between conditions for individual participants still shows mixed results, changes at the summary level move in the improving direction for performance and effort, despite adverse impacts of the AR display for other subscales.

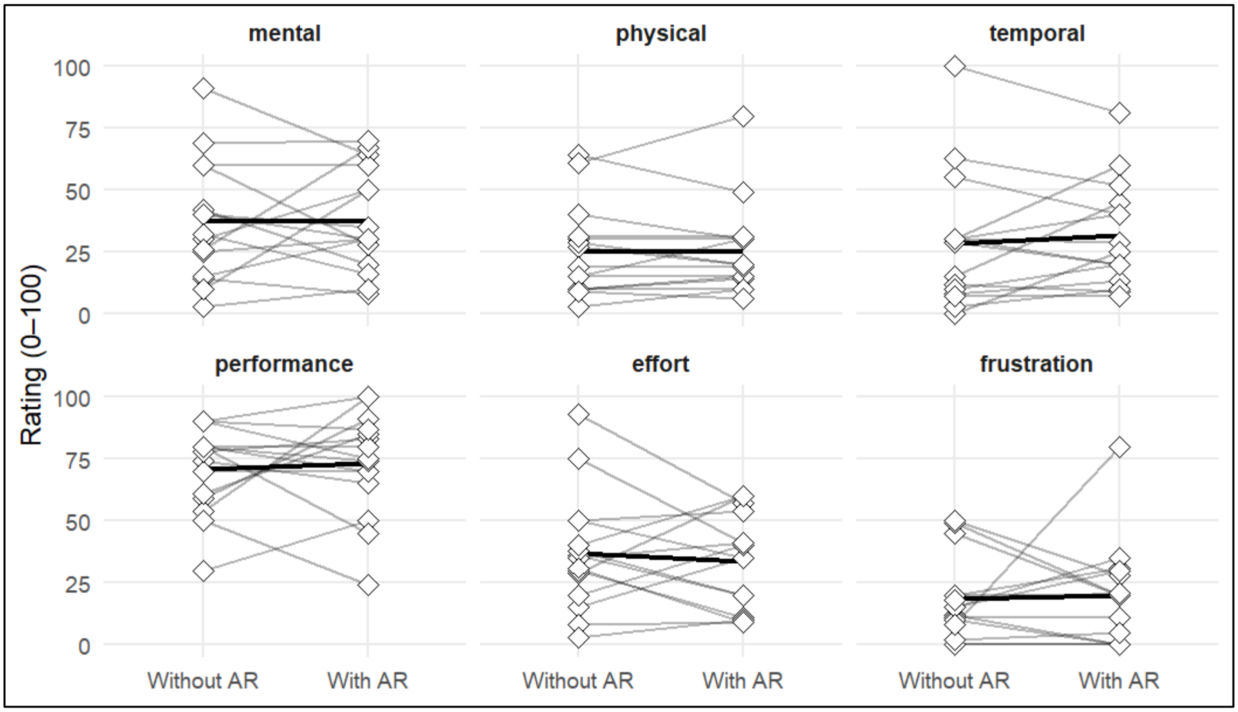


Figure 19: Normal Flight TLX Subscales

Self-assessed situational awareness yielded similar findings, as no significant difference exists between conditions. A Wilcoxon test for the three SART sub-categories – demand, supply, and understanding, yields p-values of 0.176, 0.115, and 0.154, respectively. Therefore, the difference between AR and No AR conditions fails to reach significance at $\alpha = 0.05$. The results for individual subscales are included in Figure 20, including the mean value for each subscale by condition. The results suggest that the presence of the AR display may have slightly adversely affected demand, but modestly improved supply and understanding. Between the AR and No AR conditions, the results suggest a slight decrease in arousal and an increase in information quantity.

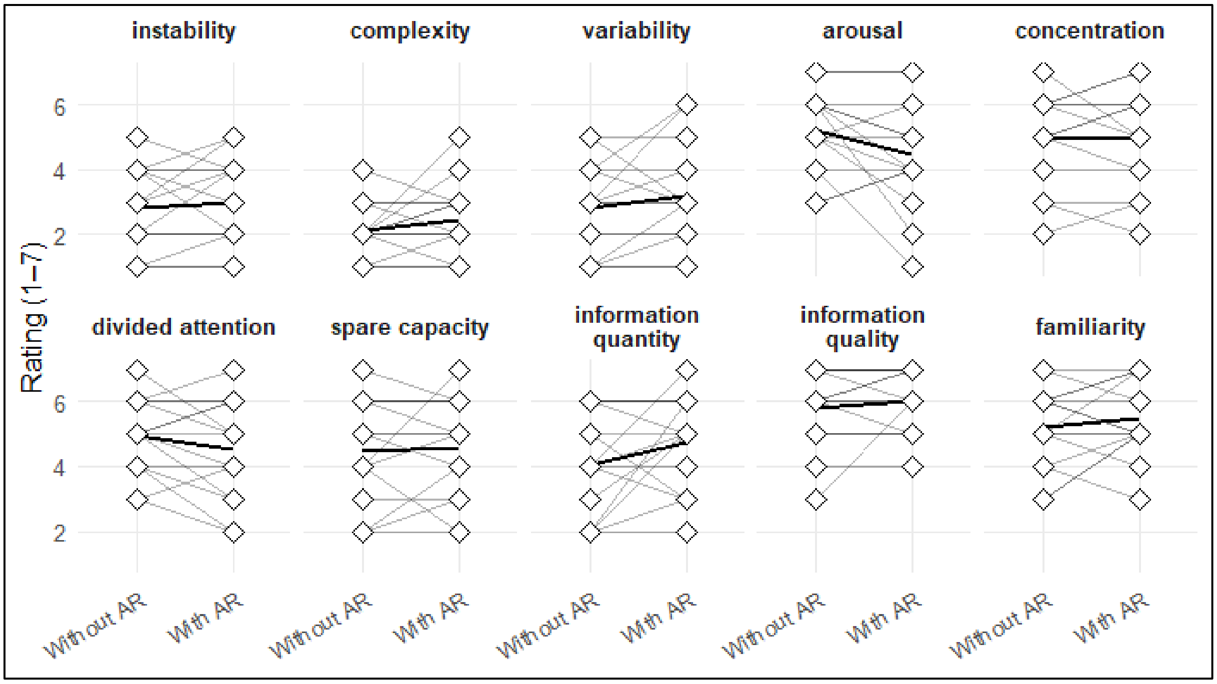


Figure 20: Normal Flight SART Subscales

The complete (unabridged) results of the statistical testing, assessing differences by condition, for workload and situational awareness are included in Appendix D. This section includes both the composite results and the subscales comprising both TLX and SART. For all measures assessing elements of workload and situational awareness, there is no significant difference between AR conditions. We rely on nonparametric methods for these analyses, since the self-assessed data lies on an ordinal, discrete scale. While mean values are valuable for visualization, the statistical assessment of results relies only on median values.

Despite having no statistically significant effect on participants' flight performance, workload, and situational awareness, the AR display did seem to affect how participants interacted with the checklist. Participants reported the percentage of items from the checklist (on the AR display, or on paper for the No AR condition) that they looked at, as well as the percentage of items that they found useful. In this comparison, we found a 13.1% increase in checklist items viewed, as well as a 5.3% increase in checklist items considered useful. A Wilcoxon test of mean values to assess whether a true difference may exist yielded p-values of 0.025 for percentage of items looked at and 0.533 for percentage of items found useful. This result suggests that the AR display increased the rate at which participants visually attended the checklist. Since the checklist items were identical, it is unsurprising that there is no difference between the percentage that participants found useful. The change in mean checklist usage and perceived usefulness between conditions is shown in Figure 21.

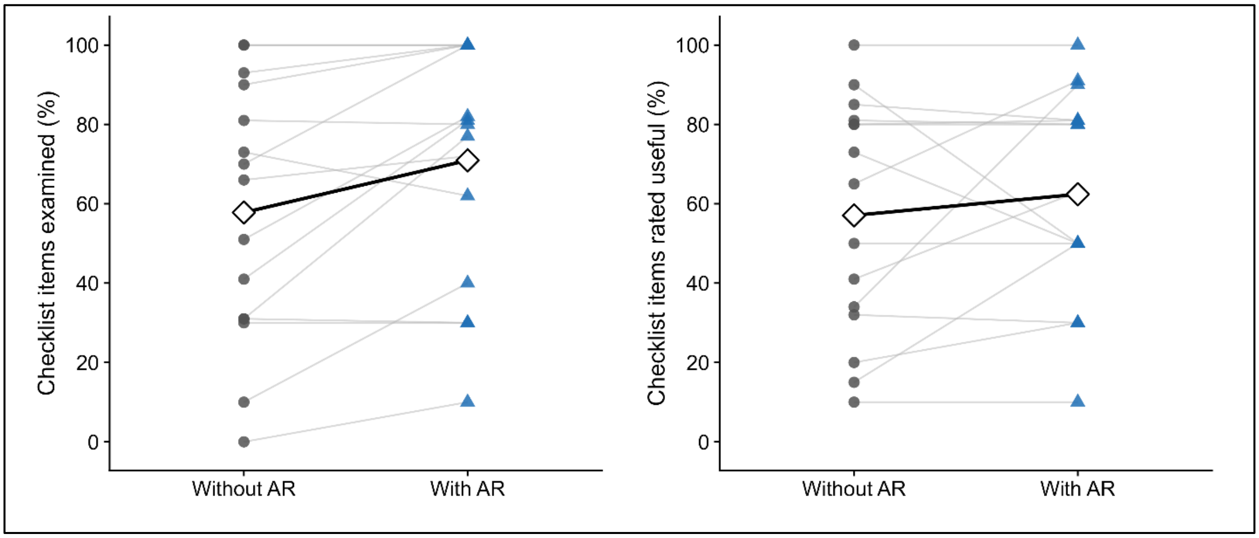


Figure 21: Normal Flight Checklist Usage and Usefulness

5.8.2 Emergency Flight Condition

We proceed in the analysis by considering the emergency flight condition. Since this portion of the study used a between-subjects design, the statistical power is considerably reduced. In addition, the number of participants in each group was uneven, further reducing the power of statistical analyses. Nevertheless, we can use the same techniques as the normal condition to compare objective flight performance, workload, situational awareness, and checklist usage measures by AR display condition. Upon the loss of engine power, participants could deviate from the original flight plan to any of the available runways. The landing locations and outcomes are shown in Table 5. Participants without the AR display were notably more consistent in landing on Runway 12, which was the closest to their starting location. Given the briefed instructions to “land as soon as conditions permit,” the decision to land on Runway 12 represents a very safe choice of landing surface. It required the least distance and amount of turning (only 90 degrees) to land, helping to preserve energy in the critical engine-out emergency scenario. Considering the variability in outcomes when participants used AR compared to the condition without AR, it does appear that the AR glasses may have affected pilot decision-making, though the sample size and methodology do not support more systematic analysis.

Table 5: Emergency Condition Landing Outcomes

Outcome	Without AR	With AR	Implications
Landing on 12	6	3	Safe (least distance and turning)
Landing on 3R	1	3	Moderately Safe (pilots familiar from past trials, but farther and 180-degree turn required)
Crash / Other Surface Landing	0	1	Critical failure
Landing on 21L	0	1	Moderately Safe (shorter distance to land than 3R, but difficult maneuvering required)

Next, we consider objective flight performance based on the variability of flight parameters. Without engine power, the glideslope will inevitably be steeper than the 3-degree standard and the airspeed slower than the normal flight, though the variability of the parameters remained within the participants' control. A Wilcoxon test was used to assess the extent to which a difference may exist between the AR and No AR groups for the standard deviation of each flight parameter. The tests suggest no significant difference exists in flight performance between the AR (n=8) and No AR (n=7) groups of participants. Results of the statistical test for glideslope ($p = 0.772$), airspeed ($p = 0.183$), and vertical velocity ($p = 0.603$) do not reach the $\alpha = 0.05$ level of significance. The results of these tests are in Table 6. Figure 22 provides a visualization of the flight performance data, showing the distribution of standard deviation values between groups. An analysis of central tendency, as with the normal flight condition, is provided in Appendix D, Table D5. This table indicates a very minor trend; participants showed a tendency to fly slightly slower (-2.12 knots) and steeper (0.88 degrees) in the AR group.

Table 6: Variability of Flight Performance Parameters, Emergency Flight

Flight Parameter	**No AR Median *[IQR]***	**AR Median *[IQR]***	**W**	***p***	**Cohen's *d***
SD Glideslope (deg)	3.50 *[3.07, 4.57]*	3.50 *[3.11, 3.93]*	31	0.772	0.55
SD Airspeed (kts)	10.66 *[7.88, 12.01]*	7.22 *[6.12, 8.29]*	40	0.183	0.93
SD Vertical Velocity (fpm)	400.70 *[390.74, 565.88]*	434.70 *[235.46, 463.37]*	33	0.603	0.68

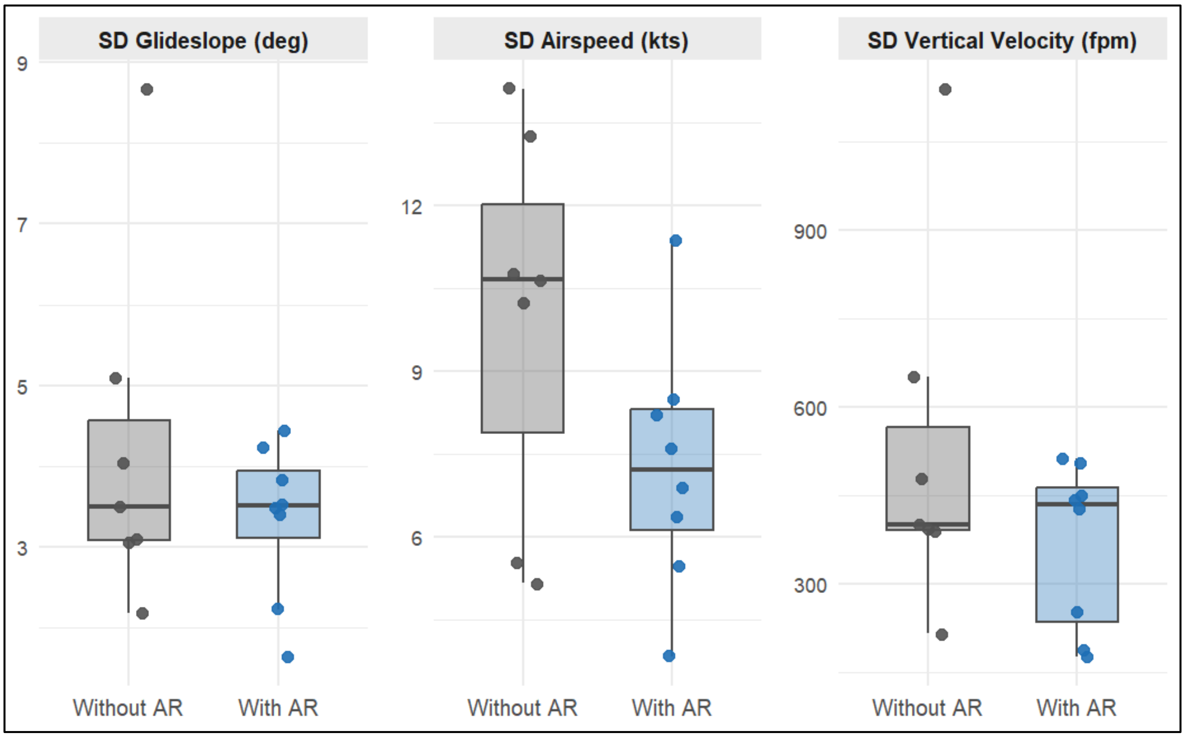


Figure 22: Variability of Flight Performance Parameters, Emergency Flight

Next, we consider the potential effects of the distractor AR display on participant self-reported workload and situational awareness. In the emergency condition, we note that situational awareness is lower (SART < 20) and workload is higher (Raw TLX > 50) compared to the previous trials (Normal Flight Conditions: Raw TLX < 40, SART > 25). This difference provides an initial confirmation that the presence of the engine failure in the simulator affected participants' cognitive demands. It also suggests that the emergency condition was sufficiently different from the normal condition, in accordance with the intent of the study design. The distributions of the TLX and SART composites are provided in Figure 23. The workload for participants with the AR display seems to be consistently higher, on average, than the No AR group; mean workload for the No AR group is 51.1/100, compared to 61.1/100 for the AR group. While mean self-assessed situational awareness seems to be comparable between the AR and No AR groups (mean values of 18.7 and 18.9, respectively), the variability is lower, raising the possibility that the AR display may have contributed to consistent (but low) awareness levels.

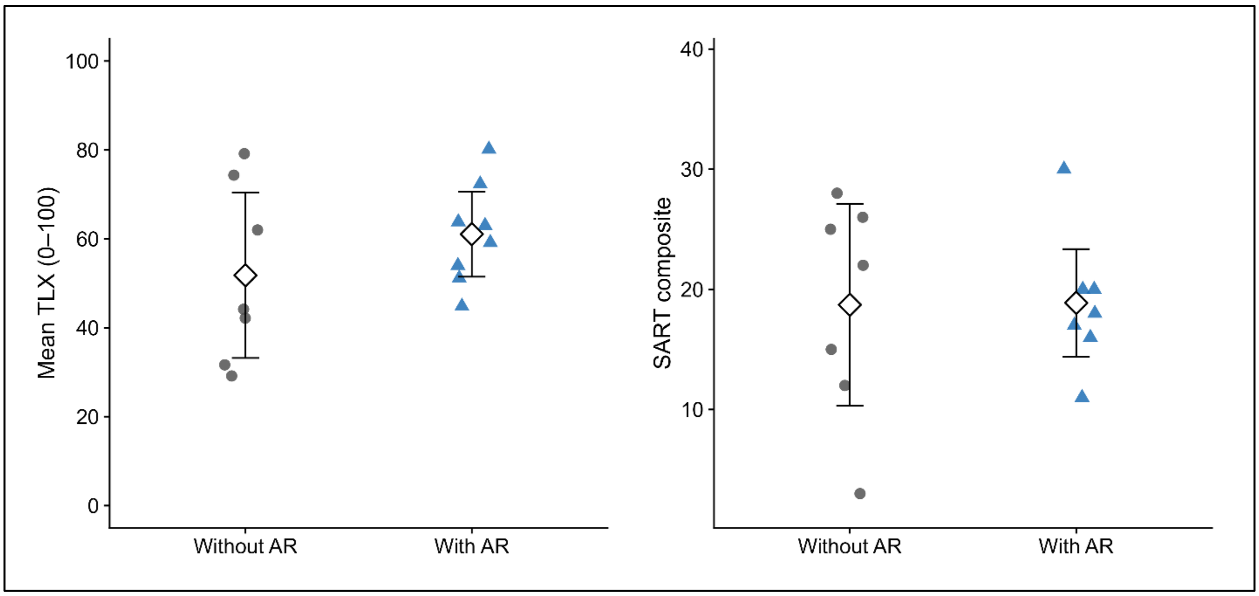


Figure 23: Emergency Flight Overall TLX (left) and SART (right)

To understand the impacts of the AR display on workload more clearly, we analyzed the associated subscales for each composite measure. Since the study in this stage uses a between-subjects design, we have less statistical power to evaluate differences. Nevertheless, we plot the distributions for each experimental group for each subscale in Figure 24. Wilcoxon tests for the various subscale items (Appendix D) show no significant difference between conditions. Both the mental workload and performance subscales show relatively large effect sizes (Cohen's *d* values of -0.824 and 0.982, respectively), but fail to reach significance ($p = 0.132$ and $p = 0.173$, respectively). With a small sample size, the finding of no statistically significant difference is unsurprising. Observationally, the AR display seems to increase mental, physical, and temporal workload, while decreasing performance, increasing effort, and increasing frustration. These trends suggest that the group assigned the AR display found the system counterproductive in their efforts to land safely. Despite the lack of statistical power to indicate a difference between the groups, results trend towards a detrimental influence from the AR display.

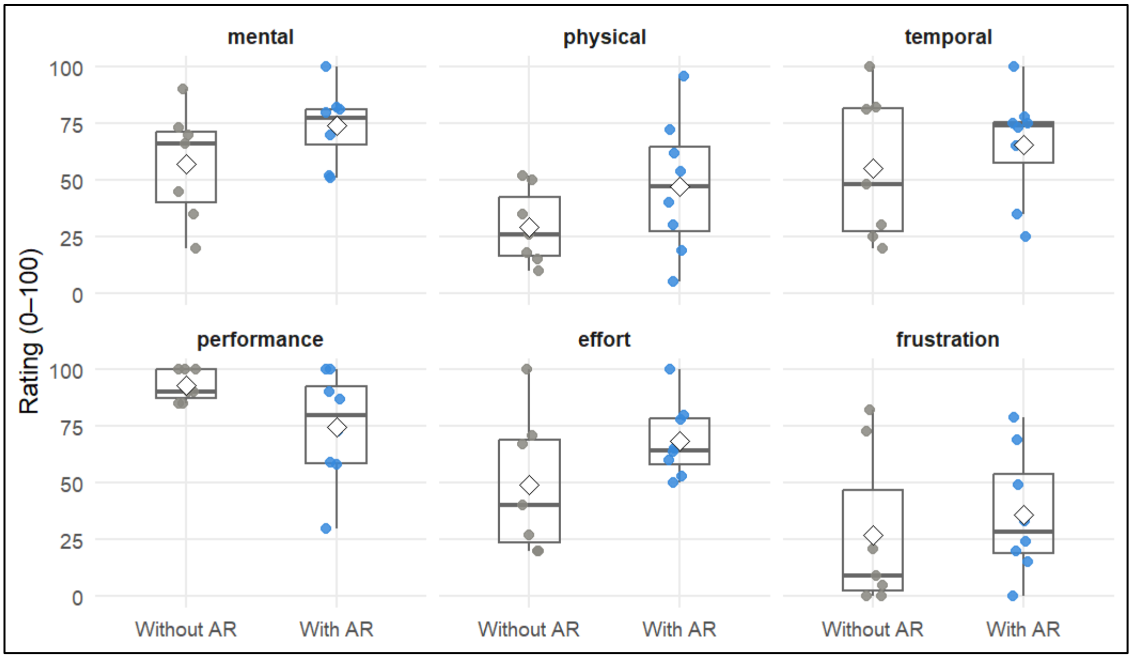


Figure 24: Emergency Flight TLX Subscales

The subscales associated with the SART indicate a similar impact to the AR display group in terms of situational awareness. The subscale composites (supply, demand, and understanding) fail to reach significance, with a Wilcoxon test yielding p-values of 0.323, 0.382, and 1.000, respectively. Even at the highest level of granularity of the SART, the ten subscales fail to reach significance, suggesting that no significant difference exists in situational awareness factors between groups. The subscales between groups are shown in Figure 25. While not significant, there may be a marginal impact from the AR display on situational awareness, observed through several trends. Participants in the AR group tended perceive the situation to be slightly more unstable, complex, and variable than counterparts in the No AR group, based on the corresponding SART subscales. Notably, the AR group showed slightly lower median arousal and higher median concentration, despite an observable increase in median divided attention and decrease in median spare mental capacity. Information quality and quantity, as well as situation familiarity, was generally unchanged. These results suggest a potential trend in that the AR display may have posed a challenge to maintaining situational awareness while providing no benefit in terms of median information quantity.

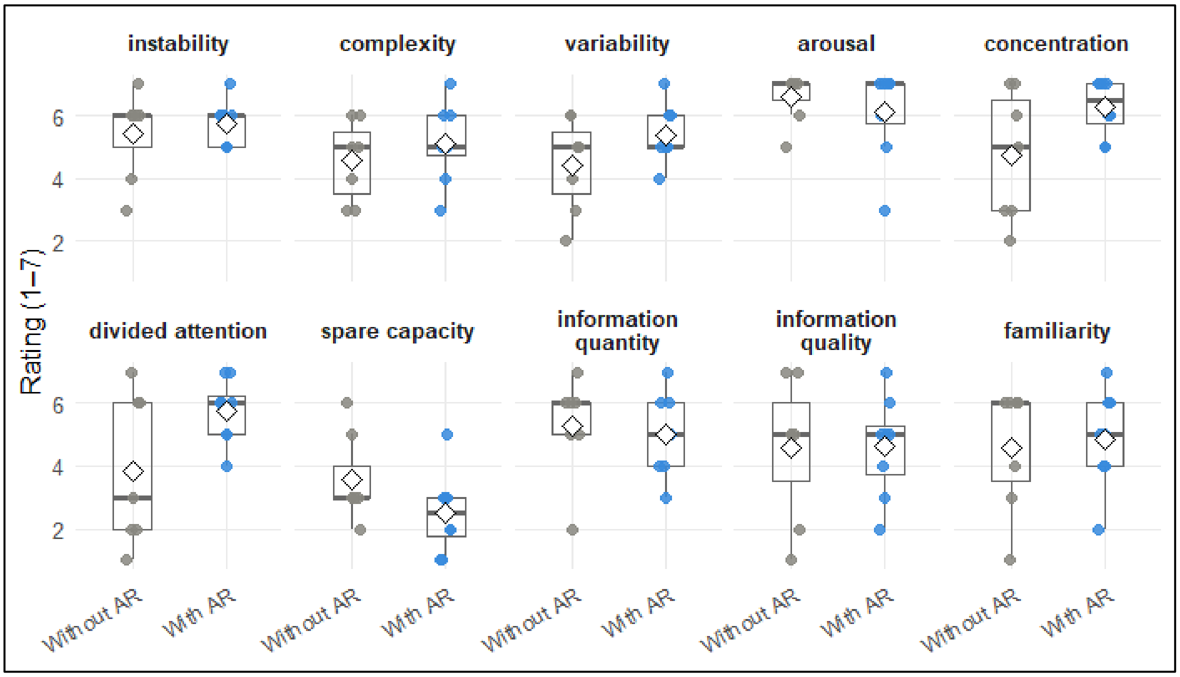


Figure 25: Emergency Flight SART Subscales

At the end of the study, after the simulator flights, we asked participants several survey questions relating to the use of the AR display. First, given that eight of 15 participants experienced a form of automation failure with AR in the emergency condition, we assessed their self-reported trust in the system. Figure 26 shows how participants in each experimental group reported the extent to which they trusted the AR system. The error bars show the interquartile range. The results reflect a substantial, but non-significant ($p = 0.139$), decline in trust for participants that experienced incorrect AR cues in the emergency condition. Still, trust was moderate for all participants, with only two participants reporting complete trust in the AR system.

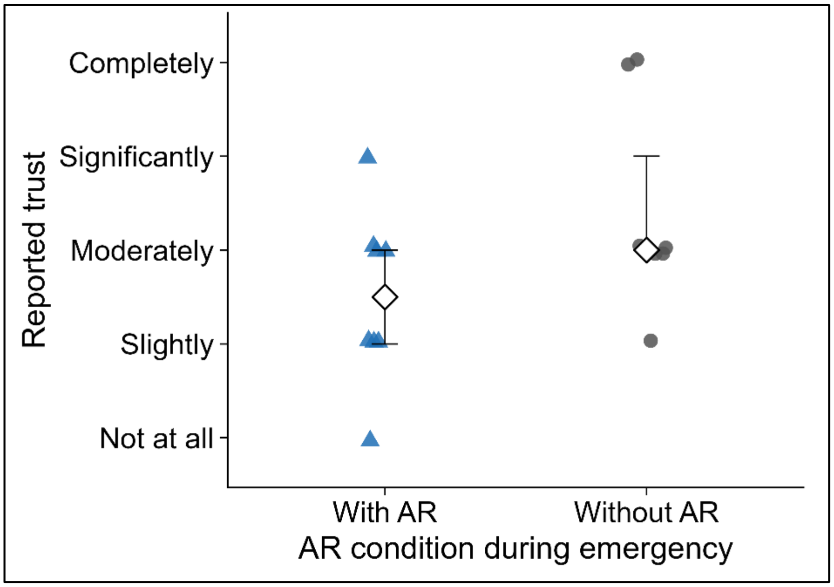


Figure 26: Reported Trust by AR vs Without AR Groups

Several questions assessed the perceived usability and usefulness of the AR system at the end of the study, which are included in Appendix C. The results are depicted in Figure 27. Though the survey questions required responses from verbal options, these were translated into a 1-5 scale for analysis, with the most favorable option as 5 and the least favorable as 1. Participants indicated their preference for using the system on a normal (non-emergency) flight, which was generally neutral or negative. Participants felt that the AR display may have improved their detection of deviations from their intended flight path, e.g. being faster or slower than their intended airspeed. Approximately half of participants felt the system could be useful for real-world flying operations, but fewer felt that it would be helpful for pilot training. Most participants felt that the AR display did not significantly affect their scan pattern or interfere with the efficacy of their visual scan (ability to acquire important flight information). Trust, as previously reported, varied widely by participant and condition.

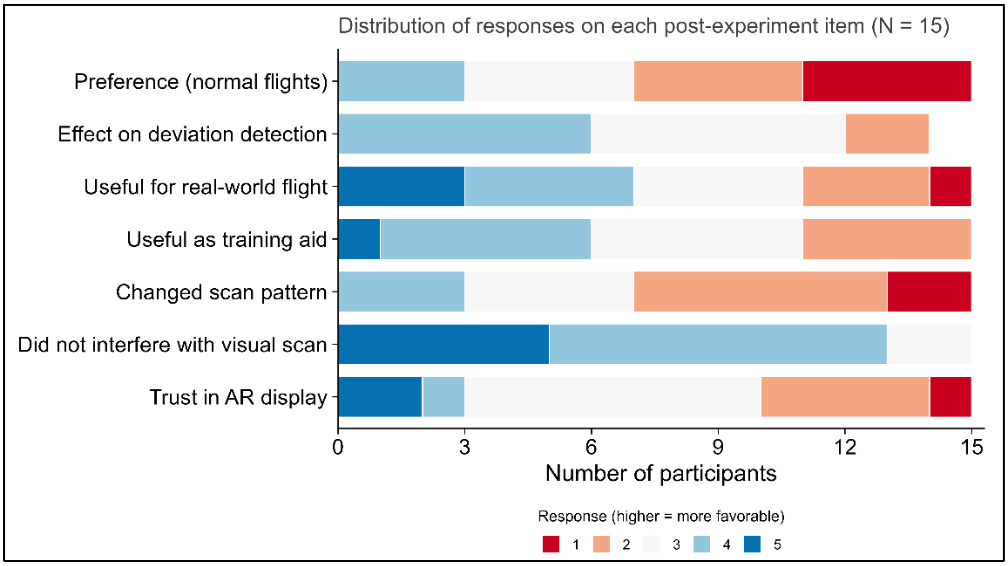


Figure 27: Subjective Perception of AR System Responses

To assess the effect of experiencing an AR display automation failure during the emergency condition, we also separated the responses by experimental group. These results are shown in Figure 28. In general, responses were notably more favorable towards the AR system with participants who were not exposed to the irrelevant and incorrect AR cues in the emergency condition (No AR group).

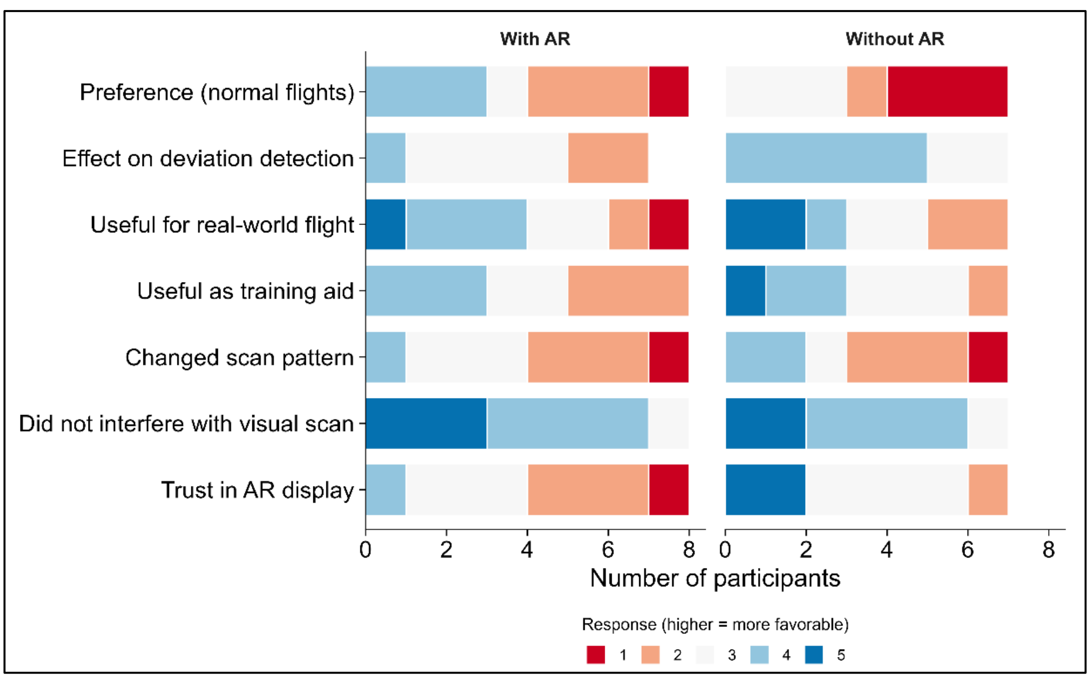


Figure 28: Subjective Perception by Experimental Group

5.9 Discussion

The goal of the multimodal approach in this study, including both objective and subjective data, was to holistically assess the effect of the AR display on flying performance. We seek to understand both positive and negative impacts from simple AR visual cues on the factors which affect pilot performance and satisfaction. Later, we note limitations on the study design that could inform future research efforts.

5.9.1 Checklist Usage

To begin, we observed a significant increase in the quantity of checklist items attended. Given that checklist non-adherence is a noted flight safety issue (Degani & Wiener, 1991), this finding may be meaningful. Participants found a similar proportion of the checklist items to be useful as the paper version, so the net increase in usage could support flying safety. One might argue that the percentage of checklist items viewed should be calibrated with the checklist items found useful, so increasing the quantity that pilots view could be counterproductive. The downside of increased checklist usage is that pilots may be spending less time on potentially time-critical tasks (Risukhin, 2005). We believe that pilots are unlikely to maintain perfect awareness of all checklist items over the course of a flight when not relying on a reference. Given the criticality of checklist adherence for flight safety (Section 3.1.3), and the fact that only critical items are included on the checklist to begin with, a higher rate of checklist usage can be assumed to be an advantage (Dismukes & Berman, 2010).

5.9.2 Flight Performance

In both normal and emergency situations within our study, pilot performance appears to be relatively unaffected by the AR display. While this result suggests that pilot skill may have been robust to AR interference for the emergency condition, in which the AR display was an intentional distractor, this finding also reveals that the display did not improve performance in the normal flight condition. Since the AR display was intended to support pilots in performing the flying task, this finding is informative and contradictory to prior research. It does not support the finding of Rowen et al.

(2019), in which a supplementary AR display improved maritime navigation performance. Several factors could contribute to this difference between findings. Most notably, the task of boat captains is distinct from that of a pilot. Among these differences is the pace; events in an aircraft typically must happen more quickly, so pilots may act more quickly without reviewing information presented on the AR display. In addition, the AR display implementations provided different information. The display used by Rowen et al. provided real-time position and motion data, essentially consolidating existing ship instrumentation with the AR display. This information more directly supports manual control performance. In contrast, our AR display implementation provided access to checklist items to support safety.

In evaluating the lack of significant differences in flight performance between the AR and No AR conditions, we may also consider the difficulty of the prescribed task. While more complete and representative than the flying task in the previous study, it was still somewhat simple. Participant #13 described the flying task as “rudimentary,” reflecting the routine nature of landing an aircraft. While several factors made the task more challenging, including landing at an unfamiliar location, the basic task being assessed was well-known to participants. This apparent familiarity is reflected by relatively high estimates of performance (a median value of 75/100 with AR and 78/100 without) in the TLX questionnaire. Despite the high self-reported performance, objective performance was not consistently high. Figure 13 showed that flight parameter variability changed considerably across the tasks. For instance, the median standard deviation of glideslope in the final approach phase (below 500 ft) is nearly as large as the desired value itself (2.84 degrees for No AR, with the ideal glideslope as 3 degrees). For the participants’ familiarity with the flying task, performance still varied considerably.

As previously mentioned, the presence of the AR display graphics did not significantly impact our measures of performance. In the normal flight condition, performance was slightly lower while using the AR device compared to the baseline. Participants were slightly more variable in each flight parameter, as seen in Table 4, representing potentially decreased stability. This trend may suggest that the demands of attending the AR display conflicted with those of controlling the aircraft. This conclusion is supported by the observed increases in mental demand, physical demand, and temporal demand. Insufficient situational awareness may also affect pilots' manual control performance, as control inputs may be delayed and reactive to deviations, rather than proactive (Stanton et al., 2001). The assessment of situational awareness further supports the slight observed decline in performance, as the median SART composite score declined by 2 points with the introduction of the AR display (Figure 18).

In an unexpected turn, the trend in flight performance of the AR display group for the emergency condition was marginally higher than the baseline. Participants were less variable in airspeed control (-3.44 knots), though very slightly worse in vertical speed variability (+34 fpm). One explanation for this behavior might be that the increased demand on participants in contending with the incorrect AR cues produced more effortful, focused flying. This conclusion would be supported by the observed increase in the median composite TLX score of +16.12 from the No AR to AR condition, reflecting trend of increased workload. While non-significant ($p = 0.272$), this trend has a medium effect size (Cohen's $d = 0.577$). The Yerkes-Dodson Law asserts that, for a given task, there is a level of workload that results in optimal performance (F. Zhou et al., 2021). Increasing workload tends to increase task performance to a point, then decrement performance as workload continues to climb. It is possible that the additional task of consciously processing AR cues increased workload closer to an optimal level, resulting in marginally improved performance. This finding is not in support of incorrect AR cues, of course. Rather, it reflects the fact that a failure of the AR display system seems to have influenced pilot behavior, as seen in the trend of reduced variability. In a real-world situation, which would likely bring higher stress than a simulation, the AR display may increase workload to the point of severe performance decrements. Still, it is intriguing that performance in one area was improved with an intentionally adverse intervention in this study.

5.9.3 Cognitive Factors

The impact of the AR display on pilot workload and situational awareness did not support the original hypotheses. Workload tended to increase with use of the AR device in both the normal and emergency conditions. Situational awareness with the AR display tended to decrease in the normal flight, yet tended to increase slightly in the emergency flight. These mixed results diverge from existing literature in the augmented reality space. For these comparisons, we will primarily compare results from the normal flight, since the addition of an emergency to assess the impact of automation failure in the AR display is not as well-studied.

First, the finding of a slight, non-significant increase in workload reflects the mixed results found in similar studies. Li et al. (2022) found no significant difference in TLX composite or subscales from a traditional checklist to a voice-controlled AR checklist, but found a decrease in physical demand, temporal demand, and frustration. While the finding of no difference matches our results, we found an increase in each of the noted subscales. The AR display implementation used by Li et al. was similar in the information displayed, but took an interactive approach (allowing for voice/gesture control) compared the fully automated system in our research. Maibach et al. (2022) found an increase in workload from pilots using a helmet-mounted AR display, especially among more experienced pilots. The present study relied on relatively young, inexperienced pilots and effects of experience were not specifically observed, so it is possible that the different experience levels can explain differences in our results. While not relying on a NASA TLX scale for workload assessment, Tran et al. (2018) noted that use of a supplemental AR device may decrease mental stress at the expense of physical discomfort. The Tran et al. study used spatial cues to guide participants through cockpit actions, again differing from our implementation. We observed both mental and physical demand increases, but physical demand remained lower than mental in both flight conditions.

Though pilots in our study generally experienced an increase in workload with use of the AR system, greater experience with the system may have reduced the cognitive demand. Specifically, Moesl et al. (2023) noted that experience with AR systems for flight training decreased workload compared to the first exposure. This finding aligns with several participant comments in our study. Participant #15 noted that the AR system “took some getting used to,” while participant #13 suggested a more thorough briefing about the capabilities of the AR system. The briefing used is included in Appendix C. These statements suggest that further use with the AR display may have trended towards a workload improvement, as originally hypothesized. As participants built familiarity with the system, they may have been able to rely on it more consistently.

In several cases, existing literature shows mixed impacts of supplementary AR systems on situational awareness. Maibach et al. (2022) noted a net decrease in assessed situational awareness, while Li et al. (2022) saw no difference in any SART composite or subscale measures. Given that pilots develop situational awareness with time and experience (Prince et al., 2007), a disruption of their typical cockpit information processing would logically represent a setback in awareness. Nevertheless, we hypothesized that the AR display would assist in situational awareness since pilots would have access to information without looking away from the environment or instruments. Rowen et al. (2019) found an increase in SA from a head-worn AR display, but in the maritime domain. In another example of a display providing real-time position and motion information, Yuan, Ng, Li, et al. (2024) saw improved SA through an AR HUD. Since the display implementation used in this research did not provide information that would necessarily improve pilots’ spatial awareness of their surroundings, it is understandable that the display would not affect (or slightly harm) their overall situational awareness.

5.9.4 Attention and Visual Scan

It is worth noting the effect of the AR display on attention, especially in terms of distraction. For safe and effective flying, pilots must direct their attention purposefully within the cockpit to the most important task or information source at a given time (Gray et al., 2016). Selective attention, which may guide and filter the information that pilots will process, can be disrupted by a salient outside stimulus (Wickens & Carswell, 2012).

In post-experiment feedback, participant #10 stated that the AR system “became a distraction because it was not aware of the engine failure,” reinforcing the challenge of navigating incorrect AR cues. Participant #2 noted that the AR cues “distracted [them] a lot at the beginning of the emergency,” leading them to “ignore it after a couple of seconds” and wish for “an option to just turn it off.” These comments align with the previous discussion about the impact of the AR display in the emergency condition. It seems that the cues, as intended through the research design, functioned as distractors, disrupting the pilots in accomplishing a safety-critical landing. Despite the lack of significant difference in flight performance in the emergency condition between groups, this feedback illuminates the potential downsides of automated AR displays, which are inherently susceptible to failure.

To handle the surplus of information in the cockpit, pilots rely on purposeful visual scanning strategies. By directing visual attention to specific areas for a short duration, pilots can maintain awareness of the environment and aircraft status (Ziv, 2016). Compared to a paper checklist on or near the pilot’s lap, the AR system was intended to reduce the burden on pilots by more closely consolidating visual information. Nevertheless, the critical information for aircraft control, like airspeed, altitude, and bank, must remain the focus of the pilot’s visual search. To assess the impact of the AR display, participants reported the extent to which the AR system affected or interfered with their visual scan. Zero pilots reported that the AR system had a significant impact on their ability to visually acquire important information, while two reported a “moderate” impact and thirteen reported “slight” or “no” impact. This result is positive, indicating that neither correct nor incorrect AR cues were unacceptably obstructive.

While pilots retained access to important flight data (since the AR system did not occlude it), results suggest that the AR display did affect their typical visual scan patterns. All but two participants reported a change to their visual scan behavior from the No AR condition, with three reporting a “significant” change. Since the AR display inherently introduces new information into the visual field, a finding of no change would indicate that the participants did not view the cues at all. The result of changed visual scan reinforces the finding of increased checklist usage, showing that participants did change their information searching behavior with the AR system present. To assess the extent to which the system may have affected participants' visual scan, we can consider the assessments of usability that participants offered. We attempted to triangulate the pilots' gaze behavior, since the capability to directly measure it was not available to us.

5.9.5 Preference and Perceived Utility

Participant preferences between the traditional paper checklist and AR system, as well as opinions on the potential for such systems in operational and training use cases, were mixed. The majority of participants, across both experimental groups, indicated that they would prefer not to use the AR system on a typical flight. This finding may be attributed to several factors. First, the functionality of the AR system was very limited. The display provided no real-time flight data, which would be more versatile and dynamic than the checklist cues. Further, the display did not allow for customization or interaction, confining participants to one specific user experience. Second, the display provided information that participants may have memorized already. The training syllabus specifically requires the checklist items presented to be completed as a "flow," or memorized sequence. By using the same checklist procedures as the Auburn University School of Aviation, we ensured that the material was technically sound and accurate, but created redundancy with participants' previous training. For this reason, the AR display may have failed to augment the cognition of participants in this instantiation.

Among the most counterintuitive results was the increased preference for the AR display on a typical flight by participants that had experienced irrelevant cues on the emergency flight. This finding may reflect the fact that increased familiarity with or exposure to the system made participants more optimistic about its potential. Still, similar proportions of each experimental group assessed the potential for using the AR system in flight operations and flight training positively. In both cases, less than half of participants saw the AR checklist system as potentially "very" or "extremely useful."

Given the simplicity and constraints of the example that participants saw, this result is not surprising. A more advanced or interactive AR system may have more clearly demonstrated the merits of the technology.

Still, the lack of enthusiasm surrounding the technology should not be ignored, especially since existing literature has largely reported the opposite. Li et al. (2020) found that participants assessed a similar AR display system as more usable, preferring it to a paper checklist. However, the system used in the Li et al. work showed multiple items at a time and allowed participants to indicate when they completed an action. This additional functionality may have supported assessments of usability. Though the current study did not assess differences by gender, Schaffernak et al. (2020) identified gender as potentially influencing acceptance of AR systems. Yuan, Ng, Li, et al. (2024) further emphasize the importance of pilot acceptance for widespread adoption of the technology.

In comparing the negligible findings of this study with positive outcomes in existing literature, we may conclude that the specific information we chose to display through the AR medium was not optimal. In other words, the pilot checklist may not be the best information to display through an AR system. In considering the more complete picture, there appears to be no improvement from paper to AR-based checklists in workload, situational awareness, or flight performance across several studies (Li et al., 2020, 2022). AR displays that provided spatial, motion, or location information, in comparison, found performance or cognitive benefits (Rowen et al., 2019; Yuan, Ng, Lau, et al., 2024). This suggests that the pilot checklist may not be the ideal information to display in the AR format. Flight data and navigation information remain useful even in the case of an in-flight emergency, while a checklist may not. Pilots must maintain awareness of the aircraft's speed, altitude, and location in order to make safe real-time decisions (Loukopoulos et al., 2003). In addition, AR is unique in its ability to provide spatial information relative to the real world (Azuma, 1995), a feature which our AR checklist did not utilize. Thus, the decrements in workload and SA seen with our AR system in the emergency condition may have been exacerbated by displaying checklist information, not flight data.

5.9.6 *Trust and Overreliance*

In considering the impacts of automation failure in the study, perhaps the most direct effects can be seen in participants' assessments of system trustworthiness. Participants assigned to use the AR display for the emergency condition experienced incorrect cues that suggested infeasible or unsafe courses of action. Since the cues repeated the same instructions from a normal flight, participants quickly recognized that the AR system "did not immediately give the correct info" (participant #14). This system behavior led to a marked decline in average trust in the system, as seen in Figure 26. Trust is difficult to regain once lost, bringing serious implications for long-term adoption and use (Dzindolet et al., 2003). Even participants that experienced the system performing reliably did not have perfect trust, as evidenced by five of seven participants in the No AR condition reporting that they trusted the system only "slightly" or "moderately." This result may be explained by their limited exposure to the AR system, as trust depends on both familiarity and system reliability, which require sustained use over time to build and assess (Sheridan, 2019). The novelty of the system alone may affect trust, which could be a subject of future investigations.

Though the experimental design did not have any specific means of assessing it, overreliance is an important consideration in evaluating the AR display system. Overreliance, or the behavioral or cognitive tendency to depend excessively on an automated system, is widely recognized in literature (Inagaki & Itoh, 2013; J. D. Lee & See, 2004; Romeo & Conti, 2026). As participant #1 insightfully noted, student pilots in training with access to automated AR systems may "learn to rely on it fully instead of just memorizing the checklists, and then what happens if that device becomes unavailable to them[?]." Since other participants (#2,13) noted that the sequential AR cues were "useful" in not "having to look down and check your checklists," the potential for pilots to over-depend on guidance from the system is logical. Even an advanced system could not adequately replace sound judgement and proactive information seeking. This study used only a perfectly reliable system for the normal flight and a perfectly unreliable system, with all cues incorrect or irrelevant, for the emergency flight. To assess the extent to which participants follow incorrect cues, incorporating easily measurable incorrect cues within a stream of correct ones would allow a clear assessment (Kelley et al., 2024; Warden et al., 2023). Future research could examine the characteristics of overreliance with supplemental AR systems in safety-critical environments, like the cockpit.

5.9.7 Limitations

While we were able to extend our understanding of AR displays in the cockpit through this work, the study had several limitations worth noting. A key limitation of the evaluation is the reliance on a flight simulator environment for data collection. From a safety perspective, the flight simulator is preferable; we intended to identify errors and degraded performance which would cause risk in a real aircraft. Additionally, since simulators are widely used in training, it is likely that participants will take the task seriously. Simulators are a widely-accepted to be a reasonable proxy for actual flying, as evidenced by the fact that pilots can log actual flight hours in simulators (Federal Aviation Administration, 2024b).

Nevertheless, a simulator may not fully capture the dynamics of a real flight environment. The simulator lacks both physical elements, like real motion and sound, and the mental element of real consequences. Flight simulators intend to model true flight dynamics to the extent that training transfers into real-world flying (Malmquist et al., 2023). As a result, flight simulators do not necessarily present all the sensory and perceptual elements that pilots must process (Robinson & Mania, 2007). In other words, since they are not designed for assessments of human information processing, they may not represent fully the true dynamic of flying. As a result, there is a risk of misunderstanding or underestimating the effects of the new display type on pilot behavior. This tradeoff is necessary, however, for the assurance of safety in completing the evaluation. In addition, we have mitigated this risk by not relying on the results as absolute representations of performance. We instead compare results between conditions, considering only the relative changes, which may reasonably indicate what the difference might be in a real aircraft environment.

Another limitation of the simulator is specific to a weakness of augmented reality. The distance to the simulator screen (for optical accommodation) is only a few feet, whereas looking out of the front window to the horizon in real life may be many miles. It is likely that there are differences in switching to an AR display between a simulated horizon a few feet away and a real horizon miles away. However, this impact is unavoidable and likely minimally impactful to the research. While Arefin et al. found a minor decline in performance per additional meter of focal distance (2022), the visual search task assessed in that study was far more intensive than ours. Nevertheless, the change in focal depth to the AR display in the simulator is not representative of a real flying scenario. We must accept the limitations of the simulator environment to gain access to emergency conditions that would be unsafe to replicate in real life.

In an ideal implementation of this experiment, eye tracking data would be collected to differentiate where pilots are directing their attention at any given time. Unfortunately, it was not technically feasible for us to integrate the AR glasses and an eye-tracking device. The display and processing elements of the AR glasses take up the space where an eye-tracker would typically be positioned. Furthermore, eye tracking cannot be mounted to the front of the glasses, since it would block the line of sight from the glasses. An external eye-tracking device (like a mounted camera or sensor) may struggle to differentiate glances between AR projections and real-world objects, since they overlap in the visual field. The lack of objective gaze data means that our analysis of gaze behavior must rely on participant feedback, which may be unreliable given that participants are deliberately subjected to a high-workload situation. In the emergency condition with AR, participants reported very low spare mental capacity, casting doubt about their ability to notice and recall their own gaze behavior.

As mentioned in Section 5.2.2, the timing of AR cue presentation was approximate, requiring adjustments if participants changed their flight paths or speed.

We did not record the specific amount of time an AR cue was presented, nor the time between cues. Therefore, this dynamic of the Wizard-of-Oz approach may have introduced inconsistency into the experimental procedure. We intended for participants to see the same cues for the same proportion of time, which served as a basic guide for standardization. Deviations in the cue presentation timing may have resulted in differences in pilot performance. Future work might use screen-recording software and a processing algorithm to detect differences in AR sequence timing, allowing this data to be matched to participant performance at different stages of the flight.

In the same vein, we did not have the means or design to collect data on checklist compliance. A video of the simulator cockpit, for instance, could have been tied to the sequence of AR cue presentation to assess whether pilots actually completed the checklist items. While we suspect that compliance would be higher with an AR system, given the significant increase in checklist items viewed, we cannot decisively draw this conclusion. This may be a valuable finding, supporting similar implementations of AR. As discussed in Section 5.9.5, we suspect that flight data might be more valuable than procedural or checklist cues in the AR format. Empirical data on pilot compliance, compared to the paper checklist condition, would be informative in substantiating or refuting this claim.

The between-subjects approach for the emergency condition required tradeoffs. Notably, by analyzing the performance under emergency conditions between subjects, we lost statistical power. Only half of all participants completed a landing with AR cues and an emergency flight condition. Unfortunately, this dynamic is necessary for understanding the effect of AR in abnormal conditions. Without a baseline condition for the emergency scenario, the effects of AR and the emergency would be entirely confounded. With both emergency conditions performed by every participant, behavior would be significantly different on the second flight because participants would likely anticipate the failure. The only way to capture the unique emergent behavior from the emergency condition was by only presenting it once, necessitating a between-subject design.

Another limitation that has arisen as a theme of the preceding analysis is participant experience and familiarity with the AR system. In designing the experiment, we decided not to allow a practice period with the device for two reasons. First, no interaction was required with the AR system; participants did not need to learn any new skill in order to perform the task. Second, the practice session might build familiarity with the flying task more than with the AR system, reducing the variability between conditions. Too much practice might also limit the potential for the AR to assist, since the flying task and sequence of checklist items could be completed based on learned cues within the simulation. Based on *post hoc* analysis, this decision may have brought unintended consequences due to participants' lack of familiarity with the system. Several participants noted that the AR system's functionality was "unexpected" (participant #11) or "took some getting used to" (participant #15). Further, the fact that participants with more exposure to the AR system (in the form of the emergency flight) rated their preference for it more favorably suggests that increased exposure may have improved participants' ability to use the system. In the same vein, the experiment was conducted only during a single session, limiting the participants' exposure to the task and AR system. Future work could assess the effects of repeated use on perceived usability and trust, to include any effects of gender, age, or flight experience.

6 Conclusion

The central goal of this research was to extend our understanding of the effects of implementing AR technology into the cockpit from a performance and safety standpoint. Two studies explored specific properties of AR displays that could reasonably affect a pilot in the execution of their duties. Study 1 examined the impact of AR occlusion on basic flight maneuver performance, finding that pilot performance was robust to occlusion in the visual field. Study 2 assessed the effects of an AR-based checklist system functioning as an aid in a normal flight and a distractor in an emergency. In this study, negative AR experiences showed the potential to decrease flying stability, increase measures of workload, decrease measures of situational awareness, and harm system trust. Further, the AR-based checklist system did not seem to benefit pilots in normal flight conditions, showing no improvement in flight performance and mixed effects on workload and situational awareness. While the AR system led to significantly increased self-reported checklist utilization, this finding further reflects a substantial change in pilot information processing and attentional control behavior, an original concern for AR use in the safety-critical cockpit environment.

Through both studies, participants seemed somewhat robust to AR display interventions, whether positive or negative. In the case of visual field occlusion in Study 1, pilots adjusted visual scan and information sourcing to execute maneuvers at a consistent level of performance. While obstructing the view of the flight instruments may have marginally decreased stability, pilots were able to compensate for the presence of a substantial AR occlusion. In Study 2, pilot flight performance was not improved by AR cues designed to be helpful, nor deteriorated by AR cues designed to be misleading and distracting. The AR system was somewhat impactful to the performance-mediating cognitive factors in the time-limited emergency flight condition, trending towards an increase in workload and decrease in situational awareness. This finding suggests that, in line with our original research framing, AR may adversely affect pilots when their composure and competence is most critical – in emergencies.

6.1 Contributions

This work, considered as a whole, makes several contributions to the research space surrounding augmented reality in aviation. In general, research in the space has been generally characterized by optimism, focusing on an idealistic upside for supplementary AR systems in aviation. The message of current literature tends to particularly stress the potential for benefits from AR systems (Li et al., 2020, 2022; Tran et al., 2018; Yuan, Ng, Lau, et al., 2024; Yuan, Ng, Li, et al., 2024). This work recognizes the potential for detrimental impacts of AR systems in an already complex, demanding environment. Far from being antithetical to efforts of innovation, we offer a new line of inquiry to identify impacts on safety and performance that may not be reflected in studies that rely on ideal flight conditions and perfectly reliable systems. To our knowledge, Study 2 represents the first assessment of AR systems in emergency flight conditions. It also represents the first investigation of automation failure in a cockpit-based head-worn AR display system. While the results do not show a dramatic breakdown in pilot ability or flight safety, the assessments of workload and SA, as well as participant feedback, suggest some impact of the AR system on participants.

The systematic evaluation of both the presence of and content within the AR display system represents a further contribution of this work. By assessing the impact of AR occlusion of the real world in Study 1, we isolated visual obstruction from the information that was presented via AR display, such as that in Study 2. Though the flying tasks were slightly different between the two studies, this methodological approach helped to reduce the range of possible confounds in Study 2. Since pilot performance was robust to a much larger visual obstruction in the first study, we have reason to believe that visual occlusion due to the small graphical cues in the second study minimally impeded pilots in Study 2. This conclusion was reinforced by the consistent survey response that the AR system did not impede pilots' ability to acquire important information in Study 2. For the potential shortcomings of AR displays in the cockpit, we offer that a small obstruction of the visual field is of comparatively low concern.

Finally, this work contributes a multi-dimensional data-gathering approach to naturalistic studies of AR in aviation. The use of objective performance data, cognitive measures, and subjective measures demonstrated our attempt to thoroughly examine pilot behavior with AR displays. This work contributes an assessment of objective flight performance for an AR checklist system, which did not, to our knowledge, exist before. Within the assessment of objective flight performance, we collected data on multiple aircraft parameters relevant to the specific flying task. Subjective data measures included trust, perceived utility, preferences, and usage. Given that pilot adoption of the technology may be entirely voluntary as it matures, these assessments are especially relevant for the extending the current state of the research space. The use of several modalities of data improved our ability to identify even subtle trends, helping to provide a foundation for future research.

6.2 Implications and Future Work

Despite the tempo of innovation with AR technology, research regarding human readiness to use AR in real-world operations remains nascent. This work may motivate future comprehensive evaluations of AR systems in the cockpit, assessing the potential for the technology to enhance or degrade performance and safety. The central implication of this work is that AR technology does not have a clear, consistent, or significant impact on pilots, but its effects remain non-negligible. This finding is, in some ways, a more dangerous implication than a consistent decline in performance with the AR system. Recurrent errors in design that pilots can correct immediately tend to be ignored, though they indicate a shortcoming in system design (Amalberti & Wilbaux, 1994). Pilots can adapt to issues with an AR display system, especially in simple, low-workload flight situations, like the steep turns of Study 1. When the situation becomes more dynamic and demanding, like the emergency condition of Study 2, the effects of the AR system may also become more prevalent. This observed trend should impel further research into human readiness with respect to supplementary AR displays in safety-critical domains, as negative effects may be masked by operator skill and adaptability, only to emerge in the most demanding circumstances.

The investigation of safety implications of AR displays lends itself to discussion of regulations surrounding the technology. As mentioned in Section 3.4.1, no outright ban exists for AR displays, though any equipment installed in aircraft must meet FAA standards. Most notably, the technology must "not adversely affect the safety of the airplane or its occupants" when functioning "normally or abnormally" (Federal Aviation Administration, 2024a). Based on the findings of this research, simple AR-based checklists may not meet the criteria for FAA certification. This finding reinforces the idea that, in technology development and regulation, AR technology should not be treated as just another display. The information processing, human behavior, and safety implications associated with AR are distinct, necessitating more thorough research.

While this work did not attempt to validate a specific AR device or system, it helped to establish the need for more research to definitively surmount the regulatory standards for AR displays.

To build a complete understanding of the consequences of using AR systems in the cockpit, this work could be extended in several ways. First, empirical research on the type of content to project through an AR display could add value. Current research, including our own, has relied on logic, experience, and existing information materials to decide the specific information that pilots will see in the AR display. It may be helpful to definitively understand the advantages and disadvantages of conveying/portraying procedural, spatial, navigation, or other information through the AR medium. Next, future work may examine how flight experience or experience using an AR system affects pilot behavior. These factors may significantly mediate the extent to which pilots rely on the AR system, affecting the extent to which the display could help or hurt their performance. In a similar vein, multi-session studies could help to reveal whether repeated use of AR systems improves pilot performance and perception of their usability. Research is also needed to develop guidelines for the use of supplemental AR systems, including requirements for training to prevent overreliance. These areas of research may provide a more complete understanding of the benefits and detriments of supplementary AR displays in the cockpit, laying the foundation for widespread use.

References

Abbott, K. H. (2023). Chapter 14—Automated systems in civil transport airplanes: Human factors considerations. In J. R. Keebler, E. H. Lazzara, K. A. Wilson, & E. L. Blickensderfer (Eds.), *Human Factors in Aviation and Aerospace* (Third, pp. 373–397). Academic Press. https://doi.org/10.1016/B978-0-12-420139-2.00020-4

AC 25-11B - Electronic Flight Displays. (n.d.). Retrieved December 3, 2025, from https://www.faa.gov/documentlibrary/media/advisory_circular/ac_25-11b.pdf

Airplane Flying Handbook. (2021). Federal Aviation Administration. https://www.faa.gov/regulations_policies/handbooks_manuals/aviation/airplane_handbook

Alarcon-Aneiva, L. J., & Fala, N. (2025). Oscillation and amplitude based unsupervised flight maneuver scoring: A data-driven pilot expertise measure. *23rd International Symposium on Aviation Psychology*, 108. https://ir.library.oregonstate.edu/downloads/9880w0536#page=117

Amalberti, R., & Wilbaux, F. (1994, January 1). *Advanced automated glass cockpit certification: Being wary of human factors*. Human Factors Certification of Advanced Aviation Technologies. https://ntrs.nasa.gov/citations/19950028365

Anand, U. (2025). The Unfaltering Algorithm vs. Adaptive Human Cognition: A Comparative Analysis of AI Autopilot and Human Pilot Performance in Simulated and Real-World Flight Emergencies. *IOSR Journal of Electronics and Communication Engineering, 1*, *20*(6), 36–45.

Arefin, M. S., Phillips, N., Plopski, A., Gabbard, J. L., & Swan, J. E. (2020). Impact of AR Display Context Switching and Focal Distance Switching on Human

Performance: Replication on an AR Haploscope. *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 571–572. https://doi.org/10.1109/VRW50115.2020.00137

Arefin, M. S., Phillips, N., Plopski, A., Gabbard, J. L., & Swan, J. E. (2022). The Effect of Context Switching, Focal Switching Distance, Binocular and Monocular Viewing, and Transient Focal Blur on Human Performance in Optical See-Through Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics*, *28*(5), 2014–2025. https://doi.org/10.1109/TVCG.2022.3150503

Aukstakalnis, S. (2016). *Practical Augmented Reality: A Guide to the Technologies, Applications, and Human Factors for AR and VR*. Addison-Wesley Professional.

Azuma, R. T. (1995). A Survey of Augmented Reality. *Presence: Teleoperators and Virtual Environments*, *6*(4), 355–385.

Bailey, R. E., Kramer, L. J., Kennedy, K. D., Stephens, C. L., & Etherington, T. J. (2017). An assessment of reduced crew and single pilot operations in commercial transport aircraft operations. *2017 IEEE/AIAA 36th Digital Avionics Systems Conference (DASC)*, 1–15. https://doi.org/10.1109/DASC.2017.8101988

Baraas, R. C., Imai, F., Yöntem, A. Ö., & Hardeberg, J. Y. (2021). Visual Perception in AR/VR. *Optics and Photonics News*, *32*(4), 34. https://doi.org/10.1364/OPN.32.4.000034

Baron, S. (1988). Pilot Control. In *Human Factors in Aviation* (pp. 347–385). Academic Press. https://doi.org/10.1016/B978-0-08-057090-7.50017-5

Bates, M. J., Colwell, C. D., King, R. E., Siem, F. M., & Zelenski, W. E. (1997). *Pilot Performance Variables.* (No. ALCFTR19970059; Crew Systems Directorate, Human Engineering Division). United States Air Force Armstrong Laboratory. https://apps.dtic.mil/sti/html/tr/ADA326600/

Bezmalinovic, T. (2024, February 16). Optofidelity measured the passthrough latency of Quest 3 & Vision Pro. *MIXED Reality News*. https://mixed-news.com/en/quest-3-vision-pro-passthrough-latency-measurement/

Biberman, L. M., & Alluisi, E. A. (1992). Pilot Errors Involving Head-Up Displays (HUDs), Helmet-Mounted Displays (HMDs), and Night Vision Goggles (NVGs). *Institute for Defense Analyses*, Article IDAP2638. https://apps.dtic.mil/sti/html/tr/ADA250719/

Bolton, M. L., Biltekoff, E., & Humphrey, L. (2022). The Level of Measurement of Subjective Situation Awareness and Its Dimensions in the Situation Awareness Rating Technique (SART). *IEEE Transactions on Human-Machine Systems*, *52*(6), 1147–1154. https://doi.org/10.1109/THMS.2021.3121960

Bonanni, L., Lee, C.-H., & Selker, T. (2005). Attention-based design of augmented reality interfaces. *CHI ’05 Extended Abstracts on Human Factors in Computing Systems*, 1228–1231. https://doi.org/10.1145/1056808.1056883

Braarud, P. Ø. (2021). Investigating the validity of subjective workload rating (NASA TLX) and subjective situation awareness rating (SART) for cognitively complex human–machine work. *International Journal of Industrial Ergonomics*, *86*, 103233. https://doi.org/10.1016/j.ergon.2021.103233

Brzeska, K., Urzędowski, M., & Kozuba, J. (2024). Classic Analog Pilot Instruments and "Glass Cockpit" in the Context of Pilot's Preferences and Confidence in Their Use During Flight. *Aviation and Security Issues*, *5*(1), 15–27. https://doi.org/10.55676/asi.v5i1.57

Burke, K. A., Wing, D. J., & Haynes, M. (2016). Flight Test Assessments of Pilot Workload, System Usability, and Situation Awareness of Tasar. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*(1), 61–65. https://doi.org/10.1177/1541931213601014

Cahill, J., Callari, T. C., Fortmann, F., Suck, S., Javaux, D., Hasselberg, A., Stoeve, S. H., & van Doorn, B. A. (2017). Adaptive Automation and the Third Pilot: Managing Teamwork and Workload in an Airline Cockpit. In L. Longo & M. C. Leva (Eds.), *Human Mental Workload: Models and Applications* (pp. 161–173). Springer International Publishing. https://doi.org/10.1007/978-3-319-61061-0_10

Carroll, M., Wilt, D. F., Sanchez, P., & Carstens, D. (2019). Examining Pilot Decision Making in Information-Rich Cockpits. *20th International Symposium on Aviation Psychology*, 73–78.

Causse, M., Mercier, M., Lefrançois, O., & Matton, N. (2025). Impact of automation level on airline pilots' flying performance and visual scanning strategies: A full flight simulator study. *Applied Ergonomics*, *125*, 104456. https://doi.org/10.1016/j.apergo.2024.104456

Cessna 172 Skyhawk Standardized Maneuvers Manual. (2018, August 20). Auburn University Professional Flight.

Clamann, M., & Kaber, D. B. (2004). Applicability of Usability Evaluation Techniques to Aviation Systems. *The International Journal of Aviation Psychology*, *14*(4), 395–420. https://doi.org/10.1207/s15327108ijap1404_4

Cooper, A. (1999). *The inmates are running the asylum*. Sams.

Cowan, N. (2010). The Magical Mystery Four: How is Working Memory Capacity Limited, and Why? *Current Directions in Psychological Science*, *19*(1), 51–57. https://doi.org/10.1177/0963721409359277

Crawford, J., & Neal, A. (2006). A Review of the Perceptual and Cognitive Issues Associated With the Use of Head-Up Displays in Commercial Aviation. *The International Journal of Aviation Psychology*, *16*(1), 1–19. https://doi.org/10.1207/s15327108ijap1601_1

Degani, A., & Wiener, E. L. (1991). *Human factors of flight-deck checklists: The normal checklist* (NAS 1.26:177549). https://ntrs.nasa.gov/citations/19910017830

Degani, A., & Wiener, E. L. (1997). Procedures in complex systems: The airline cockpit. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *27*(3), 302–312. https://doi.org/10.1109/3468.568739

Devean, T., & Kewley, R. H. (2009). *Overcoming Information Overload in the Cockpit: (663212010-001)*. West Point Operations Research Center of Excellence. https://doi.apa.org/doi/10.1037/e663212010-001

Dickman, J. L. (1982). Automated Performance Measurement: An Overview and Assessment. *Proceedings of the 4th Interservice/Industry Training Equipment Conference*, *1*, 153–165.

Dismukes, R. K., & Berman, B. (2010). *Checklists and Monitoring in the Cockpit: Why Crucial Defenses Sometimes Fail* (ARC-E-DAA-TN1902). https://ntrs.nasa.gov/citations/20110011145

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. https://doi.org/10.1080/10508414.2016.1226834

Dünser, A., Grasset, R., & Billinghurst, M. (2008). A survey of evaluation techniques used in augmented reality studies. *ACM SIGGRAPH ASIA 2008 Courses on - SIGGRAPH Asia '08*, 1–27. https://doi.org/10.1145/1508044.1508049

Dye, R. V. (1995). *Human performance considerations in the use and design of aircraft checklists* [Report]. U.S. Dept. of Transportation, Federal Aviation Administration, Assistant Administrator for System Safety, Office of Safety Services, Safety Analysis Division.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies, Trust and Technology*, *58*(6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

Ebbatson, M., Harris, D., Huddlestone, J., & Sears, R. (2011). Manual Flying Skill Decay: Evaluating Objective Performance Measures. In *Mechanisms in the Chain of Safety: Research and Operational Experiences in Aviation Psychology*. Ashgate Publishing, Ltd.

Eiberger, A., Kristensson, P. O., Mayr, S., Kranz, M., & Grubert, J. (2019). Effects of Depth Layer Switching between an Optical See-Through Head-Mounted Display and a Body-Proximate Display. *Symposium on Spatial User Interaction*, 1–9. https://doi.org/10.1145/3357251.3357588

Eisenberg, E., & Jensen, J. (2020). Measuring and qualifying optical performance of AR/VR/MR device displays and addressing the unique visual requirements of transparent AR/MR displays. *Optical Architectures for Displays and Sensing in Augmented, Virtual, and Mixed Reality (AR, VR, MR)*, *11310*, 211. https://doi.org/10.1117/12.2546613

Emergency or Abnormal Situation | SKYbrary Aviation Safety. (2026). SKYbrary Aviation Safety. https://skybrary.aero/articles/emergency-or-abnormal-situation

Endsley, M. R. (1995). Measurement of Situation Awareness in Dynamic Systems. *Human Factors*, *37*(1), 65–84. https://doi.org/10.1518/001872095779049499

Endsley, M. R. (2017). From here to autonomy: Lessons learned from human–automation research. *Human Factors*, *59*(1), 5–27. https://doi.org/10.1177/0018720816681350

Endsley, M. R., Selcon, S. J., Hardiman, T. D., & Croft, D. G. (1998). A Comparative Analysis of SAGAT and SART for Evaluations of Situation Awareness. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *42*(1), 82–86. https://doi.org/10.1177/154193129804200119

Equipment, 14 C.F.R. § 25 Subpart F (2024). https://www.ecfr.gov/current/title-14/part-25/subpart-F

Erickson, A., Kim, K., Bruder, G., & Welch, G. F. (2020). *A Review of Visual Perception Research in Optical See-Through Augmented Reality*.

Gabbard, J. L., Mehra, D. G., & Swan, J. E. (2019). Effects of AR Display Context Switching and Focal Distance Switching on Human Performance. *IEEE Transactions on Visualization and Computer Graphics*, *25*(6), 2228–2241. https://doi.org/10.1109/TVCG.2018.2832633

Ganeswaran, P., & Stevens, R. (2024). Breaking the near-vision barrier with focus tunable lenses: Addressing vergence-accommodation conflict in AR, VR, and MR systems. *Optical Architectures for Displays and Sensing in Augmented, Virtual, and Mixed Reality (AR, VR, MR) V*, *12913*, 103. https://doi.org/10.1117/12.3002570

German, E. S., & Rhodes, D. H. (2016). *Human-model interactivity: What can be learned from the experience of pilots with the glass cockpit?* Conference on Systems Engineering Research.

Goulet-Pelletier, J.-C., & Cousineau, D. (2018). A review of effect sizes and their confidence intervals, Part I: The Cohen’s d family. *The Quantitative Methods for Psychology*, *14*(4), 242–265. https://doi.org/10.20982/tqmp.14.4.p242

Grabowski, M., Rowen, A., & Rancy, J.-P. (2018). Evaluation of wearable immersive augmented reality technology in safety-critical systems. *Safety Science*, *103*, 23–32. https://doi.org/10.1016/j.ssci.2017.11.013

Gray, R., Gaska, J., & Winterbottom, M. (2016). Relationship between sustained, orientated, divided, and selective attention and simulated aviation performance:

Training & pressure effects. *Journal of Applied Research in Memory and Cognition*, *5*(1), 34–42. https://doi.org/10.1016/j.jarmac.2015.11.005

Gruen, R., Ofek, E., Steed, A., Gal, R., Sinclair, M., & Gonzalez-Franco, M. (2020). Measuring System Visual Latency through Cognitive Latency on Video See-Through AR devices. *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 791–799. https://doi.org/10.1109/VR46266.2020.00103

Guest, W., Wild, F., Vovk, A., Lefrere, P., Klemke, R., Fominykh, M., & Kuula, T. (2018). A Technology Acceptance Model for Augmented Reality and Wearable Technologies. *Journal of Universal Computer Science*, *24*(2), 192–219. https://oro.open.ac.uk/71033/

Hadjiosif, A. M., Dajaj, Y., Ranjan, T., & Smith, M. A. (2025). Tiny visual latencies can profoundly impair implicit sensorimotor learning. *Scientific Reports*, *15*(1), 16084. https://doi.org/10.1038/s41598-025-98652-2

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology* (Vol. 52, pp. 139–183). North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

He, J., McCarley, J. S., Crager, K., Jadliwala, M., Hua, L., & Huang, S. (2018). Does wearable device bring distraction closer to drivers? Comparing smartphones and Google Glass. *Applied Ergonomics*, *70*, 156–166. https://doi.org/10.1016/j.apergo.2018.02.022

Herbeck, S. R., Murdoch, M. J., & Thorstenson, C. A. (2024). Adjusting Transparency Toward Optimizing Face Appearance in Optical See-Through Augmented

Reality. *Journal of Perceptual Imaging*, *7*(0). https://doi.org/10.2352/J.Percept.Imaging.2024.7.000404

Hertzum, M. (2021). Reference values and subscale patterns for the task load index (TLX): A meta-analytic review. *Ergonomics*, *64*(7), 869–878. https://doi.org/10.1080/00140139.2021.1876927

Hicks, J. S., Durbin, D. B., Morris, A. W., & Davis, B. M. (2014). *A Summary of Crew Workload and Situational Awareness Ratings for U.S. Army Aviation Aircraft* (Army Research Laboratory). Defense Technical Information Center. https://doi.org/10.21236/ADA606506

Hilton, B. (2012). *Comparing the Effects of Simulated, Intelligent Audible, Checklists and Analog Checklists in Simulated Flight* [Dissertation, Western Michigan University]. https://scholarworks.wmich.edu/cgi/viewcontent.cgi?article=1114&context=dissertations

Hoff, K. A., & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Hughes, D. (2024, June 19). Universal Avionics' head-wearable enhanced vision is now on 737NG. *Aerospace Testing International*. https://www.aerospacetestinginternational.com/features/universal-avionics-head-wearable-enhanced-vision-is-now-on-737ng.html

Iacobucci, D., Popovich, D. L., Moon, S., & Román, S. (2023). How to calculate, use, and report variance explained effect size indices and not die trying. *Journal of Consumer Psychology*, *33*(1), 45–61. https://doi.org/10.1002/jcpy.1292

Inagaki, T., & Itoh, M. (2013). Human's Overtrust in and Overreliance on Advanced Driver Assistance Systems: A Theoretical Framework. *International Journal of Vehicular Technology*, *2013*(1), 951762. https://doi.org/10.1155/2013/951762

Jipp, M. (2016). Reaction Times to Consecutive Automation Failures: A Function of Working Memory and Sustained Attention. *Human Factors*, *58*(8), 1248–1261. https://doi.org/10.1177/0018720816662374

Kantowitz, B. H., & Campbell, J. L. (1996). Pilot Workload and Flightdeck Automation. In *Automation and Human Performance* (1st ed., pp. 117–136). CRC Press.

Kelley, B., Raikwar, A., Wickens, C. D., McMahan, R. P., Ortega, F. R., & Clegg, B. A. (2024). Addressing Human Factors Related to Artificial Intelligence Integrated Visual Cueing. *2024 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, 209–213. https://doi.org/10.1109/ISMAR-Adjunct64951.2024.00051

Kim, S., Nussbaum, M. A., & Gabbard, J. L. (2019). Influences of augmented reality head-worn display type and user interface design on performance and usability in simulated warehouse order picking. *Applied Ergonomics*, *74*, 186–193. https://doi.org/10.1016/j.apergo.2018.08.026

Klostermann, M., Conein, S., Felkl, T., & Kluge, A. (2022). Factors Influencing Attenuating Skill Decay in High-Risk Industries: A Scoping Review. *Safety*, *8*(2), 22. https://doi.org/10.3390/safety8020022

Knabl, P., Schmerwitz, S., & Ernst, J. (2014). Event expectancy and inattentional blindness in advanced helmet-mounted display symbology. *Proceedings of the Human Factors and Ergonomics Society Europe Chapter 2014 Annual Conference*.

Koc, B. (2024). Augmented Reality and Virtual Reality Solutions in Aviation Industry. In *Smart and Sustainable Operations Management in the Aviation Industry*. CRC Press.

Kruijff, E., Jr. LaViola, J. J., & Poupyrev, I. (2011). *3D User Interfaces: Theory and Practice* (D. A. Bowman, Ed.). Addison-Wesley Professional.

Kyle, A. R., Rouser, B., Paul, R. C., & Jurewicz, K. A. (2025). Quantifying Pilot Performance and Mental Workload in Modern Aviation Systems: A Scoping Literature Review. *Aerospace*, *12*(7), 626. https://doi.org/10.3390/aerospace12070626

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863. https://doi.org/10.3389/fpsyg.2013.00863

Lambooij, M., Fortuin, M., Heynderickx, I., IJsselsteijn, W., Fortuin, M., Heynderickx, I., & IJsselsteijn, W. (2009). Visual Discomfort and Visual Fatigue of Stereoscopic Displays: A Review. *Journal of Imaging Science and Technology*, *53*, 1–14. https://doi.org/10.2352/J.ImagingSci.Technol.2009.53.3.030201

Landman, A., Groen, E. L., van Paassen, M. M. (René), Bronkhorst, A. W., & Mulder, M. (2017). Dealing With Unexpected Events on the Flight Deck: A Conceptual Model of Startle and Surprise. *Human Factors*, *59*(8), 1161–1172. https://doi.org/10.1177/0018720817723428

Landry, S. J. (2012). Human Factors and Ergonomics in Aviation. In *Handbook of Human Factors and Ergonomics* (pp. 1667–1688). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118131350.ch61

Lee, B., Yoo, C., Jeong, J., Lee, B., & Bang, K. (2020). Key issues and technologies for AR/VR head-mounted displays. *Advances in Display Technologies X*, *11304*, 1130402. https://doi.org/10.1117/12.2551400

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lewis, J. R. (2012). Usability Testing. In *Handbook of Human Factors and Ergonomics* (pp. 1267–1312). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118131350.ch46

Li, W.-C., Bord, T., Zhang, J., Braithwaite, G., & Lone, M. (2020). Evaluating System Usability of Augmented Reality in Flight Operations. *Contemporary Ergonomics and Human Factors*. https://publications.ergonomics.org.uk/uploads/Evaluating-System-Usability-of-Augmented-Reality-in-Flight-Operations.pdf

Li, W.-C., Braithwaite, G., & Yu, C. (2014). The Investigation of Pilots' Eye Scan Patterns on the Flight Deck during an Air-to-Surface Task. In D. Harris (Ed.), *Engineering Psychology and Cognitive Ergonomics* (pp. 325–334). Springer International Publishing. https://doi.org/10.1007/978-3-319-07515-0_33

Li, W.-C., Zhang, J., Court, S., Kearney, P., & Braithwaite, G. (2022). The influence of augmented reality interaction design on Pilot's perceived workload and situation awareness. *International Journal of Industrial Ergonomics*, *92*, 103382. https://doi.org/10.1016/j.ergon.2022.103382

Liu, J., Gardi, A., Ramasamy, S., Lim, Y., & Sabatini, R. (2016). Cognitive pilot-aircraft interface for single-pilot operations. *Knowledge-Based Systems*, *112*, 37–53. https://doi.org/10.1016/j.knosys.2016.08.031

Lone, M., & Cooke, A. (2014). Review of pilot models used in aircraft flight dynamics. *Aerospace Science and Technology*, *34*, 55–74. https://doi.org/10.1016/j.ast.2014.02.003

Loukopoulos, L. D., Dismukes, R. K., & Barshi, I. (2003). *Concurrent task demands in the cockpit: Challenges and vulnerabilities in routine flight operations*. NASA Flight Cognition Laboratory.

Maibach, M.-J., Walko, C., & Kalra, A. (2022). *Evaluation of a head-mounted display and advanced flight control laws for helicopter ship deck landing*. https://hdl.handle.net/20.500.11881/4391

Malmquist, S., Carstens, D. S., & Dahlstrom, N. (2023). Almost Like the Real Thing – The Hidden Limits in Flight Simulation and Training. In *Human Factors in Simulation and Training* (2nd ed.). CRC Press.

Marangunić, N., & Granić, A. (2015). Technology acceptance model: A literature review from 1986 to 2013. *Universal Access in the Information Society*, *14*(1), 81–95. https://doi.org/10.1007/s10209-014-0348-1

Marsh, A. K. (2016, November 5). *GA simulators reaching for airline quality*. https://www.aopa.org/news-and-media/all-news/2016/may/11/general-aviation-simulators-reaching-for-airline-quality

Maybury, M. T. (2012). Usable advanced visual interfaces in aviation. *Proceedings of the International Working Conference on Advanced Visual Interfaces, AVI '12*, 2–3. https://doi.org/10.1145/2254556.2254558

McKendrick, R. D., & Cherry, E. (2018). A Deeper Look at the NASA TLX and Where It Falls Short. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*(1), 44–48. https://doi.org/10.1177/1541931218621010

McRuer, T. (1995). *Pilot-Induced Oscillations and Human Dynamic Behavior* (NASA Contractor Report No. 4683). NASA.

Milgram, P., & Kishino, F. (1994). A Taxonomy of Mixed Reality Visual Displays. *IEICE Transactions on Information and Systems*. https://www.semanticscholar.org/paper/A-Taxonomy-of-Mixed-Reality-Visual-Displays-Milgram-Kishino/f78a31be8874eda176a5244c645289be9f1d4317

Minaskan, N., Pagani, A., Dormoy, C.-A., Andre, J.-M., & Stricker, D. (2021). A Study of Human-Machine Teaming For Single Pilot Operation with Augmented Reality. *2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, 397–402. https://doi.org/10.1109/ISMAR-Adjunct54149.2021.00091

Mitchell, J., Vermeulen, L. P., & Naidoo, P. (2009). *Flying glass: A qualitative analysis of pilot perceptions of automated flight-decks after 20 years*. http://hdl.handle.net/2263/14744

Mixon, T. R., & Moroney, W. F. (1982). *An Annotated Bibliography of Objective Pilot Performance Measures* (NAVTRAEQUIPCEN IH-330). Department of Operations Research, Naval Postgraduate School.

Moens, L. L., Lydon, S., Cucurachi, S., O'Connor, P., Sauter, T. C., Töndury, G.-A., & Manser, T. (2026). Measuring Situation Awareness: A Meta-Review Across Domains. *Human Factors*, *68*(5), 632–672. https://doi.org/10.1177/00187208251412110

Moesl, B., Schaffernak, H., Vorraber, W., Braunstingl, R., & Koglbauer, I. V. (2023). Performance, Emotion, Presence: Investigation of an Augmented Reality-Supported Concept for Flight Training. *Applied Sciences*, *13*(20), 11346. https://doi.org/10.3390/app132011346

NASA Human Systems Integration Division. (2020). *TLX @ NASA Ames—Home*. NASA Task Load Index. https://humansystems.arc.nasa.gov/groups/tlx/index.php

Nguyen, T., Lim, C. P., Nguyen, N. D., Gordon-Brown, L., & Nahavandi, S. (2019). A Review of Situation Awareness Assessment Approaches in Aviation Environments. *IEEE Systems Journal*, *13*(3), 3590–3603. https://doi.org/10.1109/JSYST.2019.2918283

O'Hare, D., Wiggins, M., Batt, R., & Morrison, D. (1994). Cognitive failure analysis for aircraft accident investigation. *Ergonomics*, *37*(11), 1855–1869. https://doi.org/10.1080/00140139408964954

Parasuraman, R., Bahri, T., Deaton, J. E., Morrison, J. G., & Barnes, M. (1992). *Theory and Design of Adaptive Automation in Aviation Systems* (No.

NAWCADWAR9203360). Article NAWCADWAR9203360. https://apps.dtic.mil/sti/html/tr/ADA254595/

Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Peng, X., Niu, Q., Liang, Y., Luo, Y., Lu, N., & Li, X. (2025). Effects of unexpected event urgency and flight scenario familiarity on pilot trainees performance and stress responses. *Frontiers in Physiology*, *16*. https://doi.org/10.3389/fphys.2025.1599122

Pilot Logbooks, 14 CFR § 61.51 (2024). https://www.ecfr.gov/current/title-14/part-61/section-61.51

Pradhan, P., Rostami, M., Kamoonpuri, J., & Chung, J. (2023). The State of Augmented Reality in Aerospace Navigation and Engineering. In *Applications of Augmented Reality—Current State of the Art*. IntechOpen. https://doi.org/10.5772/intechopen.1002358

Prince, C., Ellis, E., Brannick, M. T., & Salas, E. (2007). Measurement of Team Situation Awareness in Low Experience Level Aviators. *The International Journal of Aviation Psychology*, *17*(1), 41–57. https://doi.org/10.1080/10508410709336936

Proctor, R. W., & Proctor, J. D. (2012). Sensation and Perception. In *Handbook of Human Factors and Ergonomics* (pp. 57–94). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118131350.ch3

Rahman, M. M., & Schulmeyer, N. P. (2025). *XDF Stream Selector: A Python Toolkit for Multi-Modal Data Stream Processing* (Version 1.0.0) [Jupyter Notebook]. Aviation/Engineering Operations aNalysis. https://github.com/aeon-lab/stream_selector

Ramachandran, K. M., & Tsokos, C. P. (2015). Chapter 12—Nonparametric Tests. In *Mathematical Statistics with Applications in R (Second Edition)* (pp. 589–637). Academic Press. https://doi.org/10.1016/B978-0-12-417113-8.00012-6

Rantanen, E. M., Talleur, D. A., Taylor, H. L., Bradshaw, G. L., Emanuel, T. W., Lendrum, L., & Hulin, C. L. (2001). Derivation of pilot performance measures from flight data recorder information. *Proceedings of the 11th International Symposium on Aviation Psychology*.

Risukhin, Dr. V. (2005). Checklist Usage as an Independent Variable in Student Pilot Task Performance Assessment. *2005 International Symposium on Aviation Psychology*, 635–641. https://corescholar.libraries.wright.edu/isap_2005/111

Robinson, A., & Mania, K. (2007). Technological research challenges of flight simulation and flight instructor assessments of perceived fidelity. *Simulation & Gaming*, *38*(1), 112–135. https://doi.org/10.1177/1046878106299035

Rohacs, J., Jankovics, I., & Rohacs, D. (2019). Less-skilled pilot decision support. *Aircraft Engineering and Aerospace Technology*, *91*(5), 790–802. https://doi.org/10.1108/AEAT-12-2017-0269

Rolland, J., & Hua, H. (2005). Head-Mounted Display Systems. *Encyclopedia of Optical Engineering*, 1–13. https://doi.org/10.1081/E-EOE-120009801

Rolland, J. P., Holloway, R. L., & Fuchs, H. (1995). Comparison of optical and video see-through, head-mounted displays. *Telemanipulator and Telepresence Technologies*, *2351*, 293–307. https://doi.org/10.1117/12.197322

Romeo, G., & Conti, D. (2026). Exploring automation bias in human–AI collaboration: A review and implications for explainable AI. *AI & SOCIETY*, *41*(1), 259–278. https://doi.org/10.1007/s00146-025-02422-7

Ross, P. (2004). Human Factors Issues of the Aircraft Checklist. *Journal of Aviation/Aerospace Education & Research*, *13*(2). https://doi.org/10.58940/2329-258X.1553

Rovira, E., Pak, R., & McLaughlin, A. (2017). Effects of individual differences in working memory on performance and trust with various degrees of automation. *Theoretical Issues in Ergonomics Science*, *18*(6), 573–591. https://doi.org/10.1080/1463922X.2016.1252806

Rowen, A., Grabowski, M., Rancy, J.-P., & Crane, A. (2019). Impacts of Wearable Augmented Reality Displays on operator performance, Situation Awareness, and communication in safety-critical systems. *Applied Ergonomics*, *80*, 17–27. https://doi.org/10.1016/j.apergo.2019.04.013

Rubio, S., Díaz, E., Martín, J., & Puente, J. M. (2004). Evaluation of Subjective Mental Workload: A Comparison of SWAT, NASA-TLX, and Workload Profile Methods. *Applied Psychology*, *53*(1), 61–86. https://doi.org/10.1111/j.1464-0597.2004.00161.x

Rudisill, M. (2000, January 1). *Crew/Automation Interaction in Space Transportation Systems: Lessons Learned from the Glass Cockpit*. NASA Langley Research Center. https://ntrs.nasa.gov/citations/20040086769

Santos, S., Parraca, J. A., Fernandes, O., Villafaina, S., Clemente-Suarez, V. J., & Melo, F. (2022). The Effect of Expertise during Simulated Flight Emergencies on the Autonomic Response and Operative Performance in Military Pilots. *International Journal of Environmental Research and Public Health*, *19*(15), 9141. https://doi.org/10.3390/ijerph19159141

Sarbach, A., & Weber, T. R. (2025). Next-Generation Navigation: Evaluating the Impact of Augmented Reality on Situation Awareness in General Aviation Cockpits. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI '25*, 1–18. https://doi.org/10.1145/3706598.3713597

Sarter, N. B. (1994). *"Strong, silent, and 'out-of-the-loop'": Properties of advanced (cockpit) automation and their impact on human-automation interaction* [Ph.D., The Ohio State University]. https://www.proquest.com/docview/304117129/abstract/CC0B81C0B94F4380PQ/1

Schaffernak, H., Moesl, B., Vorraber, W., & Koglbauer, I. V. (2020). Potential Augmented Reality Application Areas for Pilot Education: An Exploratory Study. *Education Sciences*, *10*(4), 86. https://doi.org/10.3390/educsci10040086

Scharowski, N., Perrig, S. A. C., Aeschbach, L. F., Felten, N. von, Opwis, K., Wintersberger, P., & Brühlmann, F. (2024). *To Trust or Distrust Trust Measures:*

Validating Questionnaires for Trust in AI (arXiv:2403.00582; Version 1). arXiv. https://doi.org/10.48550/arXiv.2403.00582

Schutte, P. C., & Trujillo, A. C. (1996). Flight Crew Task Management in Non-Normal Situations. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *40*(4), 244–248. https://doi.org/10.1177/154193129604000422

Schvaneveldt, R., Beringer, D. B., Lamonica, J., Tucker, R., & Nance, C. (2000). *Priorities, organization, and sources of information accessed by pilots in various phases of flight* (DOT/FAA/AM-00/26). Federal Aviation Administration. https://doi.org/10.1037/e406222004-001

Sebok, A., & Wickens, C. D. (2017). Implementing Lumberjacks and Black Swans Into Model-Based Tools to Support Human–Automation Interaction. *Human Factors*, *59*(2), 189–203. https://doi.org/10.1177/0018720816665201

Shams, L., & Kim, R. (2010). Crossmodal influences on visual perception. *Physics of Life Reviews*, *7*(3), 269–284. https://doi.org/10.1016/j.plrev.2010.04.006

Sheridan, T. B. (2019). Individual Differences in Attributes of Trust in Automation: Measurement and Application to System Design. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.01117

Skarbez, R., Smith, M., & Whitton, M. C. (2021). Revisiting Milgram and Kishino's Reality-Virtuality Continuum. *Frontiers in Virtual Reality*, *2*. https://doi.org/10.3389/frvir.2021.647997

Skraaning, G., & Jamieson, G. A. (2024). The Failure to Grasp Automation Failure. *Journal of Cognitive Engineering and Decision Making*, *18*(4), 274–285. https://doi.org/10.1177/15553434231189375

Smith, M., Lu, F., Buzard, A., Duyar, A., Kozhokar, D., & Skarbez, R. (2025). The Visual Integration Framework: Classifying Augmented Reality Interfaces Based on Their Affect on Human Information Processing. *In Progress*.

Stanton, N. A., Chambers, P. R. G., & Piggott, J. (2001). Situational awareness and safety. *Safety Science*, *39*(3), 189–204. https://doi.org/10.1016/S0925-7535(01)00010-8

Swan, J., & Gabbard, J. L. (2005). *Survey of User-Based Experimentation in Augmented Reality*. 1st International Conference on Virtual Reality, Las Vegas, Nevada. https://www.semanticscholar.org/paper/Survey-of-User-Based-Experimentation-in-Augmented-Swan-Gabbard/025a3d634faed5e98528188ac862025e10298b03

Taylor, R. M. (1989). Situational Awareness Rating Technique (Sart): The Development of a Tool for Aircrew Systems Design. *Proceedings of the NATO Advisory Group on Aerospace Research and Development (AGARD) Symposium on Situational Awareness in Aerospace Operations, AGARD-CP-478*.

Tran, T. H., Behrend, F., Fünning, N., & Arango, A. (2018). Single Pilot Operations with AR-Glasses using Microsoft HoloLens. *2018 IEEE/AIAA 37th Digital Avionics Systems Conference (DASC)*, 1–7. https://doi.org/10.1109/DASC.2018.8569261

Van Krevelen, R., & Poelman, R. (2010). A Survey of Augmented Reality Technologies, Applications and Limitations. *International Journal of Virtual Reality (ISSN 1081-1451)*, *9*, 1. https://doi.org/10.20870/IJVR.2010.9.2.2767

Volz, K., Yang, E., Dudley, R., Lynch, E., Dropps, M., & Dorneich, M. C. (2016). An Evaluation of Cognitive Skill Degradation in Information Automation. *Proceedings* *of the Human Factors and Ergonomics Society Annual Meeting*, *60*(1), 191–195. https://doi.org/10.1177/1541931213601043

Walliser, J. C., de Visser, E. J., Wiese, E., & Shaw, T. H. (2019). Team Structure and Team Building Improve Human–Machine Teaming With Autonomous Agents. *Journal of Cognitive Engineering and Decision Making*, *13*(4), 258–278. https://doi.org/10.1177/1555343419867563

Wang, X., Love, P. E. D., Kim, M. J., Park, C.-S., Sing, C.-P., & Hou, L. (2013). A conceptual framework for integrating building information modeling with augmented reality. *Automation in Construction, Information Technologies in Safety Management*, *34*, 37–44. https://doi.org/10.1016/j.autcon.2012.10.012

Warden, A. C., Wickens, C. D., Clegg, B. A., & Ortega, F. R. (2024). Quantitative effects of overlay clutter and information access effort: Examining the scan-clutter trade-off in displays with geospatial maps. *Journal of Experimental Psychology: Applied*, *30*(4), 607–630. https://doi.org/10.1037/xap0000512

Warden, A. C., Wickens, C. D., Rehberg, D., Ortega, F. R., & Clegg, B. A. (2023). Fast, Accurate, But Sometimes Too-Compelling Support: The Impact of Imperfectly Automated Cues in an Augmented-Reality Head-Mounted Display on Visual Search Performance. *IEEE Transactions on Human-Machine Systems*, *53*(6), 1061–1072. https://doi.org/10.1109/THMS.2023.3302152

Wickens, C. D. (2002). Aviation Displays. In *Principles and Practice of Aviation Psychology*. CRC Press.

Wickens, C. D. (2007). Aviation. In F. T. Durso, R. S. Nickerson, S. T. Dumais, S. Lewandowsky, & T. J. Perfect (Eds.), *Handbook of Applied Cognition* (Second). John Wiley & Sons.

Wickens, C. D. (2013). Attention. In J. D. Lee & A. Kirlik (Eds.), *The Oxford Handbook of Cognitive Engineering* (p. 0). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199757183.013.0003

Wickens, C. D., & Carswell, C. M. (2012). Information Processing. In *Handbook of Human Factors and Ergonomics* (pp. 117–161). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118131350.ch5

Wickens, C. D., Goh, J., Helleberg, J., Horrey, W. J., & Talleur, D. A. (2003). Attentional Models of Multitask Pilot Performance Using Advanced Display Technology. *Human Factors*, *45*(3), 360–380. https://doi.org/10.1518/hfes.45.3.360.27250

Wickens, C. D., Vidulich, M. A., & Tsang, P. S. (2023). Information processing in aviation. In J. R. Keebler, E. H. Lazzara, K. A. Wilson, & E. L. Blickensderfer (Eds.), *Human Factors in Aviation and Aerospace* (Third, pp. 89–139). Academic Press. https://doi.org/10.1016/B978-0-12-420139-2.00002-2

Wild, F., Klemke, R., Lefrere, P., Fominykh, M., & Kuula, T. (2017). Technology Acceptance of Augmented Reality and Wearable Technologies. In D. Beck, C. Allison, L. Morgado, J. Pirker, F. Khosmood, J. Richter, & C. Gütl (Eds.), *Immersive Learning Research Network* (Vol. 725, pp. 129–141). Springer International Publishing. https://doi.org/10.1007/978-3-319-60633-0_11

Willett, W., Jansen, Y., & Dragicevic, P. (2017). Embedded Data Representations. *IEEE Transactions on Visualization and Computer Graphics*, *23*(1), 461–470. https://doi.org/10.1109/TVCG.2016.2598608

Wright, S., & O'Hare, D. (2015). Can a glass cockpit display help (or hinder) performance of novices in simulated flight training? *Applied Ergonomics*, *47*, 292–299. https://doi.org/10.1016/j.apergo.2014.10.017

Xiong, J., Hsiang, E.-L., He, Z., Zhan, T., & Wu, S.-T. (2021). Augmented reality and virtual reality displays: Emerging technologies and future perspectives. *Light: Science & Applications*, *10*(1), 216. https://doi.org/10.1038/s41377-021-00658-8

XREAL Air 2 Pro. (2025). XREAL US Shop. https://us.shop.xreal.com/products/xreal-air-2-pro

Yesavage, J. A., Taylor, J. L., Mumenthaler, M. S., Noda, A., & O'Hara, R. (1999). Relationship of Age and Simulated Flight Performance. *Journal of the American Geriatrics Society*, *47*(7), 819–823. https://doi.org/10.1111/j.1532-5415.1999.tb03838.x

Yu, C., Wang, E. M., Li, W.-C., & Braithwaite, G. (2014). Pilots' Visual Scan Patterns and Situation Awareness in Flight Operations. *Aviation, Space, and Environmental Medicine*, *85*(7), 708–714. https://doi.org/10.3357/ASEM.3847.2014

Yuan, X., Ng, K. K. H., Lau, C. K., & Fung, K. H. (2024). Enhancing Single-Pilot Operations With Augmented Reality HUD: Improving Situational Awareness and Performance in a Human-Centric Cockpit Design. In *AIAA AVIATION FORUM*

AND ASCEND 2024. American Institute of Aeronautics and Astronautics. https://doi.org/10.2514/6.2024-3945

Yuan, X., Ng, K. K. H., Li, Q., Yiu, C. Y., Lau, C. K., Fung, K. H., & Ng, L. H. (2024). Exploring the Human-Centric Interaction Paradigm: Augmented Reality-Assisted Head-Up Display Design for Collaborative Human-Machine Interface in Cockpit. *Advanced Engineering Informatics*, *62*, 102656. https://doi.org/10.1016/j.aei.2024.102656

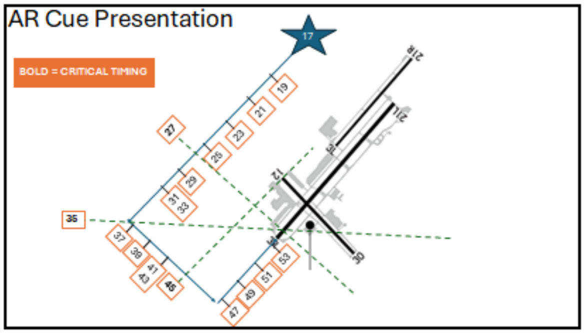
Zhou, C., Qiao, W., Hua, J., & Chen, L. (2024). Automotive Augmented Reality Head-Up Displays. *Micromachines*, *15*(4), 442. https://doi.org/10.3390/mi15040442

Zhou, F., Ji, Y., & Jiao, R. J. (2021). EMOTIONAL DESIGN. In *HANDBOOK OF HUMAN FACTORS AND ERGONOMICS* (pp. 236–251). https://doi.org/10.1002/9781119636113.ch9

Ziv, G. (2016). Gaze Behavior and Visual Attention: A Review of Eye Tracking Studies in Aviation. *The International Journal of Aviation Psychology*, *26*(3–4), 75–104. https://doi.org/10.1080/10508414.2017.1313096

Appendices

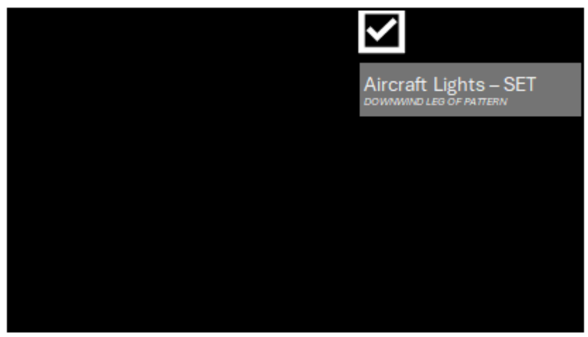
Appendix A: AR Cue Graphics and Sequence



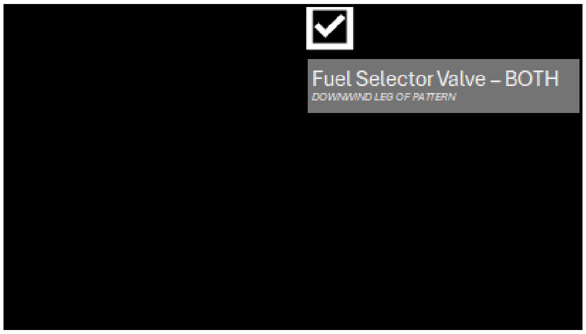
15

19

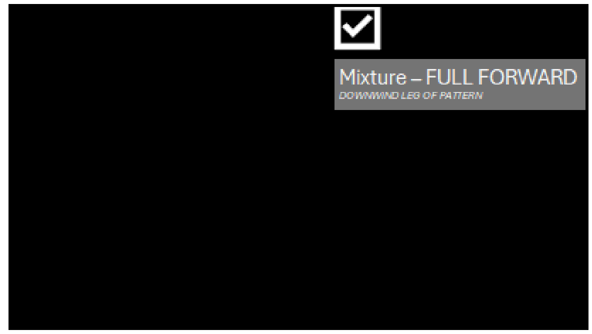
21



23



25



27

29

31

33

35

37

39

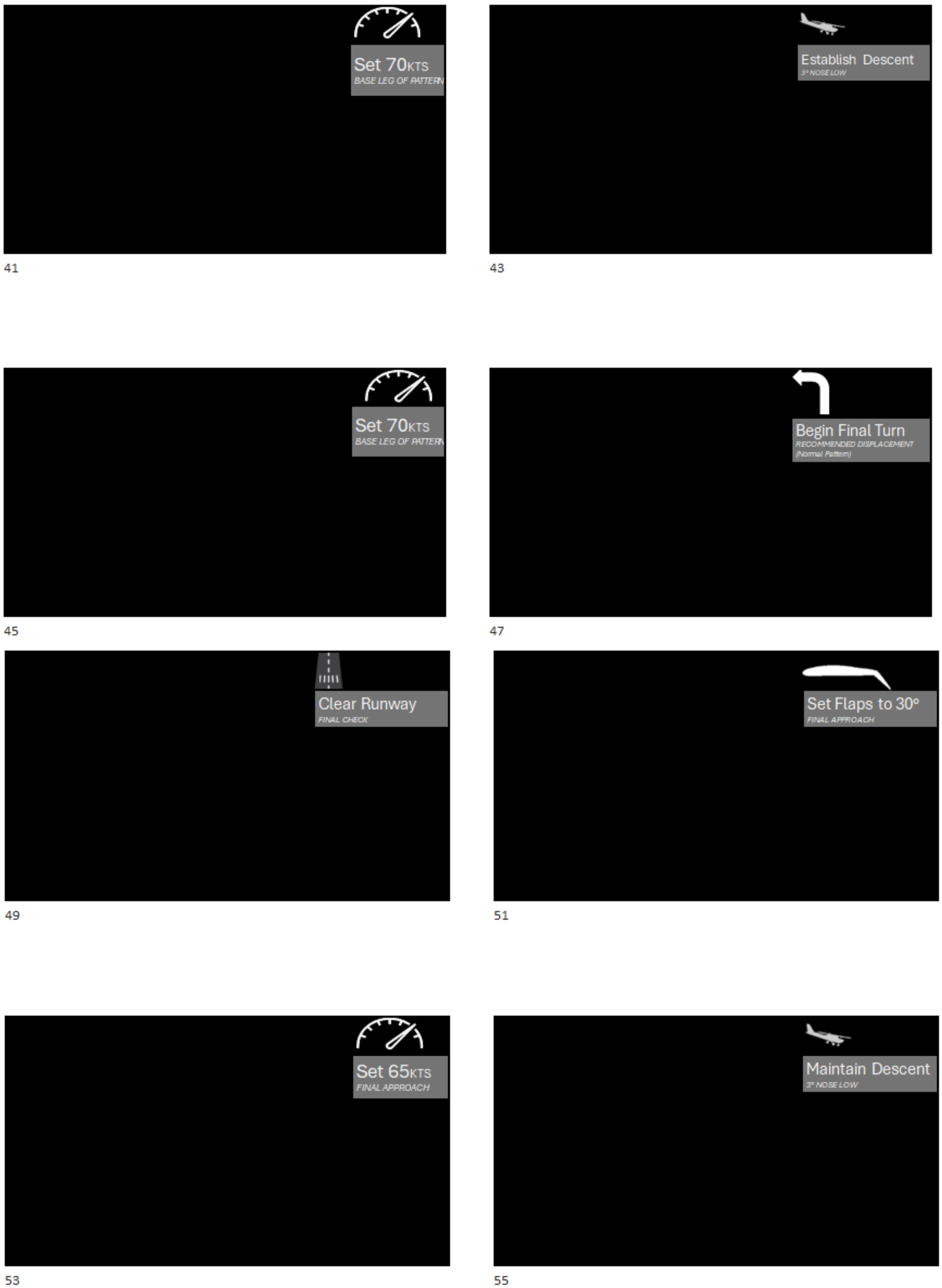


Figure A1: AR Graphics and Sequence

CESSNA 172 S NAV III SKYHAWK — NORMAL CHECKLIST

DESCENT

Flow

APPROACH BRIEFING..COMPLETE
FUEL SELECTOR VALVE..BOTH
MIXTURE..ENRICHEN
AVIONICS/FLIGHT INSTRUMENTS..CHECK/SET
LIGHTS..ON
SEATBELTS/HARNESSES..ON

DESCENT CHECKLIST..COMPLETE

DESCENT "FINAL ITEMS"

Flow

FUEL SELECTOR VALVE..BOTH
MIXTURE CONTROL..FULL FORWARD
AUTOPILOT (if installed)..OFF
CABIN PWR 12V..OFF

DESCENT "FINAL ITEMS" CHECKLIST..COMPLETE

Figure A2: Corresponding Printed Checklist

Maneuver Standards:

1. Complete the Descent Flows/Checklists.
2. At least 2 nm from the runway, enter the traffic pattern at traffic pattern altitude on a 45° entry to the downwind, maintaining a one-half mile distance from the runway on the downwind leg.
3. Set power to 2100 RPM to establish and maintain 90 KIAS. Trim as necessary.
4. Abeam the point of intended landing, set power to 1500 RPM, at or below $V_{FE10°}$ (110 KIAS), set the flaps to 10°, and establish a descent at 80 KIAS (1° nose down). Trim as necessary.
5. At the 45° point to the intended touchdown point (or at a point appropriate for wind conditions), commence a turn to the base leg while maintaining 80 KIAS.
6. On the base leg, set the flaps to 20°, and establish and maintain 70 KIAS (3° nose down). Trim as necessary.
7. Visually verify that the final approach (including the extended final and the opposite base leg) is clear, and turn final while maintaining 70 KIAS.
8. When landing is assured, set the flaps to 30°. Establish and maintain 65 KIAS (+ ½ gust factor, if applicable). Trim as necessary.
9. By 300' AFE, complete a G_UM_P check. If the approach is stabilized, call out, *"200 feet, Stabilized, Continuing."*

Figure A3: Corresponding Printed Maneuver Standards

Appendix B: Informed Consent Documentation

INFORMED CONSENT DOCUMENT

Title of research study: SafetyFest! Spring 2026

Investigator:

KEY INFORMATION

The following is a short summary of this study to help you decide whether or not you would like to participate in this research study. More detailed information is listed later on in this form.

General Information	You are being asked to take part in a research study. This research study is voluntary, meaning you do not have to take part in it. The procedures, risks, and benefits are fully described further in this consent form.
Purpose	The purpose of the study is to collect data on the performance and behavior of humans in different flight simulation scenarios involving drones and airplanes. The collected data will help us evaluate aspects of human factors and performance.
Duration & Visits	There will be a total of four short missions on our flight simulator. We expect that you will need 45-90 minutes to complete the flights, briefings, and questionnaires, depending on how many stations you are eligible for based on your prior flight experience. You will only need to visit the lab once.
Overview of Procedures	You will be asked to fly several missions in a drone and/or airplane on our simulator. You will fly these missions under various conditions, as described by the researcher overseeing your experiment. Before you begin, you will complete a demographics survey which will gather information on your experience, and each flight scenario will be followed by some questions on your workload.
Risks	This research includes minimal risk to participants, including discomfort due to wearing the VR headset and motion sickness.

Benefits	There are no direct benefits to you for participating in this study. Your data will benefit our research by helping us evaluate how people respond to different stimuli while flying. Your data may be used in future research, advancing the understanding of human factors/ergonomics in aviation safety.
Alternatives	The alternative is not to participate in this study.
Right to Withdraw from the Study	Your decision to be in this study is voluntary. If you decide to be in this study and then change your mind, you can leave the study at any time without penalty.

DETAILED INFORMATION

The following is more detailed information about this study.

Why am I being invited to take part in a research study?

You are invited to participate in a research study evaluating human performance in aviation, conducted collaboratively by Aviation Engineering and Operations Analysis (ÆON) lab. The study is led by ____________(Department of Aerospace Engineering).

How many people will be studied?

We expect about 40 people will be in this research study.

What should I know about a research study?

Someone will explain this research study to you.

Whether or not you take part in the research study is up to you. You may wish to talk to your family and friends about participating in the study.

You can choose not to take part in the study.

You can agree to take part in the study and later change your mind.

Your decision will not be held against you.

You can ask all the questions you want before you decide whether you would like to take part in the study or not.

Why is this research being done?

We aim to create a recurrent event for flight students which enables and motivates them to practice, try out new technology, and improve their flying skills in a competitive environment. This research also collects data for research, specifically to evaluate the impact of the use of the real vehicle on simulation, the usefulness of LLMs for flight feedback, the impact of AR technology on general aviation pilots, and a tailored display for high altitude operations.

What will be involved if I participate?

If you decide to participate in this research study, you will be asked to fly simulated missions on our VR and physical simulators involving a UAV and a Cessna 172, and then asked some questions about your workload and performance during the missions. You can participate in up to four research stations depending on your experience. Throughout your time in the lab, we will be recording your heart rate metrics using a Polar H10. You will get instructions on placement in the video you are about to watch, and one of the researchers will show you to a private room to place it. Stations A, B, and D are conducted in virtual reality, and you will be wearing a headset that records eye tracking information. You are welcome to ask for help with making the headset comfortable for you. Station C is conducted in our physical simulator, and you will be provided with augmented reality glasses, which you will put on just like sunglasses. If you are eligible for multiple stations, you are welcome to take breaks between them as needed.

Station A requires no experience and will show you how to fly handling quality task elements on a drone. After a short practice session, you will fly these task elements three times.

Station B requires at least 5 hours of flight experience, and will ask you to fly six steep turns to private pilot standards in our VR simulator.

Station C requires at least 30 hours of flight, and will have you flying visual approaches to land in our physical simulator. You may be provided with augmented reality glasses to wear, which will supplement your environment with additional information.

Station D requires that you have some experience with instrument approaches. You will be back in the VR simulator, flying the RNAV 27 approach into MMH.

How long will the research last and what will I need to do?
We expect that you will be in this research study for 45-90 minutes depending on the number of stations for which you qualify. You need to fly scenarios on our flight simulator setup, answer questionnaires, and wear our provided non-invasive heart rate monitor and eye tracker.

Is there any way being in this study could be bad for me?
The use of a VR headset might create a slight discomfort or motion sickness. By attending the study, you will also have to be present for a certain amount of time, which can be considered as a loss of available time. This study involves in-person research activities that carry an inherent risk for transmission of illnesses including respiratory viruses such as COVID-19, flu, and RSV. Should you feel uncomfortable, you can either ask researchers to wear a mask or elect not to participate in this research project.
The data collected today is not identifiable.

To minimize these risks, we will:
Advise and help, if necessary, on how to properly handle and adjust the VR headset and Polar H10 ECG sensor.
Interrupt the trial as soon as you voice discomfort or other concerns
Delete the original data (pupil recordings) as soon as it is converted into a de-identified form.
Store all the data on a password secured server requiring a Single Sign On (SSO).

Will being in this study help me in any way?
There is no direct benefit to you.

What happens if I do not want to be in this research?

Participation in research is completely voluntary. You can decide to participate, not participate, or discontinue participation at any time without penalty or loss of benefits to which you are otherwise entitled. Your decision not to participate or to stop participating will not jeopardize your future relations with Auburn University or the Department of Aerospace Engineering. Your alternative to participating in this research study is to not participate.

What happens if I say yes, but I change my mind later?
You can leave the research at any time; it will not be held against you. If you choose to withdraw, your data will be withdrawn as long as it is still identifiable (i.e. while it is still being recorded).

What happens to the information collected for the research?
Your privacy will be protected. Any information obtained in connection with this study will remain confidential. Information obtained through your participation may be used in further research culminating in thesis and journal publication or conference presentations. Efforts will be made to limit the use and disclosure of your personal information, including research study records, to people who have a need to review this information. We cannot promise complete secrecy. Organizations that may inspect and copy your information include the IRB. We will protect your privacy and the data you provide by keeping your research data and your identifying information separate, using instead a randomly generated participant code. Only study staff will have access to your research data. Your scheduling information (name and email address) are kept separate from the research data with no link between them.

FUTURE USE OF DATA
The data collected in the experiment will be stored in a database that is established and maintained securely and only accessible with a password and SSO. The data may be used for future research in the ÆON Lab and our collaborators, current or future.

What else do I need to know?

This research is partially funded by the National Science Foundation.

Who can I talk to?

If you have questions, concerns, or complaints, or think the research has hurt you, talk to the research team by reaching out to ________. This research has been reviewed and approved by the Auburn University Institutional Review Board (IRB). You may contact the Auburn University IRB ________________ if:

Your questions, concerns, or complaints are not being answered by the research team.

You cannot reach the research team.

You want to talk to someone besides the research team.

You have questions about your rights as a research subject.

You want to get information or provide input about this research.

Having read the information provided, you must decide whether or not you wish to participate in this research study.

Appendix C: Participant Briefing, Survey, and Pattern References

Participant Arrival (Station C)

Introduction

Welcome to Station C! In this station, you will be flying in the fixed base simulator with and without an AR headset.

- Please feel free to sit down and adjust the seat to a comfortable distance, so you can reach the pedals.

In this flight, you [WILL / WILL NOT] be wearing AR glasses.

At this station, we are investigating how an AR display could affect you in the cockpit. For the experiment, we are specifically interested in your flying performance and stability.

Please treat this as you would a real flight, even if you notice differences in the simulator from a real flight situation.

For the flight, you will be performing a normal pattern for Runway 3R at Prescott Airport in Arizona. You will start to the west and fly left traffic. You will notice that this airport has perpendicular crossing runways. Do not worry about Land-and-Hold-Short considerations. You should plan on flying this pattern, although you may deviate from the flight plan if safety dictates. In that case, land as soon as conditions permit.

For your reference, we have an airport diagram and consolidated checklist. You may recognize this as similar to that of Auburn's School of Aviation.

Figure C1: Briefing Script

Table C1: Demographic Survey Questions

How would you like to be addressed?
What is your age range?
What kind of pilot's license do you currently have?
Which ratings or endorsements do you currently have?
How many flight hours have you logged (approximately)?
How many years of flying experience do you have? (Round to the nearest year)
How often do you fly in a flight simulator?
How often do you fly in a VR flight simulator?
When was the last time you flew an aircraft?
Have you participated in our lab's events before?

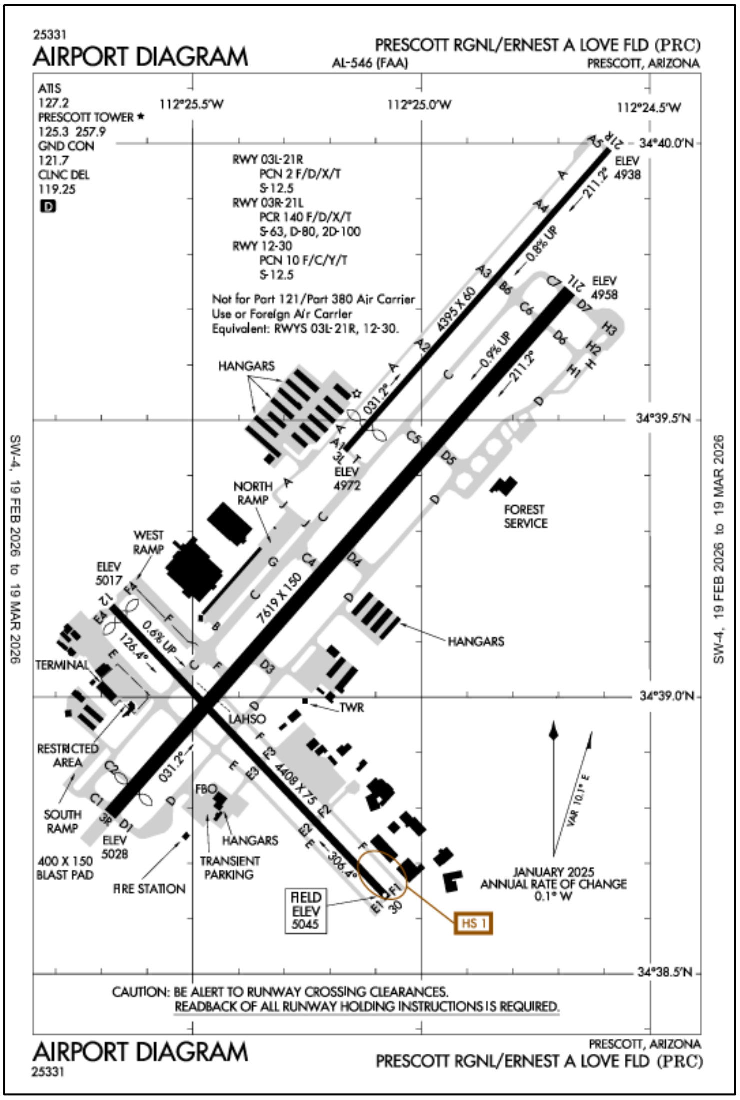


Figure C2: Airport Diagram

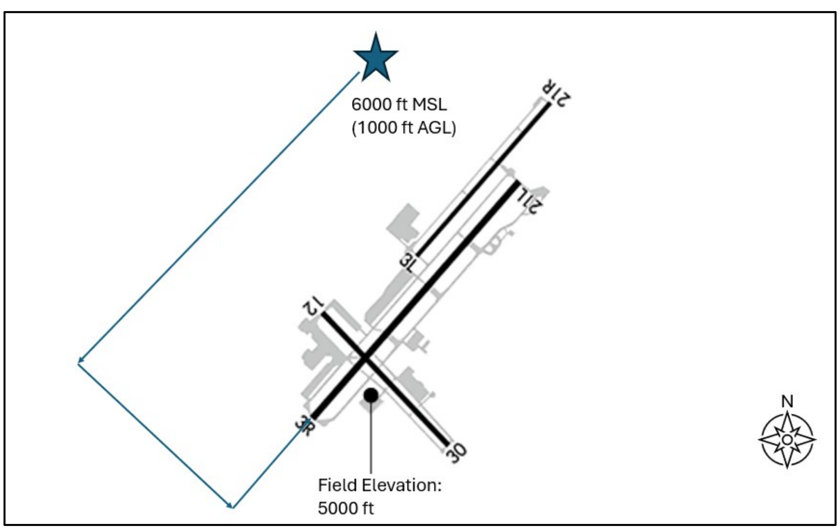


Figure C3: Pattern Mockup

Table C2: End of Survey Questionnaire

Question	Response Choices
On a normal flight (with no emergency): If given the choice, would you prefer your checklist to be presented with or without AR?	Strongly prefer without AR, Prefer without AR, No preference, Prefer with AR, Strongly prefer with AR
Compared to flying without AR, how did the AR display affect your ability to detect deviations (e.g., airspeed, altitude, bank angle)?	Much worse, somewhat worse, about the same, somewhat better, much better
To what extent do you agree with the following statement: "An augmented reality (AR) display could be useful during real-world flight operations."	Not at all useful, slightly useful, moderately useful, very useful, extremely useful
To what extent do you agree with the following statement: "An augmented reality (AR) display could be useful as a training aid during pilot instruction."	Not at all useful, slightly useful, moderately useful, very useful, extremely useful
To what extent did the AR display change your typical visual scan pattern during the flight?	Not at all, slightly, moderately, significantly
To what extent did the AR display interfere with your ability to visually acquire important flight information?	Not at all, slightly, moderately, significantly
To what extent did you trust the AR display as part of the cockpit system during the maneuver?	Not at all, slightly, moderately, significantly, completely
Is there anything else useful that you think we should know about your experience?	[Free response]

Appendix D: Additional Study 2 Data

Table D1: Central Tendency Measures for Flight Parameters, Normal Flight

Flight Parameter	No AR Median [IQR]	AR Median [IQR]	W	*p*	Cohen's *d*
Mean Glideslope (deg)	-2.79 *[-2.99, -2.74]*	-2.80 *[-3.08, -2.66]*	63	0.887	0.09
Mean Airspeed (kts)	70.34 *[66.80, 71.63]*	69.83 *[67.77, 71.66]*	44	0.379	-0.25
Mean Vertical Velocity (fpm)	-344.03 *[-370.70, -310.90]*	-347.90 *[-374.11, -319.73]*	74	0.443	0.18

Table D2: Statistical Tests for Workload and SA Subscales, Normal Flights

Self-Reported Workload and Situation Awareness Subscales During Normal Flights: Within-Subjects Comparisons

Measure	Without AR Mdn (IQR)	With AR Mdn (IQR)	*V*	*p*	*d*	*r [95% CI]*	*p* (Holm)
NASA-TLX subscales (0–100)							
Mental demand	32 (31)	30 (31)	51.5	.975	0.03	−.02 [−.53, .51]	1.000
Physical demand	19 (20.5)	20 (15.5)	28	1.000	0.04	.02 [−.51, .53]	1.000
Temporal demand	29 (21)	25 (26)	60	.660	0.21	.14 [−.41, .62]	1.000
Performance	78 (20)	75 (18.5)	48.5	.861	0.10	.07 [−.47, .57]	1.000
Effort	35 (20.5)	35 (32)	53	.712	−0.16	−.12 [−.60, .43]	1.000
Frustration	15 (11)	20 (26.5)	31	.894	0.07	−.06 [−.62, .54]	1.000
SART subscales (1–7)							
Instability	3 (2)	3 (2)	23	.516	0.18	.28 [−.28, .70]	1.000
Complexity	2 (0.5)	2 (1)	25.5	.305	0.30	.42 [−.13, .77]	1.000
Variability	3 (2.5)	3 (2)	21	.265	0.32	.50 [−.03, .81]	1.000
Arousal	5 (1.5)	4 (1.5)	10.5	.085	−0.49	−.62 [−.86, −.15]	1.000
Concentration	5 (1.5)	5 (2)	22.5	1.000	0.00	.00 [−.52, .52]	1.000
Division of attention	5 (2)	5 (2.5)	18	.177	−0.36	−.46 [−.79, .09]	1.000
Spare capacity	5 (2.5)	5 (2.5)	20.5	.774	0.11	.14 [−.41, .61]	1.000
Information quantity	4 (3)	5 (2.5)	34.5	.169	0.40	.53 [.02, .82]	1.000
Information quality	6 (2)	6 (2)	10	.572	0.21	.33 [−.23, .73]	1.000
Familiarity	5 (1.5)	5 (1.5)	43.5	.352	0.23	.32 [−.24, .72]	1.000

Note. $N = 15$. Paired Wilcoxon signed-rank tests comparing the With AR and Without AR normal flights. *Mdn* = median; *IQR* = interquartile range; *V* = signed-rank statistic; *d* = Cohen's *d*; *r* = matched-pairs rank-biserial correlation; CI = confidence interval; *p* (Holm) = Holm-corrected *p* value across the 16 subscales. No subscale reached significance.

Table D3: Statistical Tests for Workload and SA Subscales, Emergency Flight

Self-Reported Workload and Situation Awareness Subscales During the Emergency Flight: Between-Subjects Comparisons

Measure	Without AR (n = 7) Mdn (IQR)	With AR (n = 8) Mdn (IQR)	*W*	*p*	*d*	*r [95% CI]*	*p* (Holm)
NASA-TLX subscales (0–100)							
Mental demand	66 (31.5)	77.5 (15.75)	14.5	.132	−0.82	−.48 [−.81, .08]	1.000
Physical demand	26 (26)	47 (37.25)	16	.183	−0.72	−.43 [−.79, .15]	1.000
Temporal demand	48 (54)	74 (18.25)	25	.772	−0.38	−.11 [−.61, .46]	1.000
Performance	90 (12.5)	80 (33.75)	40	.173	0.98	.43 [−.15, .79]	1.000
Effort	40 (45.5)	64.5 (20.25)	17.5	.246	−0.81	−.38 [−.76, .21]	1.000
Frustration	9 (44.5)	28.5 (35.25)	21	.450	−0.29	−.25 [−.70, .34]	1.000
SART subscales (1–7)							
Instability	6 (1)	6 (1)	27.5	1.000	−0.30	−.02 [−.55, .53]	1.000
Complexity	5 (2)	5 (1.25)	21.5	.473	−0.44	−.23 [−.69, .35]	1.000
Variability	5 (2)	5 (1)	18.5	.277	−0.77	−.34 [−.74, .25]	1.000
Arousal	7 (0.5)	7 (1.25)	31.5	.679	0.37	.13 [−.45, .62]	1.000
Concentration	5 (3.5)	6.5 (1.25)	16	.166	−1.00	−.43 [−.79, .15]	1.000
Division of attention	3 (4)	6 (1.25)	16	.173	−1.05	−.43 [−.79, .15]	1.000
Spare capacity	3 (1)	2.5 (1.25)	40.5	.141	0.79	.45 [−.12, .80]	1.000
Information quantity	6 (1)	5 (2)	33.5	.551	0.20	.20 [−.39, .67]	1.000
Information quality	5 (2.5)	5 (1.5)	29	.952	−0.03	.04 [−.52, .56]	1.000
Familiarity	6 (2.5)	5 (2)	27	.952	−0.17	−.04 [−.56, .52]	1.000

Note. n = 7 (Without AR), n = 8 (With AR). Between-subjects Wilcoxon rank-sum tests. *Mdn* = median; *IQR* = interquartile range; W = rank-sum statistic; d = Cohen's d; r = rank-biserial correlation; CI = confidence interval; p (Holm) = Holm-corrected p value across the 16 subscales. No subscale reached significance.

Table D4: Statistical Tests for Workload and SA Composites, Normal Flights

Composite Workload, Situation Awareness, and Checklist Engagement During Normal Flights: Within-Subjects Comparisons

Measure	Without AR Mdn (IQR)	With AR Mdn (IQR)	*V*	*p*	*d*	*r [95% CI]*
NASA-TLX, overall	34.17 (12.08)	35.83 (10.67)	63	.887	0.09	.05 [−.48, .56]
SART composite	28 (8.5)	26 (6)	42	.529	−0.18	−.20 [−.65, .36]
Demand	9 (5.5)	8 (3)	56.5	.176	0.34	.45 [−.09, .79]
Supply	19 (4.5)	18 (5.5)	15	.115	−0.44	−.55 [−.83, −.04]
Understanding	15 (4)	16 (4.5)	35	.154	0.41	.56 [.05, .83]
Checklist items examined (%)	66 (54.5)	80 (49)	68	**.025**	0.75	.74 [.37, .91]
Checklist items rated useful (%)	65 (47.5)	63 (31)	40.5	.533	0.23	.23 [−.33, .67]

Note. N = 15. Paired Wilcoxon signed-rank tests. NASA-TLX overall ranges 0–100; SART composite = Understanding − (Demand − Supply), with Demand, Supply, and Understanding as component sums; checklist measures are percentages. *Mdn* = median; *IQR* = interquartile range; V = signed-rank statistic; d = Cohen's d; r = matched-pairs rank-biserial correlation; CI = confidence interval. Reported uncorrected. Boldface indicates $p < .05$.

Table D4: Statistical Tests for Workload and SA Composites, Emergency Flight

Composite Workload, Situation Awareness, and Checklist Engagement During the Emergency Flight: Between-Subjects Comparisons

Measure	**Without AR (n = 7) Mdn (IQR)**	**With AR (n = 8) Mdn (IQR)**	***W***	***p***	***d***	***r [95% CI]***
NASA-TLX, overall	44.17 (31.25)	61.08 (12.67)	18	.272	−0.58	−.36 [−.75, .23]
SART composite	22 (12)	18.5 (3.25)	30	.862	−0.02	.07 [−.49, .59]
Demand	15 (4)	15.5 (3.5)	20	.382	−0.61	−.29 [−.72, .30]
Supply	19 (5)	21 (3)	19	.323	−0.53	−.32 [−.73, .27]
Understanding	15 (6.5)	15.5 (5)	28	1.000	−0.02	.00 [−.54, .54]
Checklist items examined (%)	3 (8.5)	9.5 (16.25)	22	.520	−0.06	−.21 [−.68, .37]
Checklist items rated useful (%)	41 (51.5)	10 (13.25)	32	.684	0.56	.14 [−.43, .63]

Note. n = 7 (Without AR), *n* = 8 (With AR). Between-subjects Wilcoxon rank-sum tests. NASA-TLX overall ranges 0–100; SART composite = Understanding − (Demand − Supply); checklist measures are percentages. *Mdn* = median; *IQR* = interquartile range; *W* = rank-sum statistic; *d* = Cohen's *d*; *r* = rank-biserial correlation; CI = confidence interval. Reported uncorrected.

Table D5: Central Tendency Measures for Flight Parameters, Emergency Flight

Flight Parameter	**No AR Median [IQR]**	**AR Median [IQR]**	**W**	***p***	**Cohen's *d***
Mean Glideslope (deg)	-5.86 *[-6.11, -5.00]*	-4.98 *[-5.93, -4.37]*	22	0.524	-0.03
Mean Airspeed (kts)	64.19 *[63.08, 65.30]*	62.07 *[60.32, 66.36]*	34	0.524	0.36
Mean Vertical Velocity (fpm)	-669.7 *[-720.71, -613.44]*	-624.64 *[-706.75, -472.77]*	24	0.685	-0.06