Assessing Cognitive Workload in Mixed Reality Flight Simulators for Naval Aviation

Thomas Cecil, Charles Rowan, Perry McDowell
Naval Postgraduate School
Monterey, CA
thomas.cecil@usmc.mil, charles.rowan@nps.edu, mcdowell@nps.edu

Jonathan Vogl
US Army Aeromedical Research Laboratory
Fort Novosel, AL
jonathan.f.vogl.civ@health.mil

ABSTRACT

Background

The adoption of mixed-reality (MR) simulators by the Naval Aviation Enterprise (NAE) necessitates a comprehensive understanding of the impact of MR technology on human factors. Unlike traditional projected-screen simulators, MR simulators utilize head-mounted displays (HMDs) to present virtual environment visuals in close proximity to the pilot's eyes. This raises concerns regarding the distortion of natural human depth perception and distance estimation, highlighting the critical factor of stereoscopy on performance efficiency and cognitive load.

Two primary types of MR HMDs exist: optical see-through (OST) and video see-through (VST). In OST, users perceive actual objects in the physical environment, with virtual objects overlaid on a semi-transparent display within their field of view. Conversely, VST captures video from the user's viewpoint, integrating virtual objects with this video and presenting the composite imagery on screens in close proximity to the user's eyes. Notably, in OST, users must adjust their focal distance when transitioning between viewing virtual and real objects, whereas in VST, all objects are rendered on screens, eliminating the need for focal adjustments. However, no prior research has explored whether this distinction influences cognitive workload or susceptibility to cybersickness.

Significance

This study enhances the NAE's understanding of MR technology, facilitating the development of more effective MR simulators, delineating their limitations, and offering insights into integrating MR simulators into flight training curricula. The results suggest potential revisions of Naval Aviation training manuals and directives.

Methods

We assessed performance disparities among approximately thirty subjects using a version of the NASA Multiple Attribute Test Battery. Each subject engaged in tasks under three conditions with order randomly assigned:

1. OST MR (Hololens).
2. VST MR (Varjo XR-4).
3. Legacy display.

Results

Analysis indicates difference display methods create significant differences in users' cognitive workload as measured by both participants' subjective workload ratings and MATB performance.

ABOUT THE AUTHORS

Thomas Cecil is a Major in the U.S. Marine Corps and an MV-22 pilot serving as the III MEF Modeling and Simulation Officer at Camp Courtney, Okinawa, Japan. He is a graduate of the United States Naval Academy and the Naval Postgraduate School.

Lieutenant Colonel Charles Rowan, PhD, is a US Army Simulation Operations Officer serving as the Interim Director of the Modeling, Virtual Environments, and Simulation (MOVES) Institute at the Naval Postgraduate School in Monterey, CA. He is a graduate of the United States Military Academy and the Naval Postgraduate School.

Perry McDowell is a Faculty Associate for Research at the MOVES Institute. Mr. McDowell is a graduate of the United States Naval Academy and the Naval Postgraduate School.

Jon Vogl is a research psychologist at the US Army Aeromedical Research Laboratory at Fort Novosel, Alabama. Mr. Vogl is a graduate of South Dakota State University and the University of South Dakota.

Assessing Cognitive Workload in Mixed Reality Flight Simulators for Naval Aviation

Thomas Cecil, Charles Rowan, Perry McDowell
Naval Postgraduate School
Monterey, CA
thomas.cecil@usmc.mil, charles.rowan@nps.edu, mcdowell@nps.edu

Jonathan Vogl
US Army Aeromedical Research Laboratory
Fort Novosel, AL
jonathan.f.vogl.civ@health.mil

INTRODUCTION

Background

Military flight training has long relied on various simulators, including Full Flight Simulators (FFSs) with hydraulic systems and Containerized Flight Training Devices (CFTDs). These advanced simulators offer high-quality training with realistic cockpit replicas and flight optics through projector screens but are costly and require substantial space (Perry, 2023). Recognizing these limitations and high costs, the Department of the Navy (DON) emphasizes the need for innovative, cost-effective training systems to enhance naval aviation capabilities. The naval services aim to develop distributed, deployable, low-cost simulators to achieve this goal (Commander Naval Air Forces, 2021; Deputy Commandant for Aviation, 2022).

Problem Statement

Alternative mixed-reality (MR) technologies could support DON aviation plans but introduce new human factors considerations. MR simulators use head-mounted displays (HMDs) to merge virtual and real-world visuals, reducing costs by minimizing hardware and support expenses (Perry, 2023). Their smaller size allows deployment on ships or other locations. Video see-through (VST) HMDs represent one solution that uses cameras to record and merge real and virtual environments into one display. This method presents unique human factors challenges, particularly in negating natural depth perception due to vergence-accommodation conflict (VAC). VAC can cause symptoms like fatigue, headache, and difficulty focusing, affecting MR simulator effectiveness (Kennedy et al., 1993; Hua, 2017a). The extent to which VAC symptoms affect pilots in MR applications is not entirely known.

An alternative, optical see-through (OST) displays, overlays virtual objects onto a see-through lens, maintaining natural depth perception and preventing VAC-induced symptoms. However, OST displays face challenges like limited fields of view and low luminance (Condino et al., 2020). Comparing VST and OST displays can help determine the impact of these technologies on oculomotor human factors.

Research Questions

The Naval Aviation Enterprise (NAE) must better understand the impact of MR technology on human factors as naval aviation transitions to MR simulators. Unlike legacy simulators with projector screens around a cockpit, MR HMDs present visuals just centimeters from the pilot's eyes, affecting natural depth perception and distance estimation. Thus, a crucial human factor is the impact of binocular depth cues on performance efficiency.

This study examines the impact of depth-perception-related human factors in MR HMDs on a pilot's perceived workload by addressing the following research questions:

1. How is cognitive task loading affected by the limitations of 3D perception when using HMDs in MR training environments?
2. Do OST and VST HMDs result in a difference in cognitive workload in an MR training environment?

Research Approach

This research aimed to understand the impacts of various simulator displays on cognitive workload. A literature review of the technology and relevant human factors informed our methodology. We present participants with standardized task loads using a multiple-attribute task battery (MATB) developed by the United States Army Aeromedical Research Laboratory (USAARL). We developed our own tracking task (Cecil, 2024) for display with a computer monitor, OST HMD, and VST HMD. Participants completed a trial using each of these displays in randomized order. We analyzed their performance, heart rate variability, and subjective workload assessments using multiple statistical methods to determine whether significant differences exist for cognitive workload between the display groups.

Importance Of Research

This study sought to contribute to the NAE's understanding of MR HMD technology to support the naval services' implementation of these devices into their training regimens. A comprehension of MR HMD's limitations and their effects on the human user is essential to this understanding. Other studies, like those conducted by NAWCTSD (McCoy-Fischer et al., 2019; Natali et al., 2023), focused on system usability. This study augments the work of others by focusing on the user's mental workload as affected by VAC and related depth perception issues while interacting with the cockpit. Further, the study compares VST and OST HMDs, informing whether significant effects on cognitive workload result from perception of the real world alone. This study also evaluated simulator sickness symptoms with special attention to oculomotor symptoms. While not the focus of this study, the discovery of significant simulator sickness symptoms may provide context for findings of significant cognitive workload differences between display methods.

LITERATURE REVIEW

Simulators are essential in naval aviation training. The DON can enhance its capabilities with MR technology to meet future challenges and reduce costs. Unlike traditional simulators, MR applications use HMDs to immerse users in virtual environments. However, MR HMDs introduce VAC which has been linked to attentional resource depletion (i.e., cognitive workload). Understanding the link between MR HMDs and cognitive workload is crucial for effective integration into training – unintentional and/or unknown increases in workload may negatively impact training effectiveness.

Simulators and Naval Aviation

Simulators will continue to play a critical role in naval aviation training, but with constrained resources, the naval services have identified a need for innovation. The Navy and Marine Corps' visions are detailed in the Navy Aviation Plan 2030-2035 (Commander Naval Air Forces, 2021) and the 2022 Marine Corps Aviation Plan (Deputy Commandant for Aviation, 2022). Specifically, these services envision implementing fiscally responsible training solutions that improve combat readiness while also enabling distributed training with deployable simulators. While this vision does not represent a wholesale departure from legacy simulators, the services see virtual and MR simulators as the enabling technology for their visions.

As outlined in CNAF 3710, simulator use spans various applications, including NATOPS, instrument, and crew resource management (CRM) evaluations and completion of training and readiness (T&R) events. Further recognized for their cost-savings and role in mitigating risks associated with live-flight training, simulators play a crucial role in substituting flight time to maintain fundamental aeronautical skills (Department of the Navy [DON], 2022; Judy & Gollery, 2018). Their use is further governed at the service and platform level to promote safe and efficient uses of resources to build combat readiness. To enable this end, entities must appreciate the effects of human factors on the simulator's usability and operator performance.

Mixed Reality as a Solution to Modern Challenges

Although legacy simulators provide immense training value, they have shortcomings. Using projection screens to provide users with a visual representation of flight environments, legacy simulators are large and limit a pilot's field of regard. They are also expensive to purchase, operate, maintain, and update (Perry, 2023). Conversely, MR simulators often implement commercial-off-the-shelf (COTS) components versus tailor-made hardware used by legacy simulators. This results in a relatively cost-effective alternative with improved immersion and space savings offered by the HMD (McCoy-Fisher et al., 2019).

VST HMDs are the MR HMD receiving the greatest interest by the DoN (McCoy-Fischer et al., 2019; Perry, 2023). Using this type of HMD, the user perceives the real environment via cameras on the front side of the HMD (Itoh et al., 2022). The video feed of the real world is combined with the virtual flight environment via graphics shaders to provide the user with a coherent scene (Azuma, 1995). Since a user of VST HMDs does not see the real world directly, challenges of this MR technology include video resolution limiting a user's visual acuity and VAC (Geyer & Biggs, 2018; McCoy-Fisher et al., 2019).

OST HMDs are another MR technology that mitigates some of the challenges with VST HMDs. OST HMDs allow the user to see the real environment through the headset's lenses, as experienced wearing glasses. Occluding the real environment, the virtual environment is projected onto reflective, semi-transmissive beam combiners. While this simpler method does not prevent the user from directly seeing the real environment, this feature also reduces the user's immersion (Azuma, 1995).

HUMAN FACTORS AND MIXED REALITY

The human element is a crucial factor in flight simulators. MR HMDs bring specific human factors into focus, particularly the human visual system and its interaction with these devices. Therefore, it is vital to understand human depth perception and the challenges posed by MR HMD technology.

Depth Perception and Mixed Reality

Correctly perceiving 3D space is crucial for interacting with objects or training in virtual reality. Visual depth cues are classified as visual or oculomotor (Reichelt et al., 2010). Oculomotor cues include accommodation, convergence (vergence), and pupillary constriction (Reichelt et al., 2010). Vergence refers to the rotation of the eyes toward an object, while accommodation involves the eye's ciliary muscles focusing to minimize blur (Hua, 2017a). These closely coupled phenomena, along with interpupillary distance (IPD), help triangulate an object's location relative to the viewer (Itoh et al., 2022).

The HMDs used in this study are binocular, offering an autostereoscopic display with a unique view for each eye (Reichelt et al., 2010). 3D HMDs rely on monocular and binocular cues for depth perception and distance estimation. However, HMDs present challenges as they render images on a 2D screen at a fixed distance, causing a mismatch between vergence and accommodation (Inoue & Ohzu, 1997). This VAC leads to issues like fatigue and eye strain (Hua, 2017a).

Objects rendered in focus regardless of their depth eliminate the natural blur gradient, preventing the eye from accommodating different focal distances (Watt et al., 2005). This fundamental conflict affects HMD usability (Hua, 2017b; Reichelt et al., 2010). VAC impacts both OST and VST HMDs, though VST HMDs experience it when viewing both real and virtual environments (Rolland & Fuchs, 2000).

VAC and Human Factors

Understanding the human factors stemming from VAC is crucial as it is inherent in HMDs. VAC causes accommodation to respond differently to stereoscopic 3D displays than to real objects, leading to visual fatigue due to the imbalance in visual function (Inoue & Ohzu, 1997). VAC also affects vergence dynamics, resulting in increased latency and variable vergence velocity, which are indicators of visual fatigue (Vienne et al., 2014). Immersion in stereoscopic HMDs increases heterophoria, contributing to visual fatigue and challenges in binocular fusion (Wann et al., 1995). Optimal binocular fusion, depth perception, and reduced visual fatigue occur when vergence and focal distances align (Hoffman et al., 2008).

Research on binocular disorders like accommodative and convergence insufficiencies highlights similar effects. These disorders cause blurred vision, eyestrain, fatigue, and concentration difficulties (Scheiman & Wick, 2020). Studies show a link between accommodative lag and slower reaction times, greater response variability, and cognitive fatigue (Poltavski et al., 2012). Bernhardt and Poltavski (2021) found a strong correlation between symptoms of accommodative-vergence deficits, task engagement, and subjective cognitive fatigue during flight-related tasks using NASA's MATB II, supporting the idea that stressors deplete cognitive resources (Matthews et al., 2017).

However, Bernhardt and Poltavski (2021) noted the lack of objective measures of vergence and accommodation or a display method known to evoke VAC as a limitation in their study. They recommended further research to explore the effects of accommodative-vergence stress on cognitive states, suggesting the inclusion of eye tracking, heart rate variability, and subjective metrics for a more comprehensive analysis.

COGNITIVE WORKLOAD

Introduction to Cognitive Workload

Longo et al.’s (2022) definition of mental workload, Wickens et al.’s (2013) Human Information Processing (HIP) model, and Wickens’s (2008) Multiple Resource Theory (MRT) provided our foundation for understanding cognitive workload. The following are key elements of these works that informed our approach:

1. The human cognitive system possesses finite resources.
2. Circumstances and factors experienced affect the task.
3. Internal and external factors influence one’s level of attention and effort, affecting workload.

Tasks sharing commonalities across different cognitive levels compete for limited cognitive resources. For instance, the visual resources needed to read system states compete with manual control responses or vocal commands resulting from a visual scan of an aircraft’s spatial information.

Measuring Cognitive Workload

The three widely accepted categories for measuring cognitive workload are task performance, physiological, and self-reported measures (Longo et al., 2022). All three categories have benefits and limitations that warrant their combined use. Longo et al.’s (2022) survey of human mental workload provides a thorough overview of the categories. Task performance is generally used to assess an operator’s workload based on how efficiently they complete tasks. Physiological measures serve as objective surrogates of cognitive workload by analyzing an operator’s neurophysiological responses. Self-reported measures are subjective and involve an operator assessing their personal experience while completing tasks (Longo et al., 2022). This study used metrics from all three categories as discussed in the Methods section.

The study also utilized the Simulator Sickness Questionnaire (SSQ) (Kennedy et al., 1993). Although the SSQ does not measure cognitive workload, it can identify stressors that may affect cognitive vigilance and represents another important human factor affected by MR HMDs.

Driving Workload Through Standardized Tasks

This study used the United States Army Aeromedical Research Laboratory’s MATB (ARL-MATB) (Vogl et al., 2023), developed in coordination with Tennessee State University (TSU), as it proved highly customizable to meet this study’s design and is owned by the Department of Defense. The MATB facilitates the scheduling of events by the experimenter in an events file. The participant interacts with the MATB subtasks as scheduled, and their performance is captured for performance evaluation. The schedule of events allows the experimenter to subject multiple participants to a standardized task load and has proved itself an indispensable instrument in studies investigating task loading and cognitive impacts (Bernhardt & Poltavski, 2021; Rowan, 2023). The MATB subtasks include:

- System Monitoring – Participants recognize and acknowledge warning lights.
- Tracking – Participants work to keep a target centered on a reticle using a joystick control.
- Communications – Participants listen to radio transmissions for their callsign, “NASA504”, updating radios and frequencies as directed.

- Resource Management – Participants manipulate fuel pumps and negotiate pump failures and shutoffs to keep fuel tanks at target levels.

HYPOTHESES

Since simulators play such a crucial role in naval aviation training, undergoing rigorous evaluations for certification and integration into training programs, the DON’s interest in MR technology to enhance its future capabilities needs similar rigorous testing and understanding. Unlike traditional simulators, MR applications immerse users in virtual environments through innovative methods. However, MR technology introduces different physiological responses than traditional simulators such as VAC, which can cause oculomotor simulator sickness symptoms and increase cognitive workload. Understanding the links between the technology, physiological responses, and cognitive workload is essential for effectively integrating MR devices into naval aviation training.

Based upon our literature review, two hypotheses regarding MR simulators were investigated:

- Hypothesis 1: MR HMDs will result in a significantly greater cognitive workload than legacy displays.
- Hypothesis 2: VST HMDs will result in a significantly greater cognitive workload than OST HMDs.

METHODOLOGY

Overview

This study aimed to identify participants' cognitive workload differences through performance, subjective assessments, and physiological measures. A 2 x 3 repeated measures design compared two workload conditions and three display methods. USAARL's MATB was used to induce participant workload with flight-related tasks. Subjective workload was measured using CSWAG and NASA-TLX, while heart rate variability provided physiological workload data. MATB performance metrics were also analyzed. The study's results determined if different display methods in flight training significantly affect users’ cognitive workload. We conducted a pilot study to verify experimental design prior to conducting the experiment.

Participants

Thirty participants completed an institutional review board approved study (mean age = 34.5, standard deviation [SD] = 6.47). The participants included 28 males and two females. Of the participants, 28 were in the military, with time in service ranging from 1 year to 23 years (mean years of service = 11.32, SD = 5.8). Military participants, including two pilots, encompassed personnel from various occupational specialties. All participants were graduate students or civilian employees at Naval Postgraduate School (NPS).

Materials

Multiple Attribute Test Battery

This study utilized a USAARL developed MATB to drive participant workload. This version of the MATB uniquely enabled touchscreen compatibility and allowed for the display of the task battery subtasks in separate windows on three different tablet displays. The USAARL MATB also enabled the ability to set various parameters for the associated subtasks.

The study did not implement the USAARL MATB tracking task due to its nature. Instead, the researcher programmed a stand-alone tracking task in Unity. Implementing a stand-alone tracking task enabled ease of use across all three display methods. The Unity program used Lab Streaming Layer to send tracking task scores to the data collection software.

Displays

Participants experienced the tracking task using three display methods during the study. The computer monitor, or Legacy display, gave participants a tracking task akin to legacy simulator technology. This study used a Dell 24" monitor. The monitor sat behind the tablet displays and directly in front of the sitting participant at approximately eye level. The HMDs replaced the monitor during their respective trials and allowed for the display of the tracking task via a virtual display. The HoloLens 2 served as the OST HMD. The Varjo XR-4 served as the VST HMD. Due to technical difficulties and time constraints described by Cecil (2024), we used the Varjo XR-4 FE and FF models.

The study replicated a 3D cockpit employing three Galaxy Tab S8s, requiring participants to interact physically with the other MATB subtasks. Additional specifications and information about the supporting hardware and software are found in the thesis (Cecil, 2024). Figure 1 shows the experimental set up with the different displays. Of note, the HoloLens 2 looked similar to the Varjo image in Figure 1.

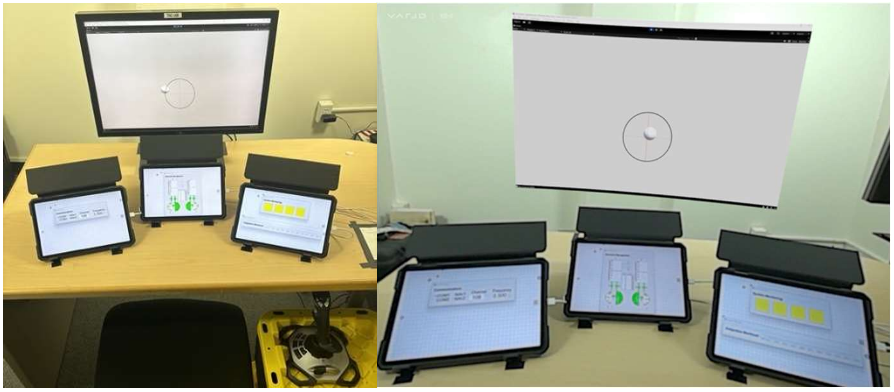

Figure 1. Set Up with Legacy Display (L) and Image through Varjo (R).

Variables

This study manipulated two independent variables: display method and workload. We used the display method as the primary independent variable to evaluate the research questions. The study consisted of three trials using three different display methods for the tracking task. We randomized the display methods to prevent any potential learning order effects. This is the only variable that changed between trials within participants. We also implemented low and high workloads during the trials to prevent underwhelming or overwhelming task loads (de Waard, 1996). Previous studies that utilized NASA's MATB-II informed our three workload parameter files (McCurry et al., 2022; Rowan, 2023). While the order of the display methods varied between participants, we kept the order of the parameter files the same.

This study's dependent variables were performance, subjective, and physiological metrics. The MATB generated performance scores for the tracking, communications, resource management, and system monitoring tasks. Subjective workload response times also served as a performance measure. Participants provided subjective workload ratings via CSWAG during the trials and NASA TLX after each trial. Finally, heart rate variability using the root mean square of successive differences provided a physiological measure of participant workload. We considered using pupil diameter as a physiological measure but elected to exclude it due to technical differences in display method brightness and the 3D nature of the test bench producing too many confounding variables for meaningful analysis. We also collected simulator sickness data (Kennedy et al., 1993) and participants' evaluation of tablet display readability, providing context to the results.

Test Procedure

Following a 15 minutes MATB training session and completion of an inventory SSQ, participants completed a 10-minute MATB trial with each tracking task display method. Participants experienced the display methods in randomized order. During a trial, the participant completed all the MATB subtasks concurrently. Following trial, participants completed a SSQ and NASA-TLX. The researchers collected demographic information via a survey after the final trial.

RESULTS

Using a MATB to drive workload conditions, this study explored cognitive workload across various display types. The analytical approach is described first. The results on cognitive workload, visual acuity, and simulator sickness follow. All analyses applied a .05 alpha level.

Analytical Approach

A statistical power analysis indicated that a multivariate analysis of variance (MANOVA) required a sample size of 29 participants to detect moderate effects of display methods on the dependent variables. The repeated measures MANOVA assessed significant differences across display methods. Discriminant analysis identified statistically significant dependent variables from the MANOVA. Univariate analysis provided further insights into individual measures across display methods. Analysis of visual acuity ratings and simulator sickness scores provides additional context for the discussion, enriching the interpretation of the study's findings.

Cognitive Workload Results

MANOVA

We performed a MANOVA to evaluate the impact of display methods on various dependent variables, including MATB performance scores, CSWAG ratings, reaction times (CSWAGrt), and heart rate variability (RMSSD). The test revealed a significant effect of the display methods on these variables, with all p-values being less than 0.001.

Discriminate Analysis

We conducted a discriminant analysis to understand the significant effects identified by the MANOVA, creating two functions to predict participant trial display groups using the original dependent variables. Function 1, explaining 89.3% of the variance with a canonical correlation of 0.513, showed a stronger correlation with display groups than Function 2, which explained 10.7% of the variance. A Wilk's Lambda test indicated that Function 1 is statistically significant (p = 0.010). The structure matrix revealed that system monitoring (SysMon), communications (COM), and tracking highly correlate with Function 1, while CSWAGrt and RMSSD correlate most with Function 2. Figure 2 illustrates the discriminate function scores for each trial, highlighting the 89.3% variance explained by Function 1 with clear lateral separation along the x-axis for the display group centroids.

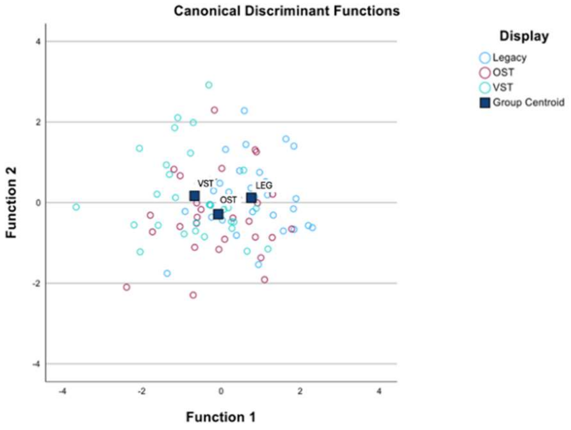

Figure 2. Discriminate Function

Classification results showed that the discriminant functions accurately attributed Legacy display trials with 70% accuracy, VST trials with 53.3% accuracy, and OST trials with 26.7% accuracy. Cross-validation yielded similar results, with 60.0% accuracy for Legacy, 50.0% for VST, and only 13.3% for OST trials, often incorrectly attributed to Legacy and VST groups.

Univariate Analysis

The univariate analysis revealed significant effects of display methods on several dependent variables, with tracking ($p < 0.001$), system monitoring ($p < 0.001$), CSWAG ($p < 0.001$), and CSWAGrt ($p = 0.002$) all showing statistically significant differences as confirmed by ANOVAs using a Holm-Bonferroni correction to reduce Type I error. Significant p values are less than 0.0071. However, resource management ($p = 0.016$), communication ($p = 0.070$), and RMSSD ($p = 0.831$) showed no significant effects. Regardless of significant effects, MR HMDs produced increased CSWAG scores and reaction times, along with reduced performance metrics compared to the Legacy display group. This trend holds true with VST HMDs performing worse than OST HMDs. RMSSD is the only variable without any noticeable differences or trends among the display methods.

NASA-TLX

Researchers conducted a separate ANOVA on NASA TLX scores, collected at the end of trials, and found significant effects of the display method (p = 0.001). Paired sample t-tests revealed significant differences between all display groups: Legacy display vs. OST HMD (t(29) = -2.588, p = 0.015), Legacy display vs. VST HMD (t(29) = -4.25, p < 0.001), and OST vs. VST HMDs (t(29) = -3.317, p = 0.002). Average NASA TLX scores, depicted in Figure 3, increase from Legacy displays (M = 58.8, SD= 10.054, n = 30) to OST HMDs (M = 61.8, SD = 7.543, n = 30) to VST HMDs (M = 67.2, SD = 10.196, n = 30).

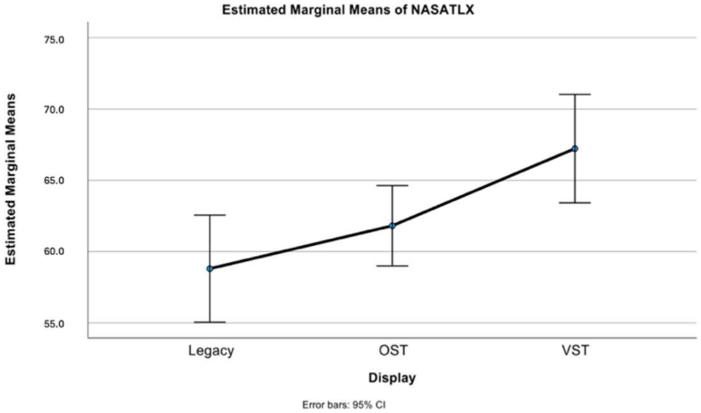

Figure 3. Mean NASA TLX Scores by Display

Simulator Sickness and Visual Acuity Ratings

After each trial, we asked participants to complete two surveys in addition to the NASA-TLX. We surveyed visual acuity ratings (VARs) (Cecil, 2024) to assess differences in a participant's ability to read the information on the tablet displays to rule out associated confounding factors such as pass-through video resolution (McCoy-Fisher et al, 2019). We also administered the SSQ (Kennedy et al., 1993) to detect oculomotor and other simulator sickness symptoms. Analysis of VARs and SSQ scores revealed significant differences between all display methods. The VST display group provided the lowest visual acuity ratings and the greatest simulator sickness symptoms, while the Legacy display group performed best. Differences in subjective ratings between the Varjo XR-4 FE and FF models were not significant, t(28) = 0, p = 1.0. Further discussion of these analyses is provided by Cecil (2024).

DISCUSSION

MR flight simulators are poised to revolutionize naval aviation training with cost-effective, deployable solutions (Commander Naval Air Forces, 2021; Deputy Commandant for Aviation, 2022; Natali et al, 2023). While MR HMDs offer numerous benefits, understanding their impact on users is essential. MR HMDs disrupt natural human depth perception, causing VAC (Hua, 2017a), a well-documented issue (Bernhardt & Poltavski, 2021; Inoue & Ohzu, 1997; Poltavski et al., 2012; Scheiman & Wick, 2020; Vienne et al., 2014; Wann et al., 1995). This study investigated how HMD limitations in 3D perception affect cognitive task loading and whether significant differences exist between legacy displays and OST and VST HMDs.

The study concluded that cognitive workload is significantly higher with MR HMDs compared to legacy display technology. MANOVA results indicated significant differences between the display methods. Discriminant functions accurately predicted group membership for Legacy and VST HMDs at 70% and 53.3%, respectively, while OST HMD group membership proved difficult to predict. ANOVAs, adjusted for family-wise error rate, confirmed significant effects of the display method on tracking scores, system monitoring scores, CSWAG ratings, and CSWAG reaction times. Plots of these variables (Cecil, 2024) showed improved performance metrics and workload ratings for legacy displays compared to the MR HMDs. Additionally, ANOVA identified significant differences in NASA TLX scores between the display groups, with paired-sample t-tests confirming significant differences with reduced cognitive workload for legacy displays versus MR HMDs.

The study also concluded that cognitive workload is significantly higher with VST HMDs than with OST HMDs. While MANOVA results supported significant differences between the three display methods, the discriminate functions struggled to distinguish OST HMDs, predicting OST group membership at only 26.7% and misclassifying

OST trials as Legacy and VST groups at rates of 33.3% and 40%, respectively. However, ANOVAs confirmed the significant effects of the display method on various performance metrics, with plots showing improved metrics and workload ratings for OST HMDs compared to VST HMDs. Paired-sample t-tests confirmed significant differences in NASA TLX scores between OST and VST HMDs.

Implications to Flight Training

This study was not specific to any aircraft and does not suggest that MR HMDs are unsuitable for fulfilling basic and instrument flight requirements per CNAF 3710 (DON, 2022). In fact, Naval Aviation Training Next projects have successfully incorporated MR HMDs into flight training (Blow, 2023; Correll, 2021; Mishler et al., 2022). However, our study found that MR HMDs are linked to higher cognitive workload, reflected in lower subtask scores, slower reaction times, and increased mental workload reported via CSWAG and NASA TLX. Naval aviation must consider these factors before fully integrating MR simulators.

First, instructional system designers should distinguish between legacy and MR flight simulators, identifying training events suited for MR simulators (McCoy-Fisher et al., 2019; Natali et al., 2023). This differentiation should be included in T&R manuals to align simulator capabilities with training goals. While we do not specify compatible mission types, tasks requiring extensive heads-down scanning may result in poorer performance and higher mental strain with MR simulators. Limited pass-through fidelity associated with VST HMDs likely contributes to increased mental workload (Natali et al., 2023), though our study did not adequately address this issue.

Second, the NAE should inform aircrew about the cognitive workload impacts of MR simulators. Although these impacts do not directly threaten flight safety, misunderstanding them can hinder effective MR simulator use. Flight instructors, evaluators, and squadron standardization personnel are the primary audiences for this information. This recommendation excludes issues like limited Field of View (FoV) and simulator sickness, which require further consideration and discussion in directives like CNAF 3710 and NAVMC 3500.

Lastly, OST and VST HMDs increased oculomotor and total SSQ scores. CNAF 3710 (DON, 2022) notes that simulator sickness symptoms, including nausea and disorientation, have been observed in various simulators. This study's findings support updating directives to caution that MR devices may cause greater simulator sickness symptoms than legacy simulators.

Limitations & Future Work

This experimental study faced technological challenges and limitations that warrant discussion. Using a MATB with aviation-similar tasks, participants judged the 3D location of touchscreen buttons rather than operating a complex cockpit with 3D switches and knobs, limiting the depth perception and distance estimation required. The study examined cognitive workload impacts driven by participants' interaction with their surroundings but did not engage the depth perception needed for 3D flight operations. Trials were 10 minutes long, potentially insufficient for participant learning to equalize performance across display methods. The participant group included few pilots, who might have mitigated MR HMD effects due to their flight experience. Technical challenges with the XR-4 FE HMD led to using the XR-4 FF model; however, a two-tailed t-test showed no significant visual acuity differences ($p = 1.0$). Despite no significant differences in visual acuity ratings, Varjo HMD limitations in reduced foveated rendering areas were noted and will be addressed in future software releases (Burwell, 2024). Finally, this thesis primarily used heart rate variability (HRV) to measure cognitive workload, but RMSSD analysis showed no significant differences between display types, suggesting HRV may be too coarse for this application.

This study demonstrated the effects of MR HMDs on cognitive workload as participants completed aviation-similar tasks. Additional work is required to understand how these effects will translate to specific aircraft and flight maneuvers over the full length of a flight sortie. Next, heart rate variability proved too coarse for our study. Electroencephalogram (EEG), galvanic skin response (GSR), and electrodermal activity (EDA) are methods surveyed by Longo et al. (2022) that may prove more effective with this application and warrant further investigation. Further, our visual acuity ratings failed to rule out visual acuity as a confounding factor, as we found significant differences between display methods. Future work should assess participants' ability to read specific texts before beginning the trial. Finally, our study found that OST HMDs performed somewhere between legacy technology and VST HMDs. Further work can be done to investigate the usability and cognitive demands of an OST HMD simulator.

CONCLUSION

This study aimed to determine if MR flight simulators are more cognitively demanding than legacy simulators. Performance metrics and participant workload ratings confirm this assertion. While only limited recommendations can be made for MR simulator use in naval aviation, the findings suggest a need for further research to understand these cognitive workload differences and their implications for MR simulator use.

ACKNOWLEDGEMENTS

The authors would like to thank the Navy Modeling & Simulation Office and NAWCTSD for supporting this research.

REFERENCES

Azuma, R. T. (1995). *Predictive tracking for augmented reality* [Dissertation, UNC-Chapel Hill].

Bernhardt, K. A., & Poltavski, D. (2021). Symptoms of convergence and accommodative insufficiency predict engagement and cognitive fatigue during complex task performance with and without automation. *Applied Ergonomics*, *90*, 103152. https://doi.org/10.1016/j.apergo.2020.103152

Blow, W. (2023, February 3). Navy training squadron set to fully adopt modernized flight program. DVIDS. https://www.dvidshub.net/news/437792/navy-training-squadron-set-fully-adopt-modernized-flight-program

Burwell, J. (2024, May 9). Re: Varjo XR-4 FE or Base Model [Email message].

Cecil, T. A. (2023). *Understanding the effects of mixed reality head-mounted displays on cognitive workload in naval aviation.* [Unpublished master's thesis, Naval Postgraduate School].

Commander Naval Air Forces. (2021). *Navy Aviation Vision 2030–2035*. United States Navy. https://media.defense.gov/2021/Oct/27/2002881262/-1/-1/0/NAVY%20AVIATION%20VISION%202030-2035_FNL.PDF

Condino, S., Carbone, M., Piazza, R., Ferrari, M., & Ferrari, V. (2020). Perceptual limits of optical see-through visors for augmented reality guidance of manual tasks. *IEEE Transactions on Biomedical Engineering*, *67*(2), 411–419.

Correll, D. (2021, May 26). Navy's new 'Project Avenger' flight training program aims to produce stronger aviators. *Navy Times*. https://www.navytimes.com/news/your-navy/2021/05/25/navys-new-project-avenger-flight-training-program-aims-to-produce-stronger-aviators/

Waard, D. de. (1996). *The measurement of drivers' mental workload* [The Traffic Research Centre VSC, University of Groningen]. https://pure.rug.nl/ws/portalfiles/portal/13410300/09_thesis.pdf

Department of the Navy [DON]. (2022). *NATOPS general flight and operating instructions manual* (CNAF M-3710.7). https://www.cnatra.navy.mil/assets-global/docs/cnaf-m-3710.7.pdf

Deputy Commandant for Aviation. (2022). *2022 Marine Aviation Plan*. United States Marine Corps. https://www.aviation.marines.mil/Portals/11/Documents/Aviation%20Plan/2022%20Marine%20Aviation%20Plan%20FINAL%20April%202022.pdf

Geyer, D. J., & Biggs, A. T. (2018). The persistent issue of simulator sickness in naval aviation training. *Aerospace Medicine and Human Performance*, *89*(4), 396–405. https://doi.org/10.3357/AMHP.4906.2018

Hoffman, D. M., Girshick, A. R., Akeley, K., & Banks, M. S. (2008). Vergence–accommodation conflicts hinder visual performance and cause visual fatigue. *Journal of Vision*, *8*(3), 33. https://doi.org/10.1167/8.3.33

Hua, H. (2017a). Enabling focus cues in head-mounted displays. *Proceedings of the IEEE, 105*(5), 805–824. https://doi.org/10.1109/JPROC.2017.2648796

Hua, H. (2017b). Optical methods for enabling focus cues in head-mounted displays for virtual and augmented reality. *Proceedings*, *Three-Dimensional Imaging, Visualization, and Display 2017*, 102190L. https://doi.org/10.1117/12.2264157

Inoue, T., & Ohzu, H. (1997). Accommodative responses to stereoscopic three-dimensional display. *Applied Optics*, *36*(19), 4509. https://doi.org/10.1364/AO.36.004509

Itoh, Y., Langlotz, T., Sutton, J., & Plopski, A. (2022). Towards indistinguishable augmented reality: A survey on optical see-through head-mounted displays. *ACM Computing Surveys*, *54*(6), 1–36. https://doi.org/10.1145/3453157

Judy, A., & Gollery, T.J. (2019). US Navy pilot competence: An exploratory study of flight simulation training versus actual aircraft training. *Journal of Applied Social Science Research and Practice, 1*(1), 4.

Kennedy, R.S., Lane, N.E., Berbaum, K.S., & Lilienthal, M.G. (1993). Simulator Sickness Questionnaire: An enhanced method for quantifying simulator sickness. *International Journal of Aviation Psychology 3*(3), 203–220.

Longo, L., Wickens, C. D., Hancock, G., & Hancock, P. A. (2022). Human mental workload: A survey and a novel inclusive definition. *Frontiers in Psychology*, *13*, 883321. https://doi.org/10.3389/fpsyg.2022.883321

Matthews, G., Warm, J. S., & Smith, A. P. (2017). Task Engagement and Attentional Resources: Multivariate Models for Individual Differences and Stress Factors in Vigilance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *59*(1), 44–61. https://doi.org/10.1177/0018720816673782

McCoy-Fisher, C., Mishler, A., Bush, Severe-Valsaint, G., Natali, M., & Riner, B. (2019). *Student naval aviation extended reality device capability evaluation.* (Technical NAWCTSD-TR-2019-001). Naval Air Warfare Center Training Systems Division.

McCurry, C. D., Mohammed, T. I., & Zein-Sabatto, M. S. (2022). Tennessee State University Multi-Attribute Task Battery (TSU-MATB). Department of Electrical and Computer Engineering, Tennessee State University.

Mishler, A., Severe-Valsaint, G., Natali, M., Cooper, T., & McCoy-Fisher, C. (2023). *Evaluating the use of virtual reality in Navy rotary training.* (Technical Report NAWCTSD-TR-2023-001). Naval Air Warfare Center Training Systems Division.

Mishler, A., Severe-Valsaint, G., Natali, M., Seech, T., McCoy-Fisher, C., Cooper, T., & Astwood, R. (2022). *Project Avenger Training Effectiveness Evaluation*. (Technical Report NAWCTSD-TR-2022-006). Naval Air Warfare Center Training Systems Division.

Natali, M., Severe-Valsaint, G., Mishler, A., Graniela, B., Wolff, R., Portoghese, R., & Cooper, T. (2023). *Project Link: Development & capability evaluation of a T-45C mixed reality trainer*. (Technical Report NAWCTSD-TR-2023-101). Naval Air Warfare Center Training Systems Division.

Perry, R. (2023, Winter). Bringing the virtual world into reality: Manned flight simulators integrate VR/MR technology for CMV-22 platform. *Naval Aviation News*, *105*(1), 22–29.

Poltavski, D. V., Biberdorf, D., & Petros, T. V. (2012). Accommodative response and cortical activity during sustained attention. *Vision Research*, *63*, 1–8. https://doi.org/10.1016/j.visres.2012.04.017

Reichelt, S., Häussler, R., Fütterer, G., & Leister, N. (2010). Depth cues in human visual perception and their realization in 3D displays. *Proceedings, Three-Dimensional Imaging, Visualization, and Display Technologies and Applications for Defense, Security, and Avionics IV*, 7690. https://doi.org/10.1117/12.850094

Rolland, J. P., & Fuchs, H. (2000). Optical versus video see-through head-mounted displays in medical visualization. *Presence: Teleoperators and Virtual Environments*, *9*(3), 287–309. https://doi.org/10.1162/105474600566808

Rowan, C. P. (2023). *Model-based assessment of adaptive automation's unintended consequences* [Dissertation, Naval Postgraduate School]. NPS Archive: Calhoun. https://calhoun.nps.edu/handle/10945/72253

Scheiman, M., & Wick, B. (2020). *Clinical management of binocular vision: Heterophoric, accommodative, and eye movement disorders* (Fifth edition). Wolters Kluwer Health.

Vienne, C., Sorin, L., Blondé, L., Huynh-Thu, Q., & Mamassian, P. (2014). Effect of the accommodation-vergence conflict on vergence eye movements. *Vision Research*, *100*, 124–133. https://doi.org/10.1016/j.visres.2014.04.017

Vogl, J., McCurry, C., Bommer, S. (2023). *USAARL MATB User Manual* [Unpublished manuscript].

Wann, J. P., Rushton, S., & Mon-Williams, M. (1995). Natural problems for stereoscopic depth perception in virtual environments. *Vision Research*, *35*(19), 2731–2736. https://doi.org/10.1016/0042-6989(95)00018-U

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(3), 449–455. https://doi.org/10.1518/001872008X288394

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2013). *Engineering psychology and human performance.* Routledge.