HUMAN PERFORMANCE USING STEREO SYMBOLOGY IN A HELMET MOUNTED DISPLAY AND ASSOCIATION WITH INDIVIDUAL STEREO ACUITY

By

Bonnie POSSELT

A thesis submitted to the University of Birmingham for the degree of

DOCTOR OF PHILOSOPHY

School of Electronic, Electrical, and Systems Engineering

University of Birmingham

Edgbaston

Birmingham, B15 2TT

2023

ABSTRACT

Binocular Helmet Mounted Displays (HMDs) are a critical part of the aircraft system, allowing information to be presented to the aviator with stereoscopic 3D (S3D) depth, potentially enhancing situational awareness and improving performance. The utility of S3D in an HMD may be linked to an individual's ability to perceive changes in binocular disparity (stereo acuity). Though minimum stereo acuity standards exist for most military aviators, current test methods may be unable to characterise this relationship. Stereo acuity was measured using three digital disparity discrimination tests along with the analogue TNO test currently used by the British military to qualify aircrew. Participants performed a simulated flying task while differentiating between two warning alerts appearing in their HMD which were presented in four different conditions of stereo depth. Both physiological measurements and subjective questionnaires were used to assess cognitive workload. The addition of S3D to a warning alert in an HMD did not improve performance. Though peripheral skin temperature decreased in all conditions, suggesting an increase in cognitive load, there was no specific relationship to S3D presentation. However, those with poorer stereo acuity reported greater levels of workload in some S3D conditions. Additionally, computer-based tests were able to measure stereo acuity to threshold in a more reliable and repeatable manner, eliminating a number of limitations seen with the TNO test. These results do not necessarily support incorporating stereo symbology in an HMD for a warning alert or altering current stereo acuity standards.

Word count: 240 **Keywords**: Stereopsis, stereo display, HMD, stereo test

The views and opinions presented herein are those of the author and do not necessarily represent the views of DoD or its Components. Appearance of, or reference to, any commercial products or services does not constitute DoD endorsement of those products or services. This research was carried out at the 711 HPW/RHBC Operational Based Vision Assessment (OBVA) Laboratory in accordance with IRB approved research protocols F-WR-2017-0095-H and F-WR-2019-0044-H.

This work is cleared for public release, Distribution A: AFRL- 2022-0871.

LIST OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

AFIC	Air Force Interoperability Council
AFVT	Air Force Vision Test
AO-V	AO vectograph
AR	Augmented Reality
A-V	Accommodation-Vergence
AVT	Automated Vision Tester
dRDS-D	Digital Random Dot Stereogram-Dynamic
dRDS-S	Digital Random Dot Stereogram-Static
DVE	Degraded Visual Environment
ECG	Electrocardiogram
EEG	Electroencephalogram
FC	Flying Class
GBS	Good Binocular Status
HITS	Highway in the sky
HMD	Head/Helmet Mounted Display
HMDS	Helmet Mounted Display System
HOTAS	Hands On Throttle And Stick
HR	Heart Rate
HRV	Heart Rate Variability
IHADSS	Integrated Helmet And Display Sight System
IPD	Interpupillary Distance
JHMCS	Joint Helmet Mounted Cueing System

JSF	Joint Strike Fighter
MBS	Moderate Binocular Status
MR	Mixed Reality
OBVA	Operational Based Vision Assessment
OLED	Organic Light Emitting Diode
RA	Research Assistant
RDS	Random Dot Stereogram
RMSD	Root Mean Square Deviation
RMSE	Root Mean Square Error
RMSSD	Root Mean Square of Standard Deviation
S3D	Stereo 3D
SA	Situation Awareness
SD	Standard Deviation
SDT	Signal Detection Theory
SST	Stereo Search Task
TNO	Toegepast Natuurwetenschappelijk Onderzoek
VR	Virtual Reality
VTA-DP	Vision Test Apparatus-Depth Perception
WSO	Weapon System Operator
ZCSBV	Zone of Clear Single Binocular Vision
4FAC	4 Forced Alternative Choice

PUBLICATIONS, PRESENTATIONS, AND AWARDS

Publications

Posselt BN, Seemiller E, Baber C, Winterbottom M, Hadley S. (2022) ‘A digital alternative to the TNO stereo test to qualify military aircrew’. Aerospace Med Human Perform 93(12) pp. 846-854.

Contributions: BP and MW-Concept and design. BP-Data collection. BP, ES, CB, MW-Data analysis, writing content and critical review. SH-Review.

Posselt BN, Seemiller E, Palmer E, Imel G, Winterbottom M, Hadley S (2021) ‘Does individual stereo acuity affect performance using stereo 3D in a Helmet Mounted Display?’ Proc. SPIE 11759, Virtual, Augmented, and Mixed Reality (XR) Technology for Multi-Domain Operations II 11790B DOI: https://doi.org/10.1117/12.2588017

Contributions: BP and MW-Concept and design. BP-Gathering data. BP, ES, EP, MW-Data analysis, writing content and critical revisions. GI-Developing flight simulation and HMD symbology. SH-Review.

Posselt BN (2021) ‘Helmet Mounted Displays in Military Aviation’ Airclues flight safety magazine, Issue 35 pp. 40-42.
https://www.raf.mod.uk/our-organisation/units/raf-safety-centre/publications/raf-safety-centre-documents/air-clues-safety-magazine-issue-35/

Posselt BN, Winterbottom M (2020) ‘Are new vision standards needed for military aircrew using 3D stereo helmet mounted displays?’ BMJ Mil Health
DOI: https://doi.org/10.1136/bmjmilitary-2020-001493

Contributions: BP-concept and design. Responsible for gathering data in the form of literature searching and reviewing source papers. BP and MW-both contributed to writing content and critical revisions.

International presentations

Invited keynote speaker, Stereoscopic Displays and Applications conference, San Francisco, USA, Jan 2023

Posselt BN, Seemiller E, Imel G, Winterbottom M, Hadley S (2022) 'Human performance using stereo symbology in a Helmet Mounted Display and association with individual stereo acuity'. Presented at AsMA, Reno, NZ. Aerospace Med and Human Perform 93(3); 255.

Seemiller E, **Posselt BN**, Winterbottom M, Blacker K, Pettijohn K, Hadley S. (2022) 'Temporally local effects of acute hypoxia exposure on stereoacuity'. Presented at AsMA, Reno, NZ. Aerospace Med and Human Perform 93(3); 205

Posselt BN (2022) 'Human performance using a stereo 3D HMD' Presented at the Royal Society of Medicine COLT military research conference.

Posselt BN, Seemiller E, Palmer E, Imel G, Winterbottom M, Hadley S (2021) 'Does individual stereo acuity affect performance using stereo 3D in a Helmet Mounted Display?' Proc. SPIE 11759, Virtual, Augmented, and Mixed Reality (XR) Technology for Multi-Domain Operations II 11790B DOI: https://doi.org/10.1117/12.2588017

Posselt BN, Van Atta A, Winterbottom M, Hadley S (2021) 'A new stereo test for military aircrew' Presented at AsMA, Denver, CO. Aerospace Med Human Perform, 96(2); 490.

Awards

Aerospace Medical Association Fellows scholarship (Dec 2022). Awarded to the AsMA presentation and manuscript judged to be of the highest scientific value, originality, quality and relevance ($2000).
https://www.asma.org/news-events/asma-news/winners-of-the-2022-asma-fellows-scholarship

Julian Ward prize (May 2022). This award is given annually by the Aerospace Medical Association to an individual for superior performance and outstanding achievement in the art and science of aerospace medicine during residency training.
http://www.asma.org/news-events/asma-news/ward-award-bonnie-posselt

Royal Society of Medicine, COLT Foundation Research and Innovation Meeting (Dec 2021). Award given to best research presentation in the lightning presentation category (£100).

Sword of Honour (Sep 2021). Awarded annually to the RAF Officer on exchange with USAF who has contributed most to relations between the two nations and their air forces. In this incidence based on the valuable vision science research conducted.
https://www.rafmaf.com/events

ACKNOWLEDGMENTS

This PhD was completely unique; the combination of an international military exchange tour, speciality training in Aviation and Space Medicine with all research conducted entirely remote from the University. I am indebted to my military leadership and in particular Dr. Steve Hadley, the lead for the OBVA laboratory, for the mountain of paperwork they endured to make it happen. Thank you to my university supervisor, Dr. Chris Baber, for being my academic rock. Also, I am eternally grateful to Dr. Marc Winterbottom and Dr. Eric Seemiller for being my day-to-day supervisors and mentors, and who taught me so much about vision science. I would be nowhere without all of my supervisors' guidance, knowledge, and endless patience. Likewise, I must pay tribute to other OBVA laboratory staff who assisted enormously; Jason Russi, Matt Ankrom, Liz Shoda, Margret Hollander, Carrie Drake, Tommy Bullock, and to the incredibly talented Geno Imel for creating the flight simulation. To Lt Col Jonelle Knapp, thank you for your advice, friendship, and keeping me sane throughout.

Undertaking a PhD in a foreign country, during a global pandemic, whilst simultaneously raising a child and completing flying training has certainly been challenging at times. None of this would have been possible without the dedicated support of my husband, Mark, who selflessly put his career on hold to allow mine to flourish; Thank you, a thousand times over. Finally, I dedicate this work to our beautiful daughters, Josephine Quimby, who was born during my first year of study, and Madeline Mae, who was born in the final year. One day, I hope they too will strive to achieve what others think is impossible.

Bonnie

1. INTRODUCTION

Complex visual display systems are becoming more common in the military operational flying environment. Stereoscopic 3-dimensional (S3D) is a feature that can be added to an aircraft's flight display to potentially enhance pilot performance, and thus the overall weapon system capability. Helmet Mounted Displays (HMDs) are one area where there is interest in the utility of S3D displays to improve operational performance and capability. S3D when used in other applications can more efficiently capture attention (Sakata *et al.*, 2017), improve search times (Reis *et al.*, 2008) and reduce response times to identify a target (Blake, Sloane and Fox, 1981), as well as reduce perceived work load (Sakata *et al.*, 2017). S3D also has the potential to declutter a display with large amounts of complex information (Kooi, 2011). However, S3D displays may adversely affect the human vision system in ways not previously considered. They could potentially cause a number of uncomfortable symptoms, such as ocular fatigue, headaches, and double vision (Mon-Williams, Wann and Rushton, 1993; Melzer and Moffitt, 2011) all of which could negatively impact operator performance and flight safety (Gavrilescu *et al.*, 2019). Furthermore, adding new flight hardware and software to support new aircraft features, such as an S3D HMD, is costly due to not only the development costs, but also rigorous reliability and hardening requirements, and extensive flight testing before it can be approved for flight. New features should be thoroughly researched to understand the potential impact on pilot and system performance, and to understand cost/benefit trade-offs. Unfortunately, the impact of design changes on human performance is often overlooked in military acquisition.

Despite dramatic advances to the cockpit, such as using multiple digital displays and optical HMDs, vision standards for military aviators have remained largely unchanged since they were first introduced in the 1920s. Current visual standards are the minimum needed to land an aircraft safely in the event of the loss of refractive correction, whilst leaving a margin for deterioration with age (Paish, 2017). The entry visual standards for aircrew in all three United Kingdom (UK) military services are set out in AP1269A (Royal Air Force, 2020). They include minimum requirements for visual acuity, stereo acuity, muscle balance, colour perception, accommodation, convergence, and refractive error. Stereo acuity is a critical element in perceiving precise depth, and as flying involves many depth related tasks, it is usually required for aviators. However, test methods currently used to measure stereo acuity and qualify military aviators may be outdated (Buzzelli, Trowell-Harris and Protsko, 1989) and other more reliable alternatives may be more suitable. The current stereo test used by the British military to select and qualify aircrew is paper-based with a number of limitations, in particular, there are concerns about its repeatability and reliability (Posselt *et al.*, 2022). Accurate and more reliable stereo tests are needed to better characterise stereo acuity in military aviators. This is important because while good stereo acuity has always been desirable in military aviators due to its links with improved depth perception and flying ability (Snyder, 1991; Wright *et al.*, 2013), good stereo acuity may become more critical with increasing use of S3D displays. This research will evaluate the current paper-based stereo acuity test, comparing it to newly developed computer-based stereo tests, enabling individual stereo acuity to be assessed against performance metrics. This research is vitally important to inform future design of HMDs using S3D displays and to ensure we have evidence-based vision standards using reliable vision tests. Additionally, new equipment must be fit for purpose and equally importantly, it must not inflict discomfort or harm upon users to whom the medical community have a duty of care. Furthermore, evidence-based vision standards are important not only to optimise performance and select operators best suited to potentially dangerous occupations, but to avoid unnecessarily excluding individuals. Outdated or non-operationally relevant vision standards along with fallible test methods are undesirable not only because it restricts the pool of eligible candidates to be selected as aircrew, but the military could be subject to breeching UK anti-discrimination legislation. Vision standards which exclude sections of the general population are acceptable under the UK Equality Act 2010, but only as long as they can be justified to be an occupational requirement (*Equality Act*, 2010).

In order to investigate the use of S3D in an HMD, along with scrutinising current military tests used to assess depth perception in aviators, the following research questions will be addressed;

1.1. Research Questions

R1: Is there any difference between the reliability and repeatability of current paper-based tests when compared to newly developed computer-based tests to measure stereo acuity?

R2: How does the addition of S3D to a warning alert task displayed in an HMD affect performance, ocular symptoms, and cognitive workload?

R3: Is there any relationship between individual stereo acuity measured with different stereo tests and performance, ocular symptoms, or cognitive workload in a warning alert task displayed with S3D in an HMD?

1.2. Helmet Mounted Displays (HMDs)

1.2.1. History and types

In its crudest form, the earliest recorded applied use of an HMD was between 1915 and 1917 with the ‘Integrated Helmet Mounted Aiming and Weapon Delivery System’ for ground-based soldiers (Marshall, 1989). It consisted of a sighting system attached to a helmet, enabling hands free, head-oriented targeting and operation of a pistol concealed within the helmet. This early version of an HMD demonstrated that a helmet need not merely be worn for ballistic protection but can be used in a number of other ways to increase capability, a function that has evolved over time. In its basic form today an HMD consists of an ‘image source and collimating optics in a head mount’ (Melzer and Moffitt, 2011). The display used in an HMD can be fully opaque, creating an immersive virtual reality (VR) environment, or it can present computer-generated visual information overlaid onto a real-world image, a form of augmented reality (AR), or some combination of the two (mixed reality: MR). In the aviation environment, HMDs are almost entirely the optical see-through variety and can be designed to present information monocularly, biocularly, or binocularly. That is to say one image is presented to only one eye (monocular), the same image is presented to both eyes (biocular) or two slightly different images are presented to each eye (binocular), as illustrated in Figure 1.

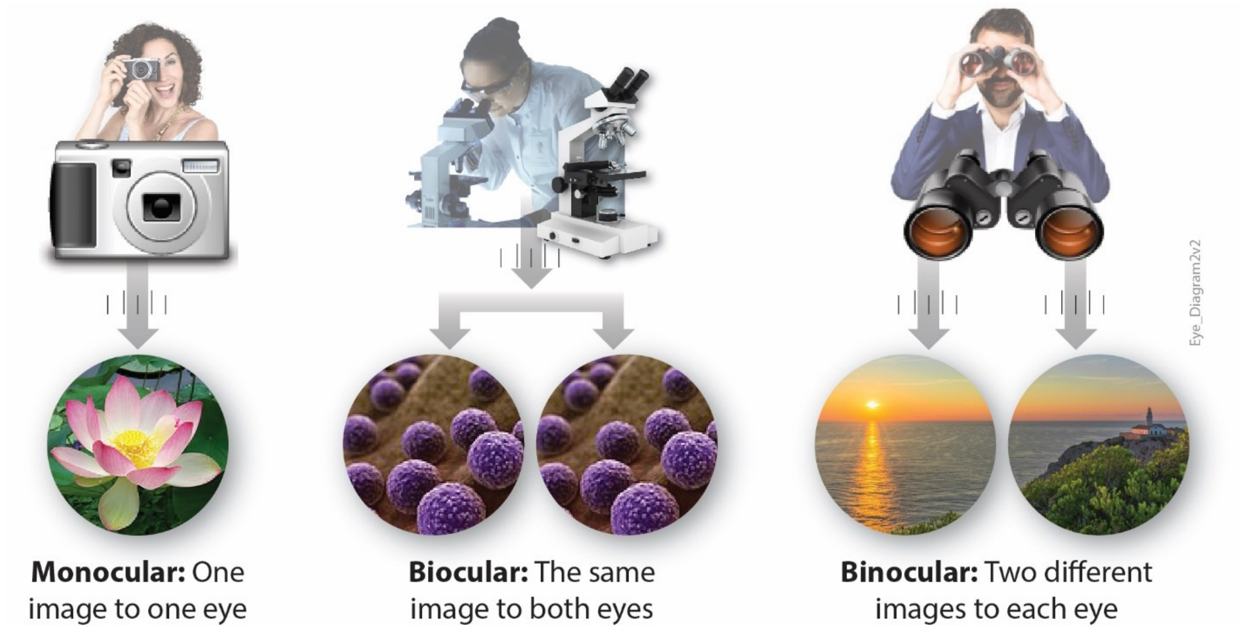

Figure 1. Three different orientations of HMDs, monocular (left), biocular (centre), binocular (right) (Posselt and Winterbottom, 2021)

Monocular systems are the simplest, least expensive and lightest option; however, they are well known to have a variety of human factors shortcomings (Bayer, Rash and Brindle, 2009). Significant issues of monocular devices include: image detection problems due to binocular rivalry and suppression, cognitive switching, ocular-motor instability, as well as asymmetrical mass distribution contributing to musculoskeletal injury (Morgan-Warren and Woodcock, 2014). Biocular systems eliminate the problem of binocular rivalry but only when suitably aligned. Because they can present binocular disparity, binocular systems alone are capable of displaying images in stereo depth and the user’s horizontal field of view can be extended wider than a biocular image if the two images are partially overlapped. They are however the most expensive option, complex to engineer, and difficult to keep aligned. Both biocular and binocular optics increase the mass of a helmet and to minimise discomfort the weight should be symmetrically distributed. There have been significant advances in recent years in optical technology;

higher resolution displays with reduced mass and weight which makes HMDs more attractive than ever, however, human factors limitations continue to be a significant issue (Keller and Colucci, 1998; Cakmakci and Rolland, 2006; Patterson, Winterbottom and Pierce, 2006; Bayer, Rash and Brindle, 2009; Melzer and Moffitt, 2011; Seemiller *et al.*, 2022).

1.2.2. HMDs in military aviation

As the earliest ground-based HMDs developed and matured, they were successfully introduced into aviators' helmets by the 1980s to augment flying performance (Melzer, 2012; Foote and Melzer, 2015). In the aviation setting, there are many operational situations in which it is critical the pilot maintain a 'head's out viewpoint' when flying, such as flying in formation, air-to-air refuelling, flying close to the ground, and when engaged in battle fighting manoeuvres. In these situations, it would pose a significant flight safety hazard if the pilot were to look back into the cockpit to read their flight instruments or information management system (Bayer, Rash and Brindle, 2009). HMDs allow information to be available to the pilot no matter the orientation of their head, and enable concurrent 'eyes out' flying operations with more efficient delivery of primary flight data (Bayer, Rash and Brindle, 2009; Ernst, Ebrecht and Korn, 2021). Information displayed can be orientated to the outside world, the cockpit, the user, or a combination of all three (Melzer, 2012). Further to keeping the pilots' 'eyes-out' of the cockpit, another capability enabling aspect of an HMD is that when coupled with a head tracking device, the pilot is able to locate a target and cue a weapon onto it without having to change the orientation of the whole aircraft, saving valuable time. However, arguably one of the most beneficial features of an HMD in an aviation environment, is that it can vastly improve a pilot's situation awareness (SA) (Bayer, Rash and Brindle, 2009). SA is the pilot's perception of their operating environment and their understanding of it in the present and projection into the future, and is a key input into pilot decision making (Endsley, 1988). By using a number of infrared, thermal, and visual sensors a virtual scene can be created to accurately represent the environment, which is especially useful when flying in a degraded visual environment (DVE), such as flying at night or in cloud. In a DVE, the HMD's sensor generated image is the virtual scene the pilot will focus on, negating the need to rely on or attend to real world visual cues. In a simulation study using an HMD with Earth-referenced imagery, specific performance benefits observed included waypoint accuracy and landing accuracy improving by 300 to 400%, as well as a 12,000% improvement in fire sector identification accuracy (Rogers, Asbury and Haworth, 1999).

Initial aviation HMDs were monocular, displaying only primary flight data to one eye of the pilot, such as the Integrated Helmet and Display Sight System (IHADSS) used by the United States (US) Army in the Apache attack helicopter (Figure 2). Contemporary HMDs used in fast jet and rotary aircraft are still largely monocular systems. Another example is the Joint Helmet Mounted Cueing System (JHMCS), which has been used for many years in aircraft such as the F-15 and F/A18 (Figure 2). Because a majority of users are right eye dominant (up to 68% (Ehrenstein, Arnold-Schulz-Gahmen and Jaschinski, 2005)), monocular systems such as IHADSS and JHMCS were designed to present information only to the right eye, making them more difficult to attend to and use by left-eye dominant individuals. Monocular systems may cause binocular rivalry and perceptual instability due to significantly different imagery and luminance presented to each eye (Patterson *et al.*, 2007), and indeed many pilots have experienced problems during training with IHADSS, such as visual discomfort, headaches, blurred vision, and disorientation (Behar *et al.*, 1990; Rash, 2008). In one survey it was reported that over 80% of Apache pilots experienced visual fatigue during a sortie using the IHADSS (Rash *et al.*, 2002), due to the dichoptic viewing conditions with only one eye able to view the outside world. For other monocular systems, like the JHMCS, each eye views the same real-world imagery with minimal occlusion/luminance differences with HMD symbology superimposed in the right eye only. In this situation, binocular fusion of the background outside window image may mitigate rivalry, but binocular rivalry may still be possible. The displayed monocular image may continue to be difficult to view due to other factors such as lack of binocular summation and masking of the virtual image by the background outside world imagery. Reduced visibility is of particular concern in low contrast conditions (Winterbottom *et al.*, 2015). Another concern with using HMDs is of cognitive shifting; Although no head movement is required to view flight data, the user must still shift their attention from the outside world to the display image. Both the outside world and display image cannot be attended to simultaneously, which can result in cognitive tunnelling (Wickens and Long, 1995; Tufano, 1997).

Figure 2. Left: Integrated Helmet and Display Sight System (IHADSS) – Wikimedia commons (https://commons.wikimedia.org/wiki/File:Integrated_Helmet_Display_Sight_System.jpg). Right: JHMCS. Image provided by Vision Systems International, used with permission.

Today, HMDs have been implemented across a number of modern aircraft with increasingly complex optical systems in both biocular and binocular orientations. They now provide not just desirable additional features but have evolved to become critical and integral parts of the aircraft's weapon and flight control systems, without which the aircraft cannot be operated. The latest generation of HMDs, such as the Collins Helmet Mounted Display System (HMDS) (Collins Aerospace, Cedar Rapids, IA, USA) using for the Joint Strike Fighter (JSF) seen in Figure 3 and the BAE Striker HMD (BAE Systems, Rochester, UK) in Typhoon in Figure 4, operate biocularly, with identical images and symbology presented to each eye. However, as they have two separate optical systems, they are capable of presenting images in a binocular fashion, i.e., in S3D, if desired. It is anticipated that in military aviation future HMDs are likely to be binocular with an optical display for each eye.

Figure 3. JSF helmet. Used with permission from Collins Aerospace

Figure 4. BAE Systems Striker II HMD. Used with permission from BAE Systems.

To summarise, HMDs used in military aviation can improve performance and SA. Consequently, HMDs are increasingly being used in military aviation and in some incidences, they are critical to operate the aircraft and weapon system. Currently most HMDs are monocular or biocular, however, it is likely they will develop to be binocular, and able to display information in S3D. For this reason, it is important to investigate human performance using S3D in an HMD, as set out in the research questions in Section 1.1.

1.3. Stereopsis and stereo acuity

If future HMDs routinely use S3D to display information, it is important to understand human binocular vision and stereopsis. Stereopsis is the ability to perceive depth, based on the difference in position of an image between the left and right retinas due to the slightly different perspective of each eye (binocular disparity). Binocular disparity is typically measured as angular difference between the retinal lock of the image in the left and right eyes. Stereo acuity is the smallest disparity that can be perceived in depth and is one way to measure binocular function. Stereo acuity is thought to be dependent on a number of factors such as stimulus duration, illumination levels, stimuli spatial frequency, contrast of stimuli in relation to individual contrast sensitivity, visual acuity, ocular motility, vergence, and accommodation (e.g. Patterson and Martin, 1992; Wilcox and Allison, 2009; Read, 2015). In the real world, when fixating on an object in depth, both eyes will move simultaneously (i.e., make a vergence eye movement) either inwards (i.e., converge) or away from each other (i.e., diverge) so that the target object will fall on each fovea with almost zero absolute disparity. The theoretical arc passing through this fixation point and the nodal points of the two eyes is called the theoretical horopter (Figure 5). All points along the horopter will have the same zero disparity, which sets a baseline from which relative depths are judged.

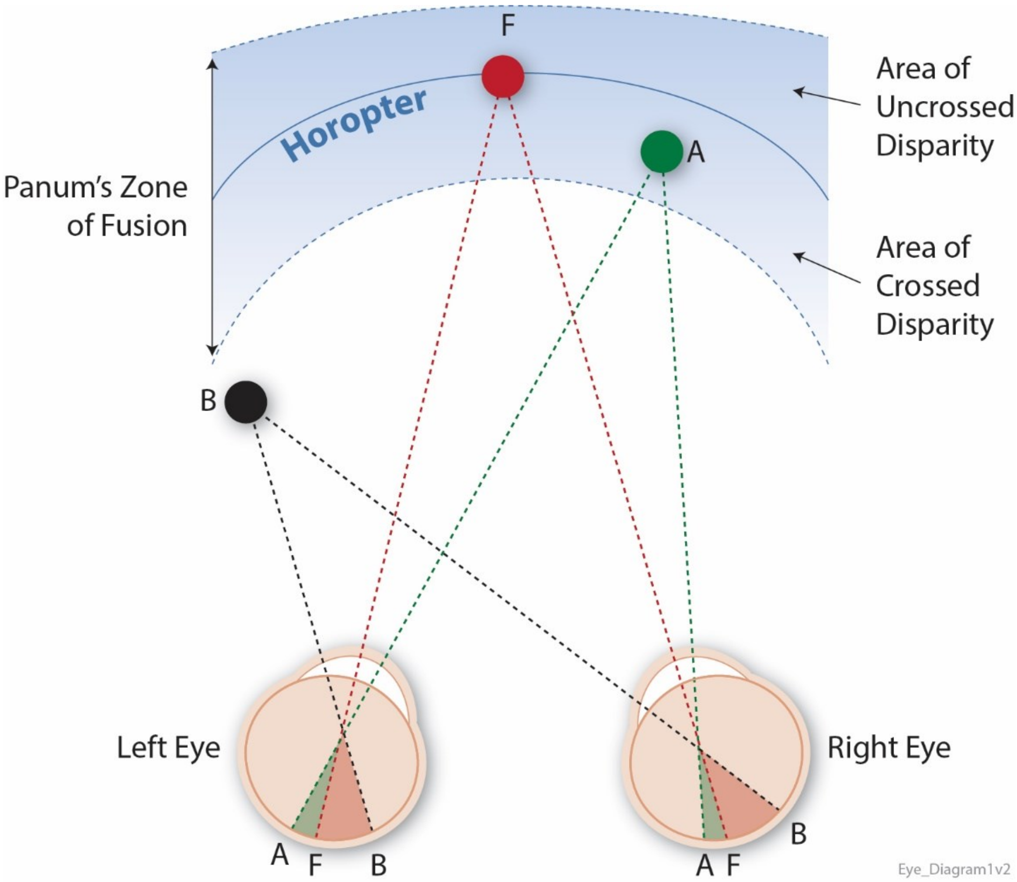

Figure 5. Horopter and Panum's zone of fusion. Target F is the object of focus, through which the horopter passes. Object A is within Panum's zone of fusion, thus it will be seen as a single image, whereas object B falls outside of the fusional area and will be diplopic.

On either side of the horopter exists a range over which disparate images can be fused to perceive a single object. Those that fall in front of the horopter are closer to the viewer and will be crossed in disparity. Those falling behind the horopter are further away and will be uncrossed in disparity. The spatial area of fused single objects is termed Panum's zone of fusion or the Zone of Clear Single Binocular Vision (ZCSBV). Within Panum's zone of fusion, where stereopsis is termed fine or patent, even very small disparities can be distinguished to give precise depth information. Objects that fall outside Panum's zone of fusion will appear diplopic, that is to say the viewer will experience double vision. Although large diplopic disparities can still contribute to depth perception in the form of coarse or qualitative stereopsis, they are less reliable depth cues (Wilcox and Allison, 2009).

Spatial positioning of the horopter and consequently Panum's zone of fusion varies constantly depending on where an individual is accommodated and their vergence postured. The size of Panum's zone of fusion varies both between individuals and within an individual depending on factors such as fatigue (Ogle, 1952; Parsons, 1984), environmental luminance, and pupil size (Ogle and Schwartz, 1959) but is roughly 10-15 arcminutes at the fovea ranging up to 70 arcminutes moving out to a peripheral angle of 6° (Ogle, 1952). However, only the middle third of Panum's zone of fusion (approximately 17.2 arcminutes crossed and uncrossed) is thought to offer comfortable binocular viewing without adverse symptoms, and is referred to as Percival's zone of comfort (Shibata *et al.*, 2011).

Up to this point targets have been described in regard to being static. Static stereopsis involves basic analysis of the absolute position of a stationary stimulus with a fixed disparity; in other words, determining if the stimulus is closer towards you or further away relative to the fixed horopter/screen. It is possible for targets displayed with static disparities to move dynamically in a combination of x- or y- axes yet fixed in the same depth plane, in a fronto-parallel manner. For the most part, when discussing stereo acuity, it is in reference to static stereo acuity. Unless specifically indicated otherwise, when discussing stereo acuity in this work, static stereo acuity is implied. However, there is another form of stereopsis; motion-in-depth stereopsis typically is when there is a changing frontal-spatial component to a S3D stimulus over time (i.e., the object moves towards or away from the viewer).

In real life there are a number of monocular cues that indicate an object is moving in depth towards a person, such as change in relative size (i.e. looming, forced perspective) and change in surface textural density and appearance (Regan and Beverley, 1978; Enright, 1987; Predebon, 1994). Changing frontal-spatial stereopsis can augment this perception of motion-in-depth yet the exact underlying neurological process is debated (Patterson, 2009b; Bonnen, Huk and Cormack, 2017). It is proposed that the perception of motion-in-depth with changing frontal-spatial stereopsis is computed by detecting interocular velocity differences, in combination with changes in binocular disparity over time (Richards and Regan, 1973; Nefs, O’Hare and Harris, 2010). Some argue that rate of changes in disparity is the more prevalent mechanism (Nefs, O’Hare and Harris, 2010), which is most effective when viewed relative to a binocularly fixed object for reference (Regan, 1993). Others contend that comparison of two monocularly viewed velocities is the more effective cue to judge speed of motion-in-depth (Harris and Watamaniuk, 1995). Regardless of the mechanism, there is evidence that motion-in-depth, or changing frontal-spatial stereopsis cues, are processed by separate neurons, compared to static position-in-depth stereopsis, and are tuned to the direction of movement (Beverley and Regan, 1973; Portfors-Yeomans and Regan, 1996; Patterson, 1999). Motion-in-depth neurons processing changing frontal-spatial stereopsis are thought to be most sensitive to directions in which targets are approaching at angles between both eyes, and there is significant adaption if the target continues to move for a few minutes (Beverley and Regan, 1973). However, in opposition to the idea of separate neuronal processing for motion-in-depth stereo, Harris *et al.* (1998) demonstrated that targets moving towards a subject in the z-axis were harder to identify than targets moving laterally in the x-axis, particularly when placed in a field of stationary distractors with binocular disparity (Harris, McKee and Watamaniuk, 1998). Harris *et al.* (1998) suggest there is no specific mechanism to detect changing S3D position, or motion-in-depth, and that left and right eye velocity signals cancel each other out, removing any binocular information. Instead, it is argued that detection of forward motion is based on many separate static binocular disparities (Cumming and Parker, 1994; Harris, McKee and Watamaniuk, 1998). Nevertheless, motion-in-depth stereopsis, can often be present in individuals unable to detect static S3D stimuli and the two measurements do not appear to correlate (Zinn and Solomon, 1985; Tittle, Rouse and Braunstein, 1988; Tidbury *et al.*, 2016). The differences between how static and motion-on-depth stereo acuity are processed are important when considering how to best present and use S3D within a display. In this work both static and motion-in-depth stereo stimuli will be used.

Static stereo acuity levels differ vastly among the general population and can vary depending on age. A visual response to disparity has been detected in infants as young as 5 weeks (Seemiller, Cumming and Candy, 2018). Stereopsis continues to develop during childhood, with coarse stereopsis maturing first by age 4, and fine stereopsis improving into adolescent years (Wilcox and Allison, 2009) Also, crossed stereo acuity develops earlier further adding to the suggestion that crossed and uncrossed stereo acuity are processed by different neurological mechanisms (Birch, Gwiazda and Held, 1982). There is some debate over what proportion of people completely lack stereopsis with estimates ranging from 5% to 30% (Richards, 1970; Ding and Levi, 2011). For those with measurable stereo acuity, there is a bimodal peak in the general population at 96 and 699 arc seconds, which appears to not be affected by age according to some studies (Hess *et al.*, 2015), but found to deteriorate over 50 years of age in others (Lee and Koo, 2005). By some estimates a further 32% of people are stereo anomalous despite otherwise normal vision (Hess *et al.*, 2015), and for these people perceived depth may be reversed (Richards, 1970, 1971), i.e. a stimuli presented with crossed disparity may be interpreted as being uncrossed, or further away from the focal plane. In stereo deficient individuals with otherwise normal vision, underlying cortical damage of some form may be the cause (Bridge, 2016). But more commonly, deficient stereopsis results from amblyopia; a disorder in which the image from one eye is not processed in one or both eyes with no known organic cause. However, amblyopia is related to ocular imbalance (strabismus) and refractive imbalance (anisometropia) between the eyes (Holmes and Clarke, 2006). There are examples of individuals who have gained or improved stereopsis in adult life with specific training (Astle, McGraw and Webb, 2011; Ding and Levi, 2011), although in such cases initial binocular vision is usually grossly abnormal.

Why is stereopsis important? As demonstrated by the significant proportion of people with deficient stereo acuity, stereopsis is not essential to daily life. One can rely solely on monocular cues, also known as pictorial or object-centred cues, to perceive depth. Such monocular cues are: relative size, interposition/occlusion, linear perspective, aerial perspective, textural gradient, atmospheric shading, luminance, height in visual field, and motion parallax (Wickens *et al.*, 2013). However, using binocular disparity depth cues can help resolve perceptual ambiguity of distance and size when using monocular cues, as monocular cues can be more easily misinterpreted (Wickens *et al.*, 2013). Binocular disparity, or stereo cues, are more effective because they are based on geometry with fewer assumptions, whereas monocular cues depend on prior knowledge and experience (Predebon, 1994; Cutting and Vishton, 1995) . For example, to estimate the distance of a tree using relative size cues, the subject must first know the height of an average sized tree, and then judge the target tree relative to their prior knowledge.

Despite not being absolutely essential to daily life, there are aspects in which stereopsis improves performance, enabling more accurate and more precise depth perception. In a study of British drivers, poor stereo acuity was associated with an increased accident rate (Davison, 1985), a finding echoed in Canadian studies of commercial truck drivers (Laberge-Nadeau *et al.*, 1996) and taxi drivers (Maag *et al.*, 1997). Another area in which stereopsis is important is sport and superior stereo acuity is evident in many professional athletes. Stereo acuity measured in professional golfers and baseball players was found to be much lower (better) when compared to the stereo acuity measured in a non-athletic group (49.42 and 53 arcseconds in golfers and basketball players vs. 94 arcseconds in non-athletic group) (Hunfalvay, Orr and Roberts, 2017). In addition, the variability of measured stereo acuity was also lower in the athletic groups

(SD of 25.27 and 18.43 vs. 54.84). Thus, these authors suggest that stereo acuity could be a predictor in the development of future athletes.

For the same reasons that it is possible to live a normal life without stereopsis, it is eminently possible to successfully pilot an aircraft without stereopsis, as evidenced by a few monocular pilots and monocular vision is allowed in trained civilian pilots of all classes, following a 6-month adaptation period (UK Civil Aviation Authority; Federal Aviation Authority, 2017). There have been inconclusive findings linking stereopsis directly to military flying performance (Wright, Gooch and Hadley, 2013). Although it is generally thought that while not absolutely necessary, stereopsis complements and enhances flying abilities (Snyder, 1991; Wright, Gooch and Hadley, 2013), and indeed, performance in a motion-in-depth stereo tracking task has been shown to positively correlate with flying ability (Kruk *et al.*, 1981, 1983). Specific to the rotary environment, height estimation is more accurate using binocular cues than monocular estimates (Hartle *et al.*, 2020). The utility of stereopsis is most apparent for closer visual ranges, but is beneficial up to 18 m (Allison, Gillam and Vecellio, 2009), making it most advantageous for situations when operating close to other aircraft or the ground, such as air-to-air refuelling, landing and take-off, taxiing, low-level flying, and formation flying. As an example, without stereopsis, landings performed monocularly involve steeper and higher descents (Grosslight *et al.*, 1978) and in some cases aviation mishaps have been attributed to visual illusions induced by monovision affecting depth perception (Nakagawara and Véronneau, 2000). Such visual illusions associated with monovision could be due to a lack of stereopsis and/or reduced contrast as a result of image blur in one eye, affecting neural processing speeds which consequently can induce a reverse Pulfrich effect illusion of motion in depth (Burge et al. 2019).

A reason why binocular vision may contribute to improved performance is that the intensity of the signal being detected is doubled; This is known as binocular summation and has been studied extensively in vision research (Blake and Fox, 1973; Blake, Sloane and Fox, 1981). A square root of 2 improvement in visual acuity was found across all spatial frequencies with binocular viewing when compared to monocular viewing (Campbell and Green, 1965). Indeed, simply having two objects to view (i.e. biocular viewing) can improve response times by 5% when compared to monocular viewing (Blake, Martens and Di Gianfilippo, 1980) and when stereopsis is present in an individual, response time to stimuli can be up to 40% faster when compared to monocular viewing (Blake, Sloane and Fox, 1981). However, the improvement in response time with viewing a S3D stimuli appears to be true only when the stimuli is within Panum’s zone of fusion. If the target is outside of Panum’s zone of fusion it will appear diplopic (termed qualitative stereopsis), and any binocular advantage will be lost. There is some evidence that targets presented in crossed disparity are easier to identify with faster reaction times observed in comparison to targets viewed with uncrossed disparity (Patterson et al., 1995). Overall, a review of binocular summation research reveals that binocular performance is superior to monocular performance for many different types of visual tasks (Blake and Fox, 1973). Additionally, pilots tend to prefer biocular/binocular HMD viewing conditions when compared to monocular conditions (Browne and Moffitt, 2009).

In summary, the benefits of stereopsis have been identified in many different aspects of daily life. It is also known that stereopsis varies among the general population, with a significant proportion being unable to process stereo cues, instead relying on monocular cues. However, due to its ability to augment depth perception, stereopsis is considered a mandatory characteristic in military aviators. Thus, a test is needed to determine if an individual possess an appropriate degree of stereopsis to qualify as a military aviator. Unfortunately, stereo acuity estimates greatly depend on the method of testing (Read, 2015), complicating the selection for the most appropriate stereo test.

1.4. Vision standards in aviation and testing

As more military organisations around the world move towards binocular HMDs and S3D displays for critical operations, ensuring aircrew have adequate stereo acuity will become increasingly important. Across all three UK military services, aircrew must meet the required entry visual standards set out in AP1269A (Royal Air Force, 2020); these standards are summarised below in Table 1, and compared with US military and civilian vision standards. Among the Five Eyes Air Force Interoperability Council (AFIC), all but Australia test their aircrew population for stereo acuity, whilst Canada tests for stereo acuity but does not enforce any stereo standard.

Table 1.Stereopsis vision standards across UK and US militaries, as well as civilian aviation organisations. Adapted from (Winterbottom *et al.*, 2014)*. AFVT = Armed Forces Vision Test, AO-V= AO-Vectograph, VTA-DP = Vision Test Apparatus-Near and Distant, WSO = Weapon System Operator, FC = Flying Class (U.S. Air Force School of Aerospace Medicine, 2014)*

	UK		**USAF**	**USN**		**USA**	**FAA**	**CAA**
	Pilot	WSO	FC I/II	FC III	Class I, II (except fixed wing), Class III (incl. UAV operators & critical flight deck personnel)	Pilot Class 1 (commercial) and 2 (private)	Class 1/2/3/4	
Stereo acuity (arcseconds)	120	N/A	40 waivable to 120 on AO-V	N/A	25 (VTA-ND) **or** 40 (Randot/Titmus/AFVT) **or** 8/8 Verhoeff No waiver	'Normal binocular Vision'	40	No standard
Test method	TNO test		AFVT	N/A	As above	AFVT Randot Titmus	None specified	

There are many stereo acuity tests used for screening and assessment for both children and adults, but they are not necessarily comparable (Read, 2015; Vancleef and Read, 2019; McCaslin *et al.*, 2020). For example, situations which encourage eye movements produce better results (i.e. lower thresholds), like the Howard-Dolman, Verhoeff, and Frisby tests, which all use real life depth stimuli (Van Ee and Richards, 2002). Lower thresholds are also seen in tests with monocular cues, such as the Randot or Titmus, and Armed Forces Vision Tester (AFVT) which all use circle contours. The presence of monocular cues can give false positive results, passing individuals with known binocular vision disorders (Cooper and Warshowsky, 1977; Fawcett and Birch, 2003). By comparison, Random Dot Stereograms (RDS) stereo tests have fewer monocular cues and are more difficult, producing higher stereo acuity thresholds (Pageau, Guise and Saint-Amour, 2009; Vancleef *et al.*, 2017). Random Dot Stereograms (RDS), as the name suggests, are comprised of small random dots whose positions differ between the two eyes. They are the only stereo tests able to isolate disparity without the use of contours or monocular cues and are often considered the best measure of stereopsis

(McCaslin *et al.*, 2020). Furthermore, the direction of disparity used in a test is an important consideration as crossed and uncrossed disparities may be processed by separate neuronal populations (Richards and Regan, 1973). Tests examining only one direction of disparity report lower levels of stereo anomaly (Hess *et al.*, 2015). For this research, as set out in Section 1.1: research questions, newly developed digital tests will be compared against the stereo test used by the UK military to assess stereopsis to qualify aircrew; the TNO test (Toegepast Natuurwetenschappelijk Onderzoek) (Figure 6), as addressed in Chapter 2: STEREO ACUITY TESTS.

1.5. S3D in a display and HMD

The benefits of stereopsis to daily life when naturally viewing the real world have been discussed. It is possible to create a S3D image digitally on a display with the potential to improve the transmission, understanding and interpretation of information. S3D permits segregation of displayed information enabling decluttering of the display (Williams, Parrish and Nold, 1994; Peterson, Axholt and Ellis, 2008). Furthermore, targets with more salient features can be created with the addition of S3D to more efficiently capture attention (Treisman, 1982; Wolfe, 1994), to improve search and response times (Nakayama and Silverman, 1986; Reis *et al.*, 2008), as well as reduce cognitive workload (Van Beurden, Kuijsters and IJsselsteijn, 2010; Russi, 2013; Sakata *et al.*, 2017), all of which will be explored in turn. In particular, S3D is thought to be most advantageous in novel complicated situations, compared to simple well learnt tasks (McIntire, Havig and Geiselman, 2012) and in degraded visual environments or scenes of great complexity where only few monocular cues are present (Wickens *et al.*, 2013).

Fundamentally, the way in which information is searched and identified within a display can be influenced with the use of S3D. A serial visual search occurs when a target must be identified within a group of distractors that have the same characteristics as the desired target (e.g., a search to find a specific letter amongst a line of text in the same font and size as the target). In this situation, the target can only be found by systematically directing attention to each object in turn. By applying a unique and distinctive feature to a target with respect to everything else being displayed, a serial search task can be transformed into a much faster detection task – becoming a parallel search task. Certain differentiating features added to targets, such as colour, shape, and size can be used (Treisman, 1982), while other properties such as Vernier offset and stereoscopic depth are suggested to have similar properties (Wolfe 1994). Indeed, Reis *et al.* (2008) demonstrated that by adding S3D to an identical target in a field of distractors, response times to identify the target were faster. However, the degree of depth separation used and individual user stereo acuity were not specified.

Nakayama and Silverman (1986) demonstrated that response times to identify a target presented in S3D are comparable to response times when other unique identifiers are used, such as motion (up vs. down) or colour (red vs. blue). When a second dimension was added to a target, (i.e., motion and colour, stereo depth and motion, or stereo depth and colour), it took longer to identify targets in the combined motion and colour condition than in either stereo condition. The authors concluded that when two identifiers of motion and colour are used, the task reverts to being a slower serial search task. However, when S3D is used in conjunction with another identifier, either motion or colour, reaction times remain similar to those reported with only one unique identifier. So, whilst colour is an effective unique identifier on its own and many different colours can be used to code a greater number of categories (Treisman and Gelade, 1980), S3D is just as powerful for detection (Kooi, 2011). With regard to the optimum S3D disparity needed to transform a serial search task into a faster detection task, there is some evidence that the larger the difference in S3D disparity between a target and its distractors, the more efficient the process becomes (Wolfe, 1994). Such improved performance could be expected with larger simulated S3D separations used, that is until the image falls outside of Panum's zone of fusion and appears diplopic (Shibata *et al.*, 2011).

In an aviation environment the ability of S3D to transform a search task into a detection one could be of operational importance. By changing a serial search task into a parallel detection one, the use of S3D could make the viewing, identifying, processing, and acting upon newly displayed information more efficient. For example, it is imperative that a pilot respond to a failure of their oxygen or cabin pressurisation system and take appropriate action to avoid hypoxia. If no action to a potential hypoxia situation is taken, the pilot could become incapacitated, which would result not only in mission failure, but also potentially loss of the aircraft and even human life. In high workload situations, phenomena such as change blindness, when a stimulus changes but goes unnoticed by the observer (Attwood *et al.*, 2018), are a real risk and of concern (Wickens et al., 2009). Thus, it is essential that critical information be displayed in such a way for the pilot to be aware of the situation as soon as possible. This requires a suitable warning symbol to alert the pilot to their situation. In order to be as noticeable as possible and to minimise the risk of change blindness, an alarm should be as salient as possible (Wickens *et al*., 2013). The greater the salience of an object, the more it will stand out from its background, and the greater chance it has of being noticed and acted upon. S3D could provide greater saliency to important displayed information, when kept within the central portion of a visual field.

Other potential benefits of S3D are improved depth perception and improved spatial awareness; there have been many studies investigating the use of S3D displays in this manner. One such example is the use S3D in a depth perception task (McIntire, Havig, *et al.* 2014). In this study the authors demonstrated an improved performance in a depth placement task using S3D compared to a 2D display. Twelve participants with stereo acuity better than 40 arcseconds as measured with the Titmus stereo test took part. Participants were required to place a virtual object at an indicated depth, matching it to a reference target within a range of up to 100 arcminutes. In general, the addition of S3D to the display improved performance of accurately placing the target object with the reference target as evidenced by a 53% reduction in mean placement error with individual improvements up to 91% (McIntire, Havig, *et al*., 2014). As an example of an aviation related situational awareness and spatial orientation task, Nataupsky and Crittenden, (1988) found faster correction responses in a S3D condition for pilot participants when they were suddenly placed in an off-nominal position from a simulated flight path. Improved performance has also been echoed in other ‘pathway-in the sky’ studies (Parrish, Busquets and Williams, 1990). In this study with five experienced helicopter pilots, performance in a central vertical tracking task improved when using an S3D display. In the S3D conditions pilots flew with reduced Root Mean Square Error (RMSE) of both pathway positioning and reduced inputs to their flying controls. However, there was no benefit to using S3D symbology in the peripheral field of view for a monitoring task (Parrish, Busquets and Williams, 1990), which is in keeping with previous reports stating that stereo acuity drops when viewing angles are greater than 6° from the fovea (Rawlings and Shipley, 1969). In addition, pilot participants preferred the S3D condition, stating that it “improved situational awareness and anticipation of future requirements” and was “easier to fly” (Parrish, Busquets and Williams, 1990).

In an alternative context, S3D has been used to create two separate depth planes to “segregate attention”. With two depth planes it is possible to reduce response times searching for and identifying a target, as described in a simulated cockpit navigation task experiment (Kooi, 2011). The use of a separate depth plane can be especially advantageous in decluttering information on a display when the viewing size is limited yet critical information needs to be highlighted. This is particularly applicable to HMDs where display space is at a premium, taking up no more than 40° x 40° in the field of view. In the examples given so far, only two separate distinct depth planes have been used. If multiple different depth planes are used, searching for a target becomes more difficult. With multiple depth planes, the use of S3D to identify a target is no longer unique and response times will increase, negating any initial benefit of using S3D (O’Toole and Walker, 1997). Indeed, increased reaction times to identify a S3D target with increasing number of distractors in other crossed and uncrossed planes is seen in experiments 1-3 (O’Toole and Walker, 1997). In these studies, nine participants were recruited from the general population, thus a representative range of stereo acuities should have been captured. However, the exact stereo acuity of their participants was not reported, so the spectrum of stereo acuities included cannot be determined; this is a significant limitation when interpreting their data.

In contrast to the advantages described above, no benefit to performance was found using S3D in a display for another simulated aviation related task. An improvement in SA was anticipated when displaying tactical information in S3D in a simulated cockpit scenario, however the opposite was found (Steiner and Dotson, 1990). Nine fighter jet pilots with “normal depth perception” participated, although the screening criterion was not clarified. A tactical situation was displayed with multiple red-coloured targets, and information presented in stereo depth for the S3D condition. Each participant was required to give a status update (“Which targets are above you?”), describe certain information of other aircraft (“What is the altitude of that aircraft?”), or analyse the situation (“What is your most immediate threat?”). Participants demonstrated significantly faster response times in the 2D condition and made 18% fewer errors when compared to the S3D condition. However, 56% of participants preferred the 3D display with 66% believing it would give them a tactical advantage. The authors themselves stated the reasons for not finding a benefit with S3D could be due to accommodation-vergence mismatch (A-V mismatch) or the effort required to focus on varying focal distances. Due to the fundamental nature of presenting images artificially, a V-A mismatch is common; the focal accommodative distance is fixed at the distance to the screen yet the vergence distance will vary depending on the location of the simulated image. In addition, Steiner and Dotson (1990) suggested that the stereo acuity of some participants may be worse than others which could have affected their results. In another study accuracy in a path tracking task did not improve with the addition of static stereoscopic disparity used in isolation, however, when combined with object motion or motion parallax depth cues, performance significantly improved, along with a reduction in visual discomfort and perceived workload (Van Beurden, Kuijsters and IJsselsteijn, 2010).

As demonstrated, there is some evidence to suggest that using S3D within a display could be helpful when searching for information, and in identifying targets. The majority of studies investigating the effects of S3D, compared to a 2D display, focus on spatial manipulation tasks or interpretation based on depth perception with few conducting research using stereo symbology in an HMD and applied aviation setting. In one of the few studies to investigate S3D within an HMD rather than a display screen, Browne *et al.* (2017) conducted a number of simulated aviation tasks using S3D symbology in an HMD, with the belief that "a depth component may attract attention or provide organization and clarity". In an effort to investigate performance with stereo symbology Browne *et al.* (2017) developed three difference experiments implementing the concept; using S3D symbology 1) for a warning symbol task, 2) to display a locator line, 3) for a Highway-in-the sky (HITS). These experiments were developed in a laboratory setting first to identify which operational tasks may benefit from the addition of S3D symbology in an HMD. Predominantly the scenarios focused on were; a situation with dense information using S3D symbology to declutter the display, and situations with strong depth and distance components. All participants were screened for stereopsis, using the Vision Assessment Corporation’s Random Dot Stereo Acuity Test with polarised glasses. Only participants with stereo acuity scores better than (lower than) 100 arcseconds were included, however, no exact threshold values were calculated. For the first warning symbol task, a clear benefit to response time was demonstrated by adding S3D to a FUEL warning symbol displayed in an HMD. Response times reduced from a median response time of 1855 ms in the 2D blinking condition to 750 ms in the most extreme S3D condition. The FUEL alert warning symbol was flashed with crossed disparity against an otherwise 2D fronto-parallel plane of distracting information, all monochrome green. The three S3D conditions used disparity settings of 8, 16, and 32 arcminutes, bringing the target information closer by 40, 120, and 170 mm respectively from a focal distance of 1 m. The authors did not, however, isolate a pure disparity only, flashing on and off, condition. One potential explanation as to why reaction times were faster when viewing the warning symbol in S3D could in fact be an unintended secondary effect; the addition of S3D enables the warning symbol to be visible continuously whilst also flashing in and out of depth with two static S3D positions. In the 2D condition this is not possible, as the initial abrupt onset captures attention, but when re-directing the person’s gaze to locate and focus on the target, it may be in an ‘off’ phase. The authors highlighted this as an area warranting further investigation.

In the second locator line target task, Browne *et al.* (2017) created a S3D locator line and measured acquisition time to identify a static target, compared to when using a non-stereo 2D locator line. Both locator lines showed an increase in search times with increasing distance to the target, but target acquisition times were significantly shorter in the S3D condition compared to the 2D locator line. When asked which condition they preferred, results were variable; some subjects preferred the 2D locator line, and others the S3D locator line. For the final HITS task, a S3D tunnel augmented the outside the window scene, geographically linked to the terrain, indicating the flight path. Subjects were assessed on their ability to fly through the frames successfully and Root Mean Square Deviation (RMSD) from the centre of each rectangle was recorded. Again, results were mixed with half of the subjects anecdotally preferring the S3D condition but this time without any overall significant improvement in flying performance demonstrated. A confounding factor to the HITS experiment could have been the reported difficulty experienced by at least half of the subjects in flying the simulated aircraft, even after significant exposure to the simulator and flight school training taught by a former fighter pilot. In summary, Browne *et al.* (2017) demonstrated different uses of stereo symbology in an HMD. While all three experiments showed promise to improve performance, only the use of S3D to augment a warning alert symbol resulted in a clearly defined, and statistically significant improvement in performance.

1.6. S3D impact on cognitive workload

HMDs in general increase the quantity and complexity of information presented to the pilot, which can add to cognitive workload, and detrimentally affect the pilot's interpretation of a scene. A potential benefit of using an S3D display is that it can reduce perceived workload, as demonstrated in a simulated surgical depth perception experiment (Sakata et al., 2017). In this study participants were assessed on three applied surgical depth perception tasks (peg transfer, navigating in space, and suturing) using a (1) conventional 2D display and (2) a S3D display viewed with polarising glasses. Performance, as measured by accuracy and speed, was significantly improved in the S3D display condition, more so at 1 m than 3 m, and perceived workload (as measured by the NASA TLX) was reduced by as much as 34% across all tasks (Sakata et al., 2017). Van Beurden *et al.* (2011) argued that workload should be used to more fully assess the utility of S3D displays and found that performance not only increased in a visual path-tracking task similar to a tele-robotic surgery task with the addition of S3D to a display, but that S3D also reduced perceived workload as measured with the NASA TLX when combined with object motion and motion parallax (Van Beurden, Kuijsters and IJsselsteijn, 2010). S3D displays have had positive effects in a number of other situations; in a simulated air traffic control task the addition of S3D to a display both improved performance and reduced perceived workload (Russi, 2013). In another study performing simulated aircraft landings using a “pathway in the sky”, heart rate did not increase as much during a S3D condition compared to a 2D version, potentially alluding to reduced pilot stress (Parrish, Busquets and Williams, 1990). In studies that were unable to find objective measurable differences in performance, many still report that subjectively participants prefer S3D conditions in precision tasks, finding them easier and more intuitive (Steiner and Dotson, 1990; Parrish and Williams, 1993; Browne, Moffitt and Wimsatt, 2017).

Cognitive workload can be measured in a number of different ways, categorised into effect on task performance, objective physiological measures and subjective ratings;

each with their own benefits and limitations (Cain, 2007). When choosing workload assessment methods, it is important to consider sensitivity, diagnostic ability, intrusiveness, transferability between applications, the time scale of use, the apparatus required, and user acceptability (Wierwille and Eggemeier, 1993; Rubio *et al.*, 2004; Cain, 2007). Integrating multiple measures is recommended to improve reliability and reduce issues that arise from using a sole measurement, such as vulnerability to noise and limited sensitivity (Wierwille and Eggemeier, 1993; Cain, 2007; Wang, Lu and Fu, 2020). Primary task performance and objective physiological measures have the benefit of recording continuously, without relying on participants' memory, but can be intrusive. Questionnaire ratings are simple to implement and are the least intrusive measure of workload. Consequently, subject questionnaires have been successfully used in a number of different settings. However, questionnaires rely on memory and are subject to individual user biases. The NASA TLX is judged to be sensitive at low levels of workload and particularly useful in applied settings (Nygren, 1991; Wierwille and Eggemeier, 1993).

For physiological measures, heart rate is one of the most frequently used objective measure of workload (Wierwille and Eggemeier, 1993). The main drawback in using heart rate as an indicator of cognitive workload is that is there is a large degree of individual variability (Wierwille and Eggemeier, 1993) and it is affected by a number of various external factors which can be difficult to control, such as emotional stress, circadian rhythm, respiration, exercise, caffeine intake, hydration and health status (Shaffer, McCraty and Zerr, 2014). Heart rate variability (HRV) measures the change in time intervals between successive heart beats, and is a better indicator of the parasympathetic nervous system reflecting heart-brain interactions (Laborde, Mosley and Thayer, 2017). HRV can be calculated from heart rate in a variety of ways. The root mean square of standard deviation (RMSSD) of eligible successive beats is the primary time domain assessment used as a measure of HRV, recorded over a 5 minute period (Shaffer, McCraty and Zerr, 2014; Shaffer and Ginsberg, 2017). A relative decrease in HRV, indicated by lower RMSSD, indicates lower parasympathetic activation and decreased vagal tone which suggests a physical or cognitive stressor is present (Laborde, Mosley and Thayer, 2017). However, the belief that HRV is a good indicator of cognitive workload is not universal; Some argue that HRV lacks the sensitivity required to differentiate between different levels of workload and mental stress (Nickel and Nachreiner, 2003).

Another physiological measure of workload is peripheral skin temperature. Cognitive workload reduces parasympathetic activity but increases activity in the sympathetic nervous system. Increased sympathetic activity causes peripheral vasoconstriction which reduces blood flow in the peripheral capillaries and consequently peripheral skin temperature decreases (Wallin, 1981). The relationship between peripheral skin temperature and mental activity has been illustrated in a number of studies (Or *et al.*, 2007; Novak, Mihelj and Munih, 2010; Gavrilescu *et al.*, 2019). Another method to assess cognitive workload is electroencephalographic (EEG) recordings, using scalp mounted sensors to monitor brain activity through detected cortical-evoked potentials (Isreal *et al.*, 1980; Paas, 2010). The bulky mass and volume of the sensors required to measure EEG would likely interact with an HMD, rendering it less suitable for such research. Finally, with regard to physiological measures, eye activity can be used to assess cognitive activity, measuring blink rate, pupil size, and eye tracking (Wierwille and Eggemeier, 1993). Typically blink rate and duration decrease with increased cognitive workload (Wierwille and Eggemeier, 1993), as well as dilation of the pupil with increasing task complexity (Aura *et al.*, 2021). It should be kept in mind that although thought to be a sensitive indicator of mental activity, such changes in pupil size are relatively small when compared to the pupil response to other factors such as illumination, fatigue and emotional state (Kramer, 1990). Beyond tasks requiring visual processing, other senses that could be objectively tested, are auditory, verbal, and tactile, as part of the multiple resource theory (Wickens, 2008).

As illustrated, there are a variety of ways of assessing cognitive workload. In order to obtain the best overall indication of cognitive workload, multiple assessment methods are strongly advised (Wierwille and Eggemeier, 1993). For this work a combination of subjective and objective methods were chosen to ensure a holistic approach was taken to assess cognitive workload. Based on sensitivity and availability, the methods chosen to measure cognitive workload were task performance, questionnaires, HR and HRV, as well as peripheral skin temperature.

1.7. Relationship between stereo acuity and performance using S3D in a display

Linking both performance using S3D symbology and stereo acuity, the third research question asks how does individual stereo acuity affect performance in these S3D tasks? Within a group of subjects, do those with the lowest stereo acuity thresholds perform better? Unfortunately, most studies fail to adequately characterise stereo acuity in their participants. However, for those that do, there are mixed results. For a simulated S3D surgical depth related task, stereo acuity correlated significantly with performance (Sakata *et al.*, 2017; Alhusuny *et al.*, 2021) and individuals with better stereo acuity completed the task faster. However, in both of these studies stereo acuity of participants was characterised using the Randot stereo test, which has similar limitations to the TNO stereo test and other paper-based stereo tests which are discussed in Chapter 2, Section 2.2. Thus, measured stereo acuity in these instances should be interpreted with those caveats in mind. Another study investigating stereo acuity and performance, but this time in an aviation setting, found a significant correlation between the stereo acuity and performance on a simulated aerial refuelling task using a S3D display; those with lower (better) stereo acuity thresholds performed better (Winterbottom *et al.*, 2016; O'Keefe *et al.*, 2022). However, no correlation was identified between individual stereo acuity and performance in a different simulated depth perception task to judge the distance of a helicopter rear landing wheel to a windsock and pole, although a clear performance benefit to using S3D overall was found (Winterbottom *et al.*, 2017). The benefit of using S3D in a display was more pronounced for tasks where the elevation of the helicopter tail wheel above the pole was closer (0.2 m) versus a higher elevation task (2 m). Since head tracking was used in this study, it is suggested that other cues such as motion parallax may have compensated for any deficit in stereo acuity, particularly as motion detection was significantly correlated with performance and could be a more appropriate predictor of performance (Winterbottom et al., 2017). Large differences in depth magnitude estimates made by individuals could also be explained by their experience with stereoscopic displays (Hartle and Wilcox, 2016); in that those with prior exposure to and a greater degree of familiarity with stereoscopic displays perform better than naïve participants.

1.8. Limitations of previous studies

Historically, experiments designed to examine the potential benefits of stereopsis have yielded mixed results, and it has been difficult to clearly demonstrate the operational impact of using S3D in display systems (Harrington, McIntire and Hopper, 2014; Sakata *et al.*, 2016). In many cases the anticipated benefits of S3D displays have not been recognised due to poor experimental design (Merritt, 2011). One of the most fundamental flaws in studies investigating human performance using S3D displays is the absence of accurate participant stereo testing (Sakata *et al.,* 2016). Some studies attempt to disqualify stereo blind individuals and establish a minimum stereo acuity standard for study participation, but very few define threshold values or group participants by stereo abilities (Sakata *et al.,* 2016). This is extremely important as participants who lack stereopsis will not benefit from the addition of S3D to a display, thus confounding the results, result interpretation, and any subsequent conclusions.

In addition to the factors described above, there are other examples of experimental design flaws in studies that evaluate the effect of adding S3D to display systems. For example, optimal viewing conditions may not have been enforced or displays with excessive crosstalk have been used which will detrimentally affect stereopsis perceived by a participant (Sakata *et al.,* 2016). Crosstalk occurs when a portion of an image intended to be seen by one eye leaks across to the other and results in double contours. Even as little as 5% crosstalk can detrimentally affect performance and cause viewing discomfort (Kooi and Toet, 2004). Subsequently, it is suggested crosstalk in a stereo system should be kept to less than 2% (Patterson, 2015). Alternatively, researchers may have neglected to correctly set up the left and right channel (displaying separate images to each eye) or induced excessive V-A mismatch (Merritt, 2011). Another reason could be that the systems being used for the 2D and S3D conditions are simply too disparate (Sakata et al., 2016). With advances in technology and display systems, it is now possible to display imagery with resolution such that interpretation is limited by acuity of the human eye itself and not the display. So, it is anticipated that as the image quality of display systems improves, the addition of S3D and the potential benefits it may enable will become more evident (McIntire, Havig and Geiselman, 2012). Overall, such profound oversights in the literature highlight the need to more accurately and thoroughly investigate human performance using S3D displays especially with regard to individual stereo acuity.

1.9. Adverse effects of S3D displays

Binocular HMDs using S3D symbology are not routinely used in aviation, so there is little information about any associated adverse effects. However, there is much information about adverse effects experienced with wider applications of S3D displays. S3D films have been reported to cause viewing discomfort, with symptoms such as eyestrain, double vision, headache (Kim *et al.*, 2013), nausea and disorientation in up to 50% of the subjects (Solimini, 2013). Mon-Williams, Wann, and Rushton (1993) were able to induce symptoms of blurred vision, frontal headaches, diplopia, nausea, motion sickness, and sore eyes whilst wearing a binocular virtual reality HMD within as little as 10 minutes. Viewing discomfort ratings returned to pre-testing levels within 5 minutes, along with the resolution of participant symptoms in all but two individuals who reported headaches that persisted for a few hours. However, no similar symptoms were reported after 30 minutes when this research was replicated using higher quality optics (Rushton, Mon-Williams and Wann, 1994). Indeed, it is reported elsewhere that post-viewing symptoms are significantly reduced using better aligned HMDs when compared to earlier VR HMDs after 60 minutes (Järvenpää and Pölönen, 2009). Static stereoscopic disparity used in isolation was shown to increase levels of visual discomfort, but when combined with object motion or motion parallax depth cues, visual discomfort and perceived workload significantly decreased (Van Beurden, Kuijsters and IJsselsteijn, 2010)

Symptoms of visual discomfort may well be related to individual binocular function. Causes of discomfort could originate from the exacerbation of an underlying or latent phoria, termed heterophoria. A phoria exists when there is an misalignment or imbalance between the ocular muscles and deviation from the forward aligned position becomes evident in the absence of fusion (Schroeder *et al.*, 1996). Some heterophoria is extremely common and considered normal (Parsons, 1984). When the ocular imbalance causes an eye to deviate inwards (converge), it is termed an esophoria, and when the direction is outwards (divergent) it is an exophoria. Less commonly, vertical misalignment (dipvergence) is known as a vertical heterophoria, and a cyclophoria if rotational. As it takes physical effort to maintain ocular alignment to converge on a target, single image binocular vision can break down when forced to do so over a prolonged period of time. Sheehy & Wilkinson (1989) reported that prolonged use of binocular Night Vision Goggles (NVGs) induced a shift in lateral phoria of 1.5 prism diopters in the exophoric direction, measured using the Armed Forces Vision Tester (AFVT), that persisted for up to three days. Depth perception was also reduced, measured using the Howard-Dolman test. One prism dioptre (PD) is a deflection of 1 cm at a distance of 1 m, or 0.57 degrees (34 arcminutes). It is suggested that any phoria, either induced or underlying, may detrimentally affect mission performance and cause reduced situation awareness in the cockpit (Sheehy and Wilkinson, 1989).

Other authors have also found underlying phorias to worsen when viewing S3D content. Peli (1998) assessed the effect of playing a stereoscopic video game projected through a binocular head mounted display and found that two thirds of phorias worsened, being shifted significantly in the exophoric direction by a mean of 0.9 dioptres (ranging up to 10 dioptres) after 30 minutes, which remained present even after an additional 30 minutes of rest. However, in this study phorias were measured by an optometrist and it is estimated that less than 2 dioptres cannot be reliably determined using either the Thorington, von Grafe, or Maddox rod tests (Schroeder *et al.*, 1996). A significant shift away from orthophoria at 6 m, predominantly shifted in the esophoria direction, has also been demonstrated within just 10 minutes viewing a virtual scene in a binocular HMD, using less than 50 cm apparent depth (Mon-Williams, Wann and Rushton, 1993). Only subjects with associated phorias suffered symptoms of headache, nausea, motion sickness, and sore eyes (Mon-Williams, Wann and Rushton, 1993). These findings correlate with symptom questionnaires reporting that viewing an S3D display is less comfortable than a normal biocular computer screen or a monocular display (Peli, 1998). In agreement, visual fatigue was induced using a S3D handheld game device for 30 minutes (Hirota *et al.*, 2018). After 30 minutes using a handheld game device, participants' ability to maintain binocular fusion was significantly decreased ($p<0.001$) which negatively correlated with subjective eye symptom score ($r^2 = 0.752$, $p<0.001$). As demonstrated, people with underlying visual abnormalities, such as a phorias can still appear to have normal binocular vision with measurable stereo acuity. However, such underlying visual abnormalities predispose individuals to visual discomfort in the form of eye strain and headaches, also known as asthenopia (Lambooij *et al.*, 2010). Lambooij *et al*. (2010) recruited 47 participants with stereo acuity of better than 60 arcseconds as measured with the Randot stereo test. Participants then underwent further rigorous optometric screening to classify their overall binocular ability and were sub divided into two groups: (1) those with good binocular status (GBS) and (2) participants with moderate binocular status (MBS). In the S3D display conditions, the

MBS group were slower at a reading task and reported a larger number of visual discomfort symptoms, despite having adequate stereo acuity. It has been recognised that individual differences in stereo acuity may significantly affect performance when viewing S3D content (McIntire, Havig, *et al.*, 2014) and this could have implications regarding operational performance, particularly if current tests are not able to detect and define the level of stereo anomaly.

Another prominent hypothesis regarding why people experience discomfort when using S3D displays is that S3D induces a mismatch between vergence and accommodation if the focal distance does not change (V-A mismatch) (Wann, Rushton and Mon-Williams, 1995; Patterson, 2009a; Shibata *et al.*, 2011). During normal real-world viewing conditions, the accommodation-vergence reflex is neurologically coupled with each reflex influencing the other (Judge and Cumming, 1986). It has been reported that a negative effect can be seen when any V-A mismatch is greater than 10.5 arcminutes (±0.3 dioptres) (Campbell, 1957) with viewing becoming more uncomfortable as the mismatch increases and when viewed at nearer distances (Shibata *et al.*, 2011). Other authors suggest 70 arcminutes (2.04 dioptres) should be considered as the disparity limit before comfort ratings drop significantly (Pastoor, 1993), and the image will become blurred. However, tolerance to V-A mismatch varies substantially across individuals. Using a computer based test that quantifies V-A mismatch tolerance, V-A mismatch has been shown to vary by more than two orders of magnitude (fusion range from 1.0 log arcminute to more than 3.0 log arcminute) (Winterbottom *et al.*, 2018). Furthermore, V-A mismatch tolerance was highly correlated with performance (r = 0.7,

$p < 0.001$) and reported discomfort ($r = 0.55$, $p < 0.003$) in a simulated refuelling task using a hyper-stereo S3D display (Winterbottom *et al.*, 2016), that is to say that participants with greater (better) fusion ranges performed better in the refuelling task and reported less ocular discomfort.

It is also possible that the type and amount of disparity used in a S3D display contributes to adverse symptoms experienced. In a group of 14 participants with normal stereo acuity, images displayed using crossed disparity were more comfortable to view than images displayed using uncrossed disparity at long distances, whereas images displayed using uncrossed disparity verses crossed disparity were more comfortable to view at near distances (Shibata *et al.*, 2011). In addition, individual phoria and individual size of Panum's zone of fusion correlated with symptoms of discomfort. As a recommendation, it has been suggested the maximum depth volume of S3D images should be limited to between -0.25 to +0.6 dioptres of the screen distance to avoid viewing discomfort and diplopia (Parrish and Williams, 1990). However, this advice is based on an experiment that used an S3D display placed at an unspecified distance using only four participants unscreened for stereo acuity wearing polarised glasses. Thus, caution must be taken when applying this to the spectrum of visual abilities seen in the general population. Shibata *et al.*, (2011) provide a recommendation based on the zone of clear single binocular vision identified in their research of approximately +/- 0.5 dioptres to maintain comfortable viewing conditions. In general, individuals can experience significant adverse symptoms when using S3D displays, so their design and application must be carefully considered to ensure optimal benefit and comfort for the user. A thorough review of stereoscopic systems and the parameters likely to cause adverse effects and discomfort is provided by Banks *et al.* (2012).

1.10. Additional challenges of providing S3D in an HMD.

In addition to all the factors that can cause adverse symptoms when viewing S3D displays, viewing S3D symbology in an HMD is further complicated by the need to ensure that both optics in a biocular or binocular HMD are perfectly aligned. Even if well aligned at point of manufacture, HMDs can suffer increasing misalignment over time due to mechanical wear or temperature cycling (Jamieson, 1981; Gavrilescu *et al.*, 2015). Misalignment or asymmetry between the two separate images or symbology displayed in an HMD has been reported to cause image rotation, shift, magnification or reduced resolution of one image (Kooi and Toet, 2004; Gavrilescu *et al.*, 2019). Perfect alignment in an HMD is highly desired but often not practical. Overly strict alignment requirements require unnecessarily expensive equipment, without any performance benefit, so a compromise may be necessary, but may only be possible for a specified user population.

In general humans are able to fuse images misaligned within their Panum's zone of fusion, beyond which the effort to fuse two misaligned images is too difficult and the image will appear diplopic. Fusion of a misaligned image takes effort and doing so may be uncomfortable with eyestrain and headache symptoms apparent long before fusion becomes impossible and diplopia ensues (Gavrilescu *et al.*, 2015). Additionally, over an extended period of time viewing even a mildly misaligned image is cumulatively fatiguing, with symptoms reported for limits that were previously tolerated over shorter durations (Self, 1986). Any misalignment is likely to be most easily recognised by those with good binocular vision who are able to identify small amounts of intraocular difference (Kooi and Toet, 2004; Williams *et al.*, 2018) but vertical misalignment, in particular, is universally poorly tolerated (Kooi and Toet, 2004; Williams *et al.*, 2018). Tolerated disparities may be smaller when images are viewed against a real-world scene when compared to a homogeneous background, by up to a factor of 10 (Gold and Hyman, 1970), although this was a small study with a sample size of only three participants. Tolerance of misalignment will differ based on the individual, and there is a wide variety of suggested tolerance limits (Self, 1986; Gavrilescu *et al.*, 2019). Suggested limits for military HMDs have been set at less than 3.4 arcminutes for vertical misalignment (dipvergence) between images, and less than 8.6 arcminutes of convergence (Self, 1986). More recently, it has been suggested that slightly more lenient standards can be tolerated, with no more than 10-23 arcminutes of horizonal difference, vertical difference of no more than 4-11.5 arcminutes, a rotational difference of no more than 6-12 arcminutes, and a difference in luminance of less than 30% (Patterson, Winterbottom and Pierce, 2006). These limits are more stringent for central vision (within central 15° FOV), with misalignment suggested to be kept less than 3.44 arcminutes for both convergence and divergence and less than 1.72 arcminutes of vertical misalignment in these areas (Newman and Greeley, 1997). To minimise the fatigue of fusing misaligned images, tighter standards could be imposed for prolonged use, defined as longer than 15 minutes (Newman and Greeley, 1997). In an effort to mitigate anticipated visual strain, it has also been suggested to align the HMD with the user's heterophoria (Mon-Williams, Wann and Rushton, 1993).

There is some argument that misalignment is more apparent for a see-through system, such as an HMD, when symbols are depicted in reference to real world points. If misalignment occurs in a see-through AR system, either the displayed symbol or the outside world image will appear diplopic which will be highly visible and may be more difficult for users to tolerate (Kooi and Toet, 2004). In a real-world operational setting, misalignment has been experienced by Typhoon pilots using the Striker I biocular HMD. Such discomfort is often so profound that pilots have reported manually deselecting a channel, rendering the display monocular (Williams, 2011). Other noteworthy adverse effects of HMDs briefly discussed, which are outside the scope of this work, are that any additional mass will restrict head movement, as well as adding to the mechanical load inflicted onto the head and neck of the pilot, potentially causing musculoskeletal injury (Morgan-Warren and Woodcock, 2014).

In summary, S3D displays in general can have detrimental physiological effects on the visual system as a result of V-A mismatch, misaligned images, crosstalk, and/or underlying individual phorias. These adverse physiological effects can induce a number of undesirable symptoms such as headache, nausea, disorientation, and ocular fatigue. Such symptoms, if present in an aviator could impact operational mission success and even flight safety. The degree of symptoms experienced will depend on the focal distance, whether the projected imagery is crossed or uncrossed, and on an individual's visual characteristics. There are additional factors to consider when a S3D display is within an HMD such as alignment, their see-through nature, mass, and weight distribution. These added characteristics of an HMD may exacerbate unwanted visual symptoms known to occur with S3D displays and even cause other novel undesirable effects.

1.11. Summary

It anticipated that when S3D is incorporated within a binocular HMD, it will improve human performance by making displayed information easier to understand and interpret. However, the degree of benefit offered by S3D displays may depend on the nature of the task, the context of its use and limited to certain situations. Any impact of S3D may also be influenced by individual visual abilities as well as the alignment and quality of the image or symbology itself. The use of S3D applied in aviation experiments has had mixed results. However, inadequate study design makes it difficult to fully interpret some study results, specifically with regard to the lack of stereo acuity measurement or screening for participants. Previous studies that investigated whether or not stereo acuity predicted performance in depth related tasks using S3D in a display have also had contradictory results. For a warning display task there is some evidence that S3D may be beneficial, however, there are reservations about how accurate these findings may be when examining study limitations. Specifically, there is no previous research investigating the relationship between individual stereo acuity and performance using S3D for a warning alert task simulated in an HMD. This work aims to investigate human performance using well aligned S3D symbology when applied to a warning alert displayed in an HMD whilst flying a simulated operationally relevant mission.

The aim of this research is not to suggest how S3D should be used in an HMD, but rather if S3D were to be used to display symbology, how it affects performance dependent on the visual capabilities of the user population, as defined by current military aircrew stereo acuity vision standards. Current test methods employed to qualify aircrew may have significant limitations. Principally, an inability to measure stereo acuity to threshold prevents current paper-based stereo tests from being useful when investigating any relationship between human performance and individual stereo acuity which could ultimately result in a human-machine technology mismatch. A novel aspect of this research will be to develop alternative digital computer-based stereo tests to measure stereo acuity more reliably and to threshold levels. With reliable stereo acuity scores, it will be possible to examine the ability of stereo acuity to predict performance in an aviation task using S3D in an HMD.

2. STEREO ACUITY TESTS

Work presented in this chapter draws upon published research (Posselt *et al.*, 2022). Few previous studies investigating performance using S3D displays screened their participants for stereopsis despite up to 30% of the general population being stereo anomalous (Hess *et al.*, 2015). Such an omission may hamper interpretation of results as stereo anomalous individuals will be less able to perceive disparity changes and their performance may be unaffected, or even impaired. Even fewer studies accurately measure stereo acuity in their participants. In studies that do screen for a minimum stereo acuity level, the stereo tests used are paper-based with significant limitations. The same concerns exist regarding the stereo tests currently used to select military aircrew, which are fallible and inadequate (Posselt and Winterbottom, 2021). Updated stereo acuity tests could improve the vision screening process for research and for military selection, making the process fairer, more accurate, and more reliable. Equally, having the same tests or even comparable tests employed by different countries, with aligned aeromedical policy would improve interoperability with regard to human resources, allowing each country to accept aircrew from allied partner nations without further vision testing. The ability to more reliably and accurately measure stereo acuity may also be of increasing importance when evaluating human performance using an HMD with an S3D display.

For this experiment, stereo acuity thresholds obtained using four different stereo acuity tests were analysed. The four stereo acuity tests used were: (1) the stereo acuity TNO test currently used by the UK military to qualify aircrew; (2) a static digital version of a random dot stereogram (dRDS-S) that used a pie shaped stimulus similar to the TNO test stimulus; (3) a dynamic digital random dot stereogram (dRDS-D) that also used a pie shaped stimulus; and (4) a digital Stereo Search Test (SST) which has previously been able to predict performance on a simulated air refuelling depth related task (O'Keefe *et al.*, 2022). A pilot study was conducted with 10 participants using the TNO test and static version of the RDS only (Posselt *et al.*, 2021) which informed the final study design.

2.1. Research questions

With regard to the first research question (R1) set out in Section 1.1;

- Does measured stereo acuity differ between the TNO test, two novel dRDS tests and the digital SST?
- What are the limitations of the TNO test and are alternatives more suitable?

2.2. Stereo acuity tests

Stereo tests used to qualify military aircrew, listed in Table 1. of Section 1.3, were all developed for clinical settings and largely for screening purposes. A thorough review of the USAF stereo test methods is provided by Winterbottom *et al*. (2014). None of the tests listed in Table 1. provide a true threshold measure of stereo acuity, they rely only on the ability to simply detect a static disparity with respect to zero, and most tests do not assess the person's ability to differentiate between crossed and uncrossed disparities. Thus, it is eminently possible to score well on one test whilst fail another (Leske, Birch and Holmes, 2006; Pageau, Guise and Saint-Amour, 2009). The

Operational Based Vision Assessment (OBVA) Laboratory aims to improve test methods to counter the shortfalls in analogue tests, to measure vision more reliably and to investigate relationships between vision and operational performance so that decisions regarding vision standards are rooted in evidence. Newer computer-based tests are continuously being developed and tested by the OBVA Laboratory and others (Hess *et al.*, 2016; Rodriguez-Vallejo *et al.*, 2016; Tittes *et al.*, 2019; McCaslin *et al.*, 2020). As part of the ambition to improve and modernise vision tests, three digital stereo tests are investigated here; the Stereo Search Test (SST) and two digital versions of a RDS (dRDS).

For this research, newly developed digital tests will be compared against the stereo test used by the UK military to assess stereopsis to qualify aircrew; the TNO test (Figure 6). The TNO test is a seven-page booklet of randomly paired red and green dots which should be viewed at 40 cm using red-green coloured glasses (anaglyph 3D). The individual must identify the image and the orientation of the missing segment of the circle, calling out their answer to be recorded manually by an examiner. There are two circles for each level of stereo acuity and a subject must identify both correctly to progress to the next level. The binocular disparity of targets is 480, 240, 120, 60, 30 and 15 arcseconds, although the UK military employs a version with only six pages which means the best scores that can be achieved is 60 arcseconds. Unless directly specified, the TNO test referred to in this work is the six-page version used by the UK military.

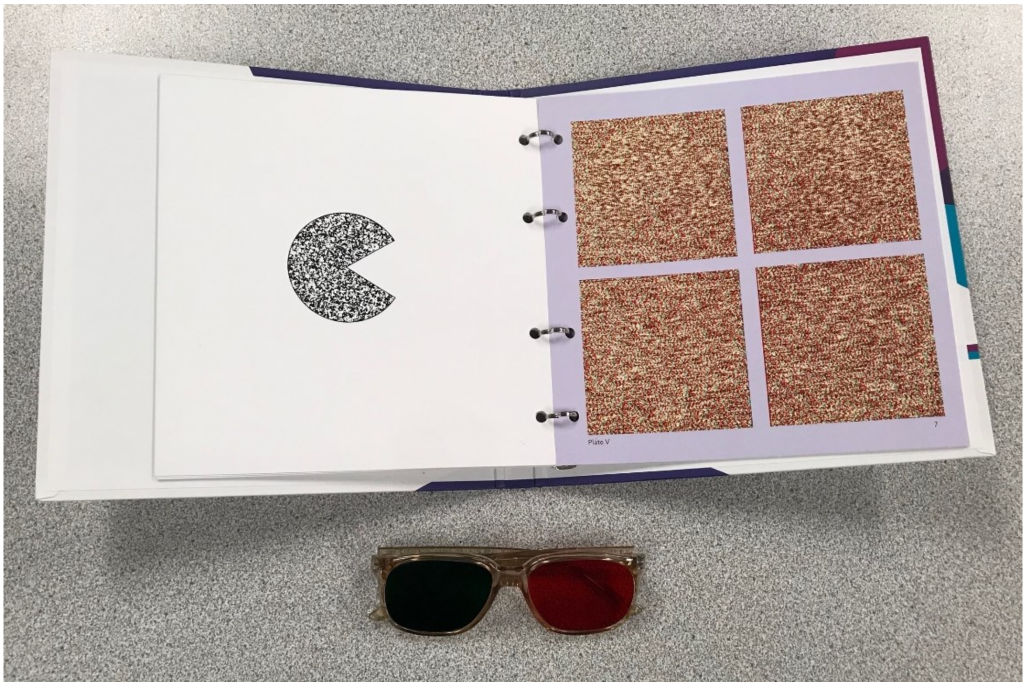

Figure 6. TNO stereo test. Photograph taken by OBVA personnel

The TNO test uses RDS stimuli, so monocular cues should not play a part in interpreting the orientation of the missing piece of circle. As an RDS test, with minimal monocular cues, it is not unexpected that subjects have higher (worse) thresholds on the TNO test compared to a stereo test which is contour based (Hall, 1982). However even taking this difference into consideration, subject performance on the TNO test can be up to 25% worse than with any other stereo tests (Momeni-Moghadam *et al.*, 2011; Antona *et al.*, 2015; Vancleef *et al.*, 2017) A possible reason why the TNO test gives higher stereo thresholds than even other RDS tests could be that the different colour filters in the glasses causes an imbalance in luminance transmittance and contrast. Significantly worse stereo acuity results were recorded with the original TNO test using red-green colour filter glasses compared to using isoluminant, grayscale, and black-white equivalent pictures (Zhang, Meng and Wu, 2021). Any reduced ability to measure stereo acuity could be exacerbated further if lighting conditions are suboptimal. Another reason as to why stereo acuity thresholds are higher using the TNO test is that it is a two-stage process, the user must first identify the circular shape and then indicate the orientation of the missing piece (Hall, 1982). In comparison, simple detection tests such as the Randot test, require the user to simply identify the sole stimuli in depth as the ‘odd one out’ (Vancleef *et al.*, 2017).

Stereo acuity can also be affected by the dot size used in an RDS test (Simons, 1981). As a method to reduce presented disparity, some stereo tests require the paper stimulus to be moved further away from the viewer, like the Random dot E stereo test. Moving the test further away has the additional effect of reducing perceived dot size, which in effect also tests visual acuity. In such situations, measured stereo acuity is worse using RDS paper tests which need to be moved further away compared to RDS tests set at a fixed distance closer to the subject (Simons, 1981). Therefore, it is possible that if smaller dots are used in an RDS test, the stimulus may be more difficult to view by those with reduced visual acuity. In support of visual acuity not being a limiting factor to interpret disparity, in studies that optimally corrected for refraction, or screened their subjects to ensure good visual acuity, similar changes in stereo acuity were observed between the TNO test and other stereo acuity tests (Lee and Koo, 2005; Antona *et al.*, 2015). In addition to the TNO test yielding higher stereo thresholds, there are concerns with using the TNO test to assess stereo acuity more generally. The TNO test has poor test re-test reliability, answers can be easily memorised, and there can be an unacceptable degree of variation between different test editions, due to flaws in the printing process (Antona *et al.*, 2015). Van Doorn *et al.* (2015) demonstrated a statistically significant difference in stereo acuity results obtained using two separate editions of the TNO test, with mean measured stereo acuities of 30 arcseconds using one edition and 60 arcseconds using a different edition ($p<0.001$). These differences were likely due to inconsistent image quality, resulting from differences in the printing process (Van Doorn *et al.*, 2014).

To address the various limitations of the TNO test outlined here, a more reliable stereo test is sought. Ideally, to assess optimal human performance, the stimulus used in a stereo test should be sufficiently bright (>30 cd/m2), with adequate contrast (Michelson contrast of a minimum of 0.1, but preferably 0.3), the image should be sharp to view and displayed for a minimum of 200 ms (Westheimer, 2013). Furthermore, the stimulus should not be crowded, and a degree of familiarity with the target should be allowed with practice trials to mitigate any significant learning affect, which can be particularly apparent with stereo acuity (Schmitt *et al.*, 2002) and specifically with a RDS test (Westheimer, 2013). The recommendations set out by Westheimer (2013) were kept in mind when designing a new digital stereo test for this research. Firstly, in order to avoid the issues caused with coloured filtered glasses, a temporally interlaced display which alternates frames to the right and left eyes viewed through active shutter glasses can be used as an effective alternative way to display a stereo stimulus, as long as the refresh rate of the screen is adequate (greater than 30 Hz) (Westheimer, 2013). Secondly, using a computer-generated random response approach removes the ability to cheat or memorise answers. Thirdly, to measure a threshold perception of a stimulus and avoid a floor effect, a psychometric function could be used. The four alternate forced choice (4AFC) adaptive procedure is a more effective way of measuring detection threshold than relying on absolute signal detection by a subject, which is biased by individual decision criteria (Leek, 2001). RDS are least likely to have any monocular cues, and while they may appear harder, they are most likely to more accurately measure stereopsis.

To that end, the ambition was to develop a stereo test using a digital 3D display and to be able to measure stereo acuity to a threshold value. For familiarity and comparison with the standard TNO test, the stimulus kept the same shape; a circle with a missing segment. To be most comparable to the TNO test, a static version of a digital RDS was developed. In addition, a version of the digital RDS test was also developed in which the dots appear and disappear with each frame refresh of the screen. Refreshing the position of RDS dots in this way is only possible when using digital technology; updating dot position dynamically between each frame refresh within the plane of depth.

In order to assess performance using a newly developed RDS digital stereo test, comparison was made with a different type of digital stereo test, the SST. The SST is part of the Automated Vision Tester (AVT); a battery of computer-based vision tests developed and validated by the OBVA laboratory (Shoda *et al.*, 2017; Winterbottom *et al.*, 2019). Using the SST to measure individual stereo acuity, it has previously been possible to predict performance in a remote vision system air refuelling simulation (O’Keefe *et al.*, 2022), yet a test similar to the standard Titmus booklet test was not able to predict performance in the same depth related task. In a further example illustrating the benefits of digital stereo testing; a computer-based stereo test similar to the SST and referred to as a dual ring stereo acuity test, predicted simulated RVS refuelling performance whereas the standard USAF stereo tests did not (Winterbottom *et al.*, 2016).

There are further examples still of the benefits to using digitalised stereo tests (Vancleef *et al.*, 2017; Tittes *et al.*, 2019); with evidence that they are better able, than standard optometric tests, to predict performance and reported discomfort using S3D displays (Winterbottom *et al.*, 2016). In one particular study, a digital version of the TNO test was more reliable than the paper TNO version (Tittes *et al.*, 2019). With a large age range (4–59 years), participants were subdivided into normal and impaired binocular vision, with and without strabismus. The paper TNO version of the test they used was the seven-page version, able to measure stereo acuity down to 15 arcseconds. For the normal control participants, most similar to our military aircrew population, median stereo acuity as measured with the digital RDS test was 51 arcseconds (range 11–161); yet, with the paper TNO test the range was greater (15–240 arcseconds). When comparing results from the paper TNO test against their digital RDS tests, counter to other studies, participants were able to achieve lower stereo acuity scores using the paper TNO test, with a more pronounced effect for finer thresholds. The threshold results obtained with the digital RDS test were calculated through maximum-likelihood fitting a normal function in Palamedes toolbox at the 62.5% correct point, with a guess rate of 25% and fixed lapse rate of 2% (Prins and Kingdom, 2018). When reassessed using an higher level 85% correct point to calculate a threshold, which is thought to be the correct point of the paper TNO test as modelled by Vancleef *et al.*, (2017), agreement between the paper TNO and digital RDS was poorer (k = 0.17, p = 0.43, n = 32). It would be expected that agreement with the digital RDS would be poorer still when compared to the UK military stereopsis vision standards as the paper TNO test used by the British military is the six-page version, thus only able to characterize stereo acuity to 60 arcseconds. Regarding reliability, Bland-Altman assessment of test re-test reliability for the digital RDS stereo test developed by Tittes *et al.* (2019) showed a mean difference of 0.04 log arcseconds and 95% limits of agreement of 0.5 log arcseconds. Of note, the digital RDS tests were less reliable for those with worse stereo acuity, with every 10-fold increase in threshold there was a 30% increase in standard error. In comparison, for the paper TNO test there was no correlation between first and second test scores and Bland-Altman analysis for test-retest showed a mean difference of 0.06 log arcseconds with wider 95% limits of agreement of 1.53 log arcseconds (Tittes *et al.*, 2019).

The ambition of Tittes *et al.* (2019) was to allow detailed quantification of stereo acuity to improve assessment and evaluation of amblyopia therapy, and they did so by creating a digital RDS that was more reliable than the paper TNO test. Its use in qualifying applicants for aircrew duties is a different application entirely but provides evidence that digital tests can provide more reliable methods for measuring stereo acuity with finer granularity of results. If aircrew were to suffer traumatic brain injury, any ocular disease or undergo treatment, a similar method would be better able to detect pathology and to more closely track recovery to potentially enable a quicker return to flying.

All of these studies highlight the substantial limitations of current paper-based tests and propose alternative digital versions to reduce or eliminate many of the problems related to paper-based stereo acuity tests. This research aims to apply digitalised stereo test principles to military stereo acuity selection standards and investigate their ability to predict operator performance when using S3D in an HMD.

2.3. Ethics

All studies were approved by the United States Air Force (USAF) Air Force Research Laboratory (AFRL) Institutional Review Board (IRB). Volunteers were provided with a written information sheet. As part of the Automated Vision Tester (AVT) battery, all participants were required to sign informed consent, indicating their permission for their unidentifiable data to be stored and used within the lab. Signed consent was obtained from each participant to collect data and keep their contact details on the OBVA database.

2.4. Apparatus

2.4.1. TNO test (UK military standard)

The TNO test used by the UK military to qualify their aircrew is a six-page version with stimuli composed of random dot stereograms (Figure 7), viewed at 40 cm using red-green coloured glasses. The subject must identify the image and the orientation of the missing segment of the circle in each box. TNO test stimuli are presented with crossed disparity and thus appear in front of the reference plane. There are two circles for each stereo acuity level and the subject must identify both correctly to progress to the next, more difficult level. The binocular disparity of targets is 480, 240, 120, and 60 arcseconds with subjects’ scores recorded manually by the examiner.

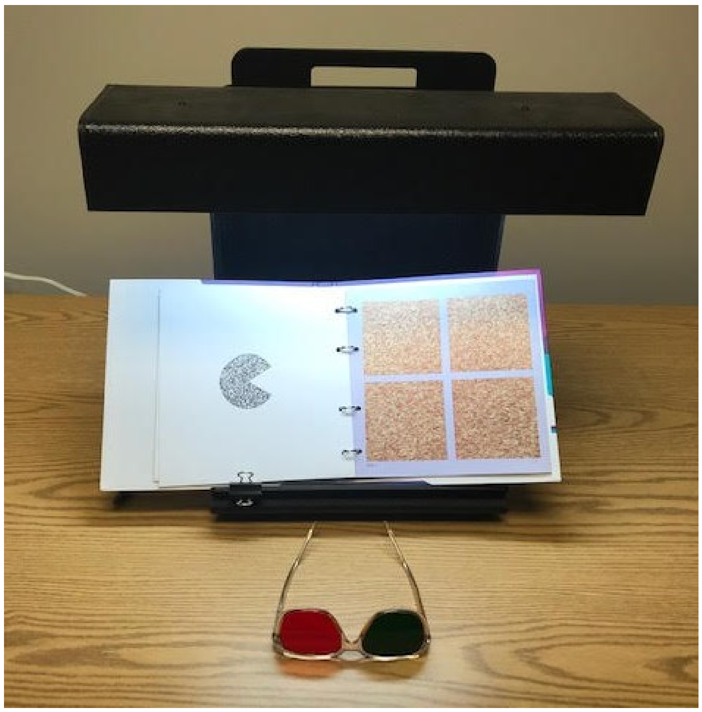

Figure 7. TNO test booklet with anaglyph glasses on an inclined stand. Photo taken by OBVA personnel.

Reflective luminance of the booklet pages under a broadband reading light (incandescent bulb) was 197 to 306 cd/m^2 measured with a handheld Konica Minolta LS-110 (Konica Minolta Sensing America, NJ, USA). Luminance decreased when viewed through the coloured lenses; to 13-23 cd/m^2 through the red lens and to 3-8 cd/m^2 through the green lens.

2.4.2. Digital Random Dot Stereogram (dRDS) tests

The stimulus used for both dRDS tests is similar in appearance to the TNO test stimulus, but displayed on a 53 x 23cm 3D vision ASUS display (ASUS, CA, USA) and uses NVIDIA 3D Vision shutter glasses synchronised via an infrared transmitter. The shutter glasses synch with the refresh rate of the screen to ensure that the two disparate images, displayed one after the other, are presented to each eye separately to create a stereoscopic image. The test is performed in a darkened room with the participants' head fixed at 1 m using a chin rest (Figure 8 and Figure 9). The stimulus is presented with crossed disparity, appearing to be forward in depth compared to the plane of the computer screen, although it is possible to create uncrossed disparities with the stimulus appearing behind the screen if desired. The perceived orientation of the missing segment of the circle is entered directly by the subject using the directional arrows on a keypad.

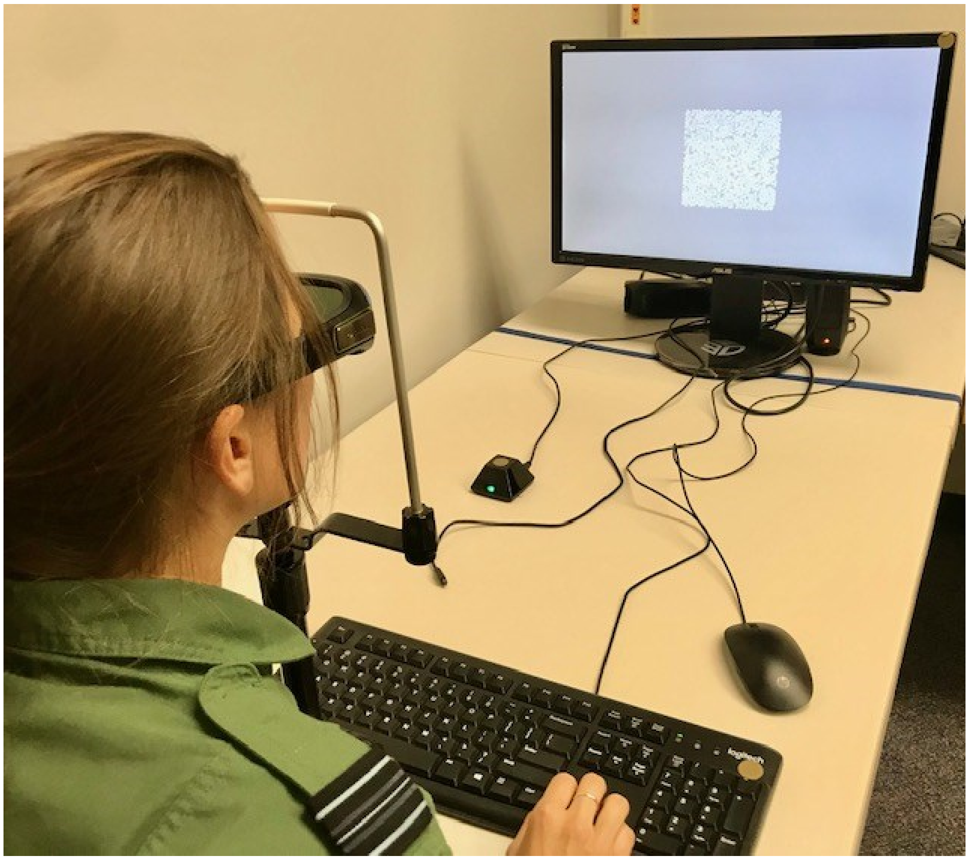

Figure 8. Apparatus for dRDS-S and dRDS-D tests. Photograph taken by OBVA personnel.

The mean luminance of the background on the stereo monitor was 49.72 cd/m^2 and 72.50 cd/m^2 on the white aspect of RDS stimulus, measured with 100% monitor brightness. When viewed through the stereo glasses luminance of the background reduces to 16.16 cd/m^2 and to 24.08 cd/m^2 of the white aspect of RDS stimulus. The temporal resolution was 120 Hz (60 Hz to each eye) and spatial resolution 1920 x 1080 pixels. The cross talk apparent by each eye with the NVIDIA active shutter glasses is illustrated in Appendix C, and is below the 5% level known to induce visual discomfort (Kooi and Toet, 2004). The RDS test was programmed in C# in visualbasic.net using Direct X and a 2.2 gamma correction applied to improve accuracy for sub-pixel shift. Sub-pixel shift is a critical capability to ensure that disparities can be simulated to match the stereo acuity of the best observers and avoid a ceiling effect (Westheimer, 2013). Each dot was defined by a Gaussian function with a sigma of 3.75 pixels and there were 4000 dots randomly placed within the box. Two versions of the dRDS were implemented: (1) a static RDS and (2) a dynamic RDS. In the static version (dRDS-S), all dots remain stationary, presented for 8 seconds. In the dynamic version (dRDS-D), the position of each dot was randomised with each frame refresh, completely eliminating any monocular cues, also presented for 8 seconds. The square box measured 14 x 14 cm and the diameter of the stimulus was 7 cm (154 arcminutes).

Figure 9. Computer screen display of the dRDS stimulus. Grey outline of circle with a missing slice is artificially overlayed to illustrate how the stimulus would appear. Photograph taken by OBVA personnel.

To create the disparate image, three regions were designated within the box: A, B, and C (Figure 10). Region A was the shape of the stimulus at its origin and all dots within it were shifted horizontally to region B and duplicated in region C. The dots in region B are presented only to the right eye and the dots in region C are presented only to the left eye, thereby creating a stereoscopic image.

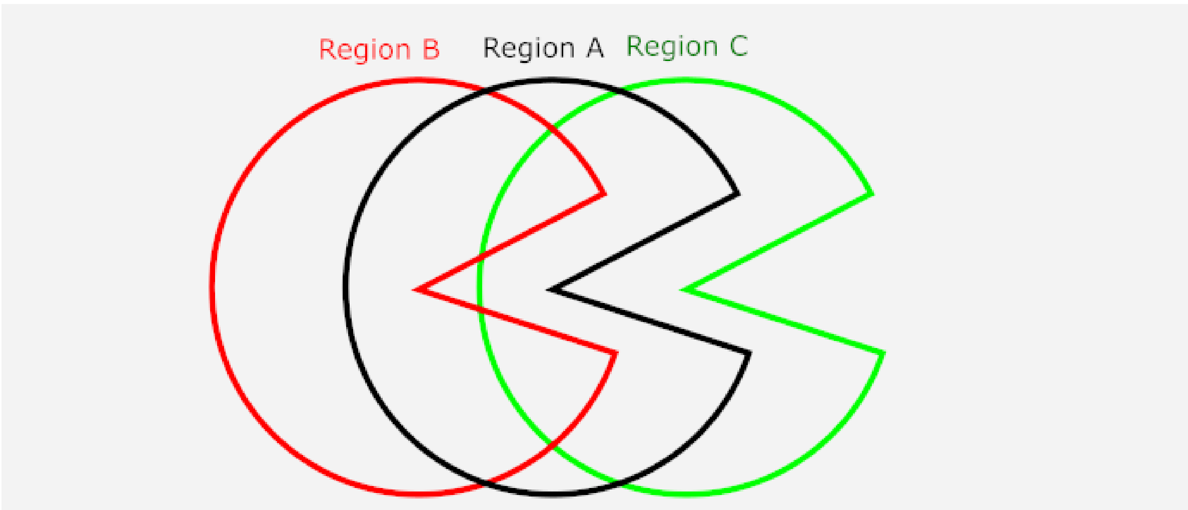

Figure 10. Dot position to create the disparate image. Image created by OBVA personnel.

Simply shifting dots from region A to regions B and C creates both gaps between dots and an overlap of dots when regions B and C are viewed binocularly relative to the background dots in the area outside of region A, which could create monocular cues. To correct for this, excess background dots (checked regions) were removed and new dots were placed in the void areas (striped region) for the opposite eye (Figure 11).

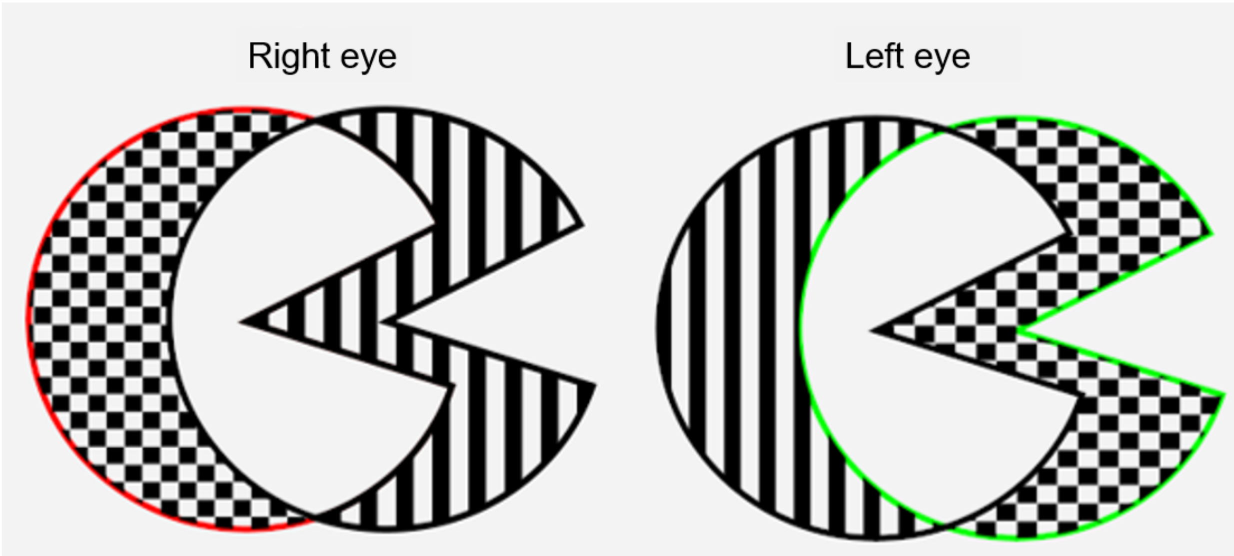

Figure 11. Result of not correcting simple horizontal shift of stimulus dots. Striped regions are areas where a void is created and in the checkered region there will be double the number of dots. Image created by OBVA personnel.

For both dRDS tests, instructions were given on the screen. Using the four alternative forced choice method (4AFC), participants were instructed to choose up/down/right/left orientation of the stimulus. The forced choice model is a more effective way of measuring a detection threshold than relying on signal detection by a subject, which is biased by individual decision criterion (Vancleef *et al.*, 2018). Using four choices is also more efficient than using two, reducing the guess rate to 25% with fewer trials needed to reach a reliable estimate of the detection threshold (Vancleef *et al.*, 2018). Each stimulus was presented for a maximum of 8 seconds, if no response was given within that time, it was logged as an incorrect response. Each participant had a 10-trial practice session using stimuli with disparities of 300 to 2000 arcseconds prior to the formal test to ensure that participants understood the test and reduce practice effects. Each dRDS test consisted of 45 trials with stereo acuity threshold, standard error, and slope estimate results displayed in an Excel spreadsheet (Microsoft, Redmond, WA, USA). The Psi paradigm (Prins and Kingdom, 2018) is an adaptive procedure that was used to determine the most informative stimulus amplitude on each trial. Psychometric function were fit to the data to estimate detection threshold (Leek, 2001) calculated at the 70.83% correct point, which is the alpha value for a Gumbell function with a guess rate of 25%. The lapse rate of the psychometric function was fixed at 2.5%, but the slope was allowed to vary to allow for greater accuracy of the threshold estimate (Prins and Kingdom, 2018). The lapse rate, which is sometimes called the finger-error rate, accounts for psychophysical errors not directly related to the observer's perception of the stimulus, such as accidentally pressing the wrong response button on the keypad. Disparity of the test stimuli was altered in 0.1 log arcsecond step sizes, based on the participant's previous prior responses (i.e., the test generally got easier if the participant answered incorrectly, but more difficult if the participant responded correctly). The design and thresholding method of the dRDS test enabled each participants' stereo acuity threshold to be measured from 5 to 8000 arcseconds.

2.4.3. Stereo Search Test (SST)

The Stereo Search Test (SST) is part of a battery of vision tests developed with industry partners to modernise vision assessment and selection in the USAF (Choi *et al.*, 2018; Winterbottom *et al.*, 2018, 2019). The four ring stimulus as illustrated in Figure 12 is performed on the same apparatus as the dRDS-S and dRDS-D (Shoda *et al.*, 2017).

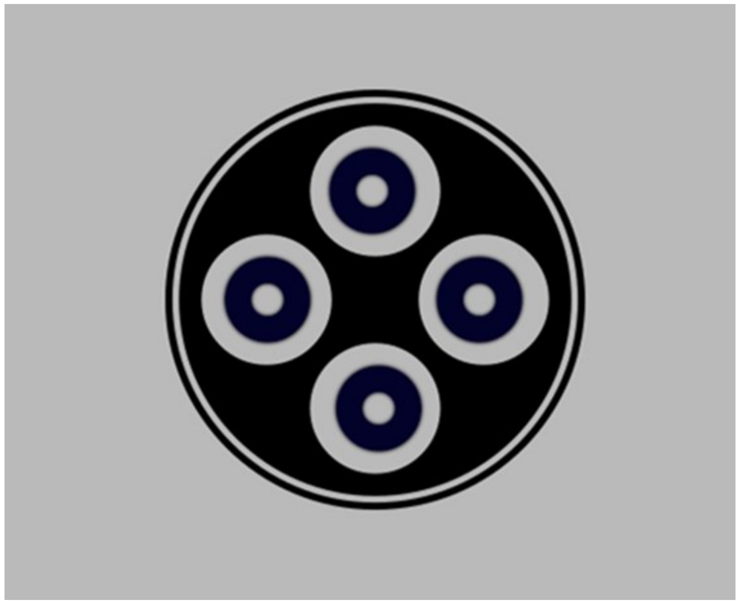

Figure 12. Stereo Search Test stimulus. Image created by OBVA personnel.

The target ring was shown with crossed disparity (in front of the reference plane), which the subject must distinguish from three distractor rings. One of the distractor rings was shown with an equal amount of uncrossed disparity (behind the reference plane), while the other two rings were displayed at the same depth as the screen’s reference plane. Thus, the observer must be able to distinguish crossed from uncrossed disparity to select the correct target ring. The SST is conducted over 40 trials. The same psychophysical procedure was used as the dRDS tests. Similar to the dRDS tests, the SST uses antialiasing to enable sub-pixel shifts in disparity (Bach *et al.*, 2001). The SST also uses monocular cue masking – randomised horizontal shifts in position of the rings to reduce the possibility that observers can rely on monocular cues to identify the crossed disparity ring. The SST has a high test-retest reliability (r = 0.94) (Winterbottom *et al.*, 2018) and a learning period is provided beforehand to minimise practice effects.

2.5. Methodology

Each participant undertook the SST as part of the AVT battery of vision tests prior to enrolling in this study. The TNO, dRDS-S and dRDS-D were taken together and repeated a second time on a separate day. The order in which the TNO and both dRDS tests were administered was randomised using Microsoft Excel software (Microsoft, Redmond, WA, USA). Verbal instructions were given for the TNO test with responses recorded by the examiner. For both the dRDS tests, written instructions were given on the screen. Each dRDS test consisted of 45 trials with stereo acuity threshold, standard error, and slope estimate results displayed in an Excel spreadsheet (Microsoft, Redmond, WA, USA).

2.6. Results

2.6.1. Demographics

Forty-five participants were recruited from the OBVA subject database. However, one participant did not complete both sessions and another participant’s data file was corrupted. A further two participants were essentially stereo blind, with scores of 4 and 3.8 log arcseconds as measured with the dRDS-S test. These scores are at the uppermost limit of the tests’ capability and are likely unreliable. Both these scores are greater than 1.96 times the standard deviation, thus these two participants are considered outliers and were excluded from further analysis with regard to assessing the stereo acuity tests. The final sample size was n = 41 (24 male, 37.5 mean age SD ± 10.1 years, range 21–70). One further participant did not complete the AVT, thus for analyses of SST results n = 40.

Recruitment was targeted to ensure participants with a wide spread of stereoacuities were included; thus, the sample population is broader than traditional military aircrew populations, who are required to meet the stereo acuity selection standard of 120 arcseconds as measured with the TNO stereo test in the UK and 40 arcseconds as measured by the AFVT for USAF.

2.6.2. Descriptive statistics

Mean and median stereo acuity values are given in Table 2 with a histogram illustrating frequency of results in Figure 13.

Table 2. Mean, median, range and standard deviation of the four stereo acuity tests

	dRDS-S	dRDS-D	TNO	SST
Mean Log arc sec (arcseconds)	1.53 (33.79)	1.78 (59.6)	1.96 (91.56)	1.37 (23.42)
Median Log arc sec (arcseconds)	1.51 (32.63)	1.78 (60.9)	1.78 (60)	1.25 (17.58)
Range Log arc sec (arcseconds)	1.10–2.24 (12.64–173.78)	1.38–2.41 (23.87–197.83)	1.78–2.38 (60–240)	0.66–2.59 (4.5–389)
SD Log arc sec (arcseconds)	± 0.25	± 0.24	± 0.22	± 0.49

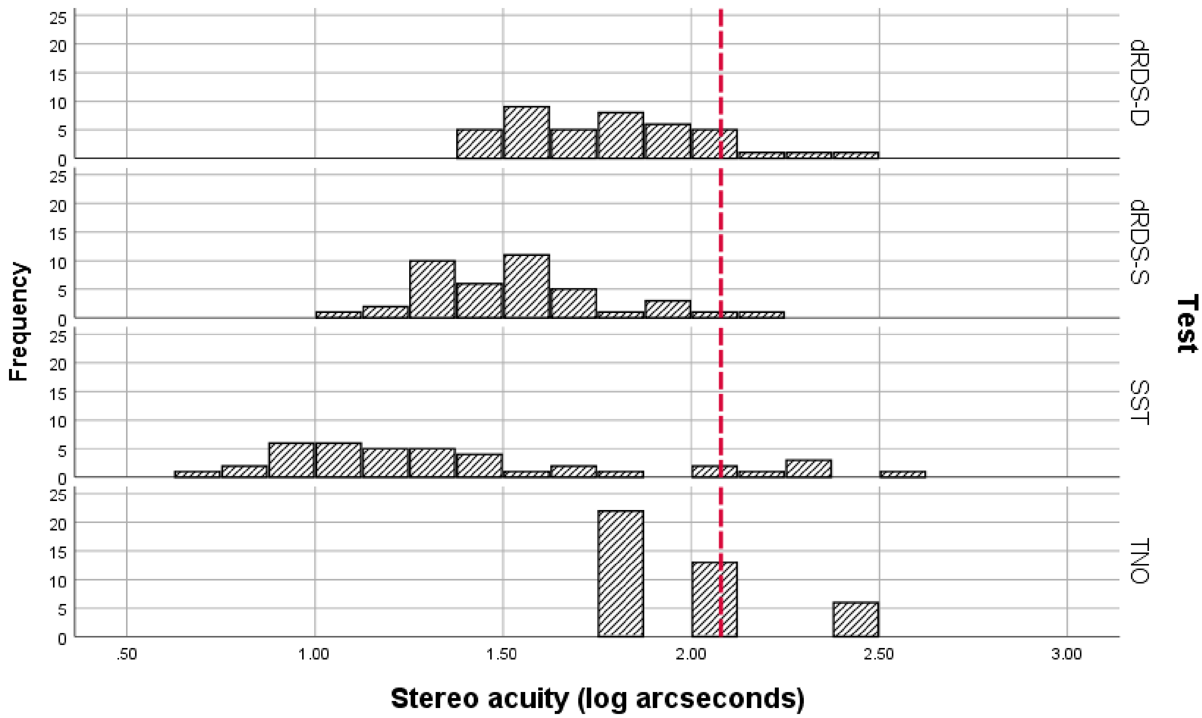

Figure 13. Histogram of stereo acuity score (log arcseconds) frequencies measured with dRDS-D, dRDS-S, SST, and TNO. Red dotted line is placed at the UK military stereo acuity standard for aircrew at 2.08 log arcseconds (120 arcseconds).

A Friedman two-way analysis of variance (ANOVA) by ranks was performed to determine if there was a statistically significant difference in the scores obtained using the four different versions of stereo acuity tests. The Friedman two-way ANOVA, a non-parametric test, was chosen because the stereo acuity scores were not normally distributed. Per Freidman's test, there was a statistically significant main effect in the distribution of stereo acuity scores for the four tests, x^2 (Chi-Squared) = 53.57, df (3), $p < .01$. Post-hoc, the Wilcoxon signed-rank test was then performed to determine where the significant differences existed between each pair of different stereoacuity tests. Using a Bonferroni-corrected p-value of .012, the only test pair that did not differ significantly was the dRDS-S compared to the SST. Results are given in Table 3.

Table 3. Wilcoxon signed rank test between pairs. Significance level $p<0.012$

Comparison	Z score	Sig.
dRDS-S vs. dRDS-D	-5.17	$p < 0.001$**
dRDS-S vs. TNO	-5.40	$p < 0.001$**
dRDS-S vs. SST	-1.90	$p = 0.06$
dRDS-D vs. TNO	-3.58	$p < 0.001$**
dRDS-D vs. SST	-3.78	$p < 0.001$**
TNO vs. SST	-4.42	$p < 0.001$**

2.6.3. Test re-test of stereo tests

Bland-Altman plots are used to quantify the reliability of a repeated test, or agreement between two measures, by comparing mean differences and calculating limits of agreement across a range of data. The closer the mean difference is to zero, the better the agreement between two measures. A smaller standard deviation of differences indicates good reliability and repeatability. Bland-Altman plots are considered a beneficial additional way to understand comparability between two measures than using simple correlation analyses alone. Correlation plots evaluate only linear association and used on their own could be misleading (Giavarina, 2015). It should be noted that while the Bland-Altman method calculates limits of agreement, it is unable to determine whether these are acceptable or not. That is a separate task entirely depending on the user's appetite for risk and need for reliability. Results for test-retest reliability for both dRDS tests and TNO as analysed with Bland-Altman plots are summarised in Table 4 and illustrated in Figure 14. The SST is taken separately as part of the AVT battery of vision tests. Consequently, not all participants completed the SST twice. The SST has been analysed separately with test re-test reliability previously reported as $r = 0.94$, $p < 0.01$ (Winterbottom *et al.*, 2018).

Table 4. Summary of test re-test of dRDS-D, dRDS-S, and TNO stereo tests. SD = Standard Deviation, LOA = Limits of Agreement

	dRDS-D	**dRDS-S**	**TNO**
Mean difference (Log arcseconds)	0.05	0.04	0.03
SD	± 0.12	± 0.19	± 0.17
LOA	± 0.23	± 0.25	± 0.34

For the dRDS-D test, the mean difference between first and second sessions was 0.05 log arcseconds (95% CI = ± 0.04) with limits of agreement ± 0.23 log arcseconds, a combined total of 0.47 log arcseconds (Figure 14, top). For the dRDS-S test, the mean difference between first and second sessions was 0.04 log arcseconds (95% CI = ± 0.04) with limits of agreement ± 0.25 log arcseconds, a combined total of 0.50 log arcseconds (Figure 14, middle). For the TNO test, twenty-seven participants achieved the same score on both attempts (66%), with some participants reporting that they were able to simply remember their answers from the previous session. With Bland-Altman analysis the mean difference between first and second sessions of the TNO test was 0.03 log arcseconds (95% CI = ± 0.05) with limits of agreement ± 0.34 log arcseconds, a combined total of 0.69 log arcseconds (Figure 14, bottom). These limits of agreement are artificially narrowed, since there are only 4 possible stereo acuity values for the RAF six-page version of the TNO test, thus, a large degree of variance has already been removed. When translated into a real-life example, subjects scoring 120 arcseconds with the TNO, may actually have a true stereo acuity varying anywhere between 70.6 to 203.8 arcseconds. These results indicate better test re-test reliability of the TNO test than previous reports of a difference of 0.06 log arcseconds and 95% limits of agreement of 1.53 log arcseconds (Tittes *et al.*, 2019). Although, Tittes et al, 2019, used a paper TNO test with an additional two levels, able to measure down to 15 arcseconds. In our subsequent analyses, the first stereo acuity score was used.

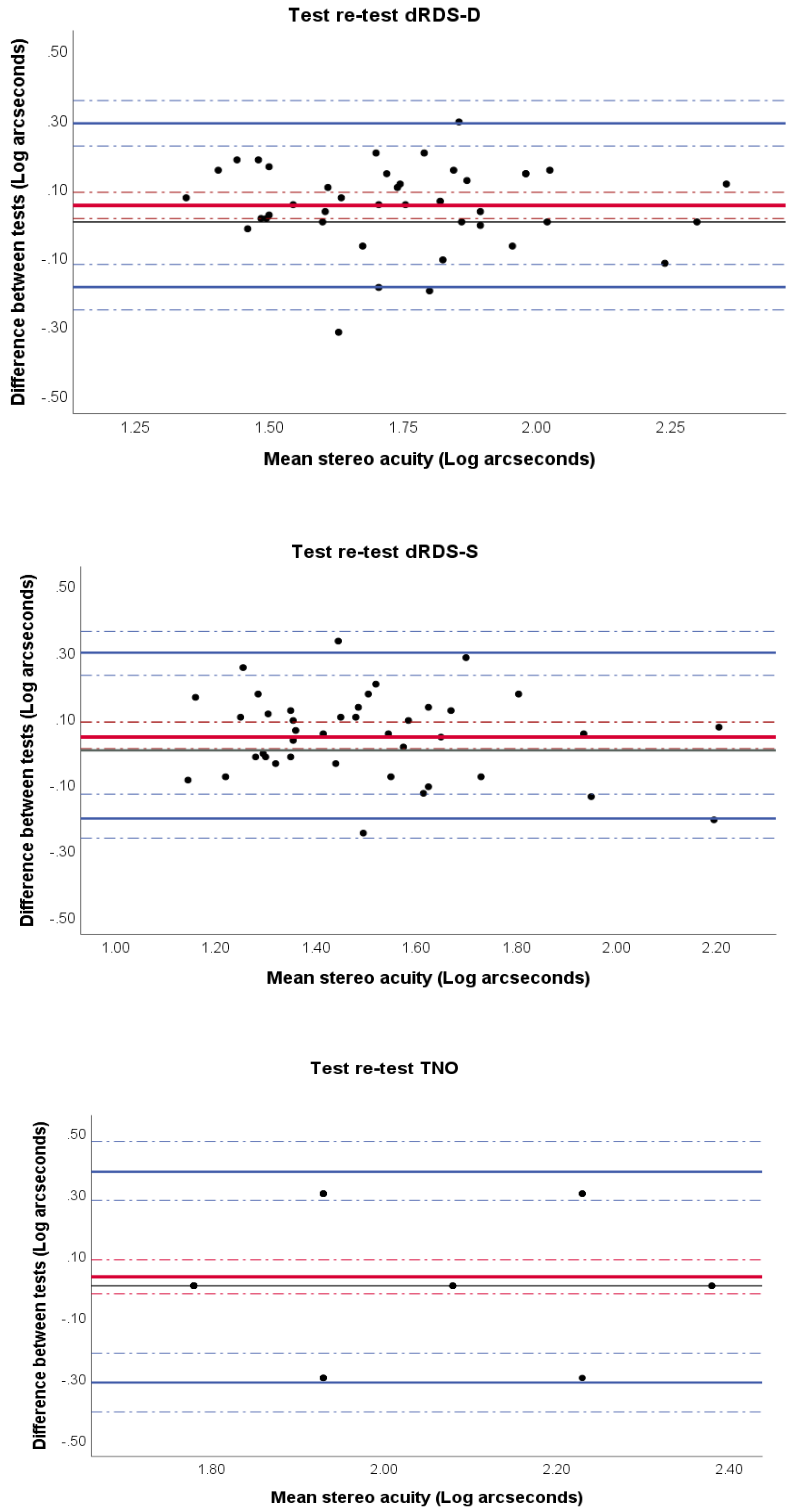

Figure 14. Bland-Altman plots illustrating the agreement between first and second test results using the dRDS-D test (top), the dRDS-S test (middle), and TNO test (bottom). Red line is the mean difference (bias), with the solid blue lines indicating the upper and lower limits of agreements. Dotted lines show 95% confidence intervals around each solid line.

2.6.4. Comparison of stereo tests

Each stereo test assesses stereopsis using a different method, giving significantly different results. It is important to compare levels of agreement between these tests, to aid interpretation and relatability. It is acknowledged that TNO data points are categorical in nature, whereas the digital tests record data in a continuous manner. However, as TNO data points measure stereo acuity, which is a continuous scale, it was thought that comparisons between the TNO test and digital stereo tests were still feasible and useful. A simple scatter plot between the two dRDS stereo tests show that they are well correlated (r = 0.73, p < 0.001), and each dRDS is in turn well correlated with the SST (SST and dRDS-D, r = 0.61, p < 0.00; SST and dRDS-S, r = 0.65, p < 0.00). There is no significant relationship between the TNO test and any of the three digital stereo tests; SST or either dRDS. These relationships are illustrated in Figure 15 and a summary of all correlation analyses is given in Table 5.

Table 5. Spearman rho correlation analyses between stereo tests

Stereo test	SST	dRDS-D	dRDS-S
SST	r = 1		
dRDS-D	r = 0.61, **p < 0.00****	r = 1	
dRDS-S	r = 0.65, **p < 0.00****	R=r = 0.73, **p < 0.00****	r = 1
TNO	r = 0.06, p = 0.71	r = 0.24, p = 0.08	r = 0.27, p = 0.08

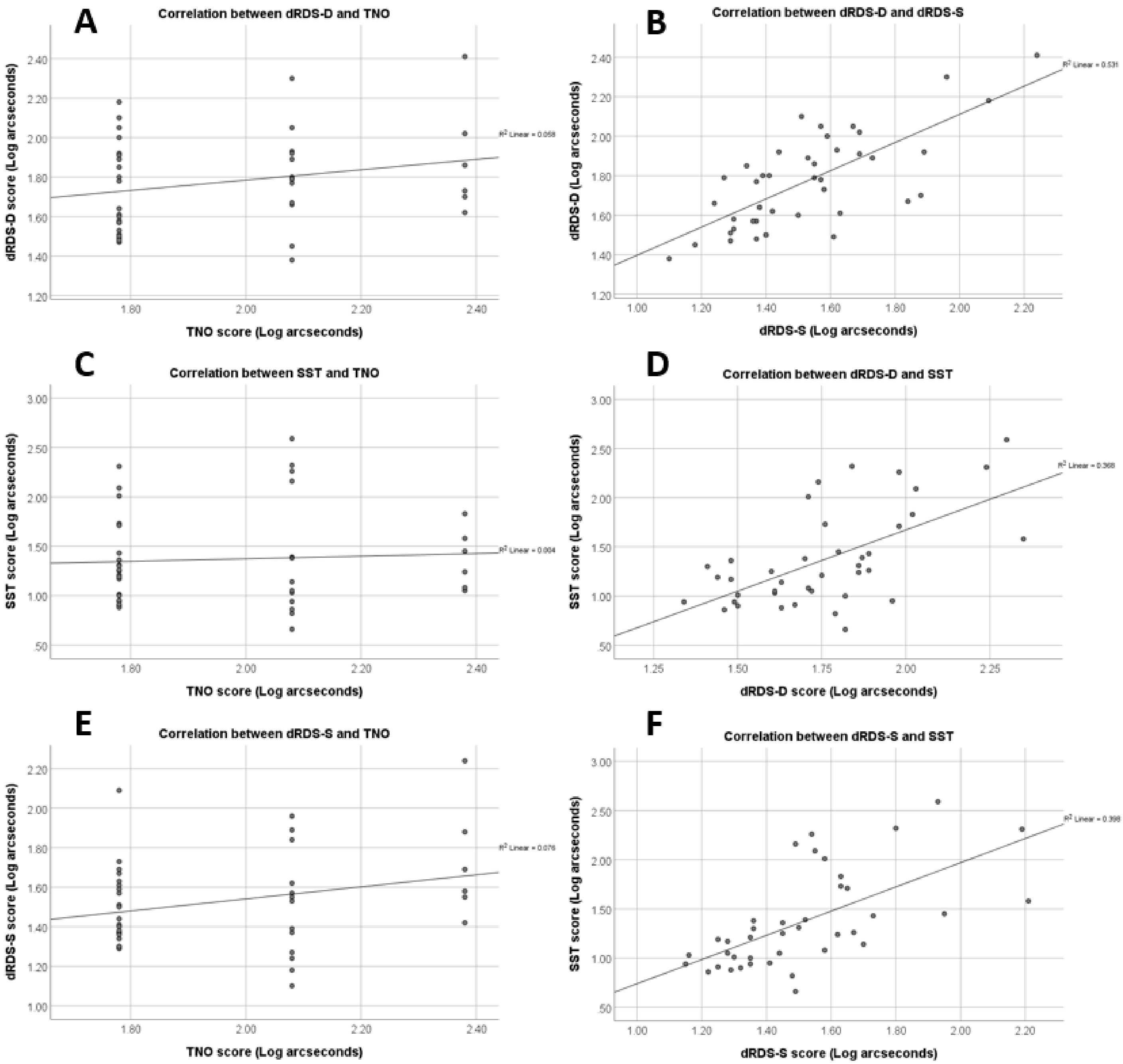

Figure 15. Scatter plots comparing stereo thresholds as measured by different tests. (A) dRDS-D against the TNO, (B) dRDS-D against the dRDS-S, (C) SST against the TNO, (D) dRDS-D against the SST, (E) dRDS-S against the TNO, (F) dRDS-S against the SST.

The relationships between stereo tests were investigated further using Bland-Altman plots. The results of Bland-Altman analysis for stereo acuity scores between the different stereo tests are illustrated in Figure 16. The mean difference between TNO test and dRDS-D is 0.11 log arcseconds (95% CI = ± 0.14), with limits of agreement ± 0.90 log arcseconds (Figure 16-A). The mean difference between the dRDS-S and dRDS-D is 0.24 log arcseconds (95% CI = ± 0.05), with limits of agreement ± 0.35 log arcseconds (Figure 16-B). The mean difference between TNO test and SST is 0.60 log arcseconds (95% CI = ± 0.17), with limits of agreement ± 1.03 log arcseconds (Figure 16-C). The mean difference between dRDS-D and SST is 0.41 log arcseconds (95% CI = ± 0.12), with limits of agreement ± 0.76 log arcseconds (Figure 16-D). The mean difference between TNO test and dRDS-S is 0.43 log arcseconds (95% CI = ± 0.08), with limits of agreement ± 0.55 log arcseconds (Figure 16-E). The mean difference between dRDS-S and SST is 0.16 log arcseconds (95% CI = ± 0.12), with limits of agreement ± 0.74 log arcseconds (Figure 16-F).

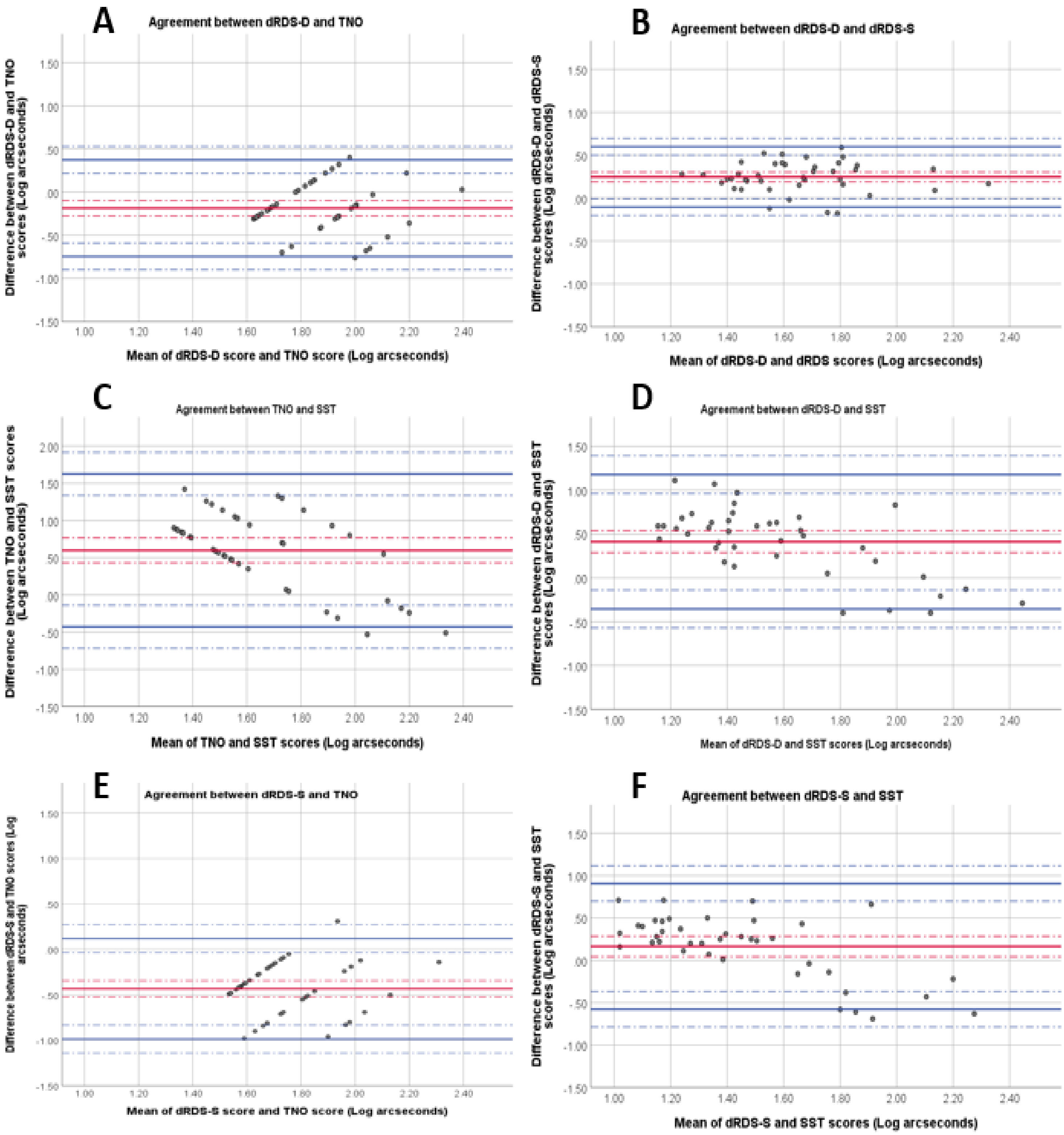

Figure 16. Bland-Altman plots comparing different stereo tests. (A) dRDS-D against the TNO, (B) dRDS-D against the dRDS-S, (C) SST against the TNO, (D) dRDS-D against the SST, (E) dRDS-S against the TNO, (F) dRDS-S against the SST.

2.7. Discussion

With a large floor effect at 60 arcseconds, the six-page paper TNO test was unable to characterise any finer degree of stereo acuity, which, by comparison, was possible using both dRDS-S and dRDS-D tests, as well as the SST. The lowest thresholds (best stereo acuity) were measured using the SST, followed by the static version of the RDS (dRDS-S), then the dynamic RDS version (dRDS-D) and the worst scores were reported using the TNO test. It should be noted that the correct point used by the OBVA laboratory for the digital tests was 70.83% whereas the correct point estimated for the TNO is higher at 85% (Vancleef *et al.,* 2017). This difference may account to a degree for underestimated thresholds measured in the digital tests. Both digital RDS tests were more reliable than the TNO test, as indicated by tighter limits of agreement. The most reliable stereo test was the dynamic RDS version (dRDS-D), which demonstrated the tightest limits of agreement.

As the computer-based tests are more precise and repeatable, they are better able to reliably identify a relationship to operational performance if one exists, as is investigated in subsequent chapters. The granularity in these digital tests is something that has been lacking in previous studies and will enable more appropriate analysis of the effects of S3D symbology and its relation to aircrew selected using current stereo acuity vision standards. Paper-based stereo tests like the TNO are simply unable to be used for correlation analyses with performance data and cannot be used to predict aviator performance. Another benefit of more precise vision screening tests is their ability to detect smaller changes in stereo acuity. The lack of sufficiently accurate and reliable tests creates a significant problem after disease or injury and more precise vision screening tests are better able to identify situations that warrant further investigation at a much earlier stage. Currently, no accurate baseline data exists to either better diagnose disease or injury requiring treatment, or to quantify recovery to support return to flying decisions. For example: traumatic brain injury and dementia are associated with worsening stereo acuity (Miller *et al.*, 1999; Mittenberg, Choi and Apple, 2000), and could potentially be detected earlier and appropriately monitored with a better stereo acuity test.

Each of the computer-based threshold tests described here eliminated limitations identified with the TNO test. Crucially, the random order of stimulus presentation makes it impossible to cheat or memorise answers, reducing the incidence of false positive results. Furthermore, there is no chromatic imbalance using active shutter glasses, and printing or illumination discrepancies are removed using a standardised computer screen. Finally, examiner interference is minimised, removing the possibility of transcription errors or human bias when giving instructions or recording results. The chief disadvantage of computer-based tests is that they require hardware resources, in the form of a computer, 3D monitor, and active shutter glasses to operate. However, given that the estimated cost to fully train a pilot on a 5^{th} generation fast jet aircraft, in which a binocular HMD is critical, is over $10 million (Mattock *et al.*, 2019), the cost of using an operationally relevant stereo test to accurately identify pilot candidates as either medically fit/unfit is miniscule and could significantly reduce the number of pilots unable to complete the intensive training programs, resulting in significant cost savings.

With regard to the significant difference in results between the digital tests, a reason why higher thresholds were recorded using the dynamic version of the dRDS when compared to the static dRDS, could be that the dRDS-D is made harder by forcing participants to recompute disparity every few frames and performance could be affected by differing dot density. Regarding the differences seen between the dRDS tests and SST, our finding that lower scores were achieved with a test using contour based feature mapping (SST), compared to a more global test requiring cross-correlation of random dots, such as with an RDS test, is in keeping with previous studies that an RDS test is less prone to monocular cues (Pageau, Guise and Saint-Amour, 2009; Vancleef *et al.*, 2017; McCaslin *et al.*, 2020). Notably, there was no significant correlation between any of the digital tests and the TNO test. An individual scoring 60 arcseconds on the TNO test could obtain a score ranging anywhere from approximately 30 to 160 arcseconds on the dRDS-S. This suggests that the TNO test does not reliably measure stereo acuity.

There is no clear answer as to which test is the best and most appropriate to use. Such a decision will depend on their ability to predict operational performance, as will be investigated here. However, regardless of whether any relationship exists between stereo acuity and performance using S3D symbology for a warning alert task in an HMD, the greater accuracy and reliability achieved using computer-based threshold tests is beneficial, as has been demonstrated by their ability to address many of the limitations associated with the TNO test. In addition to these benefits, more reliable tests able to measure to threshold support interservice and international co-operation.

Currently, the six-page version of the TNO test is unable to measure stereo acuity to the stereo acuity vision standards required by the USAF (40 arcseconds). As there are pilots from both the RAF and USAF embedded in each other's flying operations, as part of the enduring exchange programs between allied countries, it is important to have tests that are reliable, consistent, and clear. Aligning aeromedical policy and vision standards will further aid interoperability and should be advocated for. Research such as this, aimed at developing vision performance models that predict operational performance, will assist in providing evidence to set vision standards, and drive aeromedical policy which can be shared with allied nations.

2.8. Conclusion

The limitations of the paper TNO test have been clearly highlighted, with computer-based threshold tests offering feasible alternative solutions. All digital tests (dRDS and SST) are able to measure individual stereo acuity to a finer degree than the TNO test, and do so in a manner that reduces examiner interference or bias and eliminates the possibility of cheating (false positives). Of the four stereo acuity tests, the SST produced results with the lowest stereo acuity thresholds. For the two versions of the dRDS tests, the static version (dRDS-S) gave the lowest stereo acuity thresholds but the dynamic version (dRDS-D) was the most reliable with the tightest limits of agreement. While most of the computer-based stereo acuity tests produce significantly different scores, their results are well correlated. However, none of the computer-based tests correlate with the TNO test, which was expected due to its floor effect, that is its inability to measure thresholds better than 60 arcseconds. The greater granularity achieved with digital tests will enable us to investigate the relationship between stereo acuity and operational performance, which in turn will inform stereo acuity vision selection standards and future HMD requirements. Meaningful and reliable stereo acuity tests (and other threshold vision tests) will be increasingly important for military aviators for whom stereoscopic displays and HMDs are becoming more prevalent and critical to flying operations.

3. HMD WARNING ALERT TASK

3.1. Motivation

As noted in Chapter 1, there have been mixed findings on the use of S3D in a display. Based on the use of S3D in other applications, it is anticipated S3D may be beneficial when used in an HMD. However, the degree to which S3D in an HMD could be beneficial may depend on the intended task. With newer computer-based stereo tests able to more reliably measure stereo acuity to threshold, it is now possible to investigate any relationship between individual stereo acuity and effect of using S3D within a display on performance, or effect on cognitive workload. Thus, this research may assist in determining if aeromedical vision requirements, specifically stereo acuity vision standards, are appropriate for current and future HMD technology. This work is of particular relevance with the advent of 6^{th} generation fighter jets which are set to incorporate a raft of visually demanding technologies into the cockpit (Royal Air Force, 2021). Proactively designing these technologies with the human user in mind will undoubtably optimise performance and operational effectiveness.

Browne *et al.* (2017) previously demonstrated a significant benefit (reduced detection and response times) of adding static S3D to a warning alert within an HMD, during a simulated flight task (Browne, Moffitt and Wimsatt, 2017). In addition to this finding, if there is a neural mechanism sensitive to stereoscopic motion-in-depth as proposed in previous research (Regan and Beverley, 1978; Patterson, 1999), adding motion-in-depth S3D to a warning alert could be expected to affect the ability to detect and react to the alert to a different degree. Utilising a similar experimental design as Browne *et al.*

(2017), the aim of this research is to investigate how the addition of S3D, in various forms, to a warning alert task affects performance, cognitive workload, or ocular symptoms, measured subjectively and objectively. Novel to this research, individual stereo acuity will be measured to threshold with three computer-based stereo acuity tests, and for comparison, the standard RAF depth perception test. Only with an accurate assessment of stereopsis can human performance, cognitive workload, or ocular symptoms be investigated in relation to individual stereo acuity.

3.2. Research questions

With regard to the second and third research questions (R2, R3) set out in Section 1.1;

- What effect does the addition of S3D to a warning alert task displayed in an HMD have on performance, ocular symptoms, and cognitive workload?
- How does individual stereo acuity measured with different stereo tests predict performance, ocular symptoms, or cognitive workload in a warning alert task displayed with S3D in an HMD?

3.3. Ethics and recruitment

All studies were approved by the USAF AFRL Institutional Review Board (IRB), protocol number FWR20190044H (Appendix G: Participant information). Participants were recruited from a database of individuals who have taken part in previous vision studies. Volunteers older than 18 years were provided with a written information sheet and gave informed consent specific to the study. As part of the general battery of vision tests performed, all participants were required to provide informed consent, granting permission for their anonymous data to be stored and used within the laboratory. Due to the restrictions imposed on human subject research by the global COVID-19 pandemic, extra precautions were taken. Extra precautions included asking travel, symptom, and possible exposure questions to all participants prior to scheduling them, and requiring signed acknowledgement of the same questions by participants when they arrived at the OBVA Laboratory. A thorough cleaning and sanitising regime was implemented between sessions and participants, face masks were worn throughout, and social distancing guidelines were followed. Appropriate personal protective equipment (mask, gloves, coat) was worn by researchers conducting study sessions, and a detailed contact list was kept of all personnel in the OBVA Laboratory each day. Consequently, in the event of a positive case of COVID, it was possible to notify personnel who had potentially been exposed.

3.4. Flight simulation and HMD apparatus

A wide field of view flight simulator was developed, increasing the simulated flight display horizontal field-of-view from 120° used by Browne *et al.* (2017) to 270°. The flight simulation was created using five 190 x 105 cm flat panel Samsung (Samsung, Seoul, South Korea. Model No: QN85Q70RAFXZA) 4K displays (Figure 17), resulting in a 270° horizontal x 28° vertical field-of-view. The participant was seated 2.5 m from the surface of each screen panel. The “out-the-window” flight environment viewed on the Samsung displays was created using the Diamond Visionics Image Generator (IG) (Diamond Visionics, New York, USA), using five rendering machines. The HMD symbology, was developed within the Unity game engine (Unity Technologies, San

Francisco, USA) and used a further two channels, with separate inputs for the right and left eye required to create a dichoptic image. The computers used to render all seven channels operated with NVIDIA Quadro P6000 graphics cards (NVIDIA Corporation, CA, USA), that together with Quadro Sync boards synchronised the rendering of frames across each of the displays.

Figure 17. Left: Flight simulator using five 85-inch screens displaying an 'out the window' image. Right: SA-62/S HMD (SA Photonics, CA, USA). Used with permission.

Flight symbology was displayed using a SA Photonics SA-62/S HMD (SA Photonics, CA, USA), with 100% binocular overlap and a field-of-view 53° x 33°. The SA-62/S uses full colour organic light emitting diode (OLED) micro displays and a proprietary optical design to project a virtual image while keeping the size and weight of the optics to a minimum. For the purposes of this research, only green monochrome symbology was used, simulating the view through current military HMDs such as the JSF HMS or the JHMCS. The SA-62/S is more capable than other lower cost HMDs, with better pixel resolution and an Ops-Core skull head mount (Gentex, PA, USA) to ensure best fit, minimising any slippage with head movement. In addition, the eye box can be aligned precisely to individual IPD. HMD symbology as illustrated in Figure 18 was overlayed and viewed simultaneously with the out the window image on the Samsung displays.

Figure 18. Symbology displayed in HMD.

When considering any vergence-accommodation (V-A) mismatch the focal distance of the HMD and out-the-window systems were calculated in order to assess the difference between them. The focal distance to the Samsung displays was 2.5 m, or 0.4 D. By contrast the symbology displayed within the HMD was focused to 4 m, or 0.25 D. Thus, the difference between the two systems when viewed together was 0.15 D, which is within the recommended V-A mismatch tolerances of optical discomfort which should be less than 0.5 D (Shibata *et al.*, 2011). Consequently, both the HMD symbology and out-the-window display should be able to be viewed as a single clear image without causing ocular discomfort. Full operating parameters of the SA Photonics SA-62/S HMD are given in Table 6. To illustrate the benefit of using the SA-62/S, in comparison, the

Microsoft HoloLens see-through HMD has a lower resolution of 1280 x 720 pixels with a much smaller diagonal field-of view of 35 mm.

Table 6. Parameters of SA Photonics SA-62/S head mounted display

Parameter	**Value**
Diagonal FOV	62°
Horizontal FOV	53°
Vertical FOV	33°
Binocular overlap	100%
Eye box	10 mm
Eye relief	25 mm
Display resolution	1920 x 1200 pixels
IPD adjustment range	55-75 mm
Binocular weight	570g

3.5. Alignment process

For comfort, the optics were aligned as best as possible. Firstly, the SA Photonics HMD was fitted to the participant's head by the participant themself using the neck, occipital, and top adjustment knobs (Figure 19). For optimum placement of the prisms, adjustments were made fore and aft to offer adequate eye relief, followed by adjusted to the participant's IPD using the IPD adjustment knobs (SA Photonics handbook – Doc #MNL-100000-00). If required, tilt and vertical adjustments could be made and repeated to optimise clarity of the image, dependent on user feedback.

Figure 19. SA Photonics SA-62/S HMD head adjustment knobs. Used with permission

To align the optics a half cross was placed in each channel, which the user moved together, using a designated button on the joystick, to create a binocular image of a whole cross (Figure 20). To verify adequate alignment, a horizontal line was shown in each optic followed by a vertical line in each optic. If the projected line viewed in the HMD lined up with a target line displayed on the Samsung central panel, then the alignment was subjectively deemed to be sufficient. The process was repeated twice to ensure similar adjustments were made. No attempt was made to confirm veridical accuracy of alignment but all subjects self-reported perceptual alignment. Finally, the HMD was paired with a VIVE head tracker linked to three base stations (Model 2.0, HTC Vive, New Taipei City, Taiwan), which were positioned as shown in Figure 21. The efforts to counter any inherent misalignment within the HMD are described in detail in Appendix B.

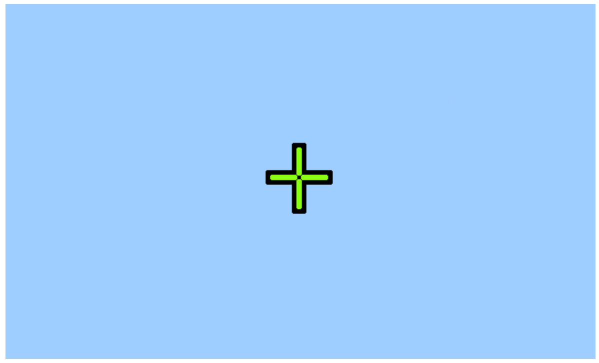

Figure 20. Display target used to align HMD optics

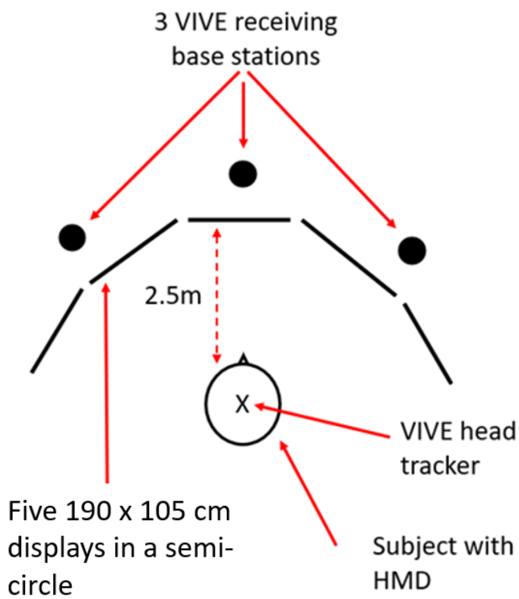

Figure 21. Arial view set up of flight simulator with location of VIVE head trackers (not to scale).

3.6. Flying task

Browne *et al.* (2017) incorporated a distracting flying task for participants in addition to the warning alert task. The format Browne *et al.* (2017) used for a flying task was a ‘highway in the sky’ (HITS) in which participants were instructed to fly through the rectangular boxes set out guiding them along a pathway in the sky. However, in order to avoid mixing the use of monocular depth cues for the HITS and binocular disparity cues in the S3D warning alert, a novel flying and navigation task was developed as the primary distracting task in this research. Flight commands of heading (HDG) and altitude (ALT) were presented in the top right corner of their HMD image (Figure 18), changing every 7 to 15 seconds and indicated by an audio tone. Using a X56 HOTAS controller (Logitech International S.A., Lausanne, Switzerland) with simple forward, back, left, and right inputs, participants were instructed to fly to a command heading and altitude, using the tracking compass indicator at the top centre and altimeter to the middle right of the display for reference (Figure 18). No thrust or power inputs were required, and the simulated speed was set at 150 knots. Heading commands would change ± 50° to 110° relative to the previous command and altitude changed ± 300 to 500 ft. The degree towards which the participant was able to achieve the command heading and altitude was calculated as a proportion from their previous position and used as a measure of their flying performance.

For our research, there were two elements to the study: 1) A flying and navigation task and 2) a warning alert symbol identification and differentiation task. The flying and navigation element of the study was emphasised as the primary task to participants. Each participant underwent a 10-minute training session at the start of their first session and were verbally coached through the flight task. An additional practice session with 20 trials of warning symbols displayed concurrently with the flying task provided extra familiarisation time. Subject performance could be roughly assessed by the researcher with real time data of deviation from instructed heading and altitude point, along with response time and accuracy of warning symbol identification. If a participant was not able to demonstrate adequate performance, extra training and practice sessions could be provided. All heading, altitude and warning symbology were skull fixed, moving with the participants' head direction. By contrast, the artificial horizon information remained fixed to the aircraft frame, visible only when looking forward.

Figure 22. A participant seated in the simulation using joystick in right hand to fly and responding to warning symbols by pulling switches with left hand.

3.7. Warning alert task

While the flying and navigation task was described to participants as their primary objective, it was actually designed to keep them engaged and distracted from the warning alert task. To improve performance assessment metrics used by Browne *et al.* (2017), in addition to a FUEL warning alert, a second similar warning alert symbol reading FLAP was introduced, turning the task from assessing simple response time, to a more complex and realistic, as well as operationally relevant, task requiring the participant to interpret and actively differentiate between two options. A warning alert symbol, reading either FLAP or FUEL, appeared to the middle left of the HMD display, 14.68° away from the command heading instructions (x-axis displacement 12.14° and y-axis 8.25°) (Figure 18). Participants were instructed to discriminate between the two warning symbols as quickly and accurately as possible by pulling one of two switches with their left hand (Figure 23), using a X56 Hands-On-Throttle and Stick (HOTAS) controller (Logitech International S.A., Lausanne, Switzerland). Both warning symbols appeared at 5 -15 second intervals in a pseudo-random manner. There were 60 trials in each condition. Response speed as well as accuracy were measured, enabling a throughput measure to be calculated (accuracy divided by cumulative response time) (Thorne, 2006).

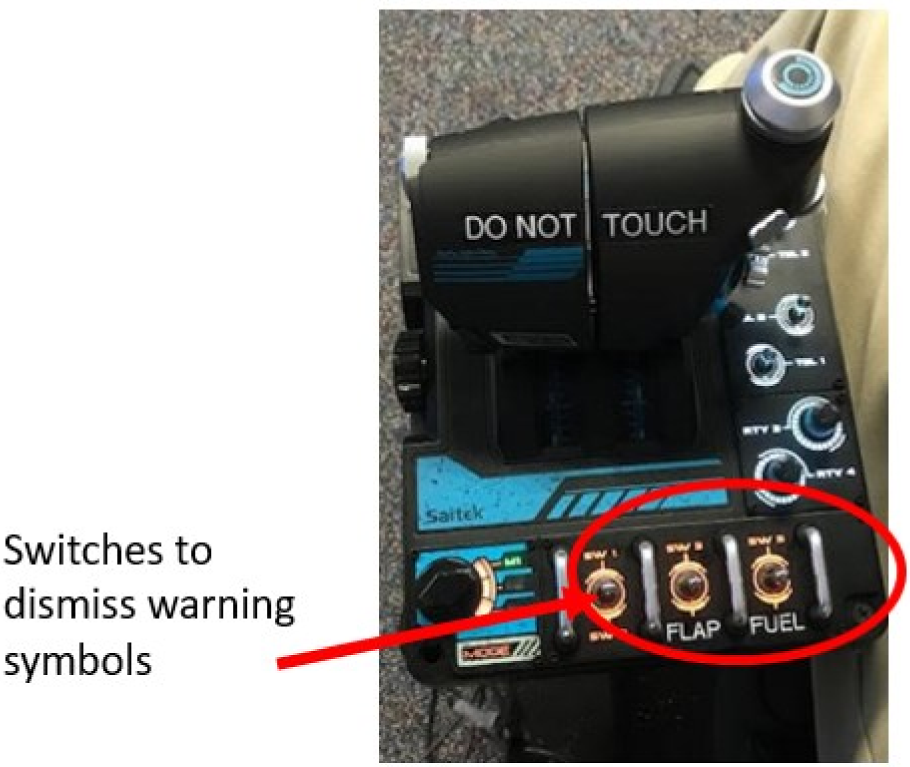

Figure 23. Switches to dismiss FLAP and FUEL warning symbols.

The FLAP symbol was always presented at the fixation plane (zero disparity) and flashed on and off at 1 Hz. The FUEL warning symbol was displayed with different disparities and presentation conditions summarised in Table 7. In the control 2D condition the FUEL symbol flashed with zero disparity at 1 Hz, in the same way as the FLAP symbol. In this condition the only way to differentiate symbols was to read the letters.

In their discussion Browne *et al.* (2017) suggested an explanation for the benefit observed when a warning alert was displayed in S3D; that the symbol was simply visible for 100% of the time, enabling a quicker response. In order to further investigate the idea that continuously being visible reduces reaction time rather than the S3D aspect itself, a condition was added in which the warning symbol flashed on and off, as with the 2D control condition, but with the ‘on’ cycle being in S3D. This condition was referred to as the intermittent condition and allowed S3D to be evaluated independently to the effect of being continuously visible. For the other two experimental conditions, the FUEL warning symbol was flashed in and out of depth with zero disparity and with 0.69° (2500 arcseconds) disparity. The FUEL alert was flashed either statically between the two disparities in the persistent condition or oscillated sinusoidally between the disparities in the motion-in-depth condition. The order conditions were performed was randomised, using a Fisher-Yates shuffle algorithm and Unity software (Unity Technologies, San Francisco, USA). The study was split into two separate sessions with the full study protocol illustrated in Figure 24. A pilot study was conducted with six participants to test procedures and equipment prior to this study; as detailed in Appendix A.

Table 7. Experimental conditions of final study protocol.

FLAP symbol presentation	**FUEL** symbol presentation
2D: Flashed on/off at 1Hz	**Control**: Flashed on/off in 2D at 1 Hz
	Intermittent: Flashed on/off with 2500 arcseconds of disparity at 1 Hz
	Persistent: Flashed in/out of depth, 2D to 2500 arcseconds of disparity at 1 Hz
	Motion-in-depth: Moving in/out of depth smoothly to 2500 arcseconds of disparity at 1 Hz

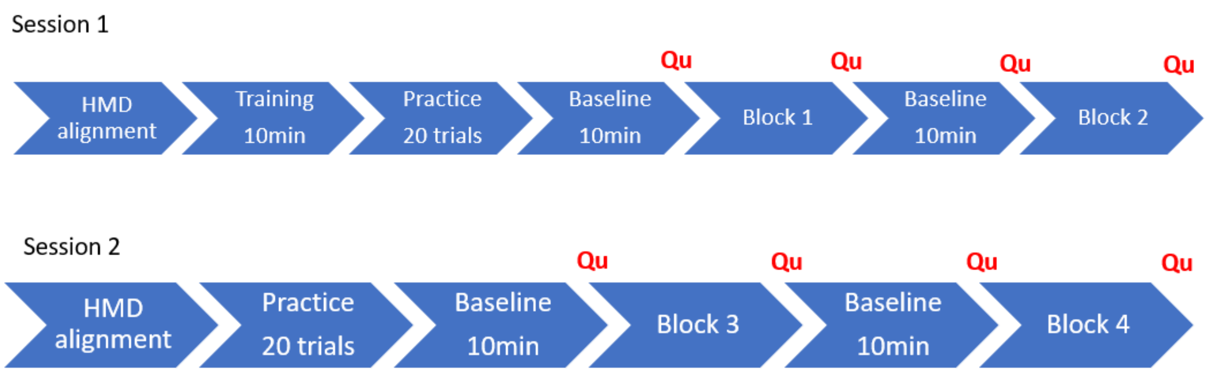

Figure 24. Final study protocol

3.8. Apparatus for physiological variables

Peripheral temperature and heart rate were recorded using a skin temperature probe (ADInstruments, 2019, Dunedin, NZ) taped to the back of the middle finger on the left hand and a 3-lead ECG (Figure 25 and Figure 26). Both dependent variables were used as an objective way to estimate cognitive workload (Gavrilescu *et al.*, 2019). Measurements were sampled at 100 Hz with PowerLab v8 and accompanying LabChart software (ADInstruments, 2019, Dunedin, NZ). Heart rate and heart rate variability were calculated from the ECG trace by computing beat-to-beat intervals (excluding ectopic beats) using the LabChart software. Normal beats were defined as those with an R-R interval of 400-1200 ms and with a complexity less than 1.8. Complexity values denote how much the detected signal beats conform to the shape and duration of the QRS complex calculated from the activity (variance of the signal segment) and mobility of the signal (ADInstruments, 2019, Dunedin, NZ). The analysis used to determine heart rate variability (HRV) was the Root Mean Square of the Standard Deviation (RMSSD) of eligible beats as an indicator of parasympathetic activity and vagal tone. A relative decrease in HRV, indicated by lower RMSSD, indicates lower parasympathetic activation and decreased vagal tone which suggest a physical or cognitive stressor is present (Laborde, Mosley and Thayer, 2017). An inter-block rest period was enforced to minimise transfer of effects from one condition to another. This also provided the time needed to gather baseline physiological data. Within-subject baseline measurements were taken over a 10-minute rest period prior to each experimental block. Data from the first five minutes of the rest period were discarded and a mean measured over five minutes in the second half was used as the baseline. Once the experiment had started, average measurements were taken every five minutes to assess changes from baseline. To minimise potential confounding effects of circadian rhythm on individual physiological variables, the second session was scheduled for a similar time of day to the first (Laborde, Mosley and Thayer, 2017).

Figure 25. Top: Temperature sensor. Bottom: 3-lead ECG sensor.

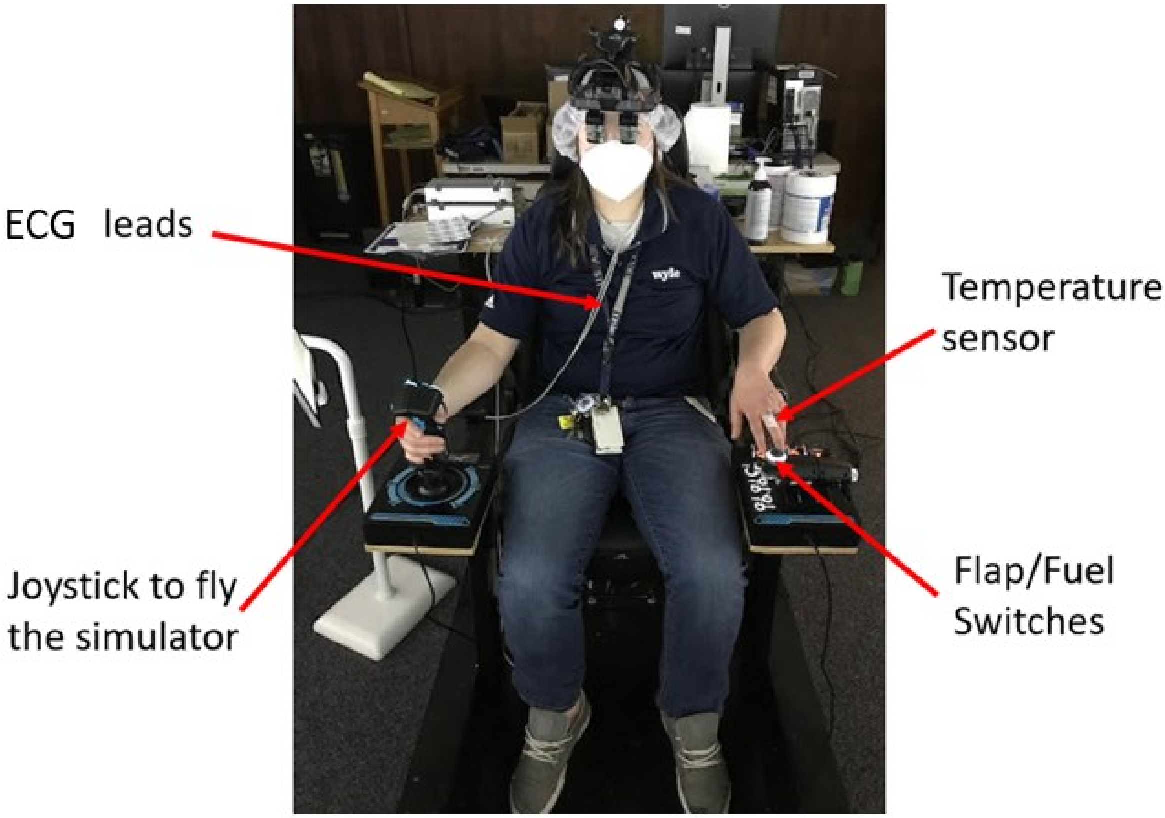

Figure 26. A participant sat with temperature and ECG sensors attached.

3.9. Questionnaire

To assess self-reported ocular fatigue symptoms and perceived cognitive workload, participants completed a questionnaire, presented to them on a Microsoft Surface Pro tablet (Microsoft, Redmond, WA). This method simplified data collection (Figure 27), and removed any interaction with the administrator in order to minimise potential bias. Five questions addressed visual symptoms (Shibata *et al.*, 2011), and were asked before and after each experimental condition. In addition, six workload-based questions were asked at the end of each condition, based on the validated NASA TLX (NASA, DC, USA) (Hart, 2006). Participants responded using a 0-10 scale, with lower scores representing less fatigue and workload, while higher scores represented increased fatigue and workload (Table 8).

Table 8. Symptom and workload questions.

Symptom questions	1. How tired do your eyes feel? *Not at all ↔ Very tired* 2. How clear is your vision? *Completely clear ↔ Very blurry* 3. How often do/did you experience double vision/images? *Not at all ↔ constantly/all the time* 4. How does your head feel? *No pain/discomfort ↔ Severe headache* 5. How does your neck and back feel? *No pain/discomfort ↔ Severe ache*
Workload questions	1. How mentally demanding did you find the task? *Easy ↔ very demanding* 2. How physically demanding did you find the task? *Easy ↔ very demanding* 3. How hurried or rushed was the pace of the task? *No rush/pressure ↔ Very rushed/frantic* 4. How successful were you in accomplishing the task? *Perfect ↔ failure* 5. How hard did you work to accomplish your performance? *No effort ↔ very hard* 6. How irritated/stressed/annoyed were you? *Content/no stress ↔ extremely stressed/annoyed*

Figure 27. A participant entering their responses to the questionnaire presented on a computer tablet.

4. HMD WARNING ALERT TASK RESULTS

4.1. Demographics

The same participants used for the stereo test study also volunteered for the warning task HMD experiments. Forty-five participants were recruited from the OBVA subject database. However, one participant did not complete both sessions and another participant's data file was corrupted; n = 43 (26 male, mean age 37.7 years ± 9.87, range 21-70).

4.2. Replicating Browne et al. (2017) study

Firstly, as this study was designed based on aspects of Browne *et al.* (2017), it was anticipated that it would be possible to replicate their findings. Browne *et al.* (2017) demonstrated a performance benefit when viewing a warning alert presented with binocular disparity (i.e., in S3D) using an HMD. Browne *et al.* (2017) screened their participants to ensure they had a stereo acuity of better than 100 arcseconds, using a Random Dot stereo acuity test booklet with polarised glasses (n=21). To replicate this sample, 36 participants with a stereo acuity better than 100 arcseconds as measured using the dRDS-D were identified in this study. Browne *et al.* (2017) measured simple response time to a single FUEL warning alert flashed in 2D and also in three other conditions with various levels of S3D, the largest disparity being 170 mm of depth. In this study, the persistent condition (i.e., flashing in and out of S3D while being continuously on) is the condition most similar to that used by Browne *et al*. (2017). With the selected subset of participants (i.e., those with stereo acuity of better than 100 arcseconds), the response times to the FUEL warning alert in the persistent condition were compared to response times in the control 2D condition, using a paired t-test. No significant difference in performance was found, (t (1,35) = 0.824, p = 0.415). This is in contrast to Browne *et al.* (2017) who found a significant difference between performance in their S3D 170 mm condition compared to their 2D control condition, using pairwise comparisons (t (1,21) = 1.88, p<0.05). Thus, Browne *et. al.'s* (2017) finding showing a reduction in response times to a warning alert symbol when displayed in S3D was not replicated in this study. A potential reason for being unable to replicate Browne *et al.'s* (2017) results, could be the difference in methodology; in that for my research the task measured choice reaction, with two warning alerts to differentiate between (FLAP and FUEL), whereas only one FUEL warning alert was used by Browne et. al (2017) in a simple reaction task. In an operational setting, there are a number of different alerts used in the cockpit that the pilot must be able to identify and differentiate between, thus the warning alert developed in the research conducted here is more realistic of the flying environment encountered by military aircrew.

4.3. Signal Detection Theory: D' and Criterion level

When assessing performance in the warning alert task, rather than simply looking at response time to the FUEL symbol, the subject's behaviour and the types of errors being made were also considered. To address this, results were analysed using a Signal Detection Theory (SDT) framework (Green and Swets, 1966). Responses were sorted into true positives (pressing FUEL when symbol reads FUEL) and false positives (pressing FUEL when the symbol reads FLAP) to calculate an observer's d' and criterion (X_c). The theory applied to this given example is illustrated in Figure 28. The blue line indicates when the FUEL signal was present and the orange line when FLAP signal was displayed. The red line is the criterion (X_c). To the right of the red line are the incidents when the FUEL switch was pulled and to the left of the line when the FLAP switch was pulled. If a participant is risk averse and pulls the FUEL switch every time regardless of whether the alert reads FLAP or FUEL in order not to miss a signal, the criterion (X_c) line will shift to the left. Consequently, there will be no false negatives but many false positive results. If only response time to FUEL is assessed in this situation the participant would score 100% for accuracy, despite getting 50% of overall responses wrong, and such a result would not be a true indication of their ability to correctly differentiate stimuli.

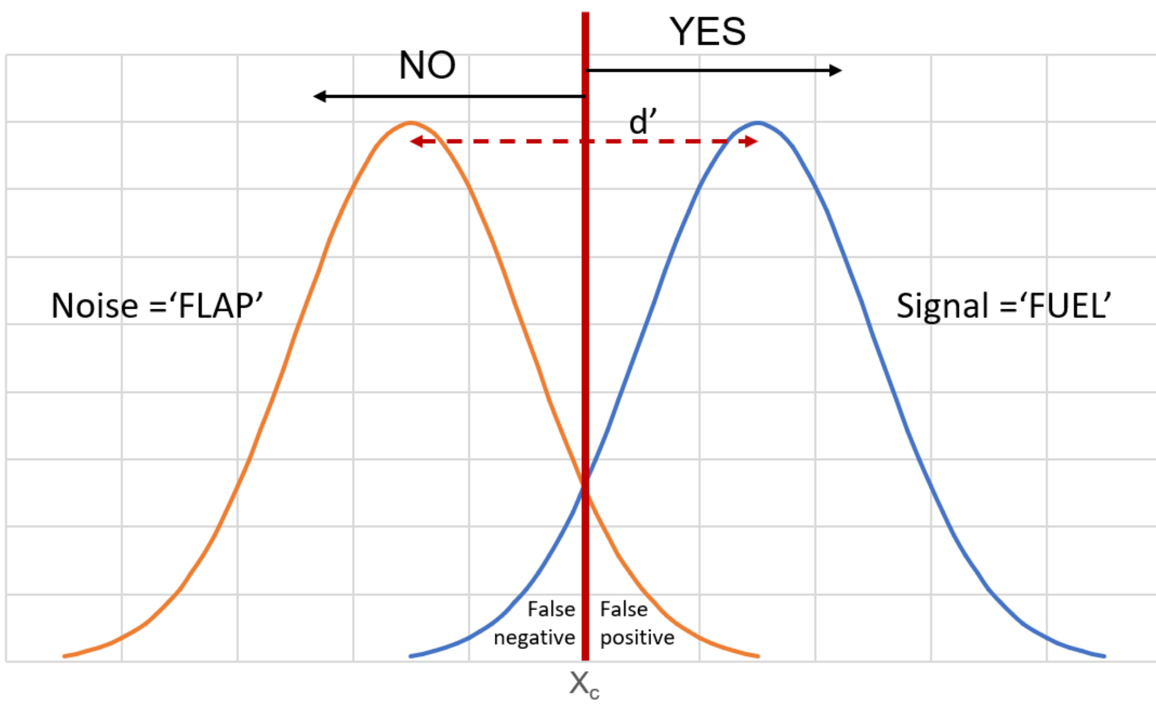

Figure 28. Graphical representation of signal detection theory. Xc = Criterion level, d’ = separation between two distributions, or sensitivity. Adapted from (Wickens *et al.*, 2013)

As part of SDT, d' is a measure of ability to discriminate between two signals, that is to say the separation of the population means of two response variables (e.g., target present and target absent) in units of pooled standard deviation. A larger d’ indicates greater separation between the mean response variables of two signals and a stronger signal. Less overlap between response variables, that is to say the larger d’ is, then the easier it is to discriminate between two stimuli, indicating greater sensitivity of a test. d’ is calculated as;

$$= z(H) - z(F)$$

Where z is the inverse of the normal distribution for the specified mean and standard deviation, H is the proportion of true positives, or ‘hits’ and F is the proportion of false positives, or incorrect responses. When d’ is interpreted with the criterion level, the sensitivity and specificity of detecting a signal can be determined. When normalised, the optimum criterion level is calculated and should be zero. If the criterion level moves to the left and becomes negative, sensitivity will increase but responses will become less specific, the opposite is true if the criterion level is more positive, with signal detection becoming less sensitive but more specific. If an observer deviates from the optimum criterion, it is called bias.

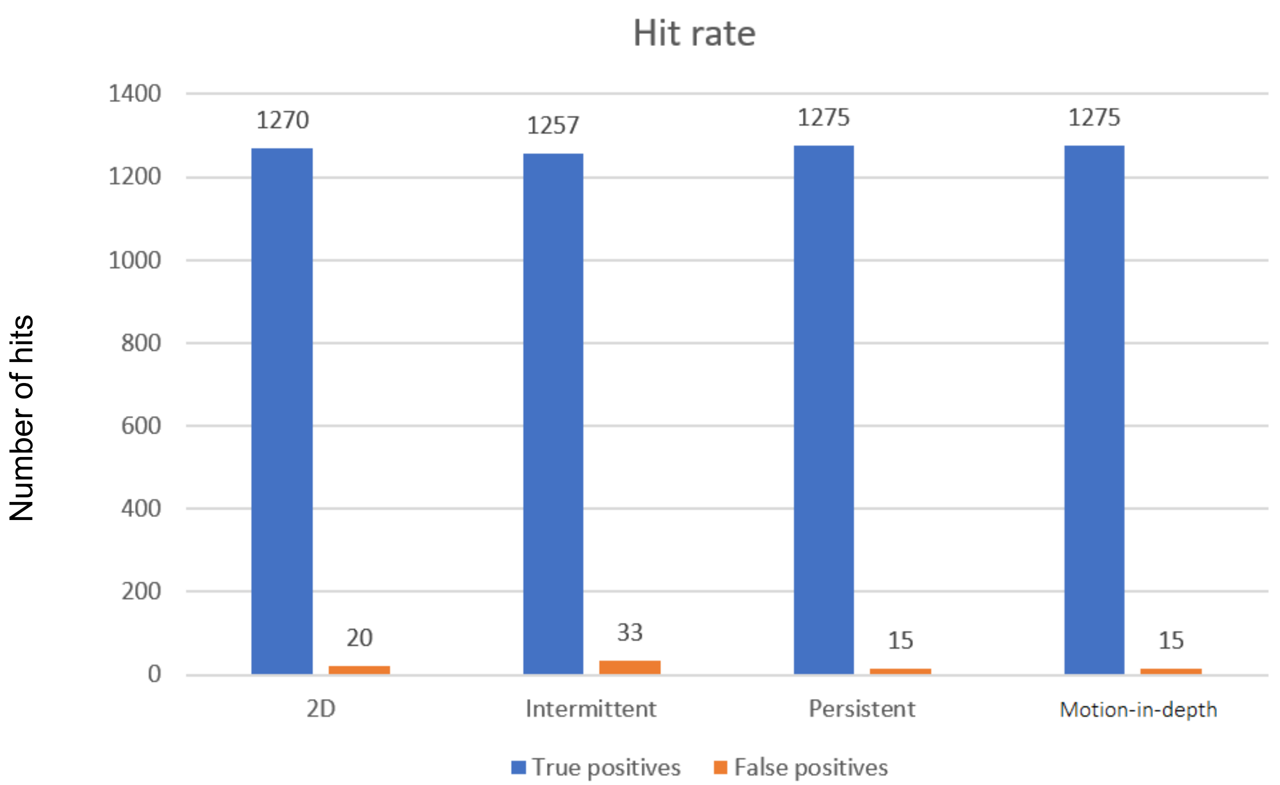

Figure 29. True positive and false positive hit rate for each condition.

As can be seen in Figure 29, the greatest number of false positive mistakes were made in the intermittent condition, with the fewest in the motion-in-depth and persistent conditions. To assess whether these differences were significant, bootstrapping was performed over 10,000 iterations, with the means and 95% confidence interval given in Table 9. There was no difference in d’ between conditions using one-way ANOVA analyses.

Table 9. Bootstrapped mean d' values for the four conditions

	2D	INTERMITTENT	PERSISTENT	MOTION
Mean d'	4.37	4.21	4.38	4.26
2.5 % percentile	4.14	3.98	4.15	4.03
97.5 % percentile	4.50	4.34	4.52	4.40

4.4. Response time

Response times that were four times longer than the mean of all response times for each alert were removed from the data set (n = 42 for FUEL and n = 47 for FLAP). Outlier data points were replaced with a mean response time specific to the participant and condition (in order to maintain a balanced data set for analysis). Mean response times to all correctly answered FUEL warning symbols are given in Table 10 and illustrated in Figure 30. No main effect between response times across conditions was identified using Friedman analysis, Chi-Square (3, 43) = 4.48, $p = 0.21$.

Table 10. Mean response times to correctly identify the FUEL warning alert.

	2D	INTERMITTENT	PERSISTENT	MOTION
Mean (seconds)	1.09	1.14	1.07	1.06
Standard deviation	± 0.42	± 0.45	± 0.36	± 0.36

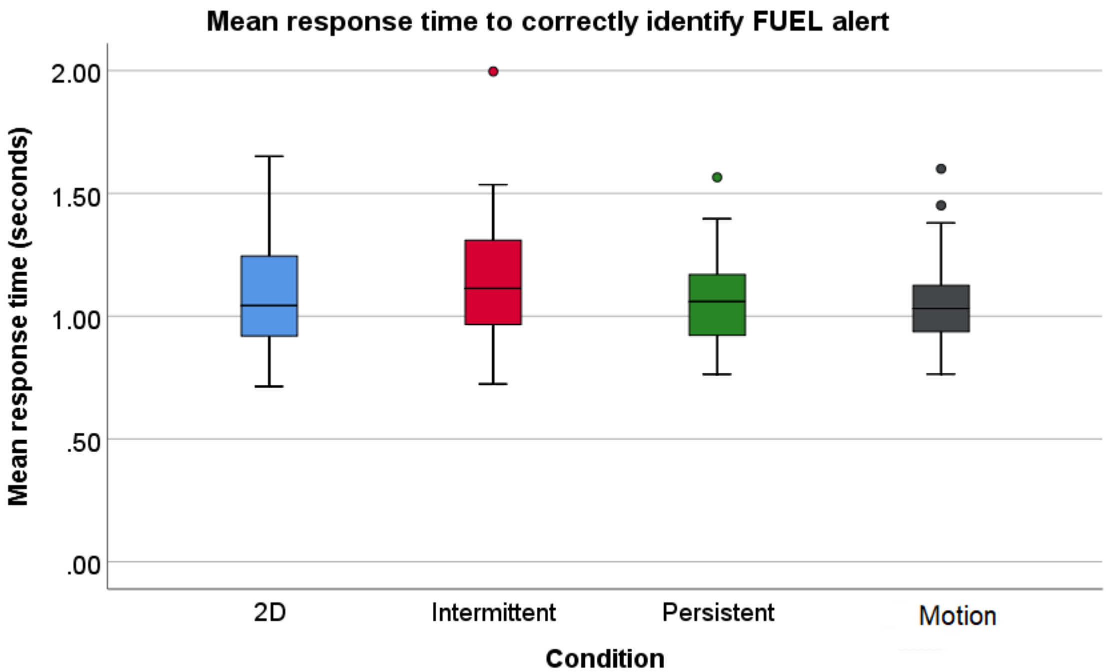

Figure 30.A box and whisker plot of response time to correctly identify the FUEL warning alert The line within each box is the median value, the box itself denotes the interquartile range and vertical bars the extreme range of values within the group.

To contrast, response times to the FLAP warning alert were also evaluated. All FLAP warning alerts were presented in 2D across all conditions, so no difference in response times was expected. Mean response times to all correctly answered FLAP warning symbols are given in Table 11. Again, no significant main effect between response times across conditions was identified using Friedman’s analysis, Chi-squared (3,43) = 5.09, p = 0.16.

Table 11. Mean response times to correctly identify the FLAP warning symbol. In each condition the FLAP alert is presented in 2D.

	2D	INTERMITTENT	PERSISTENT	MOTION
Mean (seconds)	1.15	1.16	1.13	1.17
Standard deviation	± 0.43	± 0.43	± 0.40	± 0.42

4.5. Throughput

Throughput was calculated as accuracy (percentage correctly identified) divided by mean response time (Thorne, 2006). Mean throughputs for each condition are given in Table 12 and illustrated in Figure 31.

Table 12. Mean throughput to correctly identify the FUEL warning alert in all four conditions

	2D	INTERMITTENT	PERSISTENT	MOTION
Mean	0.92	0.87	0.93	0.94
Standard deviation	± 0.19	± 0.19	± 0.16	± 0.16

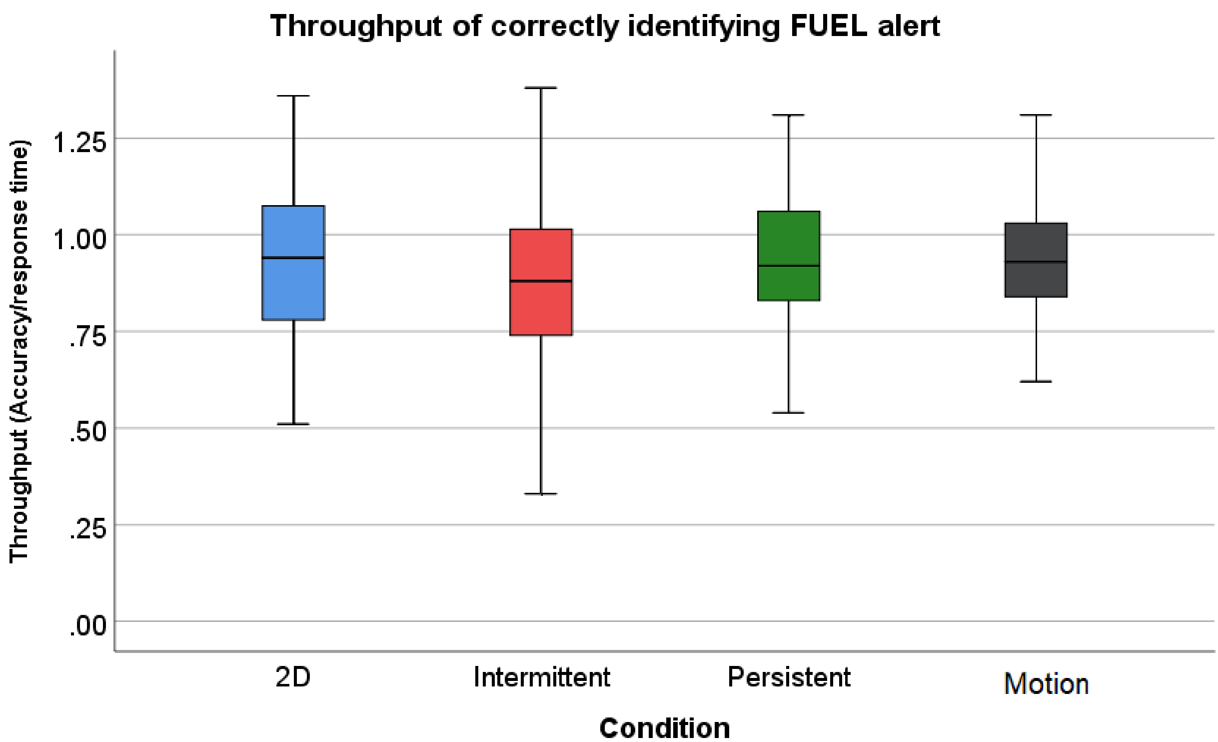

Figure 31. A box and whisker plot of throughput in identifying the FUEL warning symbol. The line within each box is the median value, the box itself denotes the interquartile range and vertical bars the extreme range of values within the group.

A significant main effect was identified using Friedman analysis, Chi Squared (3, 43) = 10.22, p = 0.017. Post hoc Wilcoxon signed ranks tests were performed with Bonferroni correction for 3-way comparison (p = 0.012) and results detailed in Table 13. There is no significant improvement in throughput between any of the experimental S3D conditions compared to the 2D control condition. The significant difference was between non-comparable conditions.

*Table 13. Wilcoxon signed ranks test comparing throughput of FUEL warning alerts between conditions. With Bonferroni correction, * denotes significance with an alpha value of <0.012*

	z score	Significance
2D vs. Intermittent	-2.04	$p = 0.04$
2D vs. Persistent	-0.37	$p = 0.71$
2D vs. Motion	-1.12	$p = 0.26$
Int vs. Persistent	-2.60	**$p < 0.01$**
Intermittent vs. Motion	-2.49	**$p = 0.01$**
Persistent vs. Motion	-0.42	$p = 0.67$

4.6. Flying performance

Flying performance was assessed by calculating the root mean square error (RMSE) of the heading and altitude flown compared to the command heading and altitude given. Both RMSEs were summed together to give a single error value per participant (Err_k) for each condition which was used as an indicator of flying performance.

$$\theta_{RMSE} = \sqrt{\frac{1}{N_{ppt}} \sum_{k=1}^{N_{ppt}} \frac{1}{n_k} \sum_{j=1}^{n_k} \left(\theta_{k,j} - \theta_{k,j}'\right)^2}$$

Where $\theta_{k,j}$ is the heading command given to participant k on trial j, and $\theta_{k,j}'$ is the mean heading achieved by participant k on trial j. $Z_{k,j}$ and $Z_{k,j}'$ below are the same but for the altitude command. N_{ppt} is the total number of participants and n_k is the number of trials performed by participant k.

$$Err_k = \sqrt{\frac{1}{2n_k}\sum_{j=1}^{n_k}\left[\frac{\left(\theta_{k,j} - \theta_{k,j}'\right)^2}{\theta_{RMSE}^2} + \frac{\left(z_{k,j} - z_{k,j}'\right)^2}{z_{RMSE}^2}\right]}$$

This way there is equal weight to both tasks and the different units of heading and altitude are accounted for. The closer the error value is to zero indicates the closer a participant was to the two-dimensional flight command target. The mean calculated flying performance values from all participants across the four conditions are given in Table 14, with the intermittent condition demonstrating the best overall flying performance, and worst in the motion-in-depth condition. However, no significant differences in flying performance between the four conditions were found with a one-way ANOVA, $F (3, 168) = 0.54$, $p = 0.65$.

Table 14. Mean flying performance metrics for the four conditions

	2D	INTERMITTENT	PERSISTENT	MOTION
Mean	1405.06	1204.75	1223.20	1492.83
Median	721.39	791.00	764.36	1057.75
Standard deviation	± 1379.68	± 1119.90	± 1230.92	± 1266.94

Seven participants identified as pilots, defined as having flown an aircraft solo previously. There were no significant differences in flying performance between the group of seven pilots when compared with flying-naïve participants, in any of the conditions, as seen in Table 15. A Welch's independent t-test was used to account for the uneven sample sizes between the pilot and non-pilot groups. As there were no significant differences in flying performance between the two groups, the group of pilots were treated no differently to the flying naïve group and all participants were analysed together in one group.

Table 15. Mean flying performance scores, with standard deviation in brackets. Independent sample t-tests (Welch's t-test) used to compare the pilot and non-pilot groups.

	2D	INTERMITTENT	PERSISTENT	MOTION
Non-pilots (n=37)	1364.20 (1422.22)	1176.88 (1178.20)	1168.19 (1206.08)	1475.84 (1257.63)
Pilots (n=7)	1615.19 (1209.96)	1348.08 (808.08)	1506.13 (1417.44)	1580.19 (1413.92)
Independent t-test	t (9.54) = 0.49, p = 0.64	t (11.64) = 0.47, p = 0.65	t (7.78) = 0.59, p = 0.57	t (7.96) = 0.18, p = 0.86

4.7. Throughput vs. Flying performance

Flying performance was plotted against performance in the warning symbol task, as calculated by throughput (Figure 32). There was a significant correlation between these two variables in all conditions (Table 16). That is to say, those with the best performance in the primary flying task (smaller RMSE) also demonstrated the best performance in secondary task (greater throughput), correctly identifying the FUEL warning alert.

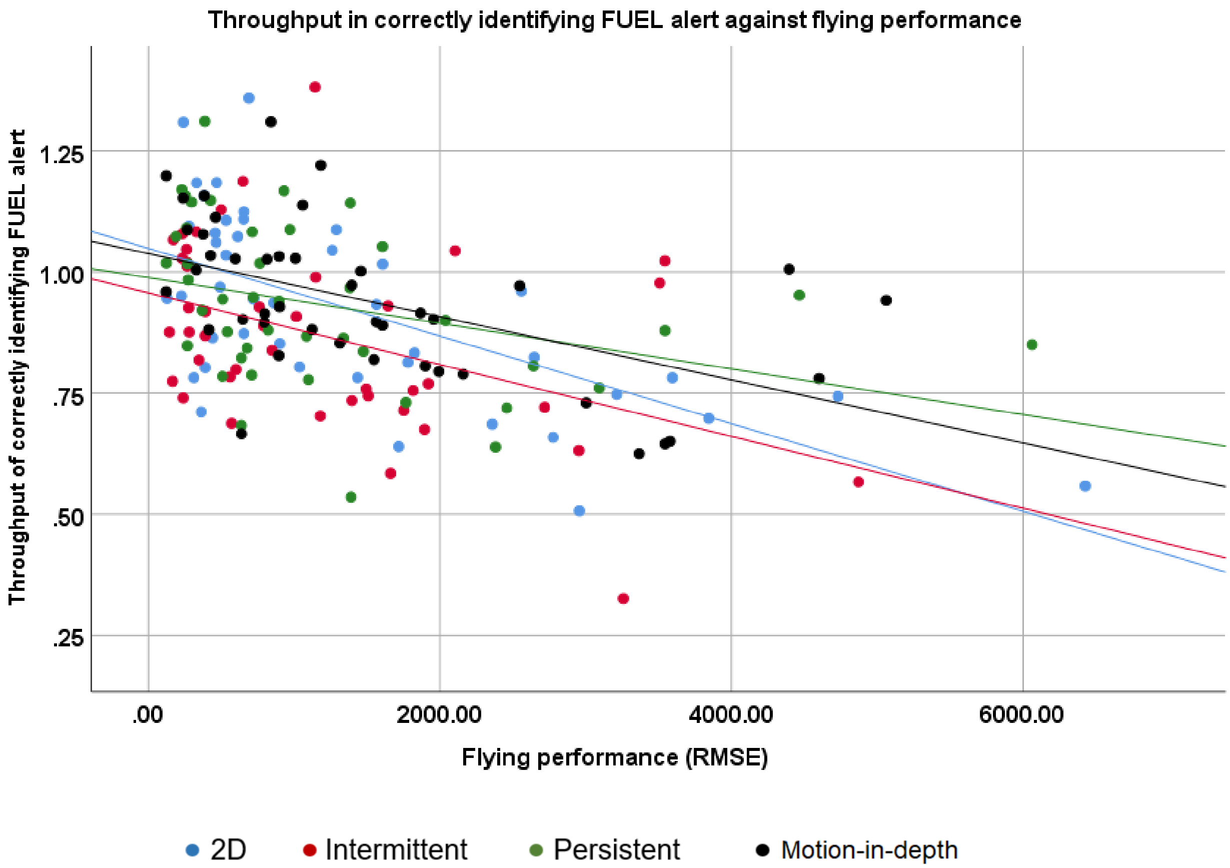

Figure 32. Relationship between performance in primary task (flying) and performance in the secondary task (throughput correctly identifying FUEL warning symbol).

*Table 16. Pearson's correlation analyses between primary and secondary task performance. * denotes significance with an alpha value of <0.05 and ** indicates <0.01*

	Correlation	Sig (2-tail)
2D	-0.64	**p < 0.00** **
Intermittent	-0.43	**p < 0.01** **
Persistent	-0.35	**p = 0.02** *
Motion	-0.52	**p < 0.01** **

4.8. Workload parameters – Questionnaire

Each participant was asked five questions relating to ocular fatigue before and after each of the four experimental conditions. After each experimental condition an additional six questions were asked, based on the NASA TLX questionnaire, to assess perceived workload. To assess for any difference in responses between conditions, a Friedman test (Friedman, 1937) was performed for each question with the results summarised in Table 17.

Table 17. Freidman test for symptom questions (SQ1-5) and workload questions (WL1-6)

Question	Chi squared X^2 (3)	Sig. (2-way)
How tired are your eyes?	3.70	p = 0.29
How clear is your vision?	10.04	**p < 0.02***
How often do/did you experience double vision?	0.81	p = 0.85
How does your head feel?	1.21	p = 0.75
How does your neck and back feel?	2.72	p = 0.44
How mentally demanding did you find the task?	2.70	p = 0.44
How physically demanding did you find the task?	1.90	p = 0.59
How hurried or rushed was the pace of the task?	1.96	p = 0.58
How successful were you in accomplishing the task?	0.87	p = 0.83
How hard did you have to work to accomplish your performance?	0.47	p = 0.92
How irritated/stressed/annoyed were you?	6.20	p = 0.10

For only one question was a significant difference between conditions identified; ‘How clear is your vision?’ A post hoc Wilcoxon sign-rank test was performed which identified the significant relationship being between intermittent and motion-in-depth conditions, $Z = -2.36$, $p = 0.018$. However, when applying a Bonferroni correction ($\alpha/3$ (df) = 0.017) this difference failed to meet significance. Furthermore, this difference was of no scientific interest as comparison of the three experimental conditions with the 2D control condition was of most importance. Changes in individual symptom scores after each condition from baseline answers and perceived cognitive workload scores are illustrated in Appendix D.

4.9. Workload parameters – Physiological

Baseline physiological measurements were recorded over 10 minutes with participants seated and at rest. Data from the first 5 minutes of the baseline period were discarded and a mean calculated over the last 5 minutes. The active experimental block was split into three 5-minute segments with mean measurements calculated for peripheral skin temperature, heart rate, and heart rate variability. For heart rate and heart rate variability, the LabChart HRV module software was used to detect valid R-R intervals (RR interval 400-1200ms, complexity <1.8).

4.9.1. Peripheral skin temperature

Changes in peripheral skin temperature across each of the four conditions are illustrated in Figure 33. In total, peripheral skin temperature decreased from individual baselines by a mean value of 2.76% in the 2D condition, by 1.65% in intermittent, by 2.12% in persistent and by 2.25% in the motion-in-depth condition.

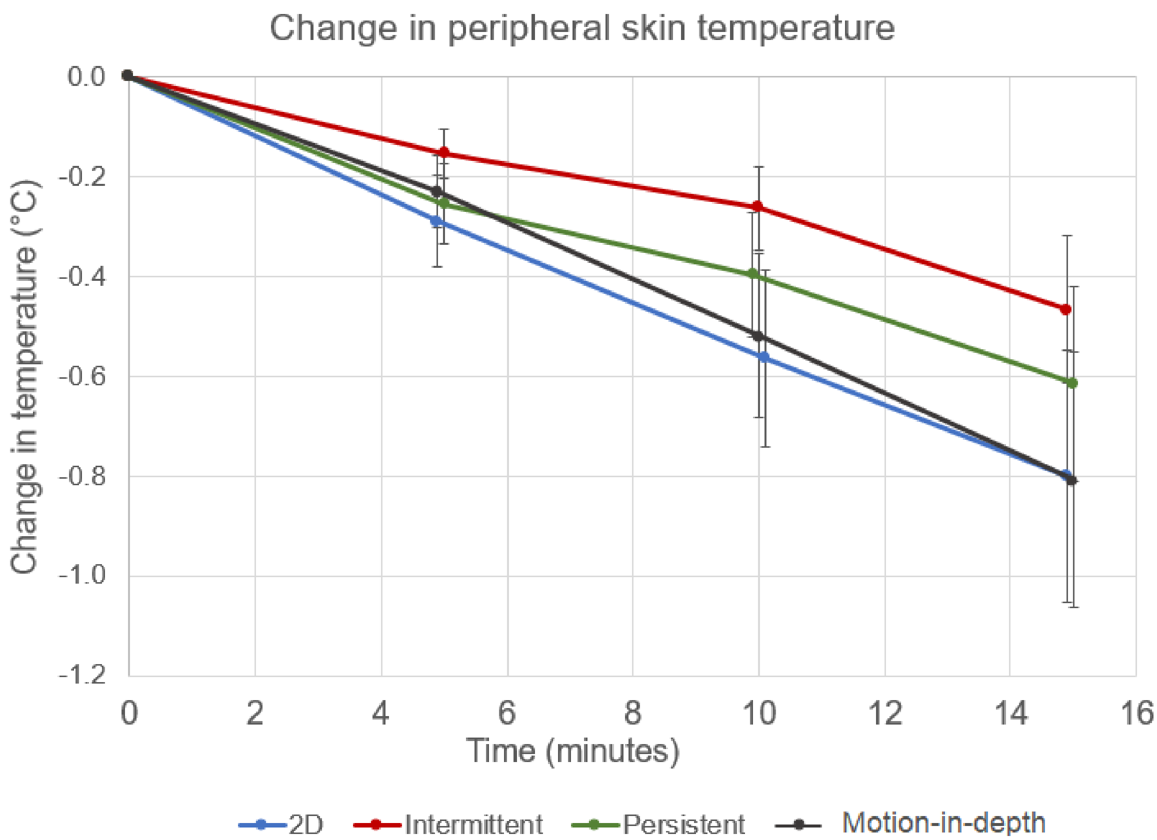

Figure 33. Change in peripheral skin temperature of all participants over time from baseline in all four conditions.

To assess differences within each condition over time, a repeated measures ANOVA with Greenhouse-Geisser correction was performed between data times points (Greenhouse and Geisser, 1959) (Table 18). If significance was found with an alpha level of <0.05, pairwise comparisons were then performed in comparison to the baseline means (Table 19).

*Table 18. Repeated measures ANOVA of mean peripheral skin temperature in three segments and to baseline value in each condition. * denotes significance with an alpha value 0.05 ** denotes significance with alpha value of 0.01.*

		Sig.
2D	F (1.49, 62.66) = 18.28	**p < 0.01****
Intermittent	F (1.68, 70.65) = 5.77	**p < 0.01****
Persistent	F (1.52, 63.80) = 13.91	**p < 0.01****
Motion	F (1.67, 70.02) = 13.84	**p < 0.01****

In each of the four conditions, peripheral skin temperature decreased significantly from baseline over time. In all of the conditions this was a highly significant decrease by the third segment.

*Table 19. Post hoc paired t-test between mean peripheral skin temperature of time segments and baseline values within a condition. * denotes significance with an alpha value 0.05 ** denotes significance with alpha value of 0.01.*

	5 min	10 min	15 min
2D	**p < 0.01****	**p < 0.01****	**p < 0.01****
Intermittent	p = 0.87	p = 0.30	**p = 0.02***
Persistent	**p = 0.03***	**p = 0.01***	**p < 0.01****
Motion	p = 0.43	**p < 0.01****	**p < 0.01****

When comparing baseline means with those in each segment, there were significant decreases in peripheral skin temperature in all conditions, when analysed with paired t-tests (Table 19). To compare differences between conditions, cumulative changes in temperature were calculated, summing the value of all mean changes in every time segment. Using paired t-tests, no significant differences were found between conditions when compared with the baseline 2D condition (Table 20).

*Table 20. Paired t-test between sum change of peripheral temperature in each condition compared to sum change in 2D condition. * denotes significance with an alpha value 0.05 ** denotes significance with alpha value of 0.01.*

	t value	Sig (2-way)
2D vs. Intermittent	-1.78	p = 0.08
2D vs. Persistent	-0.91	p = 0.37
2D vs. Motion	-0.20	p = 0.84

4.9.2. Heart rate

Changes in heart rate experienced by all participants are illustrated in Figure 34. Changes in mean heart rate within a condition were assessed using a repeated measures ANOVA with Greenhouse-Geisser correction. Heart rate reduced significantly from baseline in both the 2D and persistent conditions (Table 21).

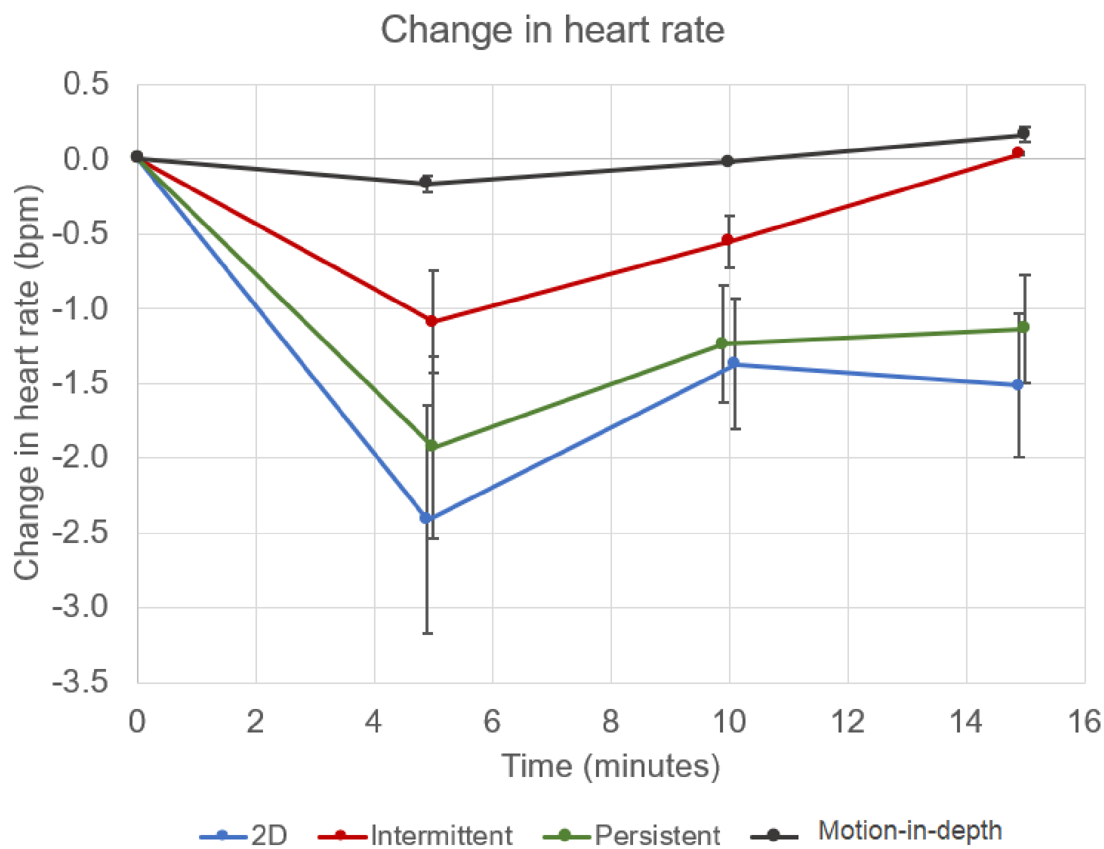

Figure 34.Change in heart rate of all participants over time from baseline in all four conditions

*Table 21. Repeated measures ANOVA of mean heart rate in the three segments and baseline values in each condition. * denotes significance with an alpha value 0.05 ** denotes significance with alpha value of 0.01.*

		Sig. (2-way)
2D	$F(1.86, 78.09) = 10.63$	**$p < 0.01$****
Intermittent	$F(1.55, 65.07) = 3.23$	$p = 0.06$
Persistent	$F(1.48, 62.10) = 7.66$	**$p < 0.01$****
Motion	$F(1.20, 50.32) = 0.26$	$p = 0.65$

If significance was found, pairwise comparisons were made between segment and the baseline measurements for that condition (Table 22). In only the 2D condition did any significant changes persist past 5 minutes.

*Table 22. Post hoc paired t-test between mean heart rate of time segments and baseline values within a condition. * denotes significance with an alpha value 0.05 ** denotes significance with alpha value of 0.01.*

	5 min	10 min	15 min
2D	**p < 0.01****	p = 0.98	**p = 0.02***
Persistent	**p < 0.01****	p = 0.13	p = 0.33

To investigate changes between conditions, paired t-tests were performed on cumulative change values and compared with those in the 2D control condition. Significant differences were found in the intermittent and motion-in-depth conditions when compared to the 2D condition (Table 23). That is to say cumulative changes of heart rate in the intermittent and motion-in-depth condition were significantly less pronounced than the decrease in heart rate seen in the 2D condition.

*Table 23. Paired t-test between sum change of heart rate in each condition compared to sum change in 2D condition. * denotes significance with an alpha value 0.05 ** denotes significance with alpha value of 0.01.*

	t value	Sig. (2-way)
2D vs. Intermittent	-2.90	**p < 0.01****
2D vs. Persistent	-0.57	p = 0.57
2D vs. Motion	-3.75	**p < 0.01****

4.9.3. Heart rate variability

HRV was derived as the RMS of the difference between successive beat intervals (RMSSD), using the LabChart HRV software. Changes in HRV in all participants is illustrated in Figure 35. To assess for significant changes within conditions, a repeated measures ANOVA with Greenhouse-Geisser correction was performed. A significant change was found only in the persistent condition (Table 24).

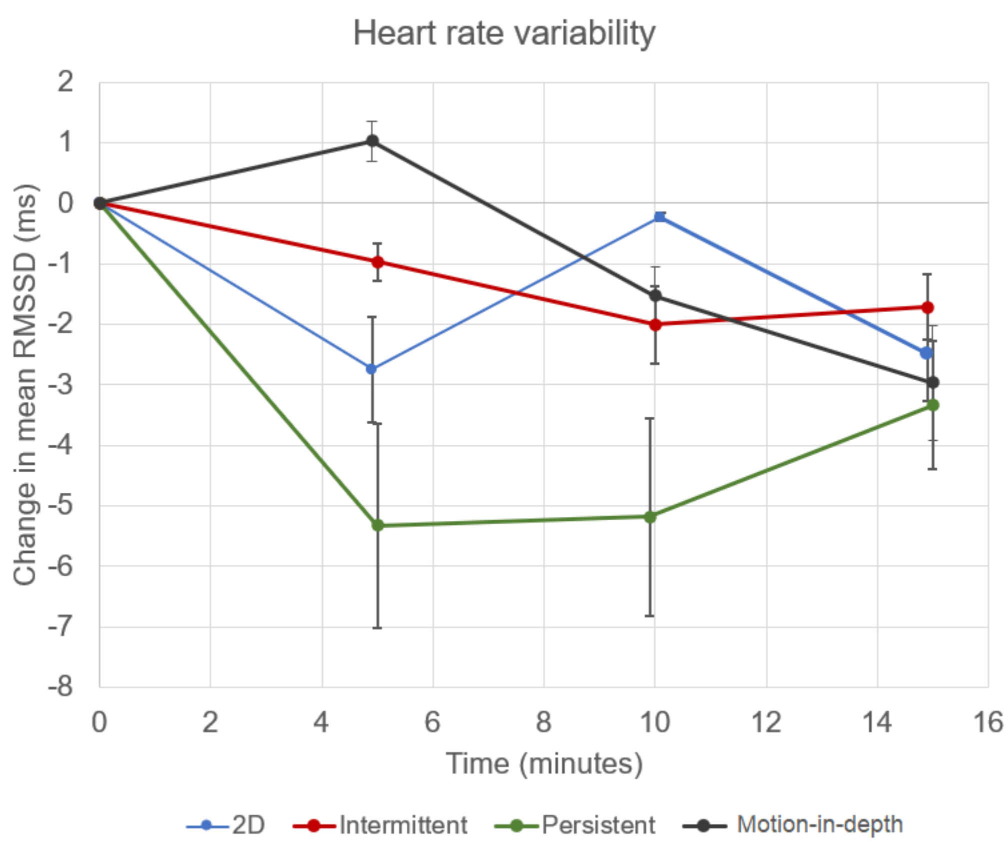

Figure 35. Change in heart rate variability (RMSSD) of all participants over time from baseline in all four conditions.

*Table 24. Repeated measures ANOVA of mean heart rate variability in the three segments and baseline in each condition. * denotes significance with an alpha value 0.05 ** denotes significance with alpha value of 0.01.*

		Sig. (2-way)
2D	F (2.10, 88.44) = 0.85	p = 0.43
Intermittent	F (2.44, 102.47) = 0.32	p = 0.76
Persistent	F (2.46, 103.470) = 5.3	**p < 0.01***
Motion	F (2.64, 111.01) = 1.42	p = 0.24

Using pairwise comparisons to baseline values, the decreases in HRV in the persistent condition were only found in the first and second segment, recovering by the third segment (Table 25).

*Table 25. Post hoc paired t-test between mean heart rate variability of time segments and baseline values within a condition. * denotes significance with an alpha value 0.05 ** denotes significance with alpha value of 0.01.*

	5 min	**10 min**	**15 min**
Persistent	**p < 0.02***	**p < 0.01***	p = 0.46

To investigate changes between conditions, paired t-tests were performed on cumulative change values and compared with those in the 2D control condition. There were no significant differences between either the intermittent, persistent, or motion-in-depth condition when compared to the 2D condition (Table 26).

*Table 26. Paired t-test between sum change of heart rate variability in each condition compared to sum change in 2D condition. * denotes significance with an alpha value 0.05 ** denotes significance with alpha value of 0.01.*

	t value	Sig. (2-way)
2D vs. Intermittent	-0.09	p = 0.93
2D vs. Persistent	1.20	p = 0.24
2D vs. Motion	-0.37	p = 0.71

4.10. Summary of results

To summarise; performance, as measured by ability to adhere to flight commands, response time and throughput to correctly identify and differentiate a warning alert, was not significantly affected by the addition of three variations of S3D to a warning alert symbol. The small differences that were observed between conditions were not significant and certainly are not operationally relevant. Significant decreases in peripheral skin temperature compared to individual baseline values were apparent, suggesting that the whole task was cognitively taxing. However, the addition of S3D to the warning alert did not further affect the cognitive burden. While there were significant differences in cumulative changes of heart rate between intermittent and motion-in-depth conditions compared to the 2D condition, the differences were that heart rate did not decrease as much. It can be argued that there were no relevant or conclusive sustained changes to heart rate or heart rate variability, the other indicators of altered cognitive workload, with the addition of S3D to a warning alert. Furthermore, there were no significant differences in participant symptom scores or workload self-assessment scores in any of the S3D conditions compared to the 2D condition.

5. ASSOCIATION WITH INDIVIDUAL STEREO ACUITY

5.1. Relationship between performance (flying performance and warning alert task) and stereo acuity

No significant relationships were identified between flying performance and measured individual stereo acuity, as illustrated in Figure 36 and summarised in Table 27. This was true regardless of the condition and the method used to measure stereo acuity. Of note, there were two individuals with very poor stereo acuity. A key aim of this work was to investigate performance using S3D in a population with a wide range of individual stereo acuity. In principle, there was no indication to discard the data from the two participants with the worst stereo acuities. For completeness, the data was analysed excluding the two individuals with the worst stereo acuities and no significant changes to the relationships with flying performance and throughput of the warning alert task were seen.

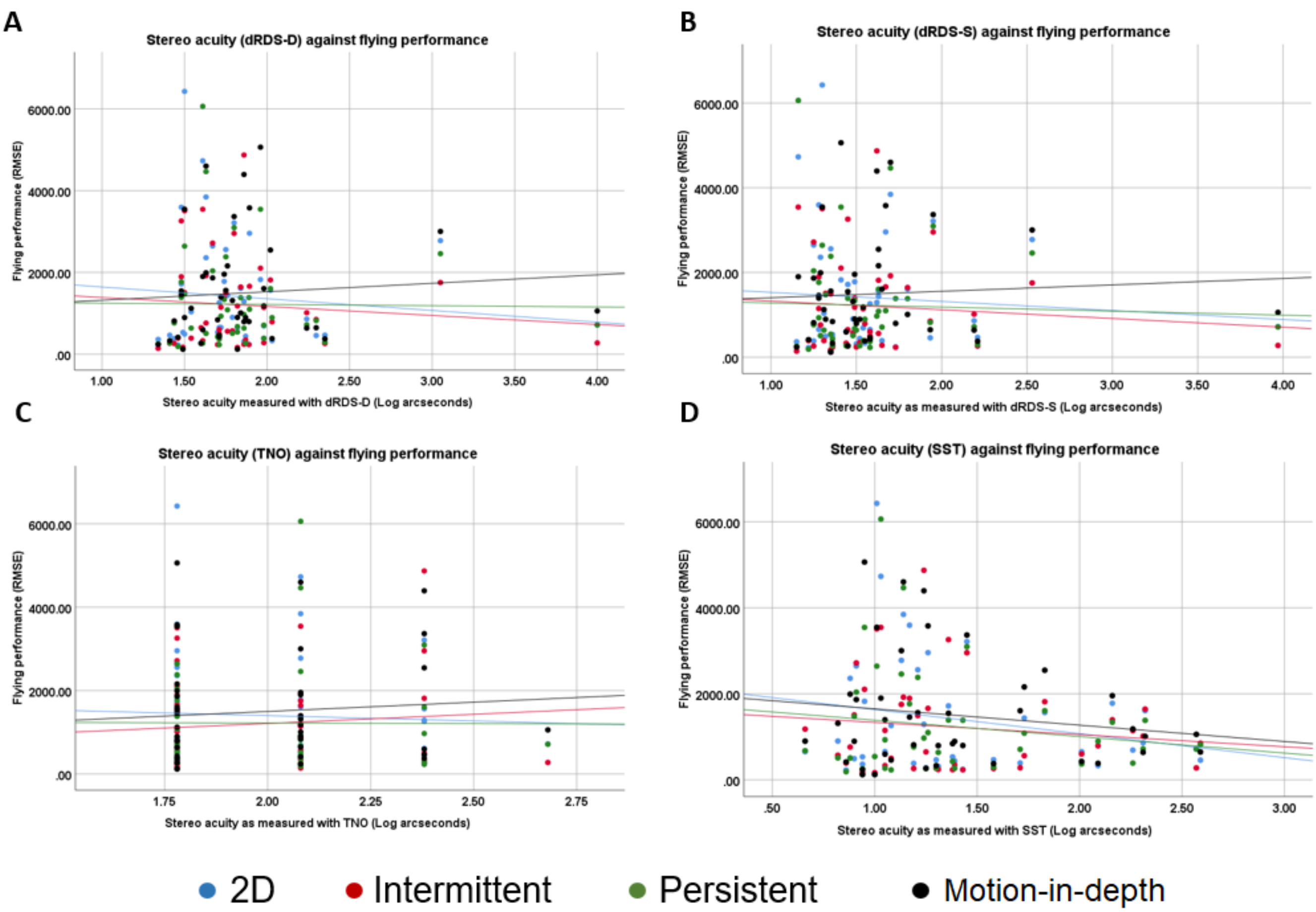

Figure 36. Relationship between performance on the primary task (flying), in all four conditions and individual stereo acuity as measured with measured with dRDS-D (A), dRDS-S (B), TNO (C), and SST (D).

Table 27. Pearson's correlation analyses between stereo acuity and performance on the primary flying task.

	dRDS-D (Corr & sig)	dRDS-S (Corr & sig)	TNO (Corr & sig)	SST (Corr & sig)
2D	-0.09, p = 0.55	-0.07, p = 0.64	-0.04, p = 0.77	-0.20, p = 0.19
Intermittent	-0.09, p = 0.57	-0.09, p = 0.58	0.09, p = 0.54	-0.13, p = 0.41
Persistent	-0.01, p = 0.94	-0.04, p = 0.82	-0.01, p = 0.97	-0.16, p = 0.32
Motion	0.07, p = 0.63	0.06, p = 0.72	0.08, p = 0.59	-0.15, p = 0.34

Secondary task performance in the warning alert task as measured by throughput was compared to individual stereo acuity as measured with the four different stereo tests (Figure 37), with correlation analyses summarised in Table 28. Two positive correlations were identified between stereo acuity and performance in the 2D condition when stereo acuity was measured using the SST, and in the persistent condition when stereo acuity was measured with the TNO test. That is to say, those with the highest stereo thresholds performed better with regard to throughput in the 2D and persistent conditions when stereo acuity was measured with the SST and TNO test respectively. These relationships are not consistent across conditions or stereo tests, and are unlikely to be true indicators of any meaningful relationship, especially as no difference to throughput was identified between conditions overall.

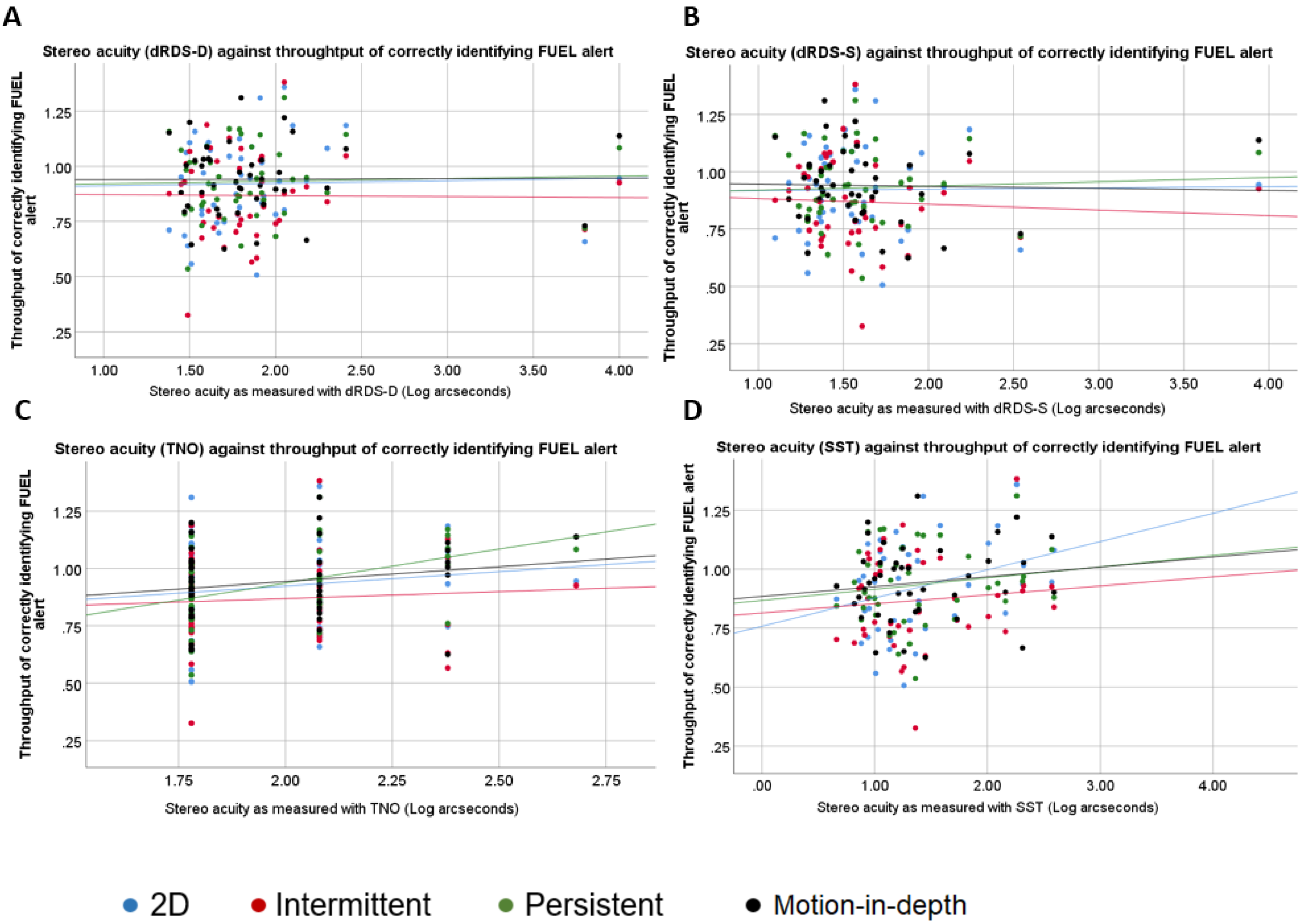

Figure 37.Relationship between performance on the secondary task (throughput correctly identifying FUEL warning symbol), in all four conditions and individual stereo acuity as measured with measured with dRDS-D (A), dRDS-S (B), TNO (C), and SST (D).

*Table 28. Pearson's correlation analyses between individual stereo acuity and performance on the secondary task (throughput correctly identifying FUEL warning symbol), in all four conditions. * denotes significance with an alpha value 0.05 ** denotes significance with alpha value of 0.01.*

	dRDS-D (Corr & sig)	dRDS-S (Corr & sig)	TNO (Corr & sig)	SST (Corr & sig)
2D	0.06, $p = 0.71$	0.04, $p = 0.79$	0.15, $p = 0.32$	0.31, **$p = 0.04$*** $R^2 = 0.10$
Intermittent	0.00, $p = 0.99$	-0.04, $p = 0.81$	0.08, $p = 0.62$	0.10, $p = 0.52$
Persistent	0.07, $p = 0.64$	0.09, $p = 0.57$	0.44, **$p < 0.01$**** $R^2 = 0.19$	0.15, $p = 0.35$
Motion	0.01, $p = 0.94$	-0.02, $p = 0.87$	0.20, $p = 0.20$	0.13, $p = 0.40$

5.2. Relationship between questionnaire results and stereo acuity

Change in individual symptom scores as well as perceived workload responses were assessed against stereo acuity as measured with the four different stereo tests; dRDS-D, dRDS-S, TNO, and SST. Using Spearman's rank test to investigate the relationships between these symptom scores and workload responses in each condition, positive correlations were found in ten incidences, listed in Table 29. Scatter plots illustrating the positive correlations found between stereo acuity and three questionnaire responses (one symptom and two workload questions) are given in Figure 38. The relationships between all other questionnaire responses and individual stereo acuity are illustrated in Appendix E. The only stereo test to not identify any relationships with individual stereo acuity in any condition was the TNO test.

*Table 29. Summary of positive correlations between questionnaire results and individual stereo acuity. *denotes significance with an alpha value 0.05 **denotes significance with alpha value of 0.01.*

Question	Stereo test	Condition	Significant correlation	R^2
How clear is your vision?	dRDS-D	Persistent	R_s (43) = 0.35, **p = 0.02***	0.03
How mentally demanding did you find the task?	dRDS-D	Motion	R_s (43) = 0.31, **p = 0.04***	0.07
	dRDS-S	Motion	R_s (43) = 0.34, **p = 0.03***	0.06
	SST	Motion	R_s (42) = 0.31, **p = 0.047***	0.05
How physically demanding did you find the task?	dRDS-D	2D	R_s (43) = 0.39, **p < 0.01****	0.12
		Persistent	R_s (43) = 0.33, **p = 0.03***	0.10
		Motion	R_s (43) = 0.47, **p < 0.01****	0.14
	dRDS-S	2D	R_s (43) = 0.33, **p = 0.03***	0.07
		Persistent	R_s (43) = 0.32, **p = 0.04***	0.07
		Motion	R_s (43) = 0.45, **p < 0.01****	0.09

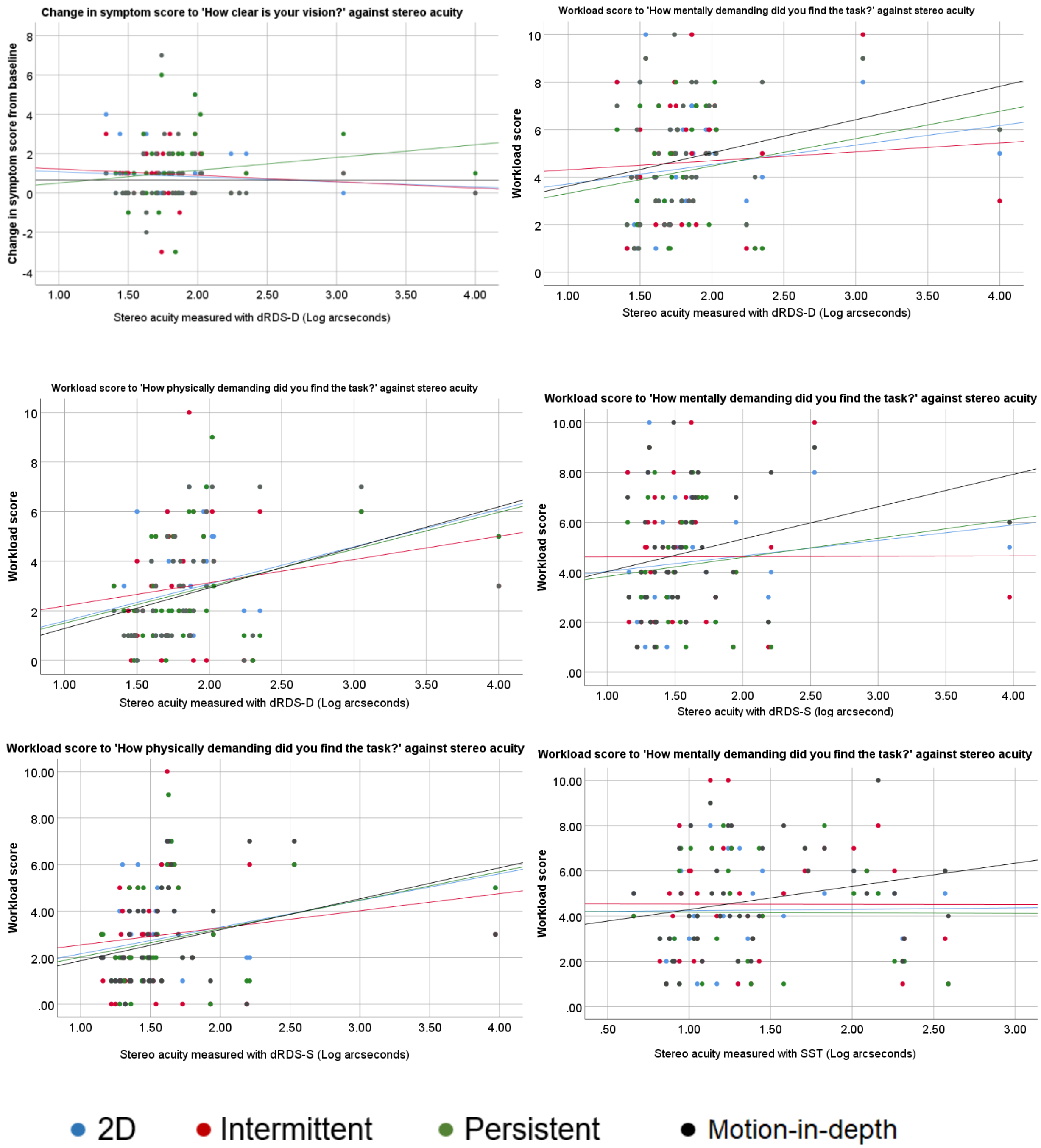

Figure 38. Significant correlations between individual stereo acuity scores and one symptom question; How clear is your vision with dRDS-D (top left), and two workload questions; How mentally demanding did you find the task? With dRDS-D (top right), with dRDS-S (middle right), and with SST (bottom right) 'How physically demanding did you find the task?' with dRDS-D (middle left), and with dRDS-S (bottom left).

5.3. Relationship between changes in physiological variables and stereo acuity

Changes in physiological variables measured (peripheral skin temperature, heart rate and heart rate variability) were assessed against stereo acuity as measured with the four different stereo tests: dRDS-D, dRDS-S, TNO, and SST. Cumulative changes to peripheral skin temperature were plotted against stereo acuity and the relationships between these two variables were analysed in the four different conditions. The cumulative changes in peripheral skin temperature across the four conditions against stereo acuity measured using dRDS-D can be seen in Figure 39; comparisons using the other stereo tests (dRDS-S, TNO, and SST) can be found in Appendix F. No relationship was identified between cumulative changes in peripheral skin temperature and individual stereo acuity, measured with any of the four stereo tests. Using Spearman rho test; these correlation analyses are summarised in Table 30.

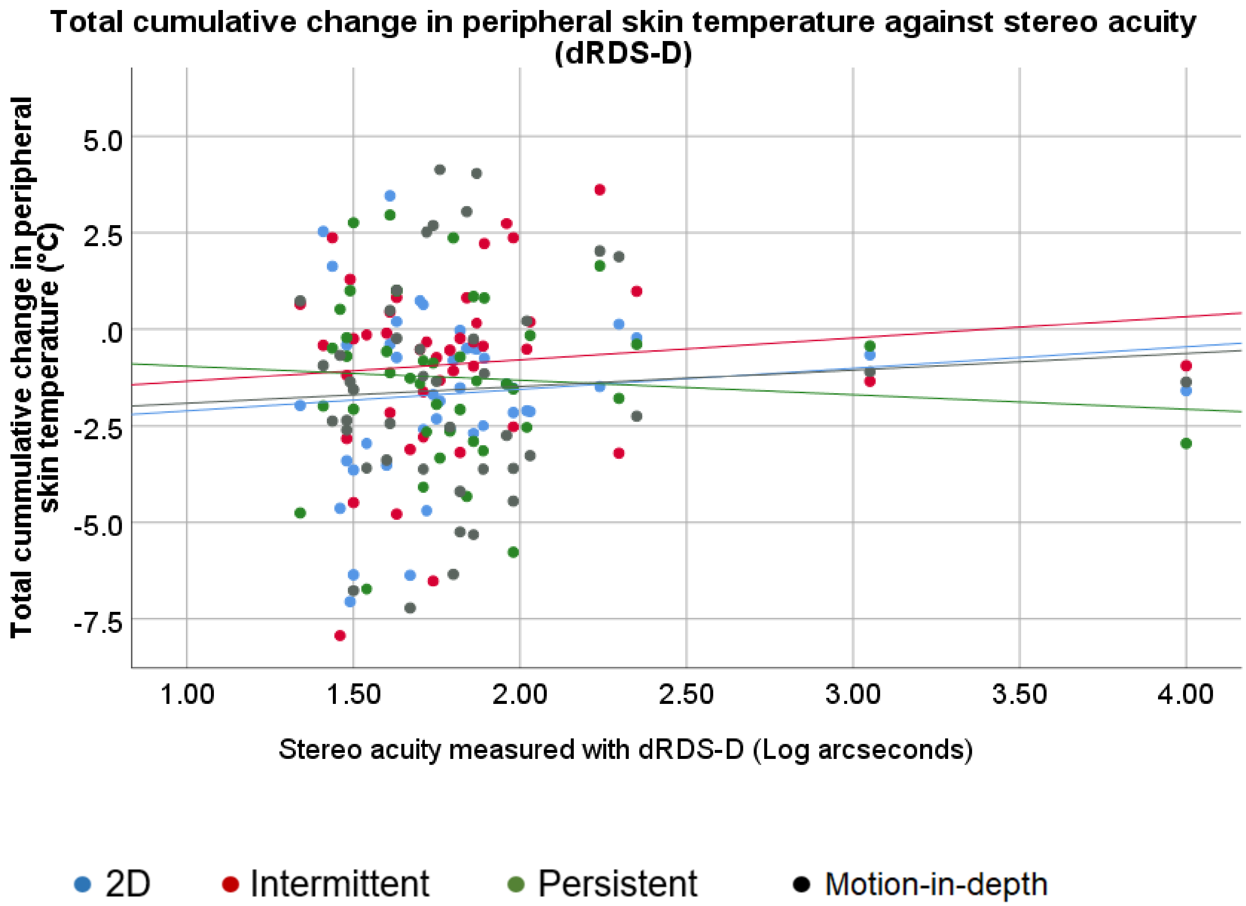

Figure 39. Scatter plot of total cumulative change in peripheral skin temperature against individual stereo acuity measured with dRDS-D in all four conditions.

*Table 30. Pearson's correlation analysis between individual stereo acuity and cumulative change in peripheral skin temperature in all four conditions. * denotes significance with an alpha value 0.05 **denotes significance with alpha value of 0.01.*

	dRDS-D Corr & Sig (2-way)	dRDS-S Corr & sig (2-way)	TNO Corr & sig (2-way)	SST Corr & sig (2-way)
2D	0.12, p = 0.46	0.13, p = 0.41	0.18, p = 0.25	0.23, p = 0.13
Intermittent	0.11, p = 0.48	0.13, p = 0.39	-0.06, p = 0.70	0.14, p = 0.37
Persistent	-0.08, p = 0.61	-0.04, p = 0.78	-0.00, p = 0.10	-0.19, p = 0.23
Motion	0.07, p = 0.65	0.10, p = 0.54	0.21, p = 0.17	0.29, p = 0.60

Cumulative changes in heart rate in the four conditions against stereo acuity measured using dRDS-D are illustrated in Figure 40; comparisons with other stereo tests can be found in Appendix F. Using Spearman rho test; these correlation analyses are summarised in Table 31. A significant correlation was identified between change in heart rate in the motion-in-depth condition and stereo acuity measured with dRDS-D; r = 0.31 (43), p = 0.04. That is to say those with worse stereo acuity, as measured with dRDS-D, demonstrated larger increases in heart rate, which may in turn suggests a greater cognitive workload. However, when the two extreme stereo acuity values are excluded, there is no longer any significant relationship between cumulative changes in heart rate and stereo acuity measured with dRDS-D; n = 41, 0.164 p = 0.30.

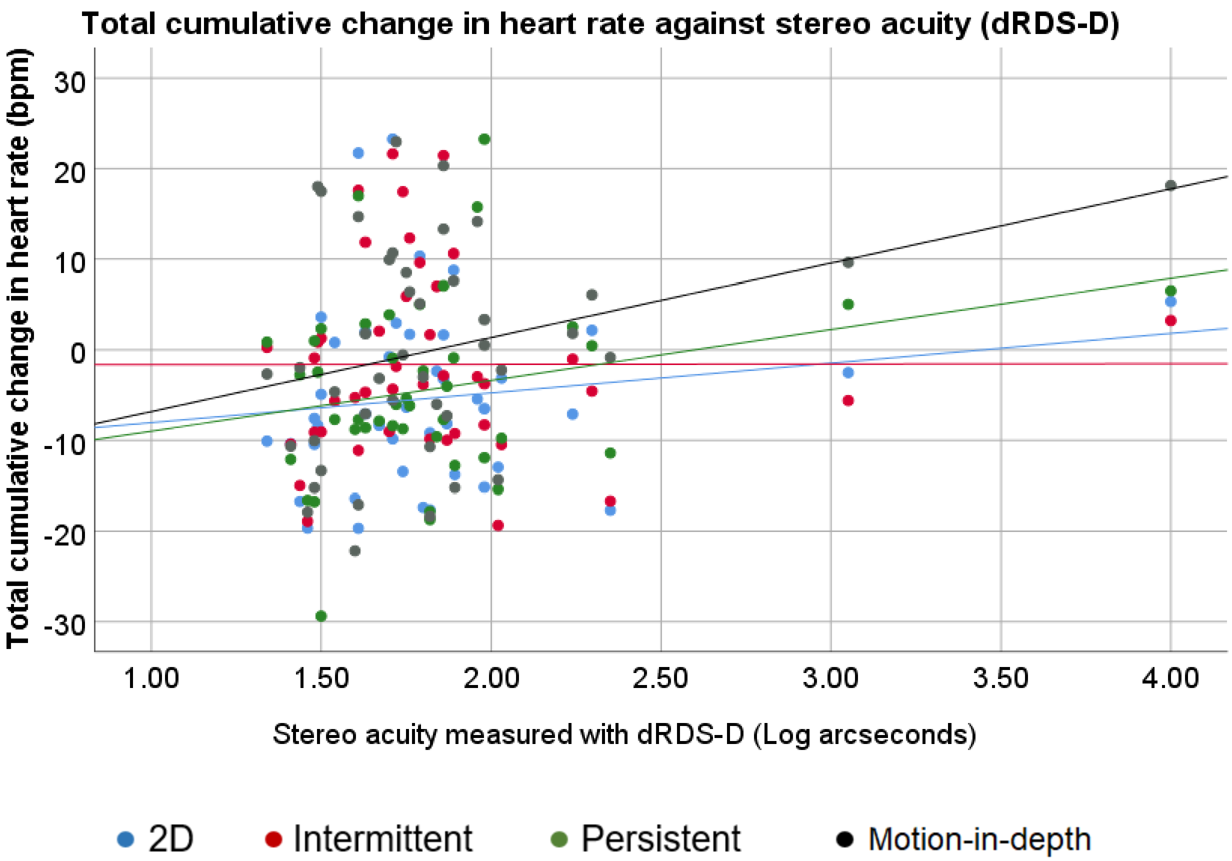

Figure 40. Scatter plot of cumulative change in heart rate variability against individual stereo acuity measured with dRDS-D in all four conditions

*Table 31. Pearson's correlation analysis between individual stereo acuity and cumulative change in heart rate in all four conditions. * denotes significance with an alpha value 0.05 ** denotes significance with alpha value of 0.01.*

	dRDS-D Corr & Sig (2-way)	**dRDS-S Corr & Sig (2-way)**	**TNO Corr & Sig (2-way)**	**SST Corr & sig (2-way)**
2D	0.15, p = 0.33	0.17, p = 0.27	0.24, p = 0.11	0.2, p = 0.91
Intermittent	0.00, p = 0.99	0.09, p = 0.55	0.220, p = 0.157	-0.00, p = 0.99
Persistent	0.25, p = 0.10	0.26, p = 0.09	0.153, p = 0.326	0.11, p = 0.48
Motion	**0.31, p = 0.04*** **R^2 = 0.10**	0.28, p = 0.07	0.228, p = 0.14	0.09, p = 0.57

Cumulative changes in heart rate variability in the four conditions against stereo acuity measured using dRDS-D are illustrated in Figure 41; comparisons with other stereo tests can be found in Appendix F. No relationship was identified between cumulative changes in heart rate variability and individual stereo acuity. Using Spearman rho test; these correlation analyses are summarised in Table 32.

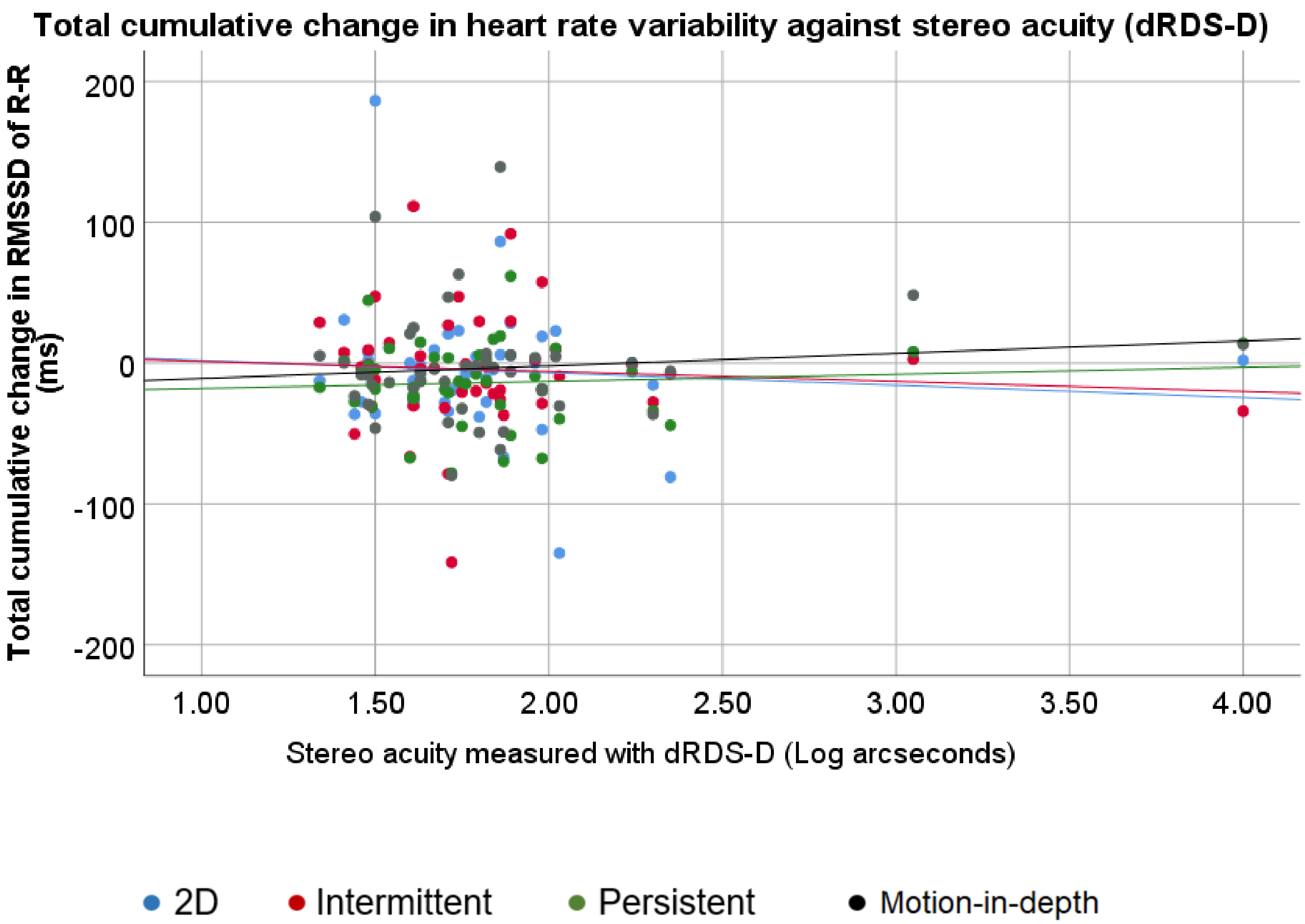

Figure 41. Scatter plot of total cumulative change in heart rate variability against individual stereo acuity measured with dRDS-D in all four conditions.

*Table 32. Pearson's correlation analysis between individual stereo acuity and cumulative change in heart rate variability in all four conditions. * denotes significance with an alpha value 0.05 ** denotes significance with alpha value of 0.01.*

	dRDS-D Corr & sig (2-way)	dRDS-S Corr & sig (2-way)	TNO Corr & sig (2-way)	SST Corr & sig (2-way)
2D	-0.09, p = 0.57	-0.05, p = 0.74	-0.18, p = 0.24	-0.19, p = 0.50
Intermittent	-0.08, p = 0.62	-0.06, p = 0.72	-0.22, p = 0.16	0.03, p = 0.87
Persistent	0.08, p = 0.61	0.16, p = 0.31	0.05, p = 0.75	0.12, p = 0.91
Motion	0.10, p = 0.51	0.08, p = 0.62	-0.21,p = 0.18	0.012, p = 0.91

5.4. Summary of results

Following correlation analyses, no significant relevant relationships were found between performance in either the primary flying task or response to the secondary warning alert task and individual stereo acuity. This was true for stereo acuity measured using digital stereo tests (dRDS-D, dRDS-S, and SST) as well as for the paper based TNO stereo test. Significant positive correlations were identified between stereo acuity (measured with dRDS-D) and three questionnaire responses (one symptom and two workload questions) but this was not universal amongst S3D conditions or stereo tests. There was a positive correlation between cumulative heart rate changes and stereo acuity in the motion-in-depth condition. That is to say those with the worst stereo acuity reported they found the task more physically and mentally demanding as well as demonstrating the greatest cumulative increase in heart rate. However, the positive correlation between stereo acuity and increase in heart rate as a measure of increased cognitive workload lost its significance when two participants with the worst stereo acuity were excluded.

6. DISCUSSION

The addition of S3D to a display has previously been shown to improve performance in a number of different tasks and applications. However, positive results have not been unanimous. In this research, the potential of S3D was investigated specifically in the operationally relevant setting of a warning alert task displayed within an HMD in a flight simulation. Whilst there has been some research performed in this area, further development was warranted, specifically with regard to investigating any relationship with individual stereo acuity as measured with more reliable and repeatable stereo tests. The lack of adequate and accurate stereo acuity screening in participants is a limitation apparent in many previous studies. A particular consideration here was to introduce disparity to the warning alert symbol in three different conditions, rather than simply S3D versus 2D. Consequently, the impact of using S3D as an alternative (i.e., intermittent condition), or as an additive (i.e., static S3D in the persistent condition, and with a continuous fronto-spatial aspect in the motion-in-depth condition) visual cue to 2D flashing was investigated. In addition to performance, the impact of S3D on ocular symptoms and cognitive workload was examined, measured using objective physiological parameters and subjective questionnaires. All measurements of performance and cognitive workload were assessed in relation to individual stereo acuity as measured with four different stereo tests, including the current stereo test used to qualify military aircrew in the U.K. Furthermore, the stereo tests themselves were scrutinised on their reliability, repeatability, and ability to predict performance in a warning alert flying task. Confidence in reliable vision tests is critical for military organisations to ensure the appropriate and fair selection of aircrew candidates, as well as supporting trained aviators throughout their careers and there is concern that current paper-based stereo tests are unable to meet these requirements and are prone to giving false positive results (Posselt *et al.*, 2022). Inappropriately selecting aircrew who lack stereopsis is of particular concern with the increasing visual demands of digital cockpits and displays used in military aviation.

6.1. Research questions reviewed

Each research question is addressed in turn and assessed as to whether it can be answered.

6.1.1. Research Question 1

- Is there any difference between the reliability and repeatability of current paper-based tests when compared to newly developed computer-based tests to measure stereo acuity?

The TNO test (six-page version) is the paper-based test currently used by the British military to measure stereo acuity as a means of assessing binocular vision, to qualify aircrew. However, it has a number of limitations and may no longer be the most suitable selection test. Furthermore, the TNO test is not able to measure stereo acuity to threshold, rendering it unable to investigate any relationship between individual stereo acuity and operational performance. For this research, in addition to the TNO test, alternative stereo tests were investigated; two computer-based digital RDS (dRDS) versions (static and motion-in-depth) of the TNO test were developed and used along with a contour circle test (SST) to measure stereo acuity. As expected, in almost every combination, there was a significant difference between tests in the stereo acuity threshold obtained. The only comparison in which results were not significantly different was between the dRDS-S and SST. All three computer-based tests correlated well with each other, but none of the digital tests correlated with the six-page version of the TNO test. Overall, the SST produced the lowest (best) stereo acuity thresholds, while the TNO test gave the worst scores. The TNO test (six-page version) is limited by a floor at 60 arcseconds, so cannot distinguish between individuals any further. All three computer-based tests produced more granular results and could measure to threshold, although they still differed. A reason why the dRDS-D produced the highest thresholds of the digital tests is that it forces participants to recompute disparity every few frames, making it harder. In addition, measured stereo acuity could be affected by a changing dot density within the dRDS-D. Both dRDS computer-based tests had tighter limits of agreement than the TNO test, indicating that they give more reliable and repeatable results. Furthermore, the computer-based tests address many of the other limitations seen with the TNO test, eliminating cheating and errors caused by printing discrepancies or intra-ocular illumination differences and removing examiner interference or bias.

In summary, significant differences were found in reliability and repeatability of computer-based tests when compared to the current paper TNO test (six-page version) used to measure stereo acuity. Based on this research, a digital alternative to the TNO stereo test to qualify military aircrew is strongly recommended (Posselt *et al.*, 2022).

6.1.2. Research Question 2

- How does the addition of S3D to a warning alert task displayed in an HMD affect performance, ocular symptoms, and cognitive workload?

In order to assess performance with the use of a warning alert symbol, a task was developed which required participants to differentiate between two similar four lettered words relevant to a flight task; FLAP and FUEL. In the control condition both alerts were flashed on and off in 2D at a rate of 1 Hz. In the other three experimental conditions, FUEL was presented in various S3D conditions (intermittent, persistent, and motion-in-depth), while FLAP was always flashed on and off in 2D. Thus, S3D could be used as an alternative or additional cue to identify the FUEL warning alert. The disparities used were chosen based on those identified as most impactful in initial pilot testing yet within levels of gross discomfort or diplopia. Performance was measured with response time and calculated as the proportion of correct responses divided by response time resulting in a measure termed throughput. When analysing mean throughput, no significant difference in response time or throughput was identified with the addition of S3D in any form, to display a warning alert.

Cognitive workload was assessed using a variety of methods, both subjective and objective. Objectively peripheral skin temperature was measured and cardiac activity recorded with an ECG, from which heart rate and heart rate variability could be calculated. For each physiological variable a within subject baseline was used for comparison, taken over 10 minutes at rest before each condition. While peripheral skin temperature did indeed significantly reduce over time, suggesting an increased cognitive workload, no additional effect was seen in any of the S3D conditions. There were no conclusive changes in either heart rate or heart rate variability to suggest increased cognitive workload and no overall differences in the S3D conditions when compared to the control 2D condition. Participants answered a subjective symptom questionnaire before and after each condition as well as a set of workload questions after each condition. There were no significant differences in the change of symptom scores or workload ratings between S3D conditions when compared to the control 2D condition.

In summary, the addition of S3D to a warning alert symbol displayed in an HMD was not found to improve performance in the task to identify and differentiate the warning alert. Furthermore, the addition of S3D to a warning alert did not affect ocular symptoms or cognitive workload, measured both subjectively and objectively.

6.1.3. Research Question 3

- Is there any relationship between individual stereo acuity measured with different stereo tests and performance, ocular symptoms, or cognitive workload in a warning alert task displayed with S3D in an HMD?

With more reliable and repeatable computer-based tests, stereo acuity was able to be measured to threshold levels in all participants. Individual stereo acuity as measured with four different stereo tests was then plotted against performance in the warning alert task, flying performance, and both subjective and objective measures of cognitive workload. No significant relevant correlation or relationship was identified between performance, either flying performance or throughput of correctly identifying a warning alert, in any of the four conditions and individual stereo acuity as measured with four different stereo tests. Regarding cognitive workload, a significant relationship was identified between cumulative heart rate and stereo acuity (measured with dRDS-D) in the motion-in-depth condition. However, when two outliers with extremely poor stereo acuity at the limits of what could be reliably measured by the stereo tests were removed from the data set, the significant correlation between cumulative heart rate in the motion-in-depth condition and stereo acuity measured with the dRDS-D became non-significant. For the questionnaire data, several significant relationships were found between stereo acuity and certain symptom questions; ‘How clear is your vision?’ in the persistent condition and for the workload questions ‘How mentally demanding did you find the task?’ in the motion-in-depth condition and ‘How physically demanding did you find the task?’ in the 2D, persistent and motion-in-depth conditions. These positive correlations were identified with a combination of different digital stereo tests. That is to say that those with the worst stereo acuity reported less clear vision, and found the task more mentally and physically demanding, suggesting they subjectively experienced a greater cognitive workload. However, these effects are not universally replicated across all S3D conditions or with all measures of cognitive workload. Thus, the idea that S3D adversely affects those with the poorest stereo acuity cannot be decisively concluded. It should be noted that the paper based TNO stereo test was unable to identify any relationships.

In summary, no relationship was found between stereo acuity and performance in a warning alert task using various presentations of S3D. Stereo acuity as measured with four stereo tests (dRDS-D, dRDS-S, TNO, SST) does not predict performance, as no change in performance was identified. However, there is some evidence to suggest that those participants with the worst stereo acuity subjectively found the task to be more mentally and physically demanding in some of the conditions.

6.2. Effect of S3D conditions

Adding disparity to an image, enabling it to be viewed in S3D, can be beneficial by more efficiently capturing attention (Sakata *et al.*, 2017). In addition, S3D can improve search times (Reis *et al.*, 2008) and reduce response times (Blake, Sloane and Fox, 1981) to identify a target. S3D can be used to declutter displayed information, aiding interpretation (Kooi, 2011) and performing a task displayed in S3D can reduce perceived cognitive workload (Sakata *et al.*, 2017). Specific to the aviation environment, improved performance has been seen in a simulated aerial refuelling task (Winterbottom *et al.*, 2016) and in estimating the distance of a helicopter landing gear (Winterbottom *et al.*, 2017) or skids above the ground (Hartle *et al.*, 2020) when using a S3D display compared to standard 2D imagery. In an HMD, reduced response times were demonstrated with the addition of S3D to display a warning alert (Browne, Moffitt and Wimsatt *et al*. 2017). Consequently, it was anticipated for this research that the addition of S3D to a warning alert could improve both reaction time and throughput to correctly identify and differentiate a warning alert, and that any effect may be correlated with individual stereo acuity. It was also anticipated that there could be differences in performance benefits seen between the types of disparity presentation, as static and motion-in-depth disparities are thought to be processed differently (Beverley and Regan, 1975). As discussed previously, some researchers have suggested that there is a neural mechanism tuned to stereoscopically-defined motion, in the same way that they are to luminance-defined motion (Portfors-Yeomans and Regan, 1996; Portfors and Regan, 1997; Patterson, 1999). On the contrary, there are authors who dismiss the idea of a specific stereoscopic motion detection system (Lu and Sperling, 1995; Harris, McKee and Watamaniuk, 1998), although there is a suggestion that those negative findings may be as a result of inadequate or inappropriate stimulus parameters (Patterson, 1999).

However, contrary to expectations, no significant improvement was identified, in either response time or throughput to correctly identify and differentiate a warning alert symbol, in any of the three different S3D conditions when compared to an equivalent 2D condition. Some objective measures of workload revealed that the task was cognitively demanding, such as the significant decrease in peripheral skin temperature seen over time. However, there was no significant difference in peripheral skin temperature changes between conditions, suggesting that in general the S3D conditions did not impact the existing cognitive burden for this particular warning alert flying task. Regarding objective measures of adverse ocular symptoms and workload using questionnaires, again there were no significant differences reported in the S3D conditions compared to the 2D condition.

A potential cause for why stereopsis may not be as effective for a warning alert task is that the visual system is simply not specifically sensitive to binocular motion differences (Harris, McKee and Watamaniuk, 1998; Patterson, 1999). Ultimately, the process to interpret retinal disparity enabling stereopsis is inherently slow and disparity detection thresholds are known to be related to the duration of stimuli exposure, increasing fourfold as exposures decrease from 1 s to 5 ms (Ogle and Weil, 1958). The principal cause for increasing thresholds with decreased exposure time is that the disjunctive eye movements of vergence are sluggish in nature, and can take as long as 800 ms (Rashbass and Westheimer, 1961). By comparison conjugate eye movements such as saccades and pursuit are much faster (Tyler *et al.*, 2012). The sluggish nature of the human stereopsis system is supported elsewhere (Harris and Watamaniuk, 1995), who also argue that consistent stereo signals are needed for detection. Regan and Beverley (1973) concur that sensitivity to stereoscopic motion-in-depth stimuli is optimal at lower frequencies, but report that motion-in-depth with frequencies up to 4 Hz can still be perceived (Regan and Beverley, 1973). However, if flashing frequency were to be reduced for an important warning symbol in the aviation environment, to allow more time needed for vergence eye movements, it could affect the alert’s operational effectiveness and should not be implemented without further investigation. With a similar argument, it is possible that a positive effect of S3D to a warning alert task could have been identified if the distracting primary flying and navigation exercise were not present. However, in an operational environment no pilot would simply be waiting for a warning alert to appear, and this approach was deemed unlikely to translate into the operational context, so was not considered further.

Browne *et al.* (2017) did find a benefit of adding S3D to a warning alert task. It is suggested that a potential explanation for the improvement in performance seen in Browne *et al.'s* (2017) S3D conditions could be that the warning alert symbol remained visible while it was flashed in and out of stereo depth, similar to the persistent condition in this research. In order to investigate the idea that a continuously visible alert could be the influencing factor, rather than S3D itself, a condition to isolate S3D (i.e., the intermittent condition) was included in the present research. In the intermittent condition the warning alert was flashed on and off, but when it was on the alert symbol was presented with disparity, in S3D. There were no differences in performance, as measured by response time or calculated by throughput, in either the intermittent or persistent condition when compared with the 2D condition, contrary to Browne *et al.* (2017). However, Browne *et al.* (2017) simply measured response time to a S3D stimulus. In comparison, the task developed in this research required identification and interpretation to differentiate between two similar symbols. Another difference between the work conducted by Browne et al. (2017) and the research here which may potentially affect results, is the fidelity and scale of the flight simulator. The equipment developed for this research consisted of a 270° horizontal field of view display, compared with the relatively smaller (120° horizontal field of view) computer screens used by Browne *et al.* (2017). However, the same HMD with the 100% binocular overlap was used for both research efforts.

6.3. Other uses for S3D in an HMD

While no benefit of adding S3D to a warning alert task has been identified in this research, flashed either statically or with motion-in-depth at 1 Hz, the potential use of S3D in an HMD should not be dismissed entirely. As has been discussed, there are many other situations where using S3D in a display has proved to be beneficial, particularly for fine depth discrimination, real depth representation, as well as spatial orientation and interpretation (McIntire, Havig and Geiselman, 2012). In an aviation environment, the use of S3D displays has been shown to improve performance when compared to a 2D display on multiple occasions; for example the use of S3D in a simulated aerial refuelling task resulted in improved performance, where performance was as assessed by a combination of the number of collisions, number of attempts, duration, and distance at attempt (Winterbottom *et al.*, 2016; O'Keefe *et al.*, 2022). Furthermore, in these particular studies, individual stereo acuity was strongly correlated with performance in the refuelling task. S3D displays were also found to be beneficial in a simulated helicopter landing task, with participants better able to discriminate the distance between the rear wheel of a helicopter and a ground target when in hover (Winterbottom *et al.*, 2017). The benefits of providing, and being able to process, binocular visual cues to more accurately estimate altitude in a simulated rotary aircraft have been replicated elsewhere (Hartle *et al.*, 2017, 2020). Performance in a terrain tracking task, simulated in a VR HMD, improved when the scene was displayed in S3D compared to 2D (Ernst *et al.*, 2019), with performance assessed as deviation from the flight path close to the surface of a featureless sea (500ft). When operating in a degraded visual environment, such as flying at night or in clouds, in desert or arctic environments, situational awareness is vastly improved with sensor generated images (e.g., night vision devices or Distributed Aperture System), which is a form of VR. Situational awareness is needed most for air-to-air refuelling, landing, formation or low-level flying. In such high risk DVE situations there may be a benefit to displaying sensor information in S3D using an HMD.

Other particularly promising applications of S3D within an HMD could be a highway in the sky (HITS) concept, providing a depth representative flight path for the pilot to follow, and a 3D locator line identifying a designated target in-depth. In addition to the warning alert task, Browne *et al.* (2017) assessed performance using both HITS and a locator line task displayed in S3D. Target acquisition time was significantly faster using a S3D locator line when compared to a traditional 2D version. In the HITS task no significant improvement in performance was demonstrated using S3D, however, anecdotally the S3D conditions were preferred by participants when compared to their 2D alternatives. Despite a lack of statistical evidence to support the use of S3D in an HMD for the task, the authors went on to advocate its potential and encouraged further investigation into this capability. Other studies show a performance benefit of using HITS displayed in S3D, but using a stereo display and not in an HMD (Parrish, Busquets and Williams, 1990). Again, this could be particularly useful in a DVE.

In summary, no benefit was identified in using S3D within an HMD to display a warning alert task. However, due to the clear demonstrated improvements in human performance of utilising binocular visual cues for certain tasks in simulated aviation environments using S3D displays; it is likely S3D could also be beneficial in an HMD when used for relative depth discrimination and spatial orientation tasks, as well as improving situational awareness. Improved situational awareness is particularly crucial when flying in a DVE when monocular cues are unreliable or absent.

6.4. Study limitations

Substantial care was taken in designing a study to investigate the effect of S3D to a warning task in an HMD. This included using a well fitted and aligned HMD, minimising V-A mismatch between optical displays, creating a representative operational flying task in a realistic flight simulator, randomising the order of conditions and blinding participants to conditions. In addition, a more appropriate performance metric was created by assessing ability to interpret and differentiate information rather than simply measuring reaction time to a stimulus. Furthermore, individual stereo acuity was more accurately and reliably characterised using a variety of newly developed computer-based stereo tests. Nevertheless, some limitations to this study are apparent.

Firstly, it could be argued that flying experience in our participants may have influenced the results. A proportion of participants did have previous flying experience (n = 7), defined as having flown a real aircraft solo; thus, their performance was compared against the rest of the flying naïve group and no significant difference was found. However, none of the participants had prior experience of using an HMD, which could potentially affect performance in a study like this. This research was conducted during the COVID pandemic, which brought novel challenges; in particular, travel limitations prevented recruiting specialist aircrew from other military bases. In spite of this, it is likely that HMD experience is of little relevance as the area of interest, being stereo acuity, fundamentally examines the visual system itself which is unlikely to be affected by ability to use an HMD, and suitable familiarisation sessions were conducted for all participants. It is noted that 11 of the 43 participants were older than 40 years of age, thus, it is highly likely these participants had a degree of presbyopia. Residual defocus related to presbyopia should not have had any noticeable effect but may have impacted the stereo tests, although stereopsis is relatively robust to defocus (Westheimer and McKee, 1980; Badcock and Schor, 1985). This research was motivated by the current vision standards used to select military aircrew and a diverse sample population was sought regarding visual abilities with a fairly equal gender ratio. However, stereo ability was not analysed according to gender, age, or ethnicity, which could be considered for future work in this area.

Regarding the HMD itself, the SA/62-A has some features which may have affected the results. The prism view-pieces are adjustable to accommodate a wide range of interpupillary distances. Thus, it was necessary to go through a multi-stage alignment process adjusting each optic in turn to create a single aligned image. Image clarity was confirmed verbally by each participant. A coarse objective measure of alignment was implemented; having the participant repeat the process twice. If the participant adjusted both optics to the same degree, it could be surmised with a degree of confidence that the participant understood the alignment process and that their viewed image was aligned producing a single clear binocular image, however, there was no way to objectively validate this. Any misalignment may affect the degree of disparity present between retinas, which could give a perception of more or less depth than intended.

Although a high-quality HMD was used in this study, with substantially better image quality and optics than a typical consumer grade augmented reality device, the HMD itself does have an inherent degree of misalignment between optics. Misalignment can detrimentally affect performance and cause visual discomfort (Kooi and Toet, 2004; Gavrilescu *et al.*, 2019; Seemiller *et al.*, 2022). The degree to how much misalignment can be tolerated in a binocular HMD varies according to different researchers (Self, 1986), but should be kept as low as possible. To counter the issue of misalignment within the optics themselves, a grid of numbered boxes was created and placed across the field of view in both optics as described in Appendix B: Inherent system misalignment. How close the boxes appeared when viewed binocularly was then assessed subjectively. Consequently, the best aligned areas could be identified, which is where the warning alert symbology was placed. The most misaligned areas were avoided, so that any misalignment between optics would be imperceptible to the user. However, misalignment could be more thoroughly addressed in future studies with more accurate measurement and warp mapping. This is particularly important if imagery, taking up more display space, were to be displayed in an HMD.

For the task itself the flying and navigation aspect was designed to keep participants' eyes attending to the top right corner of the display, to the target heading and altitude information. The idea was that this should prevent any advantage that could be gained by participants simply focusing attention at the area where the warning alert would appear, affecting response time and consequently throughput. However, as no eye tracker was incorporated into the HMD, there was no objective way to measure exactly where each participant was looking when a warning alert symbol appeared in the display. This is certainly an aspect that should be investigated for any similar future research.

Each participant was instructed to keep still during the baseline rest periods, however, some movement was noted at times. Physical movement by participants introduced noise to the ECG which may have affected beat detection and subsequent analysis. However, it is thought that any signal noise caused only minimal disruption. As no significant changes were seen in the ECG recorded measurements (HR and HRV) with the addition of S3D, it is possible that either the ECG equipment was not sensitive enough to detect any changes or that the task itself was not difficult enough to induce any changes. The ECG equipment used in this research has been used in extensively to successfully measure changes in HR and HRV in a research capacity. Thus, there is no reason to doubt the ability of this ECG equipment to detect a significant change if one exists. The simulated flying task employed in this research appeared to be sufficiently cognitively demanding, based on peripheral skin temperature changes and anecdotal participant feedback. It is likely that the warning alert task itself was not sufficiently taxing to impact cognitive workload, however, it was felt to be representative of the current operational environment. A future option could be to increase the number and complexity of other concurrent activities in a simulated flight task.

Finally, it is worthwhile highlighting that this research was conducted during the height of the global COVID pandemic, between 2020 to 2022. The pandemic was an unprecedented event, which challenged every aspect of society and made it difficult to conduct human subject research. It is testament to the professionalism of OBVA laboratory personnel and support from management that research was able to continue during this challenging time.

6.5. Alternative options to display warning alerts

Perhaps, instead of using S3D to improve performance in a warning alert task, it is worth considering other monocular methods that could prove to be more compelling and useful. Options include manipulating colour, changing the relative size of a target, or using frontal plane oscillation. While a full review of colour and colour vision is beyond the scope of this work, it is well known that using a different colour in a display can be effective to highlight information (Christ, 1975; Carter and Carter, 1988). In a search task, differentiating with colour can improve performance in target identification (Green and Anderson, 1956; Treisman, 1982; D'Zmura, 1991) and is more effective than using different shapes to differentiate targets (Christ, 1975). However, caution must be taken when experimenting with use of colour; as certain colours have their own associated meaning. For example, red is linked with 'Stop' and 'Bad/enemy', whereas green relates to 'Go' and 'Good' (Geiselman and Post, 1999; Meola, 2005). In addition, colour perception is influenced by symbol size and contrast with surrounding colours as well as individual differences in contrast and colour sensitivity (Carter and Carter, 1988). In a see-through HMD, there is the added challenge of visibility when overlaid onto the real world; certain colours desaturate more than others in high-ambient light environments, making them harder to perceive (Livingston, 2006; Foote and Hoffmann, 2018; Moffitt and Browne, 2018)

Change in relative size is an option that may best complement the use of S3D as the two occur simultaneously when viewing the real world; when an object comes towards you and retinal disparity increases along with differences in intraocular velocity, the object appears to get larger in size. Both change in size and S3D characteristics are independently able to provide a sensation of motion-in-depth, which may be more compelling when combined. Change in size or 'looming' can be a powerful cue, but also becomes ineffective at frequencies above 3 Hz (Regan and Beverley, 1979). The addition of changing size to changing disparity also improves vergence eye movements which are consequently more accurate and reliable (Erkelens and Regan, 1986), while the motion-in-depth S3D component makes motion more compelling and realistic. However, stimuli utilising these characteristics require longer duration exposures than used here to be most effective (Regan and Beverley, 1979), which may not be compatible with an operational aviation warning alert task. It is also worth considering that changing size in addition to S3D could make text harder to read. In addition, there is large variability in sensitivity across the visual field to these cues (Thompson *et al.*, 2019), so it is important to consider where on a display a target or alert using these differing characteristics should be placed. Frontal plane oscillations, that is moving from side to side in the same visual plane, could also be considered. Humans are most sensitive to frontal plane oscillations at frequencies of 1-4 Hz but can perceive motion at much higher frequencies than S3D, up to 25 Hz (Regan and Beverley, 1973). However, while it may be easier to detect motion with frontal plane oscillations, again, it may be considerably harder to read text or differentiate between items, particularly at higher frequencies.

6.6. Stereo acuity tests

In addition to no overall effect on performance being found using S3D in an HMD for a warning alert task, no relationship was identified between performance and measured stereo acuity using any of the four stereo tests. It is worth noting that those with the very worst stereo acuity did give some subjective indication they found some of the S3D tasks harder. However, the level of stereo acuity in the individuals who reported finding the task harder were far outside current military aircrew selection standards, and effects were not consistent across conditions, thus, they are not considered relevant findings.

While no performance benefit was found using S3D in this warning alert task, based on its benefit to depth related tasks, stereo acuity itself remains an important characteristic for aviators, worth continuing to mandate as a selection criterion. In addition, there may be other applications of S3D in an HMD or cockpit display that could significantly improve performance and would require adequate stereo acuity of the user to be effective. Indeed, lower stereo acuity thresholds have been shown to predict better performance using a S3D display for both an object placement task (McIntire, Wright, *et al.*, 2014) an aerial refuelling task (Winterbottom, 2015; Winterbottom *et al.*, 2016; O'Keefe *et al.*, 2022) and in other depth related surgical tasks (Sakata *et al.*, 2017;

Alhusuny *et al.*, 2021). In order to discern if any relationship exists between performance in an operational task using a stereo display HMD related operational task and individual stereo acuity, a reliable stereo acuity test is required.

This research has identified many limitations with current natural view and paper-based testing methods; all of which possess monocular cues. The presence of monocular cues predisposes a test to giving false positive results; that is to say they overestimate stereo ability (Fawcett and Birch, 2003). The ability to cheat and memorise results provides another opportunity for paper-based tests to give false positive results. In particular, the paper TNO test is unreliable, with poor repeatability (Antona *et al.*, 2015), and is likely to produce inaccurate results due to intraocular luminance differences with the anaglyph glasses (Vancleef *et al.*, 2017; Zhang, Meng and Wu, 2021) and variations in the print quality (Van Doorn *et al.*, 2014). Furthermore, the six-page version employed by the U.K. MoD has a noteworthy floor effect, so is unable to discern stereo acuity any better than 60 arcseconds. Thus, it is unable to adequately monitor changes in stereo acuity over time and cannot be used to investigate relationship with any operational performance metrics. The TNO test is not able to reliably measure stereo acuity to the degree needed to support aeromedical decisions concerning an individual's stereo ability and suitability for pilot training, and is no longer an appropriate tool for aeromedical screening in an increasingly digital cockpit environment. In stark contrast, computer-based methods are able to measure stereo acuity to threshold and offer a more reliable alternative to paper-based stereo tests. Measuring stereo acuity to threshold enables through career assessment, and the ability to track improvements or degradation over time. All digital based tests used correlated well with each other, yet the dRDS-D is likely to be the test least affected by monocular cues, recording the highest stereo thresholds of the digital tests. Using more reliable stereo tests able to measure to threshold will ensure future research in this area will produce meaningful results. Consequently, such research will be able to provide reliable evidence to inform stereo acuity standards which may then also be able to predict performance in operational tasks.

6.7. Conclusion and statement of contribution

By addressing limitations in previous studies, it has been demonstrated that the addition of disparity, or S3D, to a warning alert task in an HMD did not affect performance as measured by response time and throughput of correct responses. This was true whether the disparity used was static relative to the reference plane, or presented with motion-in-depth. It is likely that stereopsis requires more time than afforded by the 1 Hz oscillation used in this task to be effective. Furthermore, S3D did not appear to affect cognitive workload, measured both objectively and subjectively. However, there is some suggestion that those with the worst stereo acuity subjectively found the task to be harder in some of the S3D conditions.

This work adds to the known literature of both S3D technology and use of HMDs in a military aviation environment. It does not support the addition of S3D to a warning alert displayed in an HMD. Nonetheless, this is an important contribution to avoid unnecessary investment as adding new flight hardware and software to support new aircraft features, such as an S3D HMD, is costly due not only to development costs, but also rigorous reliability and hardening requirements, and extensive flight testing before it can be approved for flight. Unfortunately, the impact of design changes on human performance is often overlooked in military acquisition. This knowledge will help inform future uses of S3D in HMDs while setting out alternative display configurations to explore such as investigating the use of colour coding, or the use of S3D in a different context.

While stereo acuity measured with four different stereo tests did not predict performance in a flying and navigation as well as a warning alert task, this work identified many limitations with current stereo acuity test methods and supports the use of computer-based tests in general, which are able to measure to threshold and are more reliable. Using tests which produce reliable and repeatable results is important going forward when investigating relationships with performance metrics. The potential ability of alternative vision metrics to predict performance should be investigated as well as other types of stereo acuity tests.

6.8. Future research

There remain many other promising applications for S3D in an HMD in the aviation context, including using sensor imagery for air-to air refuelling, formation, and low-level flying in a degraded visual environment. It would also be interesting to investigate performance using S3D in a locator line or HITS task assessed against individual stereo acuity. Specific to a warning alert task displayed in an HMD, there are other manipulations that could be explored such as the use of colour, size change, and frontal plane oscillation. It is expected that these findings on the utility of a stereo display in an HMD would be applicable to other ground-based, non-aviation applications. Any future work investigating performance in a visual task should properly characterise relevant individual vision metrics with a reliable test method.

7. APPENDICES

Appendix A: Pilot study

As a way of testing procedures, to ensure suitable experimental conditions, and to establish an effective data management plan, a pilot study was conducted using six participants with a range of stereoacuities (range 1.4 to 2.4 log arcseconds, or 28 to 279 arcseconds). Seven experimental conditions were chosen to be tested as set out in Table 33. Each condition comprised of 40 trials, split 50:50 FLAP: FUEL presented in a pseudo random way, and the order was balanced between participants.

Table 33. Experimental conditions for the pilot study.

FLAP symbol presentation	**FUEL** symbol presentation
2D: Flashed on/off at 1Hz	**Control**: Flashed on/off in 2D at 1 Hz
	Intermittent 500: Flashed on/off with 500 arcseconds of disparity at 1 Hz
	Intermittent 2500: Flashed on/off with 2500 arcseconds of disparity at 1 Hz
	Persistent 500: Flashed in/out of depth, 2D to 500 arcseconds of disparity at 1 Hz
	Persistent 2500: Flashed in/out of depth, 2D to 2500 arcseconds of disparity at 1 Hz
	Motion-in-depth 500: Moving in/out of depth smoothly to 500 arcseconds of disparity at 1 Hz
	Motion-in-depth 2500: Moving in/out of depth smoothly to 2500 arcseconds of disparity at 1 Hz

In the control condition both symbols flashed on and off in 2D. In the rest of the other experimental conditions the FUEL warning symbol was flashed in three different ways. In the intermittent conditions FUEL was flashed on and off but at when on it was with a fixed disparity, at depth. In the persistent conditions the FUEL symbol alternated between two fixed disparities and in the motion-in-depth conditions the FUEL symbol oscillated sinusoidally with motion-in-depth between two disparities. In each of these three types of experimental conditions, two different disparity level were used, 0.13° and 0.69° (500 and 2500 arcseconds), resulting in seven conditions in total. General feedback was that the intermittent conditions were largely unnoticeable, the persistent conditions appeared jittery and only the motion-in-depth condition appeared to have any perceived depth. There were apparent differences between conditions but not between disparities within a condition type. In addition, the duration of each pilot study session was approximately 2 hours 45 minutes which is too long for a mentally engaging and visually stressing task, such that there was the potential for fatigue effects to impact findings. Consequently, the number of conditions was reduced to the four used in the main study as set out in
Table 7, using only one alternative disparity level of 0.69° (approx. 2,500 arcseconds) for each presentation type; intermittent, persistent, and motion-in-depth.

Appendix B: Inherent system misalignment

There is inherent misalignment present between two optics in any HMD. It was not possible in the time frame to create a unique warp map to counter this type of inherent misalignment in the system. The generic warp map supplied by SA Photonics was used, but a small degree of distortion was still present. To minimise any confounding effects of image distortion, attempts were made to characterise the inherent misalignment within the HMD headset, in order to place the warning alert symbology in the area least affected. To do this a series of numbered boxes were displayed in each optic, having been aligned using the process set out in Section 3.5: Alignment process. When viewed together, the boxes which appeared diplopic identified the areas with greatest distortion. The warning alert symbol was placed in the area with least distortion indicated by the best aligned numbered boxes (Figure 42).

Figure 42. Co-ordinated numbered boxes were placed in the same position of each optic. When the HMD was well aligned, some inherent distortion was apparent, as indicated by the diplopic boxes.

Appendix C: Crosstalk of ASUS 3D glasses

Crosstalk measurements were taken of the NVIDIA 3D vision shutter glasses (ASUS, CA, USA), with results illustrated in Figure 43 and Figure 44. On the whole, crosstalk is below the 5% level known to cause visual discomfort (Kooi and Toet, 2004).

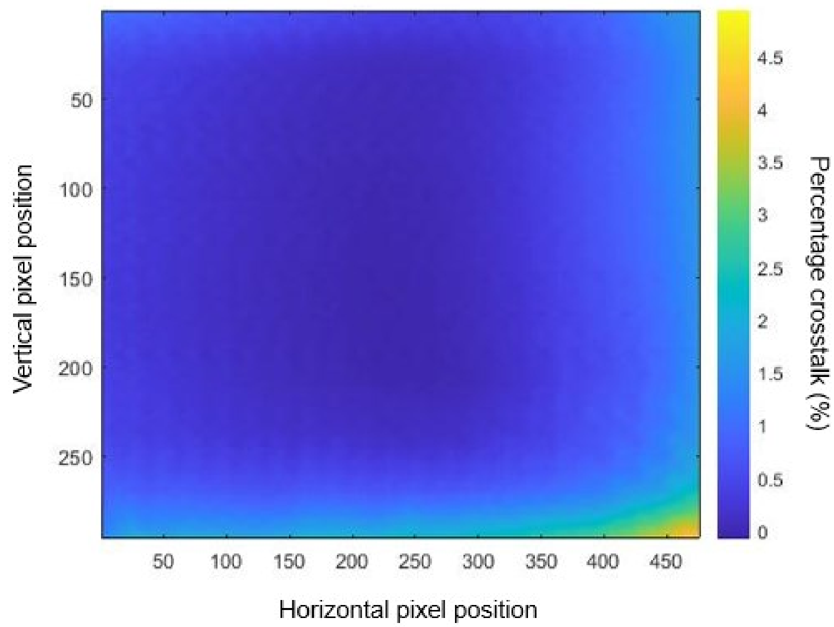

Figure 43. Crosstalk for right eye of NVIDIA 3D vision shutter glasses.

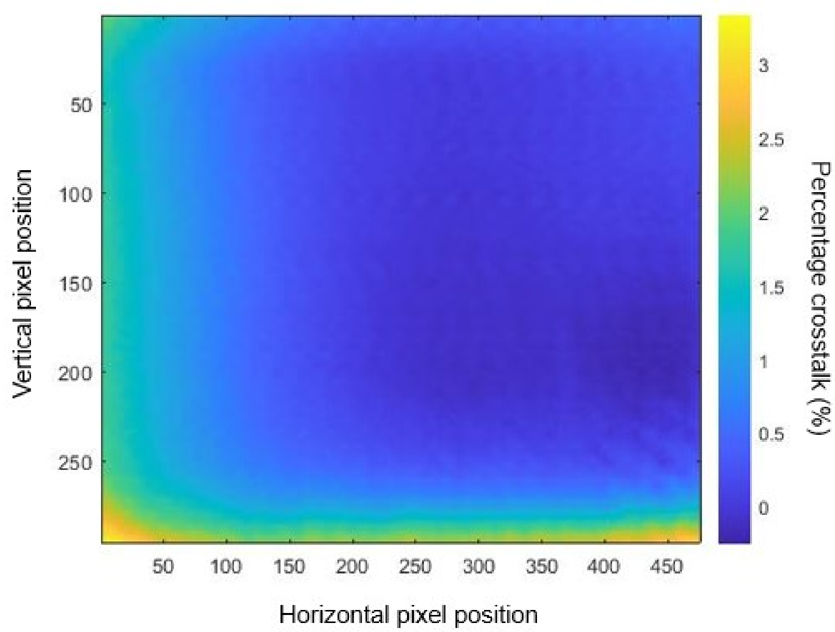

Figure 44. Crosstalk for left eye of NVIDIA 3D shutter vision glasses.

Appendix D: Questionnaire responses

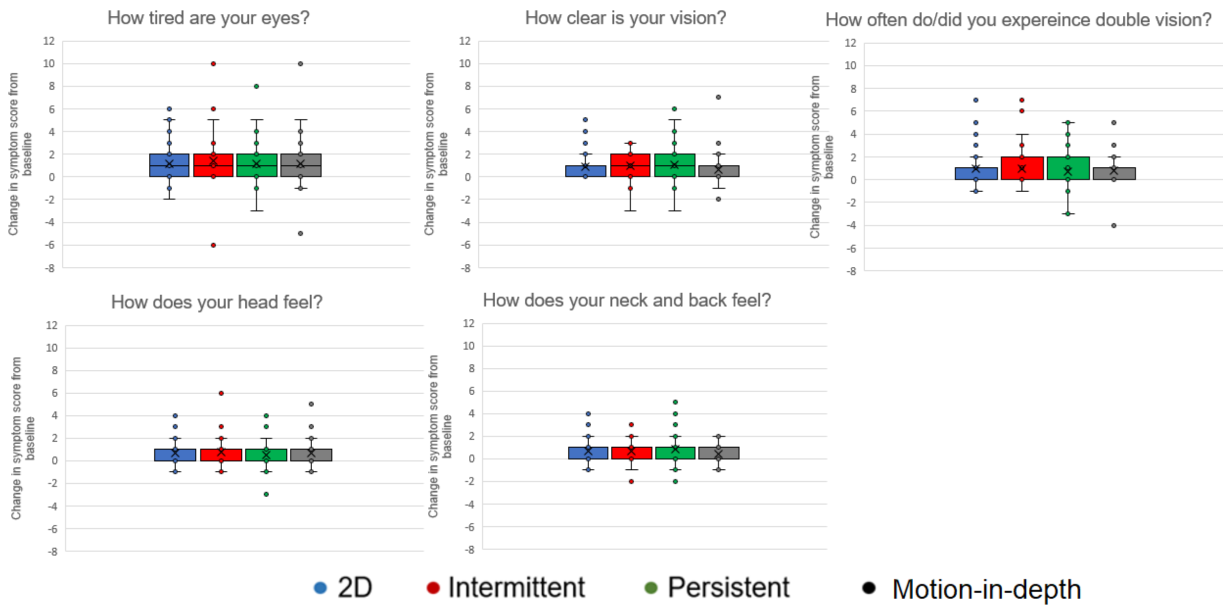

Figure 45. Change in symptoms scores after each condition from baseline value

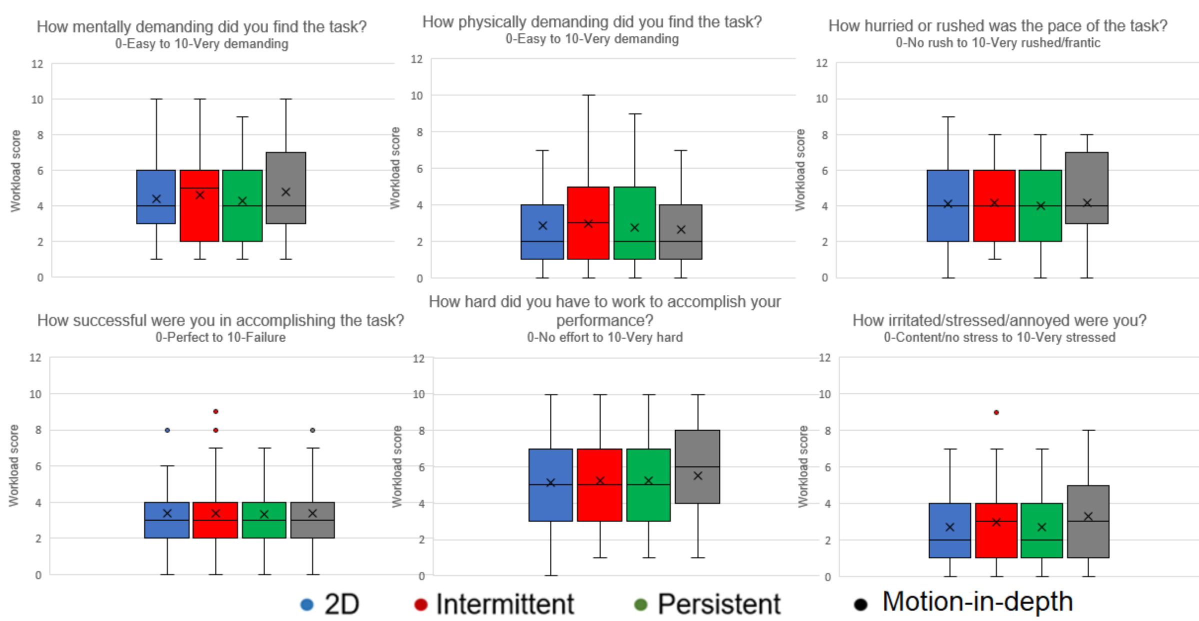

Figure 46. Workload scores after each condition.

Appendix E: Relationship between individual stereo acuity and questionnaire responses

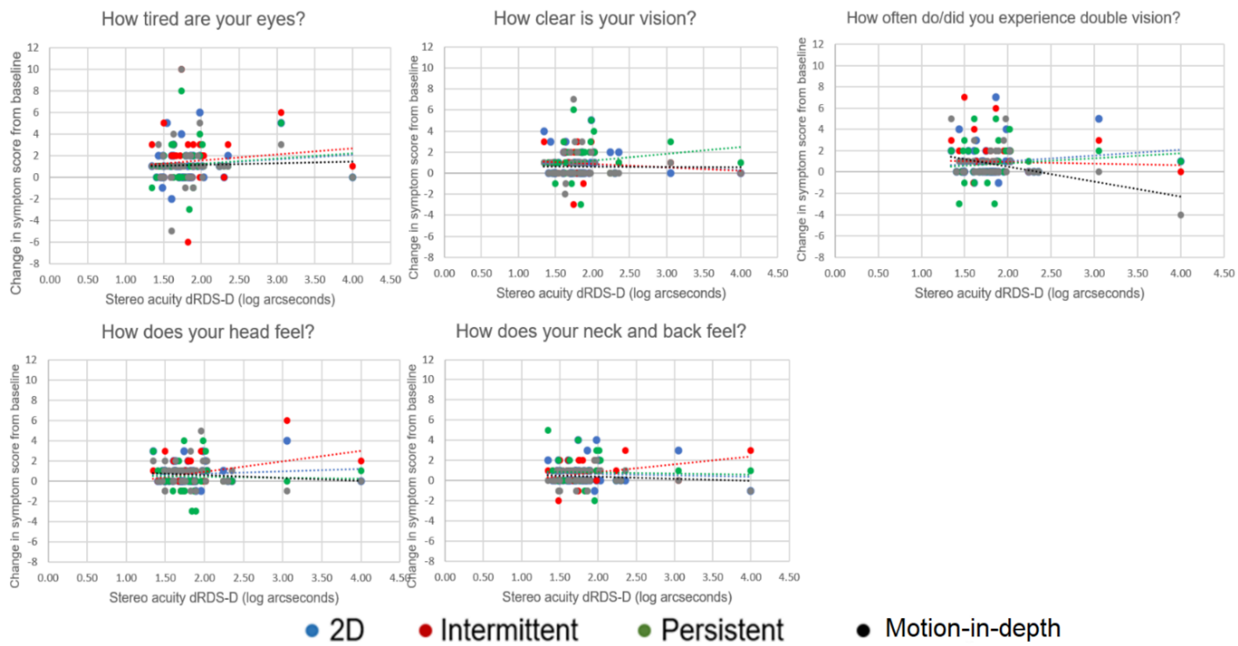

Figure 47. Individual stereo acuity against change in symptom score from baseline.

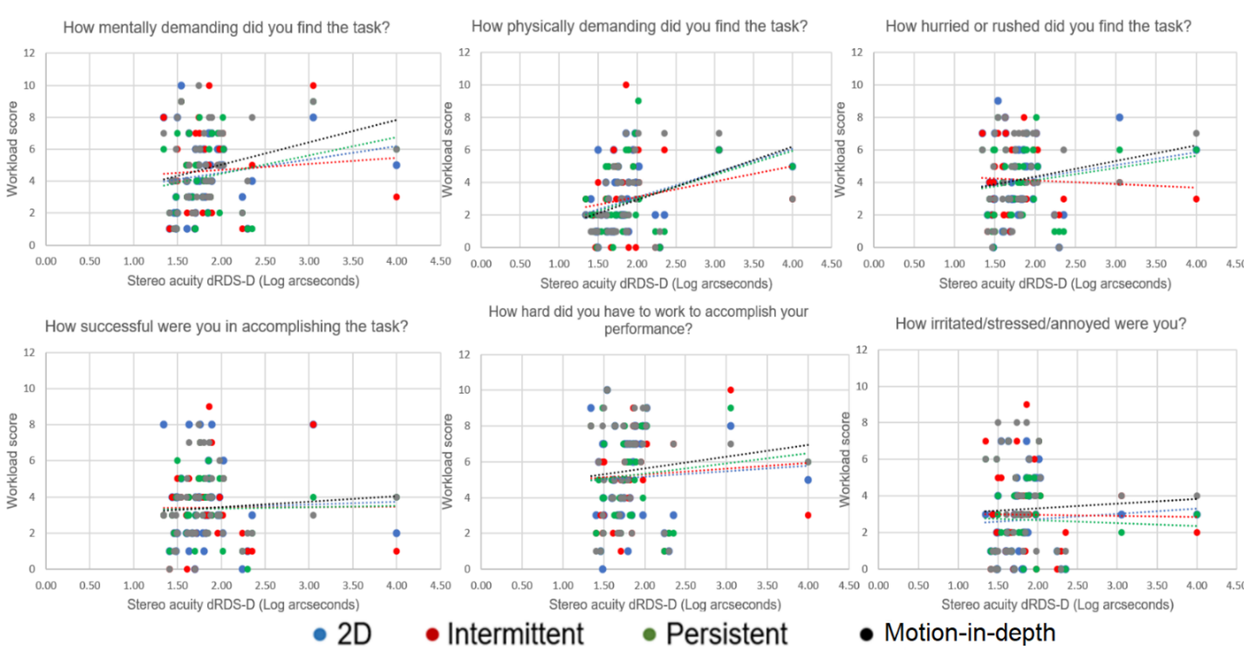

Figure 48. Individual stereo acuity against workload score.

Appendix F: Relationships between individual stereo acuity and objectively measured physiological variables

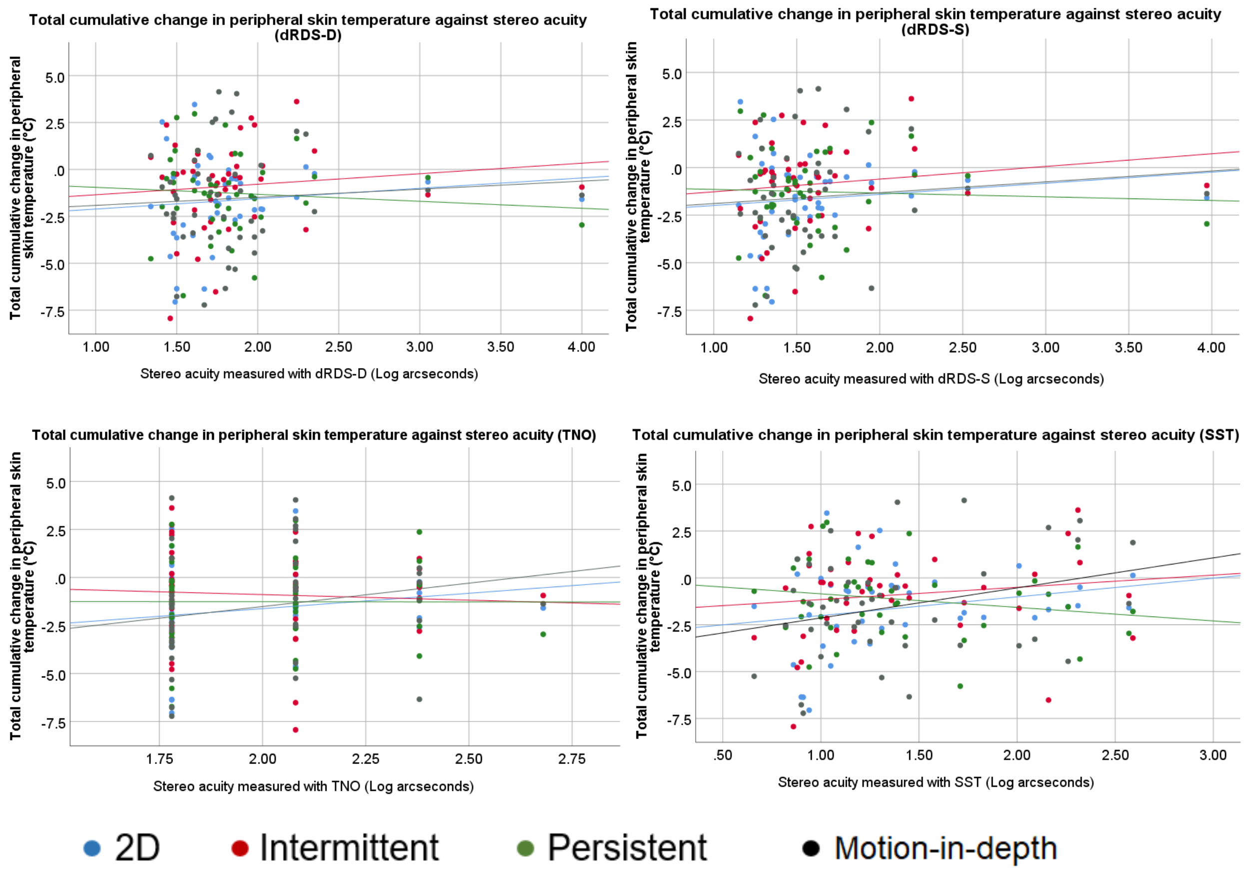

Figure 49. Scatter plots of change in peripheral skin temperature, in all four conditions and individual stereo acuity as measured with measured with dRDS-D (A), dRDS-S (B), TNO (C), and SST (D).

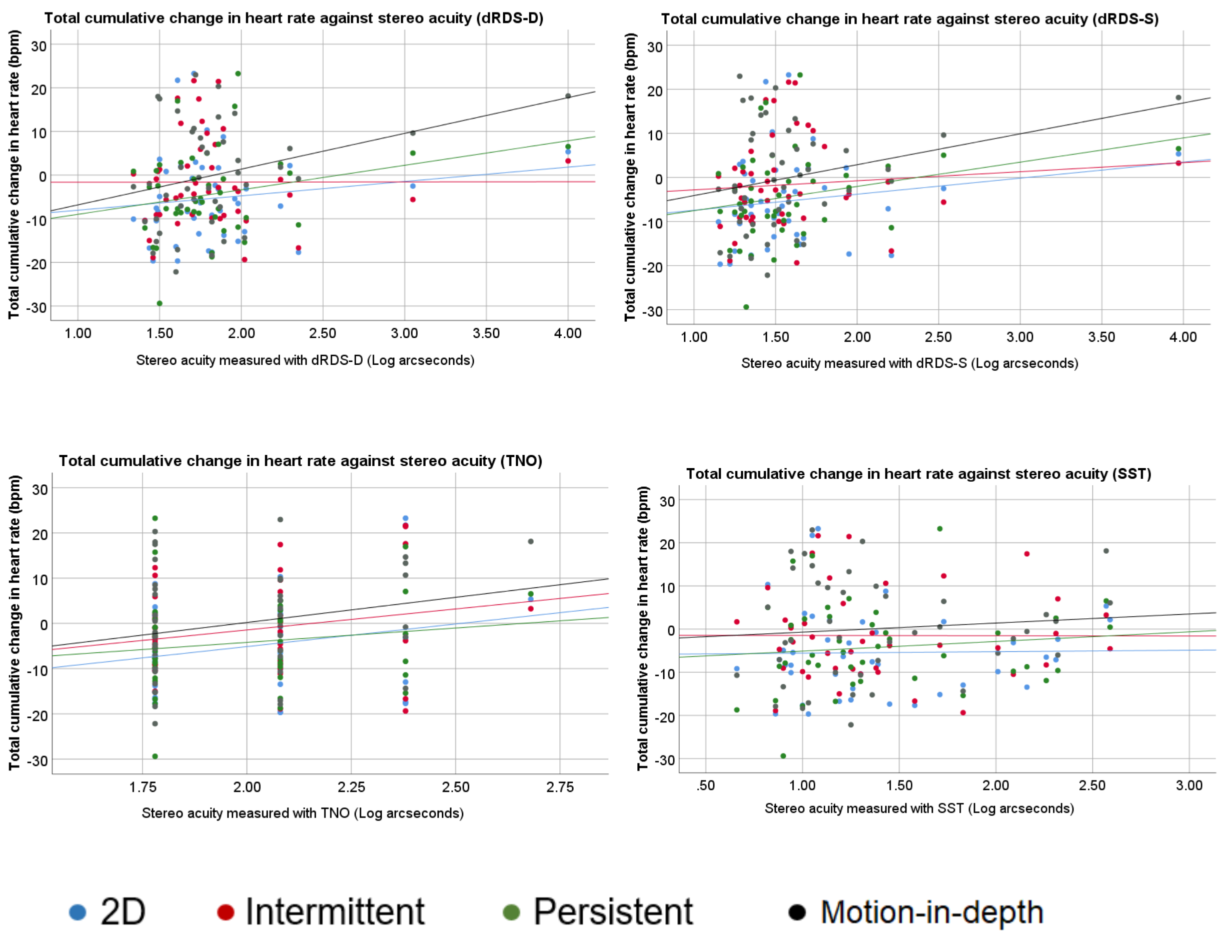

Figure 50. Scatter plots of change in heart rate, in all four conditions and individual stereo acuity as measured with measured with dRDS-D (A), dRDS-S (B), TNO (C), and SST (D).

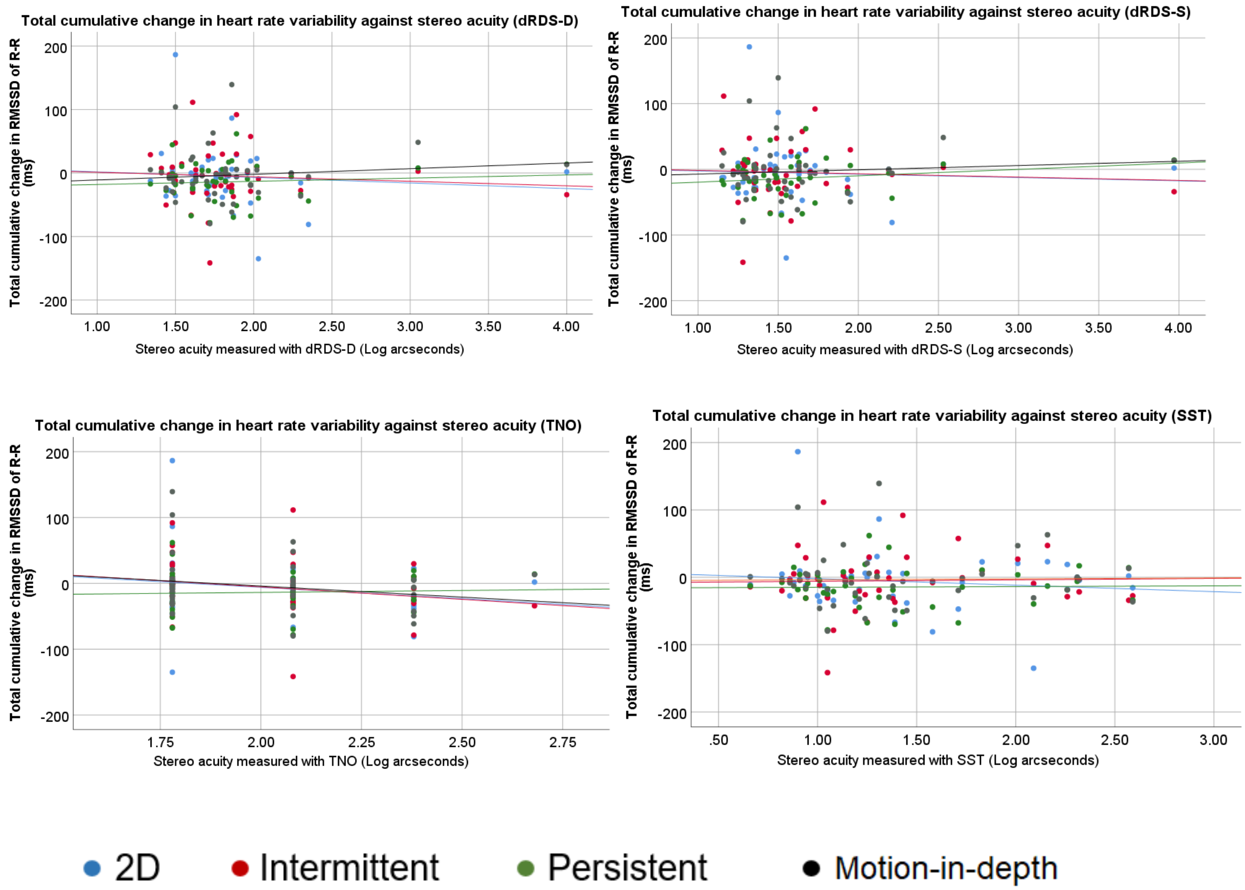

Figure 51. Scatter plots of change in heart rate variability, in all four conditions and individual stereo acuity as measured with measured with dRDS-D (A), dRDS-S (B), TNO (C), and SST (D).

Appendix G: Participation information and consent sheet

Abbreviated Consent for Research Concerning OBVA Helmet-Mounted Display Visual Performance Research

711 HPW/RHBC, Wright-Patterson AFB, OH

- Principle Investigator: Mr. Charles Bullock, DR2, Software Engineer, 711HPW/RHBC, [redacted]

You are being asked to participate in a research study.

Key study information you should know:

- The purpose of this research is to examine the relationship between existing clinical measures of visual capabilities and ocular alignment and performance on an operational task similar to that expected to be performed by an F-35 pilot using a Helmet-Mounted Display System (HMDS). The feasibility of using newly developed monitor-based measures of depth perception, visual suppression and useful field of view developed in the OBVA lab will also be examined.

- Participation in this study will consist of 6-12 sessions, ranging from 30 minutes in length to a maximum of 4 hours. You may take rest breaks at any time and are free to discontinue participation without penalty and will be paid in full, if applicable, for all time completed. If you participate in this research you will also be asked to undergo a battery of tests to evaluate your vision. Depending on your vision tests results, you may or may not be asked to perform the operational F-35 task.

- There are minimal risks involved in participating in this experiment. These risks are possible eye fatigue, eye-strain, and/or headache as a result of viewing the HMDS or computer monitors in this experiment. The frequency of occurrence of such risks is no more likely than that which you might experience when using a computer and viewing a computer monitor, or while watching television or 3D movies. The risks from using the heart rate and EKG include minor skin irritations from surgical tape attached to the hand and/or EKG pads attached to the arm.

- Complete confidentiality for military personnel cannot be promised because information bearing on their health may be required to be reported to appropriate medical or command authorities. Airmen on flight/PRP status (Personnel Reliability Program) participation in this study may reveal some reportable fact that could affect their status

(Duty Not Including Flying; Medical Evaluation Board Waiver; or permanent loss of rating). However, because most of the vision tests are either non-standard, or administered by non-medical personnel, this risk is minimal.

- You are not expected to benefit directly from this research. You will be compensated $30 per hour for your time. If you are a Federal Employee, DoD contractor, or active duty military member you may be compensated for your time, but only during "off duty" hours, which can be verified by completion of an AF Form 3902.

- Taking part in this research is voluntary. You can discontinue participation at any time without penalty or loss of benefits to which you are otherwise entitled.

Please take time to read this document and ask questions before deciding whether to take part in this research.

The researchers will take the following precautions to maintain the confidentiality of your data: In a previous protocol (FWR20170095H), you had already agreed to include your information in a re-contact registry/data repository upon agreeing to participate in the prior vision related study in which you signed a consent form. For this particular study, the researchers will not collect any identifiers linked to you. No participant identifiers will be included in any publications. Electronic data will be password-protected. Other data will be kept in a locked cabinet.

The data may be accessed by the Department of Defense for auditing purposes.

If you have questions regarding the study, contact the Principal Investigator, Charles Bullock (listed on the first page of this consent document), or Associate Investigator, Marc Winterbottom, at [redacted], or email at [redacted]. If you have questions regarding your rights as a research subject, contact the Air Force Research Labs Institutional Review Board (AFRL IRB): [redacted] or email [redacted] An IRB is an independent ethics and regulatory committee that has reviewed this protocol.

Automated Vision Test Battery Research and Development Data Registry and Subject Re-Contact Database

Informed Consent Document

For

Automated Vision Test Battery (AVT) Research and Development

Operational Based Vision Assessment (OBVA) Laboratory, 711 HPW USAFSAM/FHOH, Wright-Patterson AFB, OH

1. PRINCIPAL INVESTIGATOR

Dr. Marc Winterbottom, DR3, Sr. Research Psychologist, USAFSAM/FHOH, [redacted]

2. ASSOCIATE INVESTIGATORS

Dr. James Gaska, DR4, Senior Scientist, USAFSAM/FHOH, DSN [redacted],

Ms. Elizabeth Shoda, Contract Research Specialist, KBRwyle Laboratories, [redacted]

Dr. Eleanor O'Keefe, Contract Research Associate, KBRwyle Laboratories, [redacted]

Dr. Charles Lloyd, Contract Research Scientist, Visual Performance LLC., [redacted]

Dr. Eric Palmer, Contract Research Scientist, STI-Tech, 937-255-0333, [redacted]

Mr. Jefferson Fu, Contract Research Assistant, KBRwyle, [redacted]

Mr. Gregory Eisenhauer, Contract Research Assistant, KBRwyle Laboratories, [redacted]

Mr. Matthew Ankrom, Contract Research Assistant, KBRwyle Laboratories, [redacted]

INTRODUCTION

Many current Air Force standards and screening tests date back to the World War I and II eras, despite the fact that new technologies exist which may better and more accurately measure visual health. In addition, all of these tests are manually administered, which diminishes the accuracy and reliability of the tests, but may also limit the ability of the test to predict operational performance.

Automated Vision Test Battery Research and Development Data Registry and Subject Re-Contact Database

Furthermore, some existing standard tests may not actually test the aspect of vision they are intended to test. Current manually administered tests are based on paper charts or glass slides which have not changed since the 1940s.

The goal of this research is to create a combined data registry and subject re-contact database to assist with developing and validating a battery of automated vision tests (AVT) to supplement or replace existing US Air Force standard vision tests. Normative data will be collected, correlations between AVT tests and comparable, existing standard vision tests will be examined, test-retest reliability will be evaluated, and in some cases practice/training effects will be evaluated.

You are being asked to participate in this project to gain information that may assist the United States Air Force (USAF), and eventually other military services, advance and improve its vision testing for military members.

PURPOSE

The purpose of this research is to develop and validate a battery of automated vision tests (AVT) to supplement or replace existing US Air Force (USAF) vision test batteries, and place that data into a data registry. Additionally, the second purpose of this study is to create a subject re-contact database in order to contact subjects that have already completed the AVT protocol and offer participation in follow-on studies.

Normative data will be collected (i.e. the range of typical performance will be identified), correlations between AVT tests and comparable, existing standard vision tests will be examined, test-retest reliability will be evaluated, and in some cases practice/training effects will be evaluated. This research will help the USAF refine medical vision standards and ensure that aircrew have the visual capabilities required to perform flight tasks under what may be very challenging combat conditions.

PROCEDURES

At the beginning of the study, you will be given a short in-processing questionnaire to collect demographic information and relevant ocular medical history and experiential data. This questionnaire should take approximately 10 minutes to complete. Following the questionnaire, a number of vision screening tests will be administered to you by a professional optometrist or other vision specialist. In some cases, you may be excluded from the study if your vision does not test as normal (or corrected to normal), or if your vision *does* test as normal, as we are seeking participants with certain vision deficiencies (e.g. stereo-blind or stereo-anomalous). You will also be excluded if you fall outside the age requirements of 18 years of age or older and if you are not a U.S. Citizen. If you are a paid volunteer, you will be compensated for your time during the testing procedure. Paid volunteers include non-government civilians and Department of Defense civilians and contractors who are participating during off-duty hours. Active duty military and National Guard members/reservists on federal orders are also eligible for payment upon the completion of a valid AF Form 3902. If you are employed by the Air Force as an enlisted Airman, officer, civilian, or contractor, and are participating during duty hours, you will have to arrange for the time to participate through your supervisor or host.

Automated Vision Test Battery Research and Development Data Registry and Subject Re-Contact Database

The time requirement for this research for each volunteer is anticipated to be 1 to 6 visits of approximately 1 hour to 4 hours each. We are seeking as many as 1.000 subjects.

If you decide to participate, after completing the vision screening, you will be asked to complete one or more of the following vision tests:

Automated Vision Tests:

1) Visual Acuity Test. Computer based test where a high contrast letter C's size is varied to estimate the smallest letter C for which you can reliably identify its orientation (left, right, up, or down). A standard Windows-based PC and monitor will be used to operate the test and record your responses. You will enter your responses using a standard keyboard, game controller, or joystick.
2) Contrast Sensitivity Test. Computer based test where a letter C's contrast (i.e. visibility) is varied according to an adaptive threshold procedure to estimate the minimum letter C contrast at which you can reliably identify its orientation (left, right, up, or down). A standard Windows-based PC and monitor will be used to operate the test and record your responses. You will enter your responses using a standard keyboard, game controller, or joystick.
3) Adaptive Sensory Technology (AST) Contrast Sensitivity Test. Computer based test where a letter C's contrast is varied according to an adaptive threshold procedure to estimate the minimum contrast at which you can reliably identify the letter's C orientation (left, right, up, or down). A standard Linux-based PC will be used to generate the letter and compute the contrast. A large monitor will be used to display the test letters. You will enter your responses using a standard game controller.
4) Color Vision Test (OBVA CCT). Computer based test where a letter C's color contrast is varied according to an adaptive threshold procedure to estimate the minimum color contrast at which you can reliably identify the letter C's orientation (left, right, up, or down). The test is designed to estimate your sensitivity to 3 colors - red, green, and blue. A standard Windows-based PC will be used to generate the letters and compute contrast to estimate your just noticeable level of contrast for each color. The color test will be displayed on a standard computer monitor. Your responses will be entered using a standard keyboard, game controller, or joystick.
5) Stereo Acuity Test. Computer based depth perception test where the depth between two circular stimuli is varied to estimate your stereo acuity (i.e. estimate the smallest difference in depth you can reliably discern). The stereo acuity test will be displayed on a standard 3D video monitor. Similar to a 3D movie, you will need to wear special shutter glasses to view this 3D test. You will enter your responses using a standard keyboard, game controller, or joystick.
6) Fusion Range Test. A computer-based test designed to estimate your ability to maintain a single, binocularly fused image. A standard Windows-based PC will be used to generate the circular stimuli displayed separately to your left and right eyes. When the test begins, the left and right eye circles are aligned. The circles then move

apart, and you simply indicate when you can no longer maintain fusion (i.e. you experience double-vision), and indicate fusion break by pressing a button. The circles then move in the opposite direction until you indicate fusion recovery (i.e. the two circles appear to become one, or fuse) by pressing a button a second time. The fusion range test will be displayed on a standard 3D video monitor. Similar to a 3D movie, you will need to wear special shutter glasses to view this test. You will enter your responses using a standard keyboard, game controller, or joystick.

7) Motion Sensitivity Test. Computer-based test where the proportion of dots moving in the same direction within a field of randomly moving dots is varied according to an adaptive threshold procedure to estimate the minimum proportion at which you can reliably identify the direction (expanding, contracting, clockwise or counter-clockwise rotation). The motion test will be displayed on a standard video monitor. The test is designed to estimate your sensitivity to "motion flow". Your responses will be entered using a standard keyboard, game controller, or joystick.

8) Automated Phoria Test. An automated test using infrared cameras and shutters to measure the direction of gaze of each eye before and after the shutter blocks the fixation target. This test is intended to automate the alternating cover test used by optometrists to measure phoria (i.e. the slight muscle imbalance of the eyes). This test uses non-invasive (web-cam based) eye tracking methods developed by the OBVA Lab. Because the eye-tracker uses an infrared illuminator, there is a minimal risk of the eyes drying. To reduce the possibility of dry eyes, be sure to blink between trials, or notify the experimenter that you would like to take a break. You may discontinue the test at any time if you feel any discomfort. The infrared LEDs are very low power, and infrared illumination is unlikely to be an issue, or even noticeable at all.

9) Automated Fusion Range Test. An automated test using infrared cameras to measure eye position as two circular stimuli, presented separately to each eye, move apart in either the horizontal or vertical directions. This test is intended to automate the self-report fusion range test described above (Automated test #6), and obtain an objective measure of fusion break and recovery. This test uses the same apparatus described above for the automated phoria test. As above, the eye-tracker uses an infrared illuminator posing little risk to the eyes. There is a minimal risk of the eyes drying. As with automated test #8, be sure to blink between trials, or notify the experimenter that you would like to take a break. You may discontinue the test at any time if you feel any discomfort.

10) Ocular dominance test. An automated test using stimuli of various contrasts presented separately to each eye. A similar device as described in Automated Test numbers 8 and 9 above will be used, but does not require the use of the cameras, IR illumination, or shutters. An adaptive threshold procedure will be used to adjust either the contrast, orposition of the left vs. right eye stimuli to estimate the contribution of each eye to perceived position of the binocular stimulus.

11) Eye Movement And Intrinsic Latency test (EMAIL). An automated test measuring oculomotor features such as visual processing times and eye movement latencies. Your

reaction time for moving your eyes from a central fixation target to a peripheral target and responding to a Landolt C (left, right, up, or down) will be recorded.

Clinical Vision Tests:

1) Armed Forces Vision Tester (tests of near/far acuity, stereo acuity, and phoria). This test is similar to commonly used driver licensing vision tests.
2) AO Vectograph stereo acuity test. Similar to the stereo acuity test described above, with circular stereo acuity targets, but projected onto a screen at a far viewing distance (20 feet). Simply identify which circle appears to be popped out in depth in each row.
3) Eye lane tests of phoria and fusion range using typical vision exam equipment such as a phoropter and optotype projector. Administered by an optometrist.
4) Precision Vision contrast and acuity eye charts (illuminated using a light box).
5) Rabin CCT color test. Computer-based test where color contrast of a series of letters is selected over 5 levels according to a custom staircase procedure to estimate the minimum color contrast at which you can reliably identify orientation (left, right, up, or down). Similar to automated color test #4 above, the test is designed to estimate your sensitivity to 3 colors - red, green, and blue. A standard Windows-based laptop/netbook will be used to select contrast level and display the letters for each color. You will use a mouse/trackpad to select the correct letter on each trial.
6) Colour Assessment and Diagnosis (CAD) Test. Computer-based test where the color contrast of a moving box against a noisy background is adjusted to estimate your sensitivity to several different colors. Look carefully to identify the direction of motion (diagonally up and right, down and right, up and left, or down and left) of the moving box. A standard color monitor will be used to display the test stimuli. Your responses will be entered using a number pad.
7) Pseudoisochromatic plates (PIP) color test. The score is based on the number of correctly identified letters/numbers in a series of plates. The letters/numbers consist of colored dots presented against a background of similarly colored dots.

In-Processing Questionnaire:

1) Demographic questions including sex and date of birth.

2) Technology use questions including familiarity with technology, number of hours spent playing video games, amount of eye strain normally experienced at work, and complications associated with viewing 3D movies. These questions are relevant to our research because they can help rule out confounds in our data due to difficulties with experimental apparatus, spurious sources of eye strain, and issues viewing our monitors.

3) Vision history questions including past ocular surgeries, ocular medication usage, known color or depth deficiencies, use of medications that may affect vision, and

history of head injuries. These items are all self-report and will be kept confidential and de-identified as specified by procedures in this protocol.

4) Aviation experience questions including type(s) of aircraft, crew position(s), flying hours, history of medical waivers or disqualification for visual conditions, and whether currently in duties not to include flying (DNIF) status. Again, these items are self-report and will be kept confidential and de-identified.

Providing Contact Information:

As part of this project, you will also be asked if you would allow your contact information to be entered into a re-contact database, so that researchers can contact you for future IRB approved protocols assessing the same types of vision tests.

POTENTIAL RISKS AND/OR DISCOMFORTS

There may be some discomfort including possible fatigue, eyestrain, nausea, eye dryness, and/or headache as a result of viewing the display systems in this experiment. Infrared light from the infrared LED illuminators pose unlikely risks to your eye (cornea or retina) as the exposure is comparable to that of sunlight and, more importantly, less than a tenth of the safety limit. The frequency of occurrence of such risks is no more likely than that which you might experience when playing a 3D videogame, or watching a 3D television or movie, or being exposed to normal sunlight.

Incidental Findings

A possible incidental finding that could result during experimental participation is the potential discovery of visual dysfunction, such as poor visual acuity or color deficiency. All test data is recorded anonymously so there will be no research data file linking your vision tests to you. **Please understand that there is no plan to provide any vision testing information back to you because the testing is for research purposes only and not clinical in nature.** The only exception to this is for active duty military personnel. For active duty personnel, if a research vision test results in a reasonably obvious abnormality, the active duty subject will be informed of this and will be provided a recommendation to seek a clinical vision test through his/her primary care provider. If a research vision test results in an incidental finding of some vision status that would tend to obviously render the member unfit for duty (e.g., color blindness for some duty AFSCs), the research test result may be reported directly to the service member's supervisor or commander for further follow-up as deemed appropriate. Any instance which warrants reporting a duty limiting incidental finding to a supervisor or commander will be done with the knowledge of the military member subject and in a confidential fashion consistent with the Privacy Act. Concerning the in-processing questionnaire, all medical history questions asked, including those related to medical waivers, disqualifications, DNIFs, and other diagnoses, contain information that should already be known to the Air Force, therefore there are no additional risks from disclosure of this information, Furthermore, none of this information will be reported to your leadership as it is purely for research purposes only.

PREGNANCY RISKS

Automated Vision Test Battery Research and Development Data Registry and Subject Re-Contact Database

This research poses no additional risks to women (or fetus) who are or may become pregnant.

BENEFITS

If you agree to take part in this research study there may be no direct benefit to you beyond the educational experience of participation in a formal research project. However, the information learned from this study may someday help us to better understand visual performance and refine medical vision standards for aircrew.

COSTS

There will be no cost to you for the research study related to visits to our lab, vision tests or study procedures. However, you will be responsible for your transportation.

ALTERNATIVES TO PARTICIPATION

Your alternative is to choose not to participate in this research study. Refusal to participate will involve no penalty or loss of benefits to which you are otherwise entitled. You may discontinue participation at any time without penalty or loss of benefits to which you are otherwise entitled. Please notify one of the investigators of this study if you wish to discontinue participation.

YOUR PARTICIPATION IS VOLUNTARY

The decision to participate in this research is voluntary on your part. No one may coerce or intimidate you into participating in this program. Participate only if you want to. Ms. Elizabeth Shoda, or an associate, should adequately answer all questions you have about this study, your participation and the procedures involved. If you have any further questions, Ms. Shoda can be reached at (937) 255-5014 or elizabeth.shoda.1.ctr@us.af.mil. Ms. Shoda, or an associate, will be available to answer any questions concerning procedures throughout this study. You may withdraw from this research study at any time without penalty.

If significant new findings develop during the course of this research, which may relate to your decision to continue participate or may affect the risk involved, you will be informed. If you have any questions or concerns about your participation in this study or your rights as a research subject, please contact the Air Force Research Laboratory Institutional Review Board (AFRL IRB) at ([redacted] or [redacted] The IRB is an independent ethics and regulatory committee composed of members not associated with the research study. The IRB has reviewed this study.

COMPENSATION

If you are a paid volunteer your compensation for volunteering to participate in this research experiment will be at a competitive rate of $30 / hour issued via a VISA debit compensation card. There is no schedule or milestones required for payment. Participants completing testing at our Wright State University or University of Dayton locations may have the option of receiving either financial compensation or course credit for their participation.

You are responsible for paying any State, Federal, Social Security or other taxes on the payments you receive. You will receive a form 1099 in January of the year following your participation in this study. This form is also sent to the IRS to report any money paid to you. No taxes are withheld from the payment for your participation.

SIGNIFICANT NEW FINDINGS

You will be told by the study investigator or study staff if new information becomes available that might affect your choice to stay in the study.

CONFIDENTIALITY

Identifiable records of your participation in this study may only be disclosed according to federal law, including the Federal Privacy Act, 5 U.S.C. 552a, and its implementing regulations and the Health Insurance Portability and Accountability Act (HIPAA), and its implementing regulations, when applicable, and the Freedom of Information Act, 5 U.S.C. Sec 552, and its implementing regulations when applicable. Identifiable records may include your signed consent document and any documents related to your financial compensation for this study.

Your personal information will be stored in a locked cabinet in an office that is locked when not occupied. Electronic files containing your personal information will be password protected and stored only on a secure server. Organizations that may look at and/or copy your records for research oversight, quality assurance and data analysis include:

- The researchers named above,
- The study's Research Monitor or Consultant,
- The AFRL Wright Site IRB,
- The Air Force Surgeon General's Research Compliance office,
- The Director of Defense Research and Engineering office or
- Other IRB(s) involved in the review and approval of this protocol.

You will be identified by a code, and personal information from your records will not be released without your written permission unless required for military personnel. Information related to health and fitness for duty may be required to be reported to appropriate medical or command authorities. Complete confidentiality for military members cannot be promised. You will not be identified in any publication or in the sharing of your data about this study.

Your participation in this study may be photographed, filmed or audio/videotaped. The purpose of these recordings is for documentation purposes only, and may be used to illustrate the simulated flight task or the vision screening tests in a research publication or presentation.

After the study is completed, the non-identifiable research data may be placed in a central storage location as established by the research team. The purpose is to make study data available to other Air Force researchers and their research partners for the purposes of future health and performance research or other vision standards development activity within and outside of the military, both nationally and internationally. The data may be used for future research not yet planned by the Government, Department of Defense, Commercial entity, or educational or international organizations.

Automated Vision Test Battery Research and Development Data Registry and Subject Re-Contact Database

The shared data will not include your name or other information that can identify you. Because security breaches cannot be completely ruled out, unauthorized access to the vision screening results is a potential risk, but there is no risk of breach of privacy for you because the vision results are no linked to individuals.

FUTURE RESEARCH STUDIES

If you choose to be part of the subject re-contact database, your identifying information name, phone number, address and/or email address) will be stored separately from your vision test and questionnaire data, separated by a master log. The database that will store all data is both CAC and password protected, and only the investigators on this protocol.

Future research will be reviewed by an IRB and there will be a new consent process for those future studies. Either researchers from this study, or other researchers from 711 HPW who have IRB approval could re-contact you to let you know of similar research participation opportunities in the future.

You may update your contact information at any time or you may also change your mind at any time by contacting the Principal Investigator, Dr. Marc Winterbottom, at [redacted] or Ms. Elizabeth Shoda, [redacted] or Dr. Eleanor O'Keefe, [redacted] have your name removed from the list. If at any time the contact information for the database managers change, you will be notified.

Please indicate your permission by initialing below:

_________**I AGREE** to be re-contacted to be informed of future research studies.

_________**I DO NOT** agree to be re-contacted to be informed of future research studies.

We may wish to present some of the video/audio recordings from this study at scientific conventions or use photographs in journal publications. If you consent to the use of your image for publication or presentation in a scientific or academic setting, please sign below.

I, ______________________, agree to allow the use of photographs, video and/or audio records of my participation.

Date________________

STUDY PARTICIPATION AGREEMENT/CONSENT

Automated Vision Test Battery Research and Development Data Registry and Subject Re-Contact Database

Taking part in this research study is completely voluntary. Your signature below shows that:

- ***You agree to be in this study***
- ***The researcher has explained the study to you and you have read and understand the information you have been given***
- ***You were given the opportunity to ask questions about the study and all of your questions have been answered to your satisfaction***
- ***You understand that signing this consent does not take away any of your legal rights***

You will be given a copy of this signed consent form for your records

Volunteer Signature____**Date**_______________

Volunteer Name (printed)____

Advising Investigator Signature ______________________ **Date** _________________

Investigator Name (printed)____

Witness Signature __________________________________ **Date** _________________

Witness Name (printed)____

PRIVACY ACT STATEMENT

<u>Authority:</u> We are requesting disclosure of personal information. Researchers are authorized to collect personal information on research subjects under The Privacy Act-5 USC 552a, 10 USC 55, 10 USC 8013, 32 CFR 219, 45 CFR Part 46, and EO 9397, November 1943.

<u>Purpose:</u> It is possible that latent risks or injuries inherent in this experiment will not be discovered until sometime in the future. The purpose of collecting this information is to aid researchers in locating you at a future date if further disclosures are appropriate.

<u>Routine Uses:</u> Information may be furnished to Federal, State and local agencies for any uses published by the Air Force in the Federal Register, 52 FR 16431, to include, furtherance of the research involved with this study and to provide medical care.

<u>Disclosure:</u> Disclosure of the requested information is voluntary. No adverse action whatsoever will be taken against you, and no privilege will be denied you based on the fact you do not disclose this information. However, your participation in this study may be impacted by a refusal to provide this information.

Appendix H: Recruitment posters

Automated Vision Test Battery Research and Development

The purpose of this research is to develop and validate a battery of automated vision tests to supplement existing, manual Air Force vision tests.

We are seeking individuals with a wide range of visual capabilities. Individuals with reduced quality of vision (e.g., reduced visual acuity, depth perception, color vision, etc.) are welcome to participate.

To participate in this research you must be:

- A U.S. citizen
- 18 years old of age or older

Participation in this study involves:

- Two or more 1-hour sessions at Wright Patterson Air Force Base
- Competitive monetary compensation of $30 per hour
- The potential to participate in future research studies

AFRL IRB Approved (FWR20170095H, Version 1.xx) valid until XXX X, 201X.

Contact Information

To find out more about this study, visit www.obvalab.com or OBVA Lab on Facebook

You may also contact:

Operational Based Vision Assessment (OBVA) Laboratory, USAF School of Aerospace Medicine, 711th Human Performance Wing

Principal Investigator: Dr. Marc Winterbottom

OBVA
Operational Based Vision Assessment

Cleared, 88PA, Case # 2018-0128, 11 Jan 2018.

SEEKING PARTICIPANTS FOR A RESEARCH STUDY

USAF photo by A1C Anthony Jennings

HELMET-MOUNTED DISPLAY VISUAL PERFORMANCE RESEARCH

U.S. AIR FORCE SCHOOL OF AEROSPACE MEDICINE, WRIGHT-PATTERSON AFB

The purpose of this research is to examine the relationship between visual capabilities and performance on operational tasks similar to those expected to be performed by a pilot using a helmet-mounted display system.

We are seeking individuals with a wide range of visual capabilities. Individuals with reduced quality of vision (e.g., reduced visual acuity, depth perception, color vision, etc.) are welcome to participate.

To participate in this research you must be:

- A U.S. citizen
- 18 years of age or older

Participation in this study involves:

- An eye exam by a vision specialist
- 6-12 sessions at Wright-Patterson AFB
- Compensation of $30/hour plus completion bonuses

CONTACT INFORMATION

To find out more about this study, visit www.obvalab.com

You may also contact:

- [redacted]

Operational Based Vision Assessment (OBVA) Laboratory, USAF School of Aerospace Medicine, 711th Human Performance Wing

Principal Investigator: Dr. Logan Williams

OPERATIONAL BASED VISION ASSESSMENT

AFRL IRB Approved (FWR2019044H, Version 1.00) valid until 17 DECEMBER 2019.

Cleared, 88PA, Case # 2019-0714, 21 Feb 2019.

Appendix I: Study Protocol Sheet

PARTICIPANT ID: ____________ RA:______________ DATE:___________

SESSION: 1 2

Subject Temp:

BLOCKS:_________________ Stereo tests:_________________

Room temp:

Humidity:

Heat Index:

Signed AVT consent ☐

Written information ☐

Log entry ☐

COVID-19 Qus ☐

Sign Infection prevention Addendum ☐

Paid? Y/N

1. Before participant arrives

1.1	Put on shoe covering	
1.2	Wash hands	
1.3	Check own temp	
1.4	Put on gown, facemask and gloves	
1.5	Clean 3D & color glasses with Endust. **Clean lenses with Zeiss lens wipe.**	
1.6	Clean all input buttons with Endust	
1.7	Clean exposed chairs with all-purpose cleaner	
1.8	Clean countertops and tables	
1.9	Turn on TNO lamp	
1.10	Turn on dRDS computer and start program	
1.11	Clean flight controls and chair	
1.12	Change head tracker	
1.13	Turn on TV screens and launch Diamond Vis	

STEREO TESTS: Order_______________

2. TNO test

2.1	Subject to put on coloured glasses	
2.2	**Turn off room lights**	
2.3	Show plate – ask if can see butterfly (screening)	Y / N
2.4	Instruct to call out orientation of missing slice of circle	
2.5	Plate V - need both correct to progress Top row (480 arcsec) Down / Left Bottom row (240 arcsec) Up / Down	☐ ☐ ☐ ☐
2.6	Plate VI Top row (120 arcsec) Up / Left Bottom row (60 arcsec) Right / Right	☐ ☐ ☐ ☐
2.7	Turn off lamp & remove glasses	
	Stereo acuity	

3. dRDS test

3.1	Turn on monitor	
3.2	Turn on speakers	
3.3	Ensure chair 1m from screen	
3.4	****Turn on 3D glasses****	
3.5	Instruct to identify orientation of missing slice on circle and press directional button	
3.6	Enter ID. and ensure crossed disparity & *static* stimuli	
3.7	**Turn off room lights**	
3.8	Run demo test – confirm they can see depth	
3.9	Run test	
	Stereo acuity Log Arc sec	
	SD	
3.10	Repeat with *dynamic* stimuli	
	Stereo acuity Log Arc sec	

	SD	
3.11	Remove glasses & **place on charge**	
3.12	Turn off speakers	

HMD TASK:

Ask subject if they need to use the rest room – will be in the chair for 1.5 hrs

4. EKG + Temp

4.1	Turn on Power lab (back of box)	
4.2	Connect to laptop and start Lab Chart	
4.3	Open Template file and save as: “**Subject ID_HMD_S1/2**”	
4.4	Connect EKG leads • Black -> LEFT shoulder • White -> RIGHT shoulder • Green -> RIGHT ankle	
4.5	Connect temperature probe – metal face down Left hand middle finger, attach with tape	
4.6	Start recording	
	Shortcut comments keys • 2D start/end = Shift + 2 • Intermittent start/end = Shift + I • Persistent start/end = Shift + P • Motion-in-depth start/end = Shift + D • Baseline start/end = Shift + B	

5. Point out equipment

- Right hand:
  - Joystick to fly with
  - Crown button and flap trigger for alignment
- Left hand
  - Two switches – FLAP/FUEL. Spring loaded so will re-center.
  - Will NOT be using the throttle

6. HMD

6.1	Give subject fabric hat to wear	
6.2	Turn on head tracker -> select head tracker	
6.3	Turn on HMD - small red switch on rear black box	
6.4	Turn up luminance (press and hold top red button on rear box)	
6.5	Place HMD on subject’s head	

6.6	Lower eyepieces and remove case coverings	
6.7	Ask subject to tighten using screws on top and to the back	
6.8	Adjust eye relief as required	
6.9	Adjust tilt and horizontal position, so centered over eyes	
6.10	*Turn off room lights"	
6.11	Place **focus box** image – "Close each eye in turn. With the open eye, you need to be able to see all 4 corners of each box"	
6.12	Confirm picture is crisp and clear with all numbers/letters readable	
6.13	With Bluetooth keyboard bring up alignment reference in pictures	
6.14	Alignment • Focusing on the cross on the screen, move the two parts of the cross in your HMD together using crown button on joystick. Align over the cross on the screen for reference. When content pull flap trigger to front of joystick. • Check the two horizontal lines can line up with the horizontal line on the screen. Pull trigger flap when content • Check the two vertical lines can line up with the vertical line on the screen. Pull trigger when content.	
6.15	Repeat alignment – check similar position	
6.16	Boresight – move your head to place cross in your HMD over cross on screen & pull flap	
6.17	Enter subject ID • If S1 – randomize conditions • If S2 – check for subject	

Instructions for task: [Put up static example display – for demo purposes]

Your primary task is to fly the aircraft using the joystick in your right hand. Pushing the joystick forward, the aircraft nose will move down and you will descend.
Pull the joystick back, the aircraft nose will move upwards and you will ascend
Moving the joystick right or left will move the aircraft right or left respectively.

At the top center of the HMD image is a heading indicator of the 360degrees of a circle. At the moment it is centered on the number 10. ***Can you see this?***

To the middle right of the HMD image is a circle gauge giving your altitude/height. At the moment it indicates your current altitude is 5000ft***. Can you see this?***

In the top right corner, a heading and altitude number command will be given. In this example your heading command (HDG) is 11 and your altitude command (ALT) is 5600ft. ***Can you see this?***

Using the joystick to move, and referring to your heading indicator and altimeter, fly to the command location given. In this situation that would mean moving the joystick right to move from 10 to 11 and pulling the joystick back to climb from 5000 to 5600ft. You can do both movements at the same time. ***Does this make sense?***

When you hear a soft 'bing', this indicates that the heading and altitude commands have changed and a new flight path has been given for you to follow.

During the actual trial, when you move your head to look at the 'out the window/real world' image on the screens, the heading and altitude information will move with you. At the moment they are all fixed together for the purpose of the demo.

At periodic points throughout the task, warning symbols will appear to the middle left of your HMD image. At the moment it says 'FUEL'. Can you see this?
If the warning reads FUEL – pull the switch closest to you down.
If the warning reads FLAP – pull the middle switch down.
The FLAP warning symbol will always be presented in 2D, however, the FUEL warning symbol may appear in different formats. If you run out of fuel, you fall out of the sky so it is of the greatest importance to respond to the warning symbol and for that reason, FUEL is the switch closest to you. [This is how I remember it!]

A bell indicates a correct response, a 'eh er' indicates a wrong response.
Please enter your response as quickly as possible, but prioritize accuracy.

At this point we will do a 10-minute training session to get you familiar with flying and the navigation task. Your performance will not be assessed for this. I will talk you through it as we go.

- ➔ Talk them through the first few commands
- ➔ Review graph to ensure plateau of performance
- ➔ If not, then REPEAT the training session

Now we will do a short practice session with both the flying task and with the warning symbols. This is to get you used to where the warning symbols will appear and how to pull the switches. Your performance will not be assessed for this, so please take your time looking at the symbols and appreciate the difference between FLAP and FUEL.

Check they can see FUEL doming out towards them, like in the stereo task.

10-minute baseline -Please rest and keep your arms on the arm rest. Any movement you make interferes with the heart tracing.

7. Questionnaire

7.1	Launch from main computer. Render IP.	
7.2	Enter subject ID	
7.3	Select condition and “continue survey”	
7.4	Before study ask subject to answer ‘Pre’ symptom questions 0= Good/no symptoms, 10= Worst	

8. Protocol

Afterwards:
There were 4 different conditions

- Both symbols displayed in 2D
- Fuel symbol flashed on and off at depth
- Fuel symbol flashed in and out of depth, continuously on
- Fuel symbol moving with motion-in-depth, continuously on

Did you appreciate any difference between them?

Which condition did you find the most useful?

Room temp:

Humidity:

Heat Index:

9. After experiment

9.1	End survey	
9.2	Stop physiological data recording	
9.3	Remove HMD (place casings back on lenses)	
9.4	Turn off HMD and place carefully back in box	
9.3	Remove ECG & temperature sensors	

10. After subject leaves

10.1	Clean 3D & color glasses with discide towelettes (not the lenses)	
10.2	Clean all input surfaces with discide towelettes	
10.3	Clean all chairs	
10.4	Wash hands	
10.5	Change/remove PPE	
10.6	Complete session info sheet and place in Liz's inbox	

Appendix J: COVID-19 Human Subjects On-Boarding Procedures

Recruiting

Potential subjects will be contacted by either the research assistant (RA) or research scientist (RS) assigned to that study to determine their interest and availability in participating in the study. When a potential subject expresses willingness to participate in the study the process of on-boarding may begin.

Scheduling

Prior to scheduling, the RA or RS must verify the subject's answers to the following questions via email:

1. Do you have a fever, worsening cough, or flu-like symptoms?
2. Do you have any of the following symptoms:
    a. Cough, sore throat, or shortness of breath
    b. Difficulty breathing
    c. Fatigue, muscle or body aches
    d. Headache
    e. New loss of taste or smell
    f. Congestion or runny nose
    g. Nausea or vomiting
    h. Diarrhoea
3. Have you travelled outside of the local area (commuting distance from base) in the last 14 days?
4. Have you been in close contact with someone, including healthcare workers, suspected or having, or confirmed to have contracted, the coronavirus?

If the subject answers "yes" to any of these questions they cannot participate in the study. The exception to this are boom operators which may answer "yes" to question 3. The RA or RS will explain to the subject that if the answers to any of these questions change prior to their arrival at the lab then they **must** inform the RA immediately.

The RA or RS must keep a record of their contact with potential subjects via the "Subject Selection Tracker" to ensure compliance with COVID-19 restrictions and procedures.

Arrival to the Lab

The RA or RS will provide instructions to the participant on where to go when they arrive, based on their first scheduled session (either to building 197 or building 33). When the participant arrives, the RA or RS will begin filling out the log entry for their visit, including:

- Their name
- The date
- The time they arrived

After this information has been filled out, the RA or RS will use a no-touch thermometer to take the subject's temperature and record it on the log.

The RA or RS will again verify the subject's answers to the above COVID-19 questions and have the subject sign and date a copy of the 711HPW Infection Prevention Addendum. The signed copy of the Infection Prevention Addendum will be turned in with the Session Info Sheet and any other documents at the end of the session or at the end of the day.

At the end of the session, the RA or RS will complete the log entry, including:

- Locations (including building and room numbers) that the subject visited
- The time they left the lab

<u>Summary</u>

- Verify over email that the subject has answered "no" to the COVID-19 questions
- Keep track of this verification on the "Subject Selection Tracker"
- Upon arrival, start log entry and take temperature
- Reaffirm answers to COVID-19 questions and have participant sign and date "711 HPW Infection Prevention Addendum"
- At end of session, complete log entry

8. LIST OF REFERENCES

ADInstruments (2019) ‘Lab Chart Pro - Classification Parameters-HRV Settings’. Dunedin, NZ.

Alhusuny, A., Cook, M., Khalil, A., Treleaven, J., Hill, A. and Johnston, V. (2021) ‘Impact of accommodation, convergence and stereoacuity on perceived symptoms and surgical performance among surgeons’, *Surg Endosc*, 35(12), pp. 6660–6670.

Allison, R.S., Gillam, B.J. and Vecellio, E. (2009) ‘Binocular depth discrimination and estimation beyond interaction space’, *J Vis*, 9(1), pp. 10–10.

Antona, B., Barrio, A., Sanchez, I., Gonzalez, E. and Gonzalez, G. (2015) ‘Intraexaminer repeatability and agreement in stereoacuity measurements made in young adults’, *Int J Ophthalmol*, 8(2), pp. 374–381.

Astle, A.T., McGraw, P. V. and Webb, B.S. (2011) ‘Recovery of stereo acuity in adults with amblyopia’, *BMJ Case Reports*, 2011, pp. 10.1136/ bcr0720103143.

Attwood, J.E., Kennard, C., Harris, J., Humphreys, G. and Antoniades, C.A. (2018) ‘A comparison of change blindness in real-world and on-screen viewing of museum artefacts’, *Front Psychol*, 9, pp. 151.

Aura, C., Temme, L.A., St. Onge, P.M., McAtee, A.M., Wilson, M., Bowers, B., Baugher, K.M. and Hayes, A. (2021) ‘Operator state monitoring via pupilometry: measuring mental workload under varying luminance conditions’, *Proc. SPIE 11757, Smart Biomedical and Physiological Sensor Technology XVIII, 117570M*. pp. 78–86.

Bach, M., Schmitt, C., Kromeier, M. and Kommerell, G. (2001) ‘The Freiburg Stereoacuity Test: Automatic measurement of stereo threshold’, *Graefe’s Archive for Clinical and Experimental Ophthalmology*, 239(8), pp. 562–566.

Badcock, D.R. and Schor, C.M. (1985) ‘Depth-increment detection function for individual spatial channels’, *J Opt Soc Am*, 2(7), pp. 1211–1216.

Banks, M.S., Read, J.C.A., Allison, R.S., Watt, S.J. (2012) 'Stereoscopy and the Human Visual System'. *SMPTE Motion Imagning J.* 121(4): pp 24–43

Bayer, M., Rash, C.E. and Brindle, J. (2009) ‘Chapter 3: Introduction to Helmet Mounted Displays’, in C.E. Rash, M.B. Russo, T.R. Letowski, and E.T. Schmeisser (eds) *Helmet-Mounted Displays: Sensation, Perception and Cognition Issues*. 1st edn. U.S. Army Aeromedical Research Laboratory, pp. 47–109.

Behar, I., Wiley, R.W., Levine, R.R., Rash, C.E. and Walsh, D.J. (1990) 'Visual Survey of Apache Aviators (VISAA)'. U.S Army Aeromedical Research Laboratory, AL. Report No. 90-15.

Van Beurden, M.H.P.H., Kuijsters, A. and IJsselsteijn, W.A. (2010) ‘Performance of a path tracing task using stereoscopic and motion based depth cues’, *Proc International Workshop on Quality of Multimedia Experience, QoMEX 2010.* pp. 176–181.

Beverley, K. and Regan, D. (1973) ‘Evidence for the existence of neural mechanisms selectively sensitive to the direction of movement in space’, *Physiol. J.* 235(1), pp. 17–29.

Birch, E.E., Gwiazda, J. and Held, R. (1982) ‘Stereoacuity development for crossed and uncrossed disparities in human infants’, *Vision Res*. 22(5), pp. 507–513.

Blake, R. and Fox, R. (1973) ‘The psychophysical inquiry into binocular summation’, *Percept Psychophy.* 14(1), pp. 161–185.

Blake, R., Sloane, M. and Fox, R. (1981) ‘Further developments in binocular summation’, *Percept Psychophys.* 30(3), pp. 266–276.

Bonnen, K., Huk, A.C. and Cormack, L.K. (2017) ‘Dynamic mechanisms of visually guided 3D motion tracking’, *J. Neurophysiol.* 118(3), pp. 1515–1531.

Bridge, H. (2016) ‘Effects of cortical damage on binocular depth perception’, *Phil. Trans. R. Soc. B.* 371: 20150254. pp.1-9

Browne, M.P. and Moffitt, K. (2009) ‘Improving the Utility of a Binocular HMD in a Faceted Flight Simulator', *Proc Interservice/Industry Training, Simulation, and Education Conference*. 9178. pp. 1–11.

Browne, M.P., Moffitt, K. and Wimsatt, C. (2017) ‘Stereo Symbology Concepts for Binocular Helmet-Mounted Displays’, *Proc Image Conference*. Dayton, Ohio.

Burge, J., Rodriguez-Lopez, V. and Dorronsoro, C. (2019) 'Monovision and the misperception of motion'. *Current Biology*. 29, pp. 2586–2592.

Buzzelli, A.R., Trowell-Harris, I. and Protsko, R. (1989) ‘The Measurement of Visual Efficiency Standards for Pilots in the United States Air Force’, *Military Medicine*, 154(7), pp. 345–347.

Cain, B. (2007) 'A Review of the Mental Workload Literature'. North Atlantic Treaty Organization. Report No: RTO-TR-HFM-121-Part-II.

Cakmakci, O. and Rolland, J. (2006) ‘Head-Worn Displays: A Review’, *JDT,* 2(3). pp. 199-216.

Campbell FW. The Depth of Field of the Human Eye. *Opt Acta.* 1957; 4:157–164

Campbell, G. and Green, F.D. (1965) ‘Monocular verses Binocular Visual Acuity’, *Nature*, 208, pp. 191–192.

Carter, R.C. and Carter, E.C. (1988) ‘Color coding for rapid location of small symbols’, *Color Res App*, 13(4), pp. 226–234.

Choi, J., Kolssov, A., Mikaelian, G., Gaska, J., Williams, L., Winterbottom, M. and Van Atta, A. (2018) ‘New commercial display technologies enable threshold-based color vision screening with high cone-contrast resolution and effective cone isolation’, *Poster presentation, Association for Research in Vision and Ophthalmology*. Honolulu, HI.

Christ, R.E. (1975) ‘Review and Analysis of Color Coding Research for Visual Displays’, *Hum Factors*, 17(6), pp. 542–570.

Cooper, J. and Warshowsky, J. (1977) ‘Lateral displacement as a response cue in the Titmus Stereo test.’, *Am J Optom Physiol Opt*, 54(8), pp. 537–541.

Cumming, B.G. and Parker, A.J. (1994) ‘Binocular mechanisms for detecting motion-in-depth’, *Vis Research*, 34(4), pp. 483–495.

Cutting, J.E. and Vishton, P.M. (1995) ‘Chapter 3. Perceiving layout and knowing distances: The integration, relative potency, and contextual use of different information about depth’, in W. Epstein and S. Rogers (eds) *Perception of Space and Motion*. San Diego, USA: Academic Press, pp. 69–117.

D’Zmura, M. (1991) ‘Color in visual search’, *Vis Research*, 31(6), pp. 951–966.

Davison, P.A. (1985) ‘Inter-relationships between British drivers’ visual abilities, age and road accident histories’, *Ophthalmic Physiol Opt*, 5(2), pp. 195–204.

Ding, J. and Levi, D.M. (2011) ‘Recovery of stereopsis through perceptual learning in human adults with abnormal binocular vision’, *Proc National Academy of Sciences*, 108(37), pp. E733–E741.

van Doorn, L.L.A., Evans, B.J.W., Edgar, D.F. and Fortuin, M.F. (2014) ‘Manufacturer changes lead to clinically important differences between two editions of the TNO stereotest.’, *Ophthalmic Physiol Opt*, 34(2), pp. 243–9.

van Ee, R. and Richards, W. (2002) ‘A planar and a volumetric test for stereoanomaly.’, *Perception*, 31(1), pp. 51–64.

Ehrenstein, W.H., Arnold-Schulz-Gahmen, B.E. and Jaschinski, W. (2005) ‘Eye preference within the context of binocular functions’, *Graefe’s Archive for Clinical and Experimental Ophthalmology*, 243(9), pp. 926–932.

Endsley, M.R. (1988) ‘Design and evaluation for situation awareness enhancement’, *Proc Human Factors Society*, 32(2) pp. 97–101.

Enright, J.T. (1987) ‘Perspective Vergence: Oculomotor Responses to Line Drawings’, *Vis Research*, 27(9), pp. 1513–1526.

Equality Act 2010 (2010). UK Government. Available at: https://www.legislation.gov.uk/ukpga/2010/15/schedule/9.

Erkelens, C.J. and Regan, D. (1986) ‘Human ocular vergence movements induced by changing size and disparity.’, *J. Physiol,* 379(1), pp. 145–169.

Ernst, J.M., Ebrecht, L. and Korn, B. (2021) ‘Virtual Cockpit Instruments - How Head-Worn Displays Can Enhance the Obstacle Awareness of Helicopter Pilots’, *IEEE Aerospace and Electronic Magazine*, 36(4), pp. 18–34.

Ernst, J.M., Peinecke, N., Ebrecht, L., Schmerwitz, S. and Doehler, H.-U. (2019) ‘Virtual Cockpit: an immersive head-worn display as human–machine interface for helicopter operations’, *Opt Eng*, 58(05), pp. 051807.

Fawcett, S.L. and Birch, E.E. (2003) ‘Validity of the Titmus and Randot circles tasks in children with known binocular vision disorders’, *J. AAPOS*, 7(5), pp. 333–338.

Federal Aviation Authority (2017) 'Guide for Aviation Medical Examiners: Exam techniques and criteria for Qualification Items 31-34. Eye-Monocular vision'. Available at: https://www.faa.gov/about/office_org/headquarters_offices/avs/offices/aam/ame/guide/app_process/exam_tech/et/31-34/mv/ (Accessed: 20 December 2021).

Foote, B. and Melzer, J. (2015) ‘A history of helmet mounted displays’, *Proc SPIE, Display Technoligies and Applications for Security, and Avionics IX; and Head- and Helmet- Mounted Displays XX*, 94700T.

Foote, B.D. and Hoffmann, M. (2018) ‘Color and impact to HMD design’, *Proc SPIE 10642, Degraded Environments: Sensing, Processing, and Display*, pp. 55–61.

Friedman, M. (1937) ‘The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance’, *J Am Stat Assoc*, 32(200), p. 675.

Gavrilescu, M., Battista, J., Ibbotson, M.R. and Gibbs, P. (2015) ‘Visual fatigue induced by optical misalignment in binocular devices: application to night vision binocular devices’, *Proc. SPIE 9470, Display Technologies and Applications for Defense, Security, and Avionics IX; and Head- and Helmet-Mounted Displays XX*, pp. 249–263.

Gavrilescu, M., Karas, R., Douglass, A.G., Abel, L.A. and Gibbs, P. (2019) ‘Changes in physiological parameters induced by optical misalignment in night vision binocular devices’, *Proc SPIE 11019, Situation Awareness in Degraded Environments*, 11019F

Geiselman, E.E. and Post, D.L. (1999) ‘Helmet-Mounted Display targeting symbology color coding: An air-to-air scenario evaluation ’, *Proc International Society for Optics and Photonics*, pp. 66–75.

Giavarina, D. (2015) ‘Understanding Bland Altman Analysis’, *Biochemia Medica*, 25(2), pp. 359–372.

Gold, T. and Hyman, A. (1970) 'Visual Requirements for Head-Up Displays' *JANAIR* Contract No. N00014-66-C-0114, Office of Naval Research, Naval Air Systems Command, U. S. Army Electronics Command. Report No: 630712.

Green, B.F. and Anderson, L.K. (1956) ‘Color coding in a visual search task’, *J Exp Psychol*, 51(1), pp. 19–24.

Green, D. and Swets, J. (1966) 'Signal detection theory and psychophysics'. John Wiley, NY, USA.

Greenhouse, S.W. and Geisser, S. (1959) ‘On methods in the analysis of profile data’, *Psychometrika*, 24(2), pp. 95–112.

Grosslight, J.H., Fletcher, H.J., Masterton, R.B. and Hagen, R. (1978) ‘Monocular Vision and Landing Performance in General Aviation Pilots: Cyclops Revisited’, *Hum Factors*, 20(1), pp. 27–33.

Hall, C. (1982) ‘The relationship between clinical stereotests’, *Ophthalmic Physiol Opt*, 2(2), pp. 135–143.

Harrington, L.K., McIntire, J.P. and Hopper, D.G. (2014) ‘Assessing the binocular advantage in aided vision’, *Aviat Space Environ Med*, 85(9), pp. 930–939.

Harris, J.M., McKee, S.P. and Watamaniuk, S.N.J. (1998) ‘Visual search for motion-in-depth: Stereomotion does not “pop out” from disparity noise’, *Nat Neurosci*, 1(2), pp. 165–168.

Harris, J.M. and Watamaniuk, S.N.J. (1995) ‘Speed discrimination of motion-in-depth using binocular cues’, *Vis Research*, 35(7), pp. 885–896.

Hart, S.G. (2006) ‘Nasa-Task Load Index (NASA-TLX); 20 Years Later’, *Proc Human Factors and Ergonomics Society*, 50(9), pp. 904–908.

Hartle, B., Deas, L.M., Allison, R.S., Irving, E.L., Glaholt, M. and Wilcox, L.M. (2017) ‘The contribution of monocular and binocular cues to altitude estimation in aircrew’, *J Vis*, 17(15), pp. 42–42.

Hartle, B., Sudhama, A., Deas, L.M., Allison, R.S., Irving, E.L., Glaholt, M.G. and Wilcox, L.M. (2020) ‘Contributions of Stereopsis and Aviation Experience to Simulated Rotary Wing Altitude Estimation’, *Hum Factors*, 62(5), pp. 812–824.

Hartle, B. and Wilcox, L.M. (2016) ‘Depth magnitude from stereopsis: Assessment techniques and the role of experience’, *Vis Research*, 125, pp. 64–75.

Hess, R.F., Ding, R., Clavagnier, S., Liu, C., Guo, C., Viner, C., Barrett, B.T., Radia, K. and Zhou, J. (2016) ‘A robust and reliable test to measure stereopsis in the clinic’, *Invest Ophthalmol Vis Sci*, 57(3), pp. 798–804.

Hess, R.F., To, L., Zhou, J., Wang, G. and Cooperstock, J.R. (2015) ‘Stereo Vision: The Haves and Have-Nots’, *i-Perception*, 6(3), pp. 204166951559302.

Hirota, M., Morimoto, T., Kanda, H., Endo, T., Miyoshi, T., Miyagawa, S., Hirohara, Y., Yamaguchi, T., Saika, M. and Fujikado, T. (2018) ‘Objective Evaluation of Visual Fatigue Using Binocular Fusion Maintenance’, *Transl Vis Sci Technol*, 7(2), pp. 9.

Holmes, J.M. and Clarke, M.P. (2006) ‘Amblyopia’, *Lancet*, 367(9519), pp. 1343–1351.

Hunfalvay, M., Orr, R. and Roberts, C. (2017) ‘Evaluation of Stereo Acuity in Professional Baseball and LPGA Athletes Compared to Non-Athletes’, *Vis Dev & Rehabil*, pp. 34–42.

Isreal, J.B., Wickens, C.D., Chesney, G.L. and Donchin, E. (1980) ‘The Event-Related Brain Potential as an Index of Display-Monitoring Workload’, *Hum Factors*, 22(2), pp. 211–224.

Jamieson, T.H. (1981) ‘Thermal Effects In Optical Systems’, *Opt Eng*, 20(2), pp. 156–160.

Järvenpää, T. and Pölönen, M. (2009) ‘36.1 Advances in Near-To-Eye Display Optical Characterization’, *SID Digest of Technical Papers*. 40(1) pp. 507.

Judge, S.J. and Cumming, B.G. (1986) ‘Neurons in the monkey midbrain with activity related to vergence eye movement and accommodation’, *J Neurophysiol*, 55(5), pp. 915–930.

Keller, K. and Colucci, D. (1998) ‘Perception in HMDs: what is it in head-mounted displays (HMDs) that really make them all so terrible?’, *Proc. SPIE 3362, Helmet- and Head-Mounted Displays III*, 3362, pp. 46–53.

Kim, S.-H., Suh, Y.-W., Yun, C., Yoo, E.-J., Yeom, J.-H. and Cho, Y.A. (2013) ‘Influence of stereopsis and abnormal binocular vision on ocular and systemic discomfort while watching 3D television’, *Eye*, 27(11), pp. 1243–1248.

Kooi, F.L. (2011) ‘Chapter 10: A display with two depth layers: attentional segregation and declutter’, in Roda, C., *Human Attention in Digital Environment*. Cambridge University Press, UK. pp. 245–258.

Kooi, F.L. and Toet, A. (2004) ‘Visual comfort of binocular and 3D displays’, *Displays*, 25(2–3), pp. 99–108.

Kramer, A. (1990) 'Physiological metrics of mental workload: A review of recent progress'. Navy Personnel Research and Development Center, San Diego, CA, USA. Report No: NPRDC-TN-90-23

Kruk, R., Regan, D., Beverley, K.I., Longridge, T., Bj, H. and Ot, A. (1981) ‘Correlations Between Visual Test Results and Flying Performance on the Advanced Simulator for Pilot Training (ASPT)’, *Aviat Space Environ Med,* 52(8), pp. 455–460.

Kruk, R., Regan, D., Beverley, K.I., Longridge, T., Ot, A., Air, W. and Base, F. (1983) ‘Flying Performance on the Advanced Simulator for Pilot Training and Laboratory Tests of Vision’, *Hum Factors*, 25(4), pp. 457–466.

Laberge-Nadeau, C., Dionne, G., Maag, U., Desjardins, D., Vanasse, C. and Ékoé, J.-M. (1996) ‘Medical conditions and the severity of commercial motor vehicle drivers’ road accidents’, *Accid Anal Prev*, 28(1), pp. 43–51.

Laborde, S., Mosley, E. and Thayer, J.F. (2017) ‘Heart Rate Variability and Cardiac Vagal Tone in Psychophysiological Research – Recommendations for Experiment Planning, Data Analysis, and Data Reporting’, *Front Psychol*, 8, pp. 213.

Lambooij, M., Fortuin, M., Ijsselsteijn, W., Evans, B. and Heynderickx, I. (2010) ‘Measuring visual fatigue and visual discomfort associated with 3-D displays’, *J Soc Inf Dis*, 18(11), pp. 931.

Lee, S.Y. and Koo, N.K. (2005) ‘Change of stereoacuity with aging in normal eyes.’, *Korean journal of ophthalmology : KJO*, 19(2), pp. 136–139.

Leek, M.R. (2001) ‘Adaptive procedures in psychophysical research’, *Percept Psychophys,* 63(8), pp. 1279–1292.

Leske, D.A., Birch, E.E. and Holmes, J.M. (2006) ‘Real Depth vs Randot Stereotests’, *Am J Ophthalmol*, 142(4), pp. 699–701.

Livingston, M.A. (2006) ‘Quantification of Visual Capabilities using Augmented Reality Displays’, *Proc 5th IEEE and ACM International Symposium on Mixed and Augmented Reality*, pp. 3–12.

Lu, Z.-L. and Sperling, G. (1995) ‘The Functional Architecture of Human Visual Motion Perception’, *Vis Research*, 35(19), pp. 2697–2722.

Maag, U., Vanasse, C., Dionne, G. and Laberge-Nadeau, C. (1997) ‘Taxi drivers’ accidents: How binocular vision problems are related to their rate and severity in terms of the number of victims’, *Accid Anal Prev*, 29(2), pp. 217–224.

Marshall, G.F. (1989) ‘Back From The Past: The Helmet Integrated System Of Albert Bacon Pratt (1916)’, *Helmet-Mounted Displays*, 1116, p. 2.

Mattock, M.G., Asch, B.J., Hosek, J. and Boito, M. (2019) 'The Relative Cost-Effectiveness of Retaining Verses Accessing Air Force Pilots'. RAND Corporation.Santa Monica, CA, USA

McCaslin, A.G., Vancleef, K., Hubert, L., Read, J.C.A. and Port, N. (2020) ‘Stereotest Comparison: Efficacy, Reliability, and Variability of a New Glasses-Free Stereotest’, *Transl Vis Sci Techol*, 9(9), pp. 29.

McIntire, J.P., Havig, P.R. and Geiselman, E.E. (2012) ‘What is 3D good for? A review of human performance on stereoscopic 3D displays’, *Proc SPIE 8383, Head- and Helmet-Mounted Displays XVII: and Display Technologies and Applications for Defense, Security and Avionics VI*, 8383OX.

McIntire, J.P., Havig, P.R., Harrington, L.K., Wright, S.T., Watamaniuk, S.N.J. and Heft, E.L. (2014) 'Clinically Normal Stereopsis Does Not Ensure a Performance Benefit from Stereoscopic 3D Depth Cues', U.S. Air Force Research Laboratory, Dayton, OH, USA. Report No:AFRL-RH-WP-TP-2014-0044.

McIntire, J.P., Wright, S.T., Harrington, L.K., Havig, P.R., Watamaniuk, S.N.J. and Heft, E.L. (2014) ‘Optometric measurements predict performance but not comfort on a virtual object placement task with a stereoscopic three-dimensional display’, *Opt Eng*, 53(6), pp. 061711.

Melzer, J. and Moffitt, K. (2011) 'Head-Mounted Displays: Designing for the User'. CreateSpace Independent Publishing Platform, Charleston, SC, USA.

Melzer, J.E. (2012) ‘HMDs as enablers of situation awareness: the OODA loop and sense-making’, *Proc SPIE 8383, Head-and Helmet Mounted Displays XVII; and Display Technologies for Applications for Defense, Security, and Avionics VI,* 83830F

Meola, K. V. (2005) ‘The Psychology of Color’, *Physchology,* 3(3), pp. 41–44.

Merritt, J.O. (2011) ‘67.1: Invited Paper: Performance Gains When Using 3D Displays vs. 2D Displays’, *SID Symposium Digest of Technical Papers*, 42(1), pp. 987–990.

Miller, L.J., Mittenberg, W., Carey, V.M., McMorrow, M.A., Kushner, T.E. and Weinstein, J.M. (1999) ‘Astereopsis caused by traumatic brain injury’, *Arch Clin Neuropsychol*, 14(6), pp. 537–543.

Mittenberg, W., Choi, E.J. and Apple, C.C. (2000) ‘Stereoscopic visual impairment in vascular dementia’, *Arch Clin Neuropsychol*, 15(7), pp. 561–569.

Moffitt, K. and Browne, M.P. (2018) ‘Visibility of color symbology in head-up and head-mounted displays in daylight environments’, *Opt Eng*, 58(5).

Momeni-Moghadam, H., Kundart, J., Ehsani, M., and Gholami, K. (2011) ‘Stereopsis with TNO and Titmus Tests in Symptomatic and Asymptomatic University Students’, *J Behav Optom*, 23(2), pp. 35–39.

Mon-Williams, M., Wann, J.P. and Rushton, S. (1993) ‘Binocular vision in a virtual world : visual deficits following the wearing of a head- mounted display’, *Ophthalmic Physiol Opt*, 13(4), pp. 387–392.

Morgan-Warren, P.J. and Woodcock, M. (2014) ‘Visual aspects of advanced sensors and helmet-mounted displays’, *Aeronautical Journal*, 118(1200), pp. 155–168.

Nakagawara, V.B. and Véronneau, S.J.H. (2000) *A Unique Contact Lens-Related Airline Aircraft Accident*. Office of Aviation Medicine, Washington D.C., USA

Nakayama, K. and Silverman, G.H. (1986) ‘Serial and parallel processing of visual feature conjunctions’, *Nature*, 320(6059), pp. 264–265.

Nataupsky, M. and Crittenden, L. (1988) ‘Stereo 3-D and non-stereo presentations of a computer-generated pictorial primary flight display with pathway augmentation’, *Proc AIAA/IEEE Digital Avionics Systems Conference*. pp. 3965

Nefs, H.T., O’Hare, L. and Harris, J.M. (2010) ‘Two Independent Mechanisms for Motion-In-Depth Perception: Evidence from Individual Differences’, *Front Psychol*, 1(155) pp. 1-8.

Newman, R. and Greeley, K. (1997) 'Helmet Mounted Display Design Guide′ , Crew Systems, San Marcos, TX, USA. Technical Report No: TR-97-11.

Nickel, P. and Nachreiner, F. (2003) ‘Sensitivity and Diagnosticity of the 0.1-Hz Component of Heart Rate Variability as an Indicator of Mental Workload’, *Hum Factors*, 45(4), pp. 575–590.

Novak, D., Mihelj, M. and Munih, M. (2010) ‘Psychophysiological responses to different levels of cognitive and physical workload in haptic interaction’, *Robotica*, pp. 1–8.

Nygren, T.E. (1991) ‘Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload’, *Hum Factors*, 33(1), pp. 17–33.

O’Keefe, E., Ankrom, M., Seemiller, E.S., Bullock, T., Winterbottom, M., Knapp, J. and Hadley, S. (2022) ‘The Relationship between Vision and Simulated Remote Vision System Air Refueling Performance’, *Proc Stereoscopic Displays and Applications XXXIII. Electronic Imaging*, 34: pp. 1–6.

O’Toole, A.J. and Walker, C.L. (1997) ‘On the preattentive accessibility of stereoscopic disparity: evidence from visual search.’, *Percept Psychophys*, 59(2), pp. 202–18.

Ogle, K.N. (1952) ‘On the limits of stereoscopic vision’, *J Exp Psychol*, 44(4), pp. 253–359.

Ogle, K.N. and Schwartz, J.T. (1959) ‘Depth of Focus of the Human Eye*’, *J Opt Soc Am*, 49(3), pp. 273.

Ogle, K.N. and Weil, M.P. (1958) ‘Stereoscopic Vision and the Duration of the Stimulus’, *AMA Arch Ophthalmol*, 59(1), pp. 4–17.

Or, C.K., Duffy, V.G., Ora, C.K.L. and Duffyb, V.G. (2007) ‘Development of a facial skin temperature-based methodology for non-intrusive mental workload measurement’, *Occup Ergon*, 7, pp. 83–94.

Paas, F. (2010) ‘Using electroencephalography to measure cognitive load’, *Educational Psychol Rev*, 22, pp. 245–238.

Pageau, M., Guise, D. de and Saint-Amour, D. (2009) ‘Comparison of local and global stereopsis in children with microstrabismus’, *J Vis*, 9(8), pp. 284–284.

Paish, N. (2017) ‘Entry aircrew visual standards - Standing Committee of Aircrew Medical Standards 51’. RAF Centre of Aviaiton Medicine, UK.

Parrish, R. V., Busquets, A.M. and Williams, S.P. (1990) ‘Recent research results in stereo 3-D pictorial displays at Langley Research Center’, *Proc 9th IEEE/AIAA/NASA Conference on Digital Avionic Systems*, pp. 529–539.

Parrish, R. V and Williams, S.P. (1990) 'Determination of Depth-Viewing Volumes for Stereo Three-Dimensional Graphic Displays'. National Aeronautics and Space Adminstration, Virgina, USA. Technical paper 2999,.

Parrish, R. and Williams, S. (1993) 'Trade-Offs Arising From Mixture of Color Cueing and Monocular, Binoptic, and Stereoscopic Cueing Information for Simulated Rotocraft Flight'. National Aeronautics and Space Adminstration, Virgina, USA. Technical Paper 3268.

Parsons, J. (1984) 'Parsons’ Diseases of the Eye'. 17th edn. Edited by S.J. Miller. Edinburgh: Churchill Livingstone.

Pastoor, S. (1993) ‘Human factors of 3D displays in advanced image communications’, *Displays*, 14(3), pp. 150–157.

Patterson, R. (1999) ‘Stereoscopic (cyclopean) motion sensing’, *Vis Research*, 39(20), pp. 3329–3345.

Patterson, R. (2009a) ‘Review Paper: Human factors of stereo displays: An update’, *J Soc Inf Disp*, 17(12), pp. 987.

Patterson, R. (2009b) ‘Unresolved issues in stereopsis: Dynamic disparity processing’, *Spat Vis*, 22(1), pp. 83–90.

Patterson, R., Cayko, R., Short, G.L., Flanagan, R., Moe, L., Taylor, E., and Day, P. (1995) 'Temporal integration differences between crossed and uncrossed stereoscopic mechanisms'. *Percept Psychophys,* 57(6), pp. 891–897

Patterson, R. and Martin, W. (1992) ‘Human Stereopsis’, *Hum factors*, 34(6), pp. 669–692.

Patterson, R., Winterbottom, M., Pierce, B. and Fox, R. (2007) ‘Binocular Rivalry and Head-Worn Displays’, *Hum Factors*, 49(6), pp. 1083–1096.

Patterson, R., Winterbottom, M.D. and Pierce, B.J. (2006) ‘Perceptual issues in the use of head-mounted visual displays’, *Hum factors*, 48(February), pp. 555–573.

Patterson, R.E. (2015) 'Human Factors of Stereoscopic 3D Displays' 1st ed. Springer-Verlag, NY, USA.

Pedraza, I. (2017) ‘Overestimation of stereo thresholds by the TNO stereotest is not due to global stereopsis’, *Ophthalmic Physiol Opt*, 37(4), pp. 507–520.

Peli, E. (1998) ‘The visual effects of head-mounted display (HMD) are not distinguishable from those of desk-top computer display’, *Vis Research*, 38(13), pp. 2053–2066.

Peterson, S., Axholt, M. and Ellis, S.R. (2008) 'Managing Visual Clutter: A Generalized Technique for Label Segregation using Stereoscopic Disparity', *Proc IEEE, Virtual Reality Conference*, pp. 169–176.

Portfors-Yeomans, C. V and Regan, D. (1996) 'Cyclopean discrimination thresholds for the direction and speed of motion in depth.', *Vis Research*, 36(20), pp. 3265–79.

Portfors, C. V. and Regan, D. (1997) 'Just-noticeable difference in the speed of cyclopean motion in depth and the speed of cyclopean motion within a frontoparallel plane.', *J Exp Psychol: Hum Percept Perform*, 23(4), pp. 1074–1086.

Posselt, B.N., Van Atta, A., Winterbottom, M.D. and Hadley, S. (2021) 'A New Stereo Test for Military Aircrew', *Aerospace Med Hum Perform*, 96(6), p. 490.

Posselt, B.N., Seemiller, E., Baber, C., Winterbottom, B. and Hadley, S. (2022) 'A digital alternative to the TNO stereo test to qualify military aircrew', *Aerospace Med Hum Perform*, 93(12) pp. 846-854.

Posselt, B.N. and Winterbottom, M. (2021) 'Are new vision standards and tests needed for military aircrew using 3D stereo helmet-mounted displays?', *BMJ Mil Health*, 167(6), pp. 442–445.

Predebon, J. (1994) 'Perceived size of familiar objects and the theory of off-sized perceptions', *Percept Psychophys*, 56(2), pp. 238–247.

Prins, N. and Kingdom, F.A.A. (2018) 'Applying the Model-Comparison Approach to Test Specific Research Hypotheses in Psychophysical Research Using the Palamedes Toolbox.', *Front Psychology*, 9, pp. 1250.

Rash, C.E. (2008) 'A 25-year retrospective review of visual complaints and illusions associated with a monocular helmet-mounted display', *Displays*, 29(2), pp. 70–80.

Rash, C.E., Suggs, C.L., Mora, J.C., Pol, C. Van De, Reynolds, B.S. and Crowley, J.S. (2002) ‘Visual Issues Survey of AH-46 Apache Aviators’, U.S. Army Aeromedical Research Laboratory, AL, USA. Report No: 2002-02

Rashbass, C. and Westheimer, G. (1961) ‘Disjunctive eye movements’, *J Physiol*, 159(2), p. 339.

Rawlings, S. and Shipley, T. (1969) ‘Stereoscopic Acuity and Horizontal Angular Distance from Fixation*’, *JOSA,* 59(8), pp. 991–993.

Read, J.C.A. (2015) ‘Stereo vision and strabismus’, *Eye*, 29(2), pp. 214–224.

Regan, D. (1993) ‘Binocular correlates of the direction of motion in depth.’, *Vis Research*, 33(16), pp. 2359–2360.

Regan, D. and Beverley, K. (1979) ‘Binocular and monocular stimuli for motion in depth: changing-disparity and changing-size feed the same motion-in-depth stage’, *Vis Research*, 19(12), pp. 1331–1342.

Regan, D. and Beverley, K.I. (1973) ‘The Dissociation of Sideways Movements from Movements in Depth: Psychophysics’, *Vis Research*, 13, pp. 2403–2115.

Regan, D. and Beverley, K.I. (1978) ‘Looming detectors in the human visual pathway’, *Vis Research*, 18(4), pp. 415–421.

Reis, G., Liu, Y., Havig, P. and Heft, E. (2008) ‘Investigating the Effects of Depth on Visual Search in the Multi-Layered Display’, *Proc 20th Australasian Computer-Human Interaction Conference*. Cairns, Australlia, pp. 307–310.

Richards, W. (1970) ‘Stereopsis and stereoblindness’, *Ex Brain Res*, 10(4), pp. 380–388.

Richards, W. (1971) ‘Anomalous stereoscopic depth perception’, *J Opt Soc Amer*, 61(3), pp. 410–414.

Richards, W. and Regan, D. (1973) ‘A Stereo Field Map with Implications for Disparity Processing’, *Investig Ophthalmol Vis Sci*, 12(12), pp. 904–909.

Rodriguez-Vallejo, M., Llorens-Quintana, C., Montagud, D., Furlan, W.D. and Monsoriu, J.A. (2016) ‘Fast and reliable stereopsis measurement at multiple distances with iPad’, *arXiv preprint:* 1609.06669

Rogers, S.P., Asbury, C.N. and Haworth, L.A. (1999) ‘Evaluation of earth-fixed HMD symbols using the PRISMS helicopter flight simulator’, *Proc SPIE 3689; Helmet- and Head-Mounted Displays IV*, pp. 54–65.

Royal Air Force (2020) ‘Assessment of Medical Fitness - AP 1269A’. Ministry of Defence, UK

Royal Air Force (2021) *Team Tempest | Royal Air Force*. Available at: https://www.raf.mod.uk/what-we-do/team-tempest/ (Accessed: 8 December 2021).

Rubio, S., Diaz, E., Martin, J. and Puente, J.M. (2004) ‘Evaluation of Subjective Mental Workload: A Comparison of SWAT, NASA-TLX, and Workload Profile Methods’, *App Psychol*, 53(1), pp. 61–86.

Rushton, S., Mon-Williams, M. and Wann, J.P. (1994) ‘Binocular vision in a bi-ocular world: new-generation head-mounted displays avoid causing visual deficit’, *Displays*, 15(4), pp. 255–260.

Russi, J. (2013) 'Effects of Stereoscopic 3D Displays on Air Traffic Controller Performance'. Air Force Institute of Technology, Dayton, OH, USA

Sakata, S., Grove, P.M., Hill, A., Watson, M.O. and Stevenson, A.R.L. (2017) ‘Impact of simulated three-dimensional perception on precision of depth judgements, technical performance and perceived workload in laparoscopy’, *Br J Surg*, 104(8), pp. 1097–1106.

Sakata, S., Watson, M.O., Grove, P.M. and Stevenson, A.R.L. (2016) ‘The conflicting evidence of three-dimensional displays in laparoscopy a review of systems old and new’, *Annal Surg*, 263(2), pp. 234–239.

Schmitt, C., Kromeier, M., Bach, M. and Kommerell, G. (2002) ‘Interindividual variability of learning in stereoacuity’, *Arch Clin Exp Ophthalmol*, 240, pp. 704–709.

Schroeder, T., Rainey, B., Goss, D. and Grosvenor, T. (1996) ‘Reliability of and Comparisons Among Methods of Measuring Dissociated Phoria’, *Optom Vis Sci*, 73(6), pp. 389–397.

Seemiller, E.S., Cumming, B.G. and Candy, T.R. (2018) ‘Human infants can generate vergence responses to retinal disparity by 5 to 10 weeks of age’, *J Vis*, 18(6), pp. 17–17.

Seemiller, E.S., Imel, G., Russi, J., Posselt, B.N., Bullock, C. and Winterbottom, M. (2022) ‘Simulating global misalignment of a binocular HMD during a flying and target discrimination task’, *Proc SPIE 12125, Virtual , Augmented, and Mixed Reality (XR) Technology for Multi-Domain Operations III*, 1212506.

Self, H. (1986) 'Optical Tolerances for Alignment and Image Differences for Binocular Helmet-Mounted Displays', U.S. Air Force Research Laboratory, Dayton, OH, USA Report No: AAMRL-TR-86-019.

Shaffer, F. and Ginsberg, J.P. (2017) ‘An Overview of Heart Rate Variability Metrics and Norms’, *Front Public Health*, 5, pp. 258.

Shaffer, F., McCraty, R. and Zerr, C.L. (2014) ‘A healthy heart is not a metronome: an integrative review of the heart’s anatomy and heart rate variability’, *Front Psychol*, 5(1040), pp. 1-19.

Sheehy, J.B. and Wilkinson, M. (1989) ‘Depth perception after prolonged usage of night vision goggles’, *Aviat Space Environ Med*, 60(6), pp. 573–579.

Shibata, T., Kim, J., Hoffman, D.M. and Banks, M.S. (2011) ‘The zone of comfort: Predicting visual discomfort with stereo displays’, *J Vis*, 11(8), pp. 790–795.

Shoda, L., van Atta, A., Winterbottom, M. and Gaska, J. (2017) 'Operational Based Vision Assessment Automated Vision Test Collection User Guide'. U.S. Air Force Research Laboratory, Dayton, OH, USA. Report No: AFRL-SA-WP-SR-2017-0012

Simons, K. (1981) ‘Stereoacuity Norms in Young Children’, *Arch Ophthalmol* 99(3), pp. 439–445.

Snyder, Q.C. (1991) 'Assessment of Two Depth Perception Test to Predict Undergraduate Pilot Training Completion'*.* Air Force Institute of Technology, Dayton, OH, USA. Report No: AFIT/CI/CIA-91-057.

Solimini, A.G. (2013) ‘Are There Side Effects to Watching 3D Movies? A Prospective Crossover Observational Study on Visually Induced Motion Sickness’, *PLoS ONE*., 8(2), pp. e56160.

Steiner, B.A. and Dotson, D.A. (1990) ‘The Use of 3-D Stereo Display of Tactical Information’, *Proc Human Factors Society Annual Meeting*, 34(1), pp. 36–40.

Thompson, L., Ji, M., Rokers, B. and Rosenberg, A. (2019) ‘Contributions of binocular and monocular cues to motion-in-depth perception’, *J Vis*, 19(3), pp. 2–2.

Thorne, D.R. (2006) ‘Throughput: A simple performance index with desirable characteristics’, *Behav Res Methods*, 38(4), pp. 569–573.

Tidbury, L.P., Brooks, K.R., O’Connor, A.R. and Wuerger, S.M. (2016) ‘A systematic comparison of static and dynamic cues for depth perception’, *Invest Ophthalmol Vis Sci*, 57(8), pp. 3545–3553.

Tittes, J., Baldwin, A.S., Hess, R.F., Cirina, L., Wenner, Y., Kuhli-Hattenbach, C., Ackermann, H., Kohnen, T. and Fronius, M. (2019) ‘Assessment of stereovision with digital testing in adults and children with normal and impaired binocularity’, *Vis Research*, 164, pp. 69–82.

Tittle, J.S., Rouse, M.W. and Braunstein, M.L. (1988) ‘Relationship of Static Stereoscopic Depth Perception to Performance with Dynamic Stereoscopic Displays’, *Proc Human Factors Society Annual Meeting*, 32(19), pp. 1439–1442.

Treisman, A. (1982) ‘Perceptual grouping and attention in visual search for features and for objects.’, *J Exp Psychol: Hum Percept Perform*, 8(2), pp. 194–214.

Treisman, A. and Gelade, G. (1980). 'A feature-integration theory of attention'. *Cogn Psychol,* 12(1), pp. 97–136.

Tufano, D.R. (1997) ‘Automotive HUDs: The overlooked safety issues’, *Hum Factors*, 39(2), pp. 303–311.

Tyler, C.W., Elsaid, A.M., Likova, L.T., Gill, N. and Nicholas, S.C. (2012) ‘Analysis of human vergence dynamics’, *J Vis*, 12(11), pp. 21–21.

U.S. Air Force School of Aerospace Medicine (2014) ‘Air Force Waiver Guide’, *Online*, pp. 1–979. Available at: https://www.afrl.af.mil/Portals/90/Documents/711/USAFSAM/Air%20Force%20Waiver%20Guide_220913.pdf

UK Civil Aviation Authority (2017) Visual system guidance material . Available at: https://www.caa.co.uk/Aeromedical-Examiners/Medical-standards/Pilots-(EASA)/Conditions/Visual/Visual-system-guidance-material-GM/ (Accessed: 20 December 2021).

Vancleef, K. and Read, J.C. (2019) ‘Which Stereotest do You Use? A Survey Research Study in the British Isles, the United States and Canada’, *Br Ir Orthopt J*, 15(1), pp. 15–24.

Vancleef, K., Read, J.C.A., Herbert, W., Goodship, N., Woodhouse, M. and Serrano-Pedraza, I. (2018) ‘Two choices good, four choices better: For measuring stereoacuity in children, a four-alternative forced-choice paradigm is more efficient than two’. *PLoS ONE* 13(7): e0201366

Wallin, B.G. (1981) ‘Sympathetic Nerve Activity Underlying Electrodermal and Cardiovascular Reactions in Man’, *J Psychophysiol*, 18(4), pp. 470–476.

Wang, Z., Lu, Y. and Fu, S. (2020) ‘A Multidimensional Method for Pilot Workload Assessment and Diagnosis’, *Aerospace Med Hum Perform*, 91(12), pp. 932–939.

Wann, J.P., Rushton, S. and Mon-Williams, M. (1995) ‘Natural Problems for Stereoscopic Depth-Perception in Virtual Environments’, *Vis Research*, 35(19), pp. 2731–2736.

Westheimer, G. (2013) ‘Clinical evaluation of stereopsis’, *Vis Research*, 90, pp. 38–42.

Westheimer, G. and McKee, S. (1980) 'Stereoscopic acuity with defocused and spatially filtered retinal images', *J Opt Soc Am*, 70(7), pp. 772–778.

Wickens, C.D. (2008) 'Multiple Resources and Mental Workload On behalf of: Human Factors and Ergonomics Society', *Hum Factors* 50(3), pp. 449–455.

Wickens, C.D., Hollands, J.G., Banbury, S. and Parasuraman, R. (2013) Engineering Psychology and Human Performance. 4th edn. Pearson Education Inc, NY, USA.

Wickens, C.D. and Long, J. (1995) 'Object Versus Space-Based Models of Visual Attention: Implications for the Design of Head-Up Displays', *J Exp Psychol App*, 1(3), pp. 179–193.

Wierwille, W.W. and Eggemeier, F.T. (1993) 'Recommendations for Mental Workload Measurement in a Test and Evaluation Environment', *Hum Factors*, 35(2), pp. 263–281.

Wilcox, L.M. and Allison, R.S. (2009) 'Coarse-fine dichotomies in human stereopsis', *Vis Research*, 49(22), pp. 2653–2665.

Williams, L., Gaska, J., Winterbottom, M., Shoda, E., Keefe, E.O., Palmer, E., Bauer, M., Atta, A. Van and Hadley, S. (2018) 'Noticeable Differences for Image Misalignment in Biocular Helmet-Mounted Displays', *Proc Image Conference*. Dayton, Ohio.

Williams, S.P. (2011) 'Trial Report: Trial SHINY LID Stage 2', Flying division, Air Warfare Centre, MoD, pp. 23.

Williams, S.P., Parrish, R. V. and Nold, D.E. (1994) 'Effective use of stereoptic 3D cueing to declutter complex flight displays', *Proc SPIE 2177, Stereoscopic Displays and Virtual Reality Systems*, pp. 223–231.

Winterbottom, M., Gaska, J., Shoda, E., Keefe, E.O. and Atta, A. Van (2018) 'Operational Based Vision Assessment: Automated Vision Test Research and Reliability'.U.S Air Force Research Laboratory, Dayton, OH, USA. Report No: AFRL-SA- WP- TR-2018-0019.

Winterbottom, M., Gaska, J., Wright, S., Lloyd, C., Tey, F. and Mcintire, J. (2014) ‘Operational Based Vision Assessment Research: Depth Perception’, *J Aust Soc Aerosp Med*, 9(0704), pp. 33–41.

Winterbottom, M., Lloyd, C., Gaska, J., Williams, L., Shoda, E. and Hadley, S. (2017) ‘Investigating Aircrew Depth Perception Standards Using a Stereoscopic Simulation Environment’, *Proc International Symposium on Electronic Engineering. Stereoscopic Displays and Applications XXVIII*.

Winterbottom, M., Lloyd, C., Gaska, J., Wright, S. and Hadley, S. (2016) ‘Stereoscopic Remote Vision System Aerial Refueling Visual Performance’, *Proc Stereoscopic Displays and Applications XXVII, Electronic Imaging*, (5), pp. 1–10.

Winterbottom, M., Patterson, R., Pierce, B., Gaska, J. and Hadley, S. (2015) ‘Visibility of monocular symbology in transparent head-mounted display applications’, *Proc SPIE 9470, Display Technology and Applications for Defense, Security, and Avionics IX: and Head- and Helmet-Mounted Displays XX,* 94700Z.

Winterbottom, M.D. (2015) 'Individual differences in the use of remote vision stereoscopic displays'. Wright State University, Dayton, OH, USA.

Winterbottom, M.D., O’Keefe, E., Gavrilescu, Galholt, Kobayashi, Tsujimoto, Douglass, Shoda, E., Gibbs, Lloyd, C., Gaska, J. and Hadley, S. (2019) ‘Operational Based Vision Assessment: Stereo Acuity Testing Research and Development’, *Proc Stereoscopic Displays and Applications*

Wolfe, J.M. (1994) ‘Guided Search 2.0 A revised model of visual search’, *Psychon Bull Rev*, 1(2), pp. 202–238.

Wright, Steven, Gooch, J. and Hadley, Steven (2013) 'The Role of Stereopsis in Aviation : Literature Review'. Air Force Research Laboratory,Dayton, OH, USA. Report No: AFRL-SA-WP-TP-2013-0001.

Zhang, Y., Meng, B. and Wu, H. (2021) ‘Evaluating the Mechanism by Which the TNO Stereo Test Overestimates Stereo Thresholds’, *J Ophthalmol*, 2021, pp. 1–7.

Zinn, W.J. and Solomon, H. (1985) ‘A comparison of static and dynamic stereoacuity’, *J Am Optom Assoc*, 56(9), pp. 712–715.