The effects of increased visual information on cognitive workload in a helicopter simulator

Reilly J. Innes, Zachary L. Howard, Alexander Thorpe, Ami Eidels and Scott D. Brown

School of Psychology, University of Newcastle, Australia

Word Count: 5740.

Keywords. cognitive workload, detection response task, helicopter simulation, heads-up display.

This research was supported by a University of Newcastle Industry Linkage Pilot Grant to SB and AE. RI, ZH and AT were supported by an Australian Government Research Training Program (RTP) Scholarship. We thank Airbus and Hensoldt for their in-kind support. Correspondence concerning this article may be addressed to: Reilly Innes, School of Psychology, University of Newcastle, Callaghan NSW 2308, Australia; Email: `Reilly.Innes@uon.edu.au`

Prècis:

Cognitive workload was evaluated under different levels of heads-up display information in a helicopter simulator. Pilots completed a short flight scenario which varied across environmental and symbology conditions. Results from the DRT and flight metrics gave an indication to pilots workload and performance.

Abstract

Objective

To test the effects of enhanced display information ("symbology") on cognitive workload in a simulated helicopter environment, using the Detection Response Task (DRT).

Background

Workload in highly demanding environments can be influenced by the amount of information given to the operator and consequently it is important to limit potential overload.

Methods

Participants (highly trained military pilots) completed simulated helicopter flights, which varied visual conditions and the amount of information given. During these flights participants also completed a DRT as a measure of cognitive workload.

Results

With more visual information available, pilots landing accuracy was improved across environmental conditions. The DRT is sensitive to changes in cognitive workload, with workload differences shown between environmental conditions. Increasing symbology ap-

peared to have a minor effect on workload, with an interaction effect of symbology and environmental condition showing that symbology appeared to moderate workload.

Conclusion

The DRT is a useful workload measure in simulated helicopter settings. The level of symbology moderated pilot workload. The increased level of symbology appeared to assist pilots flight behaviour and landing ability. Results indicate that increased symbology has benefits in more difficult scenarios.

Applications

The detection response task is an easily implemented and effective measure of cognitive workload in a variety of settings. In the current experiment, the DRT captures the increased workload induced by varying the environmental conditions, and provides evidence for the use of increased symbology to assist pilots.

Manuscript Type: Research Article

Introduction

With more information present and technological advances, our ability to multitask has seemingly improved (Haapalainen, Kim, Forlizzi, & Dey, 2010), however, there is substantial literature on driver distraction and cognitive workload that suggest this is not the case (Strayer & Johnston, 2001; Strayer, Turrill, et al., 2015; Strayer, Cooper, Turrill, Coleman, & Hopman, 2017). Both added visual stimuli and seemingly useful information systems can lead to detrimental distraction due to cognitive load in drivers (Lee, Young, & Regan, 2008; Strayer et al., 2017). Here, we offer a novel and unique workload-capacity assessment of helicopter pilots. Specifically, technological advances enable rich information to be projected into pilots' heads-up displays (HUDs), but the impact of this extra information on cognitive demand is not well understood. Here we ask; can too much information be detrimental to performance? To answer this question, we tested highly qualified helicopter pilots in a flight simulator in varying environmental and HUD settings.

Cognitive demands and distractions are difficult to assess within a multitasking environment. Adding to the number of items to process (or increasing the difficulty of these items to process) causes a greater depletion of limited attentional resources (Kahneman, 1973; Townsend & Eidels, 2011). When attentional resources are low, responses are impaired and we experience a diminished ability to process and react to the demands at hand. Such is the case when completing cognitive tasks while driving – our performance is diminished in both tasks (Watson & Strayer, 2010). Here we define cognitive workload as the level of cognitive demand placed on an individual from a task/s and distraction as scenarios where the individuals attention is drawn away from the main task/s. (Lee et al., 2008)

The detection response task (DRT) adds an additional task that measures residual resources via a simple detection task. In the DRT, which is a standardised procedure

(ISO:17488, 2016), participants are asked to respond as quickly as possible to a salient stimuli, which is administered frequently, whilst performing another task. Longer response times and increased misses correspond to higher cognitive workload (Strayer et al., 2013). Reactions are impaired when people are subjected to greater task demands, leaving fewer resources to allocate to the DRT. As an example, Strayer et al. (2013) showed that DRT response times for car drivers increased with the presence of a passenger or when talking on a mobile phone (both forms of distraction), similar to the increase when performing an operation span task. The sources of cognitive load mentioned above are external to the task at hand—it is not necessary to talk on the phone while driving—but systems related to completing the task, such as a user interface, can also impose cognitive workload. In extreme cases, a user interface can undermine its intended purpose of assisting the user by presenting too much information or interrupting relevant tasks (Johnson & Wiles, 2003).

User interfaces and other information delivery systems should therefore present only as much information as a user needs in an unobtrusive way (Haapalainen et al., 2010). A complication for user interface developers is that the amount of information a user needs may change as the user's workload state changes — a level of information that may be appropriate in one context may overload the user in another. A solution to this issue is to change the amount or presentation of information in real time, based on the user's cognitive capacity. A concurrent measure of workload is one necessary step in developing these *adaptive interfaces*.

A large body of cognitive workload research is centred around distraction in driving environments, yet this research is equally critical to the understanding of human-machine interactions in aviation. Helicopter and aeroplane cockpits are both extremely demanding environments, with a plethora of interfaces delivering multiple streams of information concerning air-speed, heading, fuel, obstacles and alike. Wickens (2002) outlines interlinking

factors crucial to human interaction with aircraft, and highlights that much psychological research related to these factors has been conducted in isolation. Further, Kantowitz and Casper (2017) reference the increasing amount of technology and automation in aviation, which impacts crew workload – noting that studies of attention may assist in solving workload related problems in aviation environments. As distracted driving literature has shown, understanding the impact of this technology is vital, with the literature informing policy and technological development (Young, Hsieh, & Seaman, 2013; Strayer, Cooper, Turrill, Coleman, & Hopman, 2015). In aviation, Huttunen, Keränen, Väyrynen, Pääkkönen, and Leino (2011) and Hannula, Huttunen, Koskelo, Laitinen, and Leino (2008) both evaluated cognitive workload using the speech prosody and psychopsyhiological stress (PPS) indicators respectively. Whilst these measures are effective in assessing their related constructs, they may not be reliable indicators when assessing workload induced by technological factors. Previous work by Zimmermann et al. (2019) also aimed to assess the usefulness of additional HUD information, in the helicopter setting, with findings indicating that pilots flew more effectively under conditions of more information. Further, in the military setting where this technology is most used, landings are far more frequent and difficult, meaning that the HUD information allows a safer environment in critical scenarios. However, the measure of cognitive workload used in this study – the NASA Task Load Index – provided inconclusive results regarding cognitive workload. Evidently, with technology and automation constantly developing in avionics, literature stresses the need for evaluation of workload to ensure usability of such technology.

Some pilots and avionics developers operate as though more available information can only be beneficial, but this overlooks cognitive workload factors (Thorpe, Nesbitt, & Eidels, 2019). Inversely, the type of information given to pilots may reduce their workload if the information is more readily perceived and easily processed, such as information which

is 3D and more naturalistic (Dan & Reiner, 2017; Gerjets, Walter, Rosenstiel, Bogdan, & Zander, 2014). In the current study we use the DRT to assess the workload demands arising from changes to the environment and the way information is presented (referred to as level of symbology). As the DRT assesses cognitive workload through residual capacity, we expect results from the DRT to translate from distracted driving literature to aviation environments.

The purpose of the current study was to evaluate the effectiveness and sensitivity of the DRT in a helicopter simulator environment, by varying the difficulty (environmental factors) of simulated flight conditions. The helicopter flight task was completed in a high-fidelity flight simulator. Flight simulators are widely used and well validated training facilities (Hays, Jacobs, Prince, & Salas, 1992; Roenker, Cissell, Ball, Wadley, & Edwards, 2003), so evaluation of cognitive workload in a simulator could facilitate deeper understanding of pilots' cognitive demands. Further, we used the DRT as a tool to measure the impact of added visual information ("symbology") in a HUD on helicopter pilots cognitive workload. We compared industry standard HUD symbology to new, more information rich symbology, as well as a control condition with no symbology. For a full overview of the technology input to the HUD see Zimmermann et al. (2019). Despite the limited number of participants, we placed a high importance on ecological validity of the task, designing a flight path that emphasised a realistic scenario, and testing pilots who were highly familiar with military helicopter environments.

It was expected that more information given to pilots would result in better flight outcomes. However, it was also anticipated that more information would lead to an increase in cognitive workload, similar to results reported by Strayer, Cooper, Turrill, Coleman, and Hopman (2016), Strayer, Turrill, et al. (2015) and Strayer et al. (2019). We first hypothesized that increased symbology would increase flight performance and landing accuracy,

similar to results from Zimmermann et al. (2019). We also hypothesized that DRT response times would increase with lower visual acuity (i.e. worse simulated weather conditions). Finally, we hypothesized that DRT response times would increase with added symbology.

Method

Participants. Eight pilots with experience in helicopter simulators and 2D symbology undertook the study. All pilots were male, had over 2,000 hours flying experience and extensive simulator experience. Seven pilots were recruited from the Airbus Brisbane facility, with one recruited from Hensoldt staff. Pilots recruited from the Airbus facility were either current military personnel or involved in testing or training. It was imperative that we tested highly trained and experienced personnel to ensure that confounding variables were limited; especially related to familiarity with the large-platform helicopter equipment and the advanced heads-up display. This research was approved by the Human Research Ethics Committee at the University of Newcastle (HREC-2013-0250).

Equipment. A helicopter simulator was used as the background during data collection. Data was collected in an Airbus MRH90 Taipan Multi Role helicopter simulator. The simulator incorporated three partially overlapping screens which made up 200° x 40° field of vision. The participant sat at a radius of approximately two metres from the screen. Controls in the simulator included a collective shaft, cyclic shaft and two foot pedals. The participants were shown an electronic map and a multi-function display, which indicated altitude, ground speed, collective power and helicopter roll. Participants were also fitted with a headpiece which was placed over the participants eyes. The headpiece acted as goggles, so that the participant could still see the simulator. In conditions where symbology was added, additional information was overlaid in their visual field. The location and angle of the headpiece was tracked at high rate so that information projected into the visual field

mapped accurately and dynamically onto the visual environment.

A DRT device was used, closely adhering to ISO 17488 (2016). The DRT device included a vibrating pad, which was taped to the participant's skin near their shoulder, and a response button, which was attached to the collective shaft nearest to where the pilots thumb rested. Engström, Larsson, and Larsson (2013) provide evidence for the tactile DRT as a sensitive measure of cognitive workload, finding similar trends to the use of a visual stimulus. Furthermore, Cooper, Castro, and Strayer (2016) suggest the tactile DRT is most effective for cutting down potential visual conflicts. With an already crowded visual environment, we proposed the use of the tactile DRT to limit visual workload effects.

Stimuli and Design. Each participant completed two simultaneous tasks – the flight simulation and the DRT. For the DRT, a short stimulus was elicited via a vibration. The participant was required to respond via the response button to each iteration of the stimulus. The stimulus lasted for one second (or until the response button was pressed, whichever came first). The DRT stimulus was elicited at an interval of 3 - 5 seconds and occurred for the duration of each simulated flight. Responses entered before the onset of the next vibration stimulus were deemed "hits", and failures to respond within 2.5 seconds were deemed a "miss". Second (and subsequent) responses entered before the onset of the next stimulus, as well as responses faster than 0.1 seconds, were deemed "false alarms". Response time was measured as the time between the onset of the vibration stimulus and the pressing of the switch.

The flight simulation involved participants undertaking a predetermined flight path with multiple objectives throughout. There were two conditions of visual environment: Day and Night. In all conditions, air traffic was absent, wind speed was set at 5km/h and weather was set to have no cloud or rain. The only parameters that varied were visibility (distance in meters), time of day, dust (on or off) and FLIR (on or off). The dust parameter

related to simulated "brown-out", where simulated dust would inhibit pilots view below a certain altitude ($\sim$ 100 feet). FLIR (forward looking infrared radar) is an industry standard night vision technology, used only in the night conditions. A full summary of conditions can be seen in Figure 1.

We used three levels of HUD information; no symbology – where there was no information projected onto the pilots' HUD; 2D symbology – the generic two-dimensional information projected to the pilots' helmet (see Appendix C for more details); or conformal 3D symbology – information which appears to be overlaid onto the simulated environment, as well as the generic 2D information (see Appendix C for more details; for a full overview of Hensoldt's Sferion assistance system, see Münsterer et al. (2014)). In the 2D condition, pilots were shown industry standard symbology which included speed, heading, altitude, geographic coordinates and distance to the LZ were displayed. Figure 5 in Münsterer et al. (2014) provides a good example of standard 2D information. The 2D symbology condition was made as similar as possible to the standard heads-up display used by military helicopter pilots in modern large-platform helicopters.

In the 3D condition, symbology included the 2D information, horizon lines, ridge lines, landscape grids, highlighted obstacles and LZ virtual towers which assisted in guiding the pilot. Figure 15(d) in Zimmermann et al. (2019) and Figures 8–11 in Münsterer et al. (2014), provide good examples of the 3D symbology condition. The 3D symbology condition contained extra information, and the condition without symbology contained less. In the 3D symbology condition, all 2D symbology was shown, as well as the LIDAR (light detection and ranging laser sensor) information. This included a grid over the environment, contours, LZ information, horizon line and helicopter position. For an example of the three symbology conditions, see Appendix C.

In conditions without symbology, the headpiece remained fixed to the participants

but displayed no visual information in the Day or Dust conditions. In the Night condition without symbology, FLIR (forward looking infrared radar) information was projected in the headpiece, with no additional symbology. In the 2D symbology condition, ground speed, radial altitude, location zone distance, and helicopter position were shown, as well as basic indicators for the waypoint and landing zone (LZ).

The study used a 2x3 factorial design, with two levels of visual environment (Day or Night) and two levels of Symbology (2D or 3D). Additionally, a condition without symbology was presented in either the Day or Night environmental condition was included. The visual environment in this condition was counterbalanced across pilots. Each pilot therefore completed five conditions – four with each level of symbology and visual environment, and one of two possible no-symbology conditions.

	NO SYMBOLOGY (0D) Headpiece off	**2D MINIMAL (2D)** *LIDAR off GND SPD, RAD ALT, LZ DST, LINE	**3D MAX SYMBOLOGY (3D)** *LIDAR on All symbology.
DAY VIS = 12000 TIME = 1600 DUST = OFF	DAY 0D	DAY 2D	DAY 3D
NIGHT VIS = 12000 TIME = 2000 DUST = OFF FLIR = (ON) FLIRTIME =2000 FLIRVIS= 2400	NIGHT 0D	NIGHT 2D	NIGHT 3D

Figure 1. : Full table of experimental conditions. The table shows the 2 x 2 within subjects design with the added between-subjects conditions without symbology (shaded in grey). Each condition maintained strictly controlled simulator settings with the exception of those listed under the title. in the table "VIS" stands for the visual range (in metres); "TIME" indicates the hour of day in the simulator; "Dust" indicates whether brown out was on or off for landings; "FLIR" stands for Forward Looking Infrared Radar; "FLIRTIME" indicates the setting for FLIR time of day - a brightness setting; "FLIRVIS" indicates the visibility range setting for FLIR. Symbology conditions vary on the information which is displayed. In 0D, the headpiece is switched off (aside from FLIR on in the night condition). In 2D - Ground Speed, Radial Altitude, Landing Zones Distance and a horizon line are displayed. In 3D symbology, all of the 2D symbology with additional landing zone displays and LIDAR (Light detection and ranging) is displayed. For a further breakdown of the symbology conditions, see Appendix B. The shaded boxes show the two randomised between subjects conditions - pilots only completed one of these.

Procedure. All participants were familiar with the simulator environment, and were given instructions about the DRT. Participants were not instructed to preference either

the DRT or the flight task, but were instructed that performance was measured across both. The designated flight path was outlined to the participants. They were given several minutes of flight time to adjust to the simulator before completing a practice run on the designated path. Following this, participants were given five practice DRT trials in isolation. Participants then began the experiment. The DRT commenced as soon as the pilot lifted the collective shaft for each condition.

The flight path was identical for all six conditions. The flight path took approximately 13 minutes to complete. Pilots were given verbal instructions during the flight to inform them of the objectives. Objectives included items such as passing a specified point at a target altitude and speed (known as "gates"), as well as specific landing scenarios, for example, landing in the centre of a sand bank. The flight path was divided into six sections, each with a different requirement, such as gates or landings. The objectives for the whole flight included two landings (one of which had poor visibility), an aborted landing, and three set "gates" to pass through at target speed and altitude. Furthermore, pilots were given directions on speed and altitude for each section, as well as specific navigation instructions. For a full breakdown of the flight path, see Appendix B.

Participants each completed all 5 of the 6 conditions. The order in which the conditions were presented to participants was pseudo-randomised; the no-symbology condition was never presented first, to account for the pilots lack of familiarity with the flight path. If, during a flight trial, the participant crashed or there were any technical issues, the run was restarted. Responses in these trials were recorded separately. Participants were given breaks between flights. All flight data was recorded. DRT response times and misses were recorded.

Results

In order to include the effect of the "no symbology" condition, we treated our study as a 2x3 design, with within-subject variables of time (day or night) and mixed variables of symbology (none, 2D, 3D). Flight performance was given by a number of indicators selected after consultation with experts and aviation literature (Krueger, Armstrong, & Cisco, 1985). These indicators were measured objectively and were quantifiable, as well as remaining relevant to the task. Indicators assessed were landing data, in-flight targets and overall flight variability. The main reason to evaluate flight quality was to ensure there was no task trade-off between the flight and the DRT. DRT response time and misses were analysed. We removed 4 sets of flight data due to crashes. These crashes were generally simulator related, such as a failure to calibrate the headpiece within the environment. These flights could provide interesting insight into pilot behavior under load, however, results from this data were uninformative due to the lack of data and varying crash time points.

For the workload measure we assessed mean DRT response time and the proportion of lapses. For each metric we completed Bayesian ANOVAs for the environmental conditions, symbology conditions and the interaction. All analysis was completed using the statistical program JASP (JASP Team, 2019) using default priors. Bayesian ANOVAs operate in much the same way as traditional frequentist ANOVAs, but with a key advantage: Bayesian ANOVAs can separately identify evidence in favor of an effect vs. evidence in favor of no effect (i.e. positive evidence for the null hypothesis). This is communicated through Bayes factors (BFs), which compare the likelihood of the null hypothesis (H_0) – which assumes no difference between conditions – against the likelihood of the alternative hypotheses (H_1) – which assumes a difference between conditions. Bayesian inference has become a standard approach in many fields because of its advantages over frequentist methods (Wagenmakers,

Lee, Lodewyckx, & Iverson, 2008). For clarity, we report all BFs in the direction of showing evidence in favor of the alternative hypotheses (BF_{10}). This means that larger BFs indicate more evidence *for* a difference between conditions. BFs near 1 indicate ambiguous evidence – the likelihood of the null and alternative hypotheses are about equal – and BFs much smaller than one indicate evidence in favor of the null hypothesis. We referred to (Jeffreys, 1961) for interpretation of BF_{10}.

Flight Metrics. we assessed the accuracy of landing data by borrowing appropriate precision measures from ballistic sciences. Participants were instructed to land at a specified and marked point in the virtual environment (centre of a sand bank). We measured the absolute distance from this landing zone (LZ) to the actual landing location ("landing error") and the "circular error probable" (CEP), which is the median error radius (Nelson, 1988, p.1). We analysed landing data using CEP for each the first landing zone (LZ1) and third (LZ3). Landings at LZ2 were aborted – by design. LZ3 did not utilize any landing symbology, making it a useful control condition. At LZ1, landing accuracy (defined by median absolute distance from the defined LZ in meters) was significantly improved with 3D landing symbology. The average distance from the defined LZ was 6m (SD = 6m) with 3D symbology, compared with 40m (SD = 41m) for conditions without symbology, and 61m (SD = 65m) with 2D Symbology. A Bayesian repeated measures ANOVA showed a significant main effect of symbology ($BF_{10} = 3.01$), although evidence was ambiguous for the difference (in distance from the target) between 3D symbology and 2D symbology ($BF_{10} = 2.55$) and between 3D symbology and without symbology ($BF_{10} = 1.71$). At LZ3 there was no significant difference between levels of symbology ($BF_{10} = 0.23$). These results are depicted graphically as CEPs in Figure 2. Further to these results, landings in the 3D Symbology condition were more tightly clustered, exhibiting less variability.

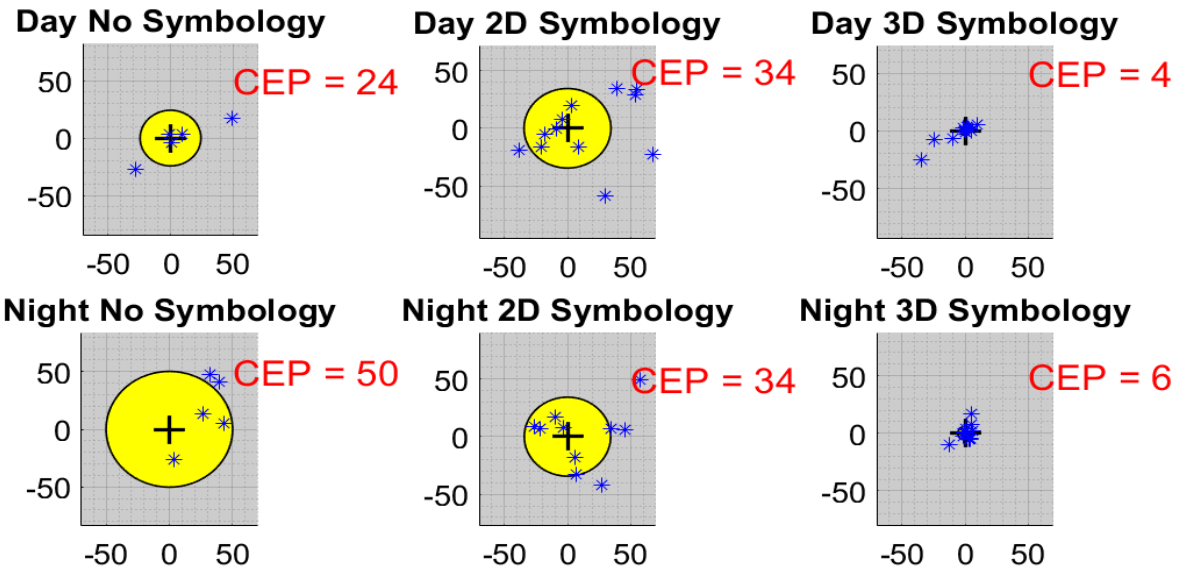

Figure 2. : CEP plots at LZ1 for each environmental condition across all levels of symbology. The cross at the centre of the circle denotes the defined landing zone, with asterisks marking the actual landings in each condition. The yellow circle marks the CEP result for each condition. The CEP value is included in the top right of each plot.

The second key performance indicator was comparison to flight targets. The first flight instruction concerned the path between waypoints LZ1 and waypoint E, which followed a river. Pilots were to maintain radar altitude of 200ft and ground-speed of 80 knots. We allowed an absolute deviation of 15 knots, and a +100, -50 ft deviation for altitude (derived in consultation with experienced military pilots). Figure 3 shows the proportion of each flight spent outside of these mission-critical parameters (altitude and speed). Bayesian ANOVAs showed a preference for the model which included the effect of symbology, environment and an interaction ($BF_{10} > 1,000$). In both measures (altitude and speed), the 2D symbology condition shows a greater proportion of time outside of the indicated boundaries ($BF_{10} > 1,000$). There is also evidence for an interaction effect between Symbology and environment ($BF_{10} > 1,000$), such that 2D symbology fares much

worse in night conditions.

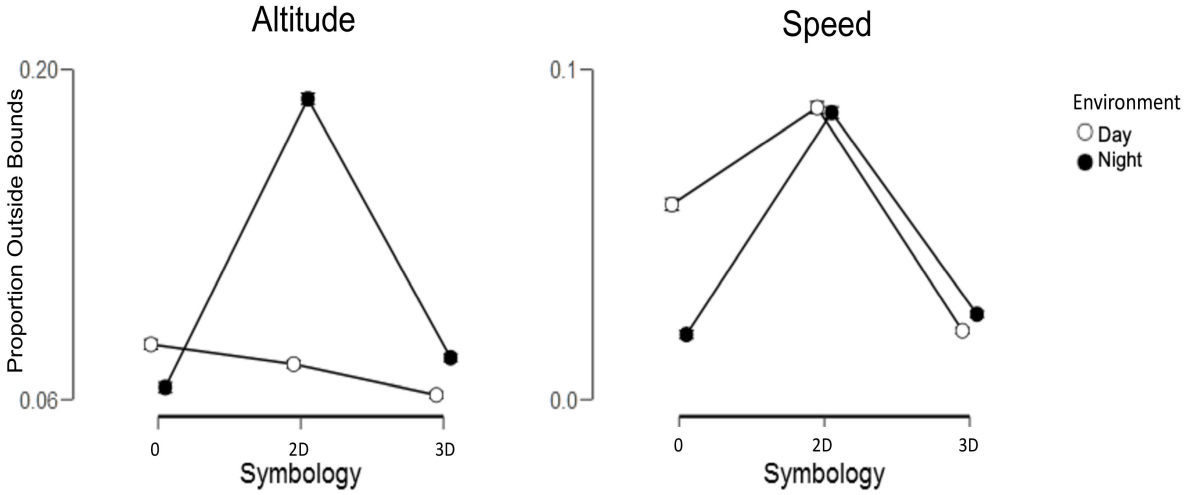

Figure 3. : Left panel; Mean proportion of time that pilots voided the mission bounds for altitude (i.e. flew above 300ft or below 150ft) across participants for the three levels of symbology. Right panel; Mean proportion of time voided the mission bounds in speed (i.e. flew above 95 knots or below 65 knots) across participants for the three levels of symbology. Error bars show the 95% confidence interval, and are too small to see due to the large amount of data – which minimised error.

At LZ2, pilots were instructed to abort landing when they approached very close (a "go around"). For this location, we analyzed minimum altitude and time below the set altitude. The target altitude was 20 feet radar altitude, with "brown-out" occurring when the pilot reached around 120 feet.

We conducted Bayesian repeated measures ANOVAs on the minimum altitude reached by pilots for the environmental and symbology conditions, which showed a preference for the model that only included the effect of symbology (BF_{10} = 24.14). The highest minimum altitude was observed in the condition without symbology (M = 33ft), which was higher than the 2D symbology (M = 22ft; BF_{10} = 11.87) and 3D symbology conditions (M = 24ft; BF_{10} = 11.346). No difference was found between the 2D and 3D

symbology conditions for this measure ($BF_{10} = 0.444$). Considering the relative distances of these altitudes from the target altitude, these results show pilots in the conditions with symbology present were able to fly closer to the target than those in the condition without symbology. A Bayesian repeated measures ANOVA on time (in seconds) spent under the target minimum altitude showed a preference for the model which included symbology ($BF_{10} = 3.70$). Pilots spent more time under the target altitude with 2D symbology (M = 1.29sec) compared to no symbology (M = 0.30sec; $BF_{10} = 2.375$) and 3D symbology (M = 0.42sec; $BF_{10} = 2.613$).

Further measures such as flight duration, flight variability across the vertical and horizontal planes were also recorded, but were not reported here, as they fail to add additional insight into flight performance over a longer distance.

From the flight performance data, it is clear that operationalizing optimal flight performance can be challenging. Whilst flight variability provided some insight into performance, and provided data across the entire course of the trial, it is not very informative about flight success and is confounded with highly-trained responses to change flight strategy in different environmental conditions. The CEP plots are limited to only a single value for each flight, yet provide a precise and objective measure of pilot's performance (at least at landing).

DRT. Mean response time was higher in the unsuccessful landing conditions than in the successful landing condition. Bayesian ANOVAs showed a strong preference for the model that included the main effect of landing for log RT ($BF_{10} = 1.588 \times 10^{10}$). Whilst informative in showing the significant increase to cognitive workload during a failed landing, we opted to remove these trials due to the high rate of misses to give a clearer assessment of DRT responses. Pilots were asked to repeat any trial where there was a crash or failed landing.

A two-way Bayesian ANOVA of log RT showed a strong preference for the model that included the effect of symbology (including the condition without symbology), visual condition and the interaction effect ($BF_{10} > 1000$). A comparison of visual conditions showed strong evidence of a difference between the High and Low visual conditions ($BF_{10} > 1000$). A comparison of symbology conditions showed ambiguous evidence of a difference between the 2D and 3D symbology conditions ($BF_{10} < 3$). A two-way repeated measures Bayesian ANOVA of misses was ambiguous, reflecting the relatively small number of missed DRT events (all $BF_{10} < 3$). Figure 4 shows log RT for 0D symbology condition in comparison with 2D and 3D symbology across High and Low visibility conditions. The interaction effect is shown here, where the difference between High and Low visibility is exaggerated in 0D symbology, and moves closer together as more symbology is added. This interaction effect suggests that symbology may moderate workload in high difficulty conditions, but may be unnecessary in low difficulty conditions.

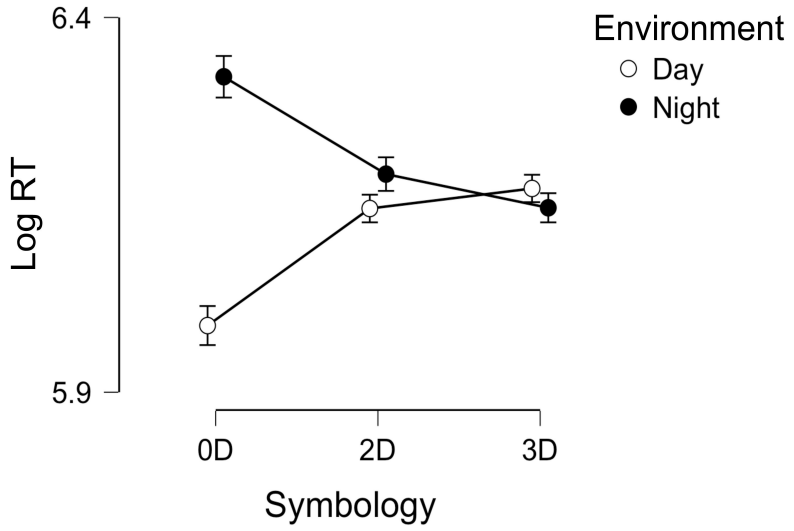

Figure 4. : Mean DRT log RT for environmental conditions across symbology conditions. Log RT is used to normalize across participants. Error bars are 95% confidence intervals.

Additionally, many prior studies have explored both main task performance and cognitive workload, however, there are limited attempts to jointly analyse these. In Figure 5 and Figure 6, we provide a novel, though rudimentary, combined analysis of both flight performance and cognitive workload. These figures show two conditions (night time with 2D and 3D symbology) for one pilot. We term this analysis a "workload heat map", where a pilot's flight path is plotted in colours that indicate their DRT response latency (calculated as a moving average response time), which is a well established proxy for their cognitive workload. Heatmaps for each pilot in each condition are included in `https://osf.io/2ntxw/`.

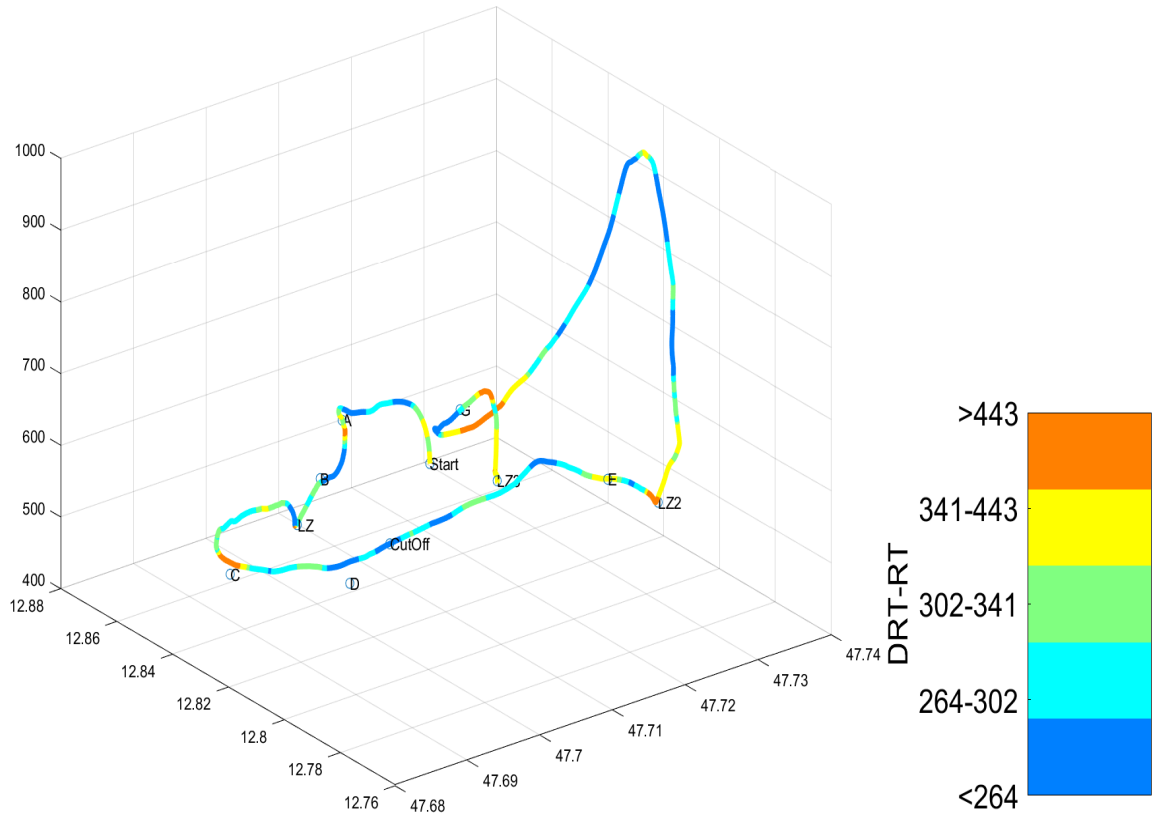

Figure 5. : Workload hit-map for the 2D Night condition for Pilot 2. The x and y axes show latitude and longitude respectively, with the z axis showing altitude. The line displays the flight path that the pilot took. Moving average DRT RT is plotted as colour across the flight for five bins.

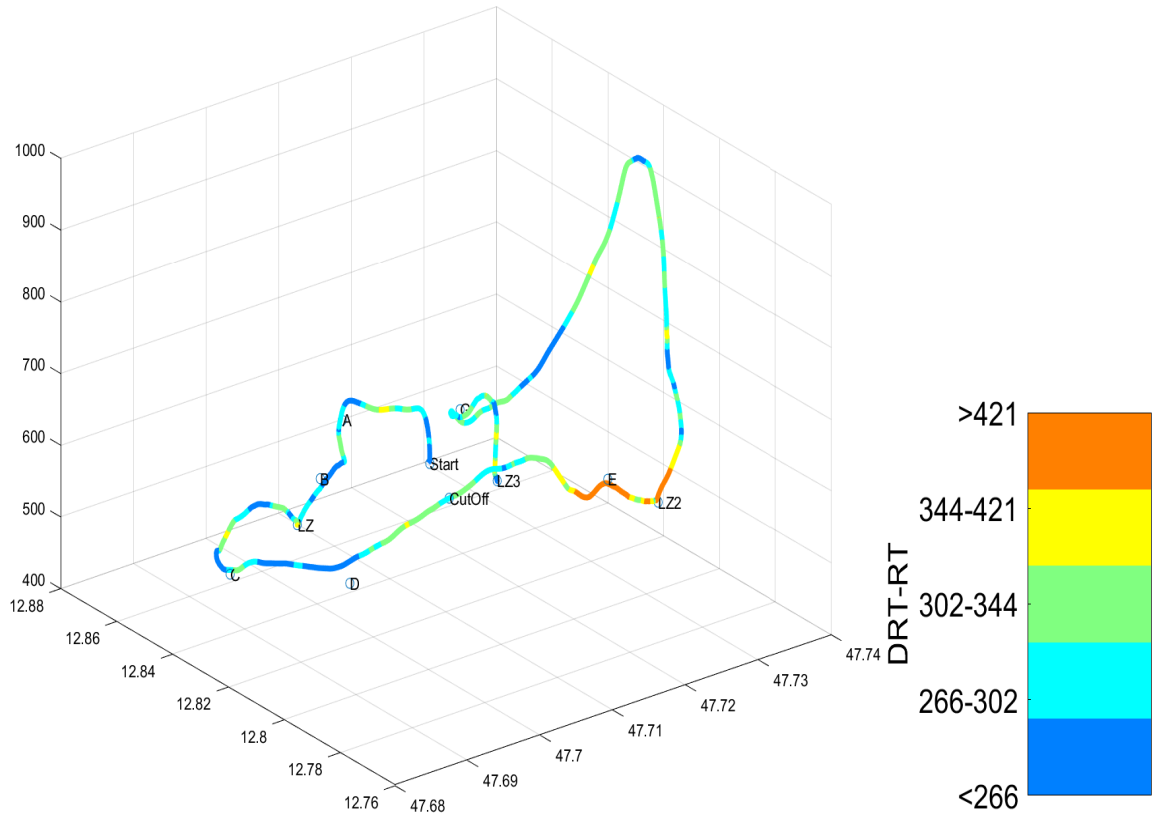

Figure 6. : Workload hit-map for the 3D Night condition for Pilot 2. The x and y axes show latitude and longitude respectively, with the z axis showing altitude. The line displays the flight path that the pilot took. Moving average DRT RT is plotted as colour across the flight for five bins.

General Discussion

Modern information systems and technological advances aim to assist operators, drivers, and pilots, but often fail to account for the impact on cognitive workload. Complex human machine interactions already present a myriad of information sources, and so before adding to these, it is important to evaluate the impact this information has on the operator (Gerjets et al., 2014; Thorpe et al., 2019). Helicopter environments are highly complex, and so adding more information to the pilots' heads-up-display could potentially prove harmful rather than helpful. The current study used the DRT to evaluate how adding information to a helicopter HUD effected pilots' cognitive workload. Further, we evaluated several flight metrics in an attempt to account for main task performance trade-off.

Results indicated that the DRT was sensitive to workload changes for environmental factors. DRT responses were slower with more difficult flight scenarios, indicating that workload increased in visual conditions of higher difficulty. Contrary to our hypotheses, DRT response times indicated that cognitive workload was relatively unaffected by additional HUD information, with no difference shown. Importantly, an interaction effect was found between symbology and visual conditions, which showed the visual condition having little effect on the 3D condition, but a greater effect on the 2D and 0D condition. This interaction effect is important for future HUD developments, as workload is moderated by symbology in various environmental conditions, which could potentially add unnecessary workload. The interaction effect between symbology and environment provides evidence of a telling story for the importance of user interface evaluation across a variety of conditions. Figure 4 shows this interaction which includes the effect of the between-subjects "no symbology" condition. Noticeably, the difference between High and Low visual conditions in the "no symbology" condition provides an insight into the utility of symbology. This gives evidence for the need for an adaptive interface, as, in clear day conditions, adding symbol-

ogy increases workload - i.e. where it is not necessary. However, in the night conditions, adding symbology actually lowers workload in comparison with "flying blind".

Unlike Coleman, Turrill, Cooper, and Strayer (2016) and Strayer et al. (2017), our results show that the additional information given to operators may not necessarily cause workload increases. Compared to driving contexts, the helicopter context appears to be much more cognitively demanding, and so this lack of difference in symbology conditions may be a result of a ceiling effect on workload. Alternatively, it may also be plausible to posit that the extremely experienced and highly trained helicopter pilots may be able to more efficiently allocate cognitive resources in order to overcome potential distractors. Research in driving literature suggests this is an unlikely explanation, with Cooper and Strayer (2008) showing no effects of practice on participants ability to overcome distractions. This finding could also be explained by the quality of the heads-up display stimuli, with 3D images more readily perceived than 2D (Dan & Reiner, 2017), meaning that although there was a greater amount of information available, it was moderated by how easily it was perceived – and how useful it might have been to the successful completion of the mission. Regardless of these contributing factors, it is clear that the extra symbology does not add additional cognitive workload to experienced pilots in the helicopter simulator.

Aside from effects of the symbology, the current study does show the usability and sensitivity of the DRT in a previously unexplored environment. The reliability and validity of the DRT has been well documented in driving environments to assess the drivers cognitive workload, however, there are limited applications in other scenarios or environments. With DRT results indicating higher levels of workload for more difficult conditions, we provide evidence that results translate across domains and show the applicability of the measure to helicopter simulator settings.

Figure 5 and Figure 6 (and further figures in `https://osf.io/2ntxw/`) show a novel

approach to jointly evaluate cognitive workload across a flight, an analysis we term "workload heatmap". This analysis shows not only the sensitivity of the DRT, but also provides scope for future analysis to track workload across the duration of a task. The "workload heat map" analysis may not be subjected to simple statistical comparison between conditions, but gives a visual reference to workload distributions across the flight path. We view this type of exploratory analysis as useful for future research (in order to manipulate workload) and for developers of adaptive interfaces.

In regards to flight performance, we analysed a variety of measures to form an objective view of flight quality. Initially we used several "gates" to assess flight performance. This is a commonly used technique in pilots' training and offers a good benchmark for pilots to achieve. The gates provide a marker of performance only for a limited number of locations during the flight, and it is difficult to draw conclusions from the specific gate-locations to the entirety of the task. Secondly, we assessed flight variability. We proposed that an optimal flight would follow a smooth trajectory, on both the horizontal and vertical planes, where sharp movements were indicative of performance lapses. However, over such a long and demanding flight path, which required constant positional adjustments, it was difficult to quantify this measure. Although flight variability gave some indication of the flight path, this was more indicative of the strategy taken (for example at night, it is advised that pilots fly lower) than of flight quality (as a good flight given the conditions may require high variability). Finally, we measured landing precision across the different environmental and symbology conditions.

Landings were a key criteria for the development of the added symbology, as it was used to assist pilots in difficult landing environments. The helicopter simulator was modelled on an Airbus MRH90, an aircraft commonly used in combat zones which require multiple and frequent landings. Consequently, the landing performance across the levels

of symbology was a key measure of flight performance, despite only offering a single value for the entirety of the flight. The CEP plots presented in Figure 2, give a clear indication of landing performance across landing zones. This analysis showed landings were more accurate for conditions of 3D symbology compared to conditions of no symbology and 2D symbology. This is a main finding for the current study, as it shows that the increased information provides assistance in difficult landing scenarios (such as at night and in brown–out). Results across environmental conditions show little variance within the 3D symbology condition, but degraded visual conditions have far greater impacts on the condition without symbology and the condition with 2D symbology. Although this metric generalizes performance across the entire flight to a single instance, it is useful when assessing the impacts of increased visual information.

Overall, results indicated that flying in degraded visual conditions led to higher cognitive workload and had a negative effect on flight performance (as indicated by flight variability and landings). Further, flight performance was unaffected by visual degradation when pilots were provided with 3D symbology. Assessing the flight performance in conjunction with the workload measure allows a more in-depth understanding of the effects of added information: 3D symbology adds no measurable workload, whilst assisting the pilots' flight performance. These results indicate that the 3D symbology may not always be useful for pilots, but is beneficial for night flying at no workload cost. These results show the effectiveness of the DRT as a cognitive workload measure outside of the driving environment and highlights the sensitivity of the DRT as a cognitive workload measurement tool for answering previously inaccessible questions.

The current study was limited by the total number of participants and a restricted stimulus set. Future studies should attempt to quantify what exact information is most useful to pilots, or whether certain symbology elements induce extra workload. Further

studies should look to assess the impact of this symbology in more advanced simulators through to real-world helicopter contexts, where the difficulty, and realism, of flying is increased. The impact across conditions of workload should also be assessed to understand whether an adaptive interface is more useful, where the level of information is updated given the difficulty of the current task.

Conclusion

The study offers a unique investigation into pilots' cognitive workload in a high-fidelity flight simulator. The analysis combines various flight metrics with simultaneous assessment of workload via the DRT. The analyses are somewhat limited by the lack of clear optimal main task performance definition and conjoint dual-task analysis. Similar to much cognitive workload literature in driving, it is often difficult to operationalize optimal main task performance, or provide a highly sensitive measure of main task performance. We have attempted to incorporate a variety of meaningful flight analysis alongside the cognitive workload measures to form a more rounded analysis of this exploratory study. The most telling results generally indicated the expected trends, with little flight path variability between conditions, but a greater effect of added information observed in landing data, where increasing the symbology consistently led to more accurate landings. Flight patterns were shown to vary between environmental conditions regardless of symbology, however landings were highly effected by symbology. Furthermore, the workload measures indicated that the increased symbology added no extra workload, and moderated workload in more degraded visual conditions.

Key Points

- Results show that the conformal 3D symbology assisted pilots landings.
- Results show that increasing the amount of information in the heads-up display accounted for no additional cognitive workload, but rather appeared to moderate workload.
- The detection response task is sensitive to workload changes in a simulated helicopter environment.
- We provide useful flight metric analysis, as well as a combined workload-flight map which may be useful in future designs and analyses.

References

Coleman, J. R., Turrill, J., Cooper, J. M., & Strayer, D. L. (2016). Cognitive workload using interactive voice messaging systems. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 60, pp. 1894–1898).

Cooper, J. M., Castro, S. C., & Strayer, D. L. (2016). Extending the detection response task to simultaneously measure cognitive and visual task demands. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 60, pp. 1962–1966).

Cooper, J. M., & Strayer, D. L. (2008). Effects of simulator practice and real-world experience on cell-phone—related driver distraction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(6), 893–902.

Dan, A., & Reiner, M. (2017). Eeg-based cognitive load of processing events in 3d virtual worlds is lower than processing events in 2d displays. *International Journal of Psychophysiology*, *122*, 75–84.

Engström, J., Larsson, P., & Larsson, C. (2013). Comparison of static and driving simulator venues for the tactile detection response task. In *Proceedings of the seventh international driving symposium on human factors in driver assessment, training, and vehicle design* (pp. 369–375).

Gerjets, P., Walter, C., Rosenstiel, W., Bogdan, M., & Zander, T. O. (2014). Cognitive state monitoring and the design of adaptive instruction in digital environments: lessons learned from cognitive workload assessment using a passive brain-computer interface approach. *Frontiers in neuroscience*, *8*, 385.

Haapalainen, E., Kim, S., Forlizzi, J. F., & Dey, A. K. (2010). Psycho-physiological measures for assessing cognitive load. In *Proceedings of the 12th acm international conference on ubiquitous computing* (pp. 301–310).

Hannula, M., Huttunen, K., Koskelo, J., Laitinen, T., & Leino, T. (2008). Comparison between artificial neural network and multilinear regression models in an evaluation of cognitive workload in a flight simulator. *Computers in biology and medicine*, *38*(11-12), 1163–1170.

Hays, R. T., Jacobs, J. W., Prince, C., & Salas, E. (1992). Flight simulator training effectiveness:

A meta-analysis. *Military psychology*, *4*(2), 63–74.

Huttunen, K., Keränen, H., Väyrynen, E., Pääkkönen, R., & Leino, T. (2011). Effect of cognitive load on speech prosody in aviation: Evidence from military simulator flights. *Applied ergonomics*, *42*(2), 348–357.

ISO:17488. (2016). *Road vehicles–transport information and control systems–detection-response task (drt) for assessing attentional effects of cognitive load in driving.* International Organization for Standardization Geneva, Switzerland.

JASP Team. (2019). *JASP (Version 0.11.0)[Computer software].* Retrieved from `https://jasp-stats.org/`

Jeffreys, H. (1961). *Theory of probability.* Oxford, UK: Oxford University Press.

Johnson, D. M., & Wiles, J. (2003). Effective affective user interface design in games [Journal Article]. *Ergonomics*, *46*(13/14), 1332-1345.

Kahneman, D. (1973). *Attention and effort.* Citeseer.

Kantowitz, B. H., & Casper, P. A. (2017). Human workload in aviation. In *Human error in aviation* (pp. 123–153). Routledge.

Krueger, G. P., Armstrong, R. N., & Cisco, R. R. (1985). Aviator performance in week-long extended flight operations in a helicopter simulator. *Behavior Research Methods, Instruments, & Computers*, *17*(1), 68–74.

Lee, J. D., Young, K. L., & Regan, M. A. (2008). Defining driver distraction. *Driver distraction: Theory, effects, and mitigation*, *13*(4), 31–40.

Münsterer, T., Schafhitzel, T., Strobel, M., Völschow, P., Klasen, S., & Eisenkeil, F. (2014). Sensor-enhanced 3d conformal cueing for safe and reliable hc operation in dve in all flight phases. In *Degraded visual environments: Enhanced, synthetic, and external vision solutions 2014* (Vol. 9087, p. 90870I).

Nelson, W. (1988). *Use of circular error probability in target detection* (Tech. Rep.). MITRE CORP BEDFORD MA.

Roenker, D. L., Cissell, G. M., Ball, K. K., Wadley, V. G., & Edwards, J. D. (2003). Speed-of-processing and driving simulator training result in improved driving performance. *Human factors*, *45*(2), 218–233.

Strayer, D. L., Cooper, J. M., McCarty, M. M., Getty, D. J., Wheatley, C. L., Motzkus, C. J., ... Horrey, W. J. (2019). Visual and cognitive demands of carplay, android auto, and five native infotainment systems. *Human Factors*, 0018720819836575.

Strayer, D. L., Cooper, J. M., Turrill, J., Coleman, J., Medeiros-Ward, N., & Biondi, F. (2013). Measuring cognitive distraction in the automobile.

Strayer, D. L., Cooper, J. M., Turrill, J., Coleman, J. R., & Hopman, R. J. (2015). Measuring cognitive distraction in the automobile iii: A comparison of ten 2015 in-vehicle information systems.

Strayer, D. L., Cooper, J. M., Turrill, J., Coleman, J. R., & Hopman, R. J. (2016). Talking to your car can drive you to distraction. *Cognitive research: principles and implications*, *1*(1), 16.

Strayer, D. L., Cooper, J. M., Turrill, J., Coleman, J. R., & Hopman, R. J. (2017). The smartphone and the driver's cognitive workload: A comparison of apple, google, and microsoft's intelligent personal assistants. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *71*(2), 93.

Strayer, D. L., & Johnston, W. A. (2001). Driven to distraction: Dual-task studies of simulated driving and conversing on a cellular telephone. *Psychological science*, *12*(6), 462–466.

Strayer, D. L., Turrill, J., Cooper, J. M., Coleman, J. R., Medeiros-Ward, N., & Biondi, F. (2015). Assessing cognitive distraction in the automobile. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(8), 1300–1324.

Thorpe, A., Nesbitt, K., & Eidels, A. (2019). Assessing game interface workload and usability: A cognitive science perspective. In *Proceedings of the australasian computer science week multiconference* (p. 44).

Townsend, J. T., & Eidels, A. (2011). Workload capacity spaces: A unified methodology for response time measures of efficiency as workload is varied. *Psychonomic Bulletin & Review*, *18*(4), 659–681.

Wagenmakers, E.-J., Lee, M. D., Lodewyckx, T., & Iverson, G. (2008). Bayesian versus frequentist inference. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Bayesian evaluation of informative hypotheses* (pp. 181–207). New York: Springer Verlag.

Watson, J. M., & Strayer, D. (2010). Supertaskers: Profiles in extraordinary multitasking ability.

Psychonomic Bulletin & Review, *17*, 479–485.

Wickens, C. D. (2002). Situation awareness and workload in aviation. *Current directions in psychological science*, *11*(4), 128–133.

Young, R. A., Hsieh, L., & Seaman, S. (2013). The tactile detection response task: preliminary validation for measuring the attentional effects of cognitive load. In *Proceedings of the seventh international driving symposium on human factors in driver assessment, training, and vehicle design* (pp. 71–77).

Zimmermann, M., Gestwa, M., König, C., Wolfram, J., Klasen, S., & Lederle, A. (2019). First results of lidar-aided helicopter approaches during nato dve-mitigation trials. *CEAS Aeronautical Journal*, 1–16.

Appendices

Appendix A: Glossary

- **Brown-out** - An instance where dust from below the helicopter is disturbed and rises to an altitude of about 120ft, thereby hampering the view for the pilot.
- **Collective lever** - controls the angle of the main rotor blades, allowing the helicopter to accelerate or decelerate.
- **Cyclic shaft** - changes the main rotors direction in order to change the direction of the helicopters movement.
- **FLIR** - Forward looking infrared radar. A sensor system that uses infrared light to see at night.
- **Ground Speed** - the speed (in knots) that the aircraft is travelling in reference to the ground
- **HUD** - Heads-up display. The information presented in the HUD is overlaid over the environment so that they do not have to shift gaze to perceive the stimulus.
- **Landing zone** (LZ) - a designated point on the map where pilots were to land. The landing zone was clearly marked in the symbology, on the map and by objects in the environment (i.e. the centre of a football field).
- **Radalt** - Radar altimeter measures altitude above the terrain that is currently beneath the aircraft.
- **LIDAR** - Light detection and ranging. A sensor system that uses pulses of laser light to measure variable distance to the ground.
- **Roll** - The degree of sideways movement in the aircraft
- **Pitch** - The degree of forward and back movement of the aircraft
- **Symbology** - The information given to pilots within their heads up display. Includes general flight metrics and more advanced environmental information.

Appendix B: Full Flight Path

Pilots were seated in the simulator and fitted with the visor and DRT's tactor patch. They were given instructions for responding to the DRT, and for completing the flight task. Three experimenters were present to collect data, with one experimenter collecting DRT data, another updating the parameters of the simulator, and a supervisor. An additional pilot was also present, navigating the participant through the flight as required. Pilots were instructed in the symbology presented in the 3D-symbology condition, and were given time to acclimate with the system. Before the flight commenced, the pilot was given a practice block of DRT trials to familiarise themselves with the stimulus and response button.

In the Day condition, visibility was set at 12,000m, time of day was set at 16:00, FLIR and dust were off. In the Night condition, FLIR was on and was set at 20:00 with FLIR visibility at 2,400m. General visibility in this condition was set at 12,000m, time of day was set at 20:00 and dust was off. In the Low Visibility, Dust condition, the dust appeared at 100m from the ground. Visibility in this condition was set at 1,200m, time of day was set at 16:00, dust was on and FLIR was off.

The flight task was divided into six sections. Way points were placed throughout the map to indicate the key points. Way points were marked on the control panel map and indicated in the symbology (for both 2D and 3D conditions). Section 1 required the pilots to take off from a designated helipad and fly to two waypoints, designated Way-point A and Way-point B. In Section 2, pilots landed at their first LZ, designated LZ 1, which was a flat sandbank. Pilots encountered brownout during this landing. Brownout began at 100ft, with a simulated brownout fully engulfing the virtual aircraft to restrict view by roughly 60ft. Section 3 was a second flight section, in which pilots followed a river through a valley to Way-points C and D, marked on two bridges along the valley, and Way-point E, marked on a church at the end of the valley. Pilots were given ideal

speed and height levels of 80kn and 200ft, and instructed to fly as close to these levels as possible during this section. Section 4 required pilots to descend to a LZ, designated LZ 2, which was marked on a triangular brown field. Pilots were instructed to "go around" or abort the landing at height of 20ft. Going below this set altitude in a real-world scenario would be potentially dangerous and could compromise mission objectives. As with LZ 1, pilots encountered brownout, which was removed when pilots cleared power lines located behind LZ 2. Section 5 was the final flight section, in which pilots ascended and descended a mountain, flying towards Way-point G nearby the take-off helipad. Section 6 was the final landing on the flight deck of a Nimitz-class aircraft carrier. The LZ, designated LZ 3, was the junction of the centre lines of the carrier's straight runway and angled runway. The full flight took approximately 13 minutes to complete. Pilots were seated in the simulator and fitted with the visor and DRT's tactile patch. They were given instructions for responding to the DRT, and for completing the flight task. Three experimenters were present to collect data, with one experimenter collecting DRT data, another updating the parameters of the simulator, and a supervisor. An additional pilot was also present, navigating the participant through the flight as required. Pilots were instructed in the symbology presented in the 3D-symbology condition, and were given time to acclimate with the system. Before the flight commenced, the pilot was given a practice block of DRT trials to familiarize themselves with the stimulus and response button. Pilots were instructed to begin the flight upon responding to the first DRT stimulus they perceived. After completing the first two sections, including landing at LZ 1, pilots were instructed to take off and continue the flight after several seconds on the ground (following standard flight procedures). They then completed the last four sections of the flight.

Appendix C: Symbology Conditions

- No symbology: In this condition, the pilot was equipped with the HUD headpiece (as shown in Figure 7, however, it was turned off so that pilots could still see the full disply with no extra visual information.

Figure 7. : An example of the simulator setup. The pilot has the head piece attached which displays the HUD information over the simulated environment. In front of the pilot are the electronic map and a multi-function display, which indicated altitude, ground speed, collective power and helicopter roll.

• 2D: In the 2D condition, pilots were equipped with the HUD headpiece which displayed several metrics in their visual field. These metrics included radial altitude, ground speed, heading, distance & direction to the landing zone. An example screenshot can be seen in Figure 8.

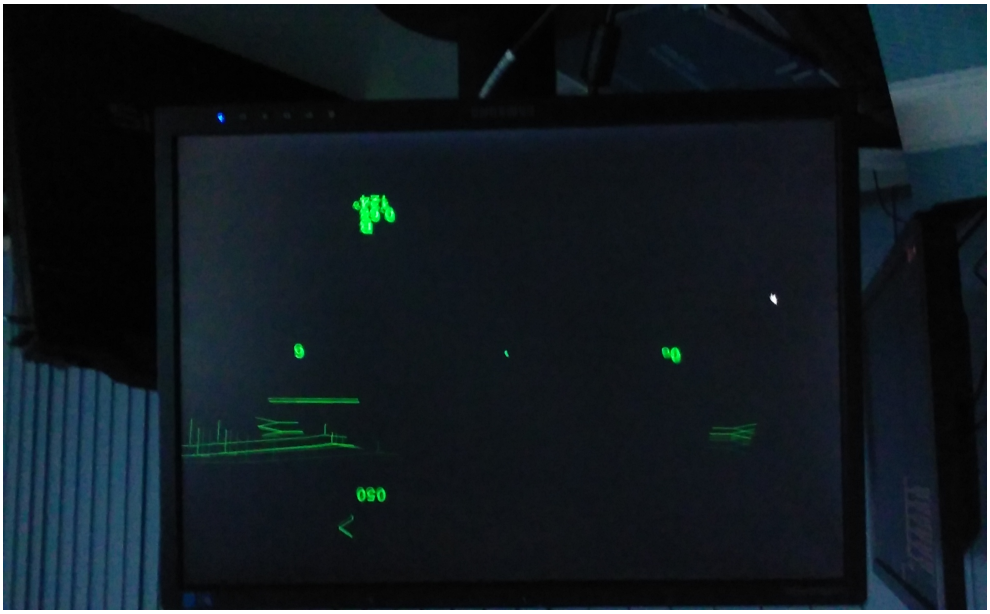

Figure 8. : An example of the projections for the 2D symbology condition. The information shown on screen was projected to the HUD in the headpiece worn by the pilot.

• 3D: In the 3D condition, pilots were equipped with the HUD headpiece which displayed several metrics in their visual field, as well as overlaying 3D visual information to the simulated environment. These metrics included radial altitude, ground speed, heading, distance & direction to the landing zone. The 3D information also given to pilots included 3D mapping of landing zones (as seen in Zimmermann et al. (2019)), flight path direction, and visual indication of obstacles (such as buildings and power lines; as shown in Figure 9).

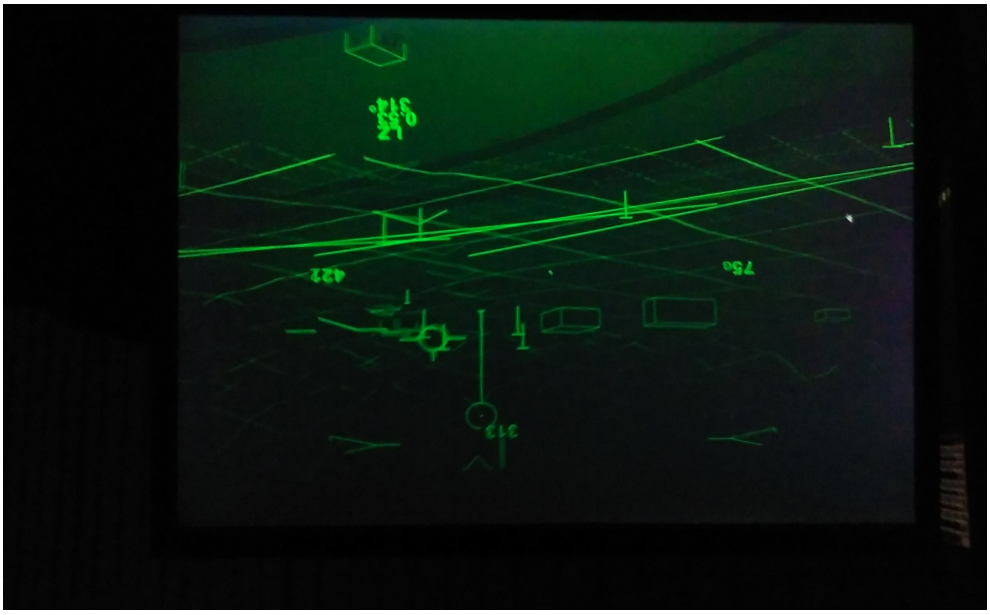

Figure 9. : An example of the projections for the 3D symbology condition. The information shown on screen was projected to the HUD in the headpiece worn by the pilot.

Reilly J. Innes, University of Newcastle, received his B Psych (hons) from the University of Newcastle in 2016.

Zachary L. Howard, University of Newcastle, received his BS in Psychology from the University of Newcastle in 2014.

Alexander Thorpe, University of Newcastle, received his B Psych (hons) from the University of Newcastle in 2016.

Ami Eidels, University of Newcastle, received his PhD in cognitive psychology from Telaviv University, Israel in 2006.

Scott D. Brown, University of Newcastle, received his PhD in cognitive psychology from the University of Newcastle, Australia in 2002.