Multivariate Analysis of Gaze Behaviour and Task Performance within Interface Design Evaluation

Blundell, J., Collins, C., Sears, R., Plioutsias, A., Huddlestone, J., Harris, D., Harrison, J., Kershaw, A., Harrison, P. & Lamb, P

Original citation:
Blundell, J, Collins, C, Sears, R, Plioutsias, A, Huddlestone, J, Harris, D, Harrison, J, Kershaw, A, Harrison, P & Lamb, P 2023, 'Multivariate Analysis of Gaze Behaviour and Task Performance within Interface Design Evaluation', IEEE Transactions on Human-Machine Systems, vol. (In-Press), pp. (In-Press).
https://doi.org/10.1109/THMS.2023.3305715

DOI 10.1109/THMS.2023.3305715
ISSN 2168-2291
ESSN 2168-2305

Publisher: Institute of Electrical and Electronics Engineers

Multivariate Analysis of Gaze Behavior and Task Performance Within Interface Design Evaluation

James Blundell, Charlotte Collins, Rod Sears, Tassos Plioutsias, John Huddlestone, Don Harris, James Harrison, Anthony Kershaw, Paul Harrison, and Phil Lamb

***Abstract*—Eye tracking technologies have frequently been used in sport research to understand the interrelations between gaze behavior and performance, using a paradigm known as vision-for-action. This methodology has not been robustly applied within the field of interface design. The present work demonstrates the benefit of employing a vision-for-action paradigm for interface evaluation. This is demonstrated through the evaluation of a novel task-specific symbology set presented on a head-up-display (HUD), developed to support pilots conduct ground operations in low-visibility conditions. HUD gaze behavior was correlated with task performance to determine whether certain combinations of gaze behavior could produce effective predictive performance models. A human-in-the-loop experiment was conducted with 11 professional pilots who were required to taxi in a fixed-base flight simulator using the HUD symbology, while gaze data toward the different HUD symbology elements was collected. Performance was measured as centerline deviation error and taxiing speed. Results revealed that appropriately timed gaze behavior toward task-specific elements of the HUD were associated with superior performance. During turns, attention toward an undercarriage lateral position indicator was associated with reduced centerline deviation ($p < 0.05$). The findings are interpreted alongside detailed posttrial user-feedback of the HUD symbology to illustrate how eye tracking methodologies can be incorporated into interface usability evaluations. The joint interpretation of these data demonstrates these novel procedures, the findings contribute to enhancing the wider domain of interface design evaluation.**

***Index Terms*—Eye tracking, head-up-display, human performance, interface design, multivariate analysis.**

I. Introduction

EYE tracking technology's capability to provide a direct link between human gaze behavior and attention has led to it being adopted across a wide range of domains, including education [1], [2] and healthcare [3]. Within the transport domain reviews of automotive [4] and aviation [5] eye tracking research highlight the valuable insights the technology has granted in understanding human information processing. Conventionally, gaze behavior is measured by dividing the visual space around the user into distinct regions of interest (ROI). In aviation research, the distribution of attention on the flight deck can be examined by defining separate ROIs for different physical cockpit elements. Such an approach has yielded findings that ∼60–70% and ∼30–40% of pilot attention is allocated to the outside scene and instruments, respectively [6], [7]. While the majority of aviation eye tracking research has involved investigating attention using ROIs that define relatively large head-down (e.g., specific cockpit displays) or head-up (e.g., the outside scene) locations [8], [9], few studies have employed smaller ROIs with sufficient spatial granularity to describe gaze behavior toward the discrete symbology elements of displays. One example being from Sarter et al. [10], who assessed pilot gaze behavior toward six ROIs within a primary flight display (PFD) during a full-mission simulation. The most fixated areas were the artificial horizon (∼25%), altitude tape (∼30%), and airspeed tape (∼20%). Consequently, the current state of eye tracking research in aviation describes pilot attention with insufficient "spatial resolution."

Analyzing eye movements is important in display design. Pilot gaze behavior is affected by onboard technologies, including color-coded avionic displays [11], traffic and weather displays [12], synthetic vision displays [13], and highway-in-the-sky (HITS) symbology [14]. Findings show these technologies bestow both positive and negative impacts on pilot attention. In the latter case, the presentation of a compelling HITS can divert attention away from critical external events [13]. This "spatial resolution" issue of aviation eye tracking research is particularly relevant to display design. Future interface optimization will require greater scrutiny to examine how varying symbology configurations affect pilot attention.

Manuscript received 26 May 2023; accepted 5 August 2023. This work was supported in part by Aerospace Technology Institute (ATI) Programme under Grant 113108, a joint Government and Industry Investment to Maintain and Grow the U.K.'s competitive position in civil aerospace design and manufacture. The programme, delivered through a partnership between the ATI, Department for Business, Energy & Industrial Strategy and Innovate U.K., addresses technology, capability, and supply chain challenges. This article was recommended by Associate Editor L. L. Chen. *(Corresponding author: James Blundell.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Coventry University Ethics Committee under Application No. P94106 and performed in line with the British Psychological Society.

James Blundell, John Huddlestone, and Don Harris are with the Institute of Clean Growth and Future Mobility, Coventry University, CV1 5FB Coventry, U.K. (e-mail: james.blundell@coventry.ac.uk; ab4919@coventry.ac.uk; ab3693@coventry.ac.uk).

Charlotte Collins, Rod Sears, and Tassos Plioutsias are with the Faculty of Engineering, Environment and Computing, Coventry University, CV1 5FB Coventry, U.K. (e-mail: charlotte.collins@coventry.ac.uk; ac3511@coventry.ac.uk; ad3903@coventry.ac.uk).

James Harrison, Anthony Kershaw, Paul Harrison, and Phil Lamb are with the BAE Systems Ltd, ME1 2XX Rochester, U.K. (e-mail: james.harrison@baesystems.com; anthony.kershaw@baesystems.com; paul.a.harrison@baesystems.com; phil.lamb666@gmail.com).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/THMS.2023.3305715.

Digital Object Identifier 10.1109/THMS.2023.3305715

Eye tracking can compliment display design through the employment of a vision-for-action paradigm [15], [16]. This involves examining the interrelations between eye movements and performance to determine whether specific gaze types are related to targeted task behaviors. Milner and Goodale [17] proposed a dorsal-lateral anatomical split within the visual cortex that can be interpreted as two independent functional modules: vision-for-perception and vision-for-action. Vision-for-action paradigms have been used extensively in sport [15], [16], and have been valuable in describing how particular gaze behaviors are associated with task expertise. In the aviation domain, Ziv [5] identified a lack of eye tracking studies employing vision-for-action paradigms.

A vision-in-action paradigm is employed in this study to investigate the benefits of a novel user-centered designed (UCD) head-up-display (HUD) taxi symbology set, developed to support low-visibility airport surface operations. Such operations are one of the most difficult phases [18] as pilots must maintain awareness of their cleared taxi route, and their position relative to the cleared route. In addition, pilots are required to monitor airport signage and markings and compare this information to the taxiway map. This is further complicated in low-visibility conditions. The current HUD symbology set was developed to provide navigational support during the above situations, when the quality of external navigational cues is degraded and where the benefit of onboard navigation aids is challenged. Previous studies [8], [19], [20] have shown pilots taxi faster and more accurately with bespoke HUD taxiing symbology versus paper charts in reduced visibility conditions. While eye tracking was implemented in one study [8], no studies have applied paradigms, such as a vision-for-action, that describe the link between attention to the discrete symbology elements and pilot performance.

The HUD can be considered as an augmented reality (AR) display, a format that has garnered significant interest in the past decade across a range of domains [21]. This is due to its capability to generate both "conformal" and "non-conformal" symbology. The primary difference between conformal and nonconformal symbology is the frame of reference, the former is geographically "scene-linked" relative to the outside scene, while the latter is located relative to the HUD's real-estate [22]. Conformal symbology facilitates cognitive processing of both the symbology and the external scene [23], [24]. However, several eye tracking studies have examined how conformal symbology can also detrimentally affect attentional mechanisms [13], [25]; the conformal symbology "locks-in" the user's attention for longer than is optimal resulting in task-relevant information in the external scene being ignored, a phenomenon known as "attentional tunneling." This study offers insights into this phenomenon as the symbology contains both conformal and nonconformal symbology.

In the current experiment, performance and eye tracking data were collected during a low-visibility simulated taxiing task. Data were analyzed using a combination of factorial and multivariate general linear mixed models (GLMMs) to interpret the relationship between these features and their combinations. Results are interpreted with emphasis on how attention to individual symbology elements is related to task performance. These findings are expanded on by integrating qualitative participant feedback of the interface. The joint interpretation of these data presents novel procedures and findings that can be applied to the wider domain of interface design evaluation.

II. Methods

A. Participants

Eleven professional pilots, holders of an Airline Transport Pilots License (ATPL), participated in the study. Average flying experience was 19.72 years (SD = 15.75) and 6011 hours on type (SD = 7121). Thirty-six percent had experience of using a HUD (4/11). One pilot was unable to take part in the debrief interview. The experiment was approved by the Coventry University Ethics Committee.

B. HUD Symbology

The symbology was designed using UCD principles [26] over a 6-month period. A workshop with three subject matter experts (SMEs) was held to establish the user requirements and task relevance of the symbology (see Table I). SMEs were HUD experienced test pilots, senior airline training captains, and a HF engineer involved in the design, development, and certification of civil aircraft. The SMEs provided input into three subsequent workshops to iteratively optimize the symbology.

Fig. 1(a) presents the HUD symbology, highlighting elements that were conformal ("scene-linked") or nonconformal. Nonconformal elements included a ground speed/throttle dial in the bottom-left corner of the display. At the top of the display was a raw data indicator showing the linear deviation of the nose wheel and main gear from the taxiway centerline (along with 10-m deviation increment markers). Located on the right of the display was a moving map, beneath which was a distance to turn dial. The conformal elements of the display included an overlay of the taxiway centerline denoting relevant routing information and hold bars representing runway hold positions. The latter hold bar symbology were omitted from the gaze analysis due to not being continuously present during trials. Nonconformal representations of the conformal symbology were also provided on the moving map element. More detailed descriptions of the symbology are provided as supplementary material. The total viewing area of the HUD was 30° x 22.5° of visual angle (VA).

C. Simulator Facility

A fixed-wing simulator running X-plane 11 Professional (Laminar Research), running the flight model of a Boeing 737 type aircraft, was employed. The simulator was equipped with a $180° \times 40°$ collimated projection system enabling participants to experience the equivalent real-world depth perception required for accurate perception of HUD conformal symbology. Each participant was seated in the left-seat position with the tiller located to their left. A custom, BAE Systems, data logging program was developed to interface with the flight model and simulator environment to drive the HUD symbology (60 Hz) and retrieve relevant X-Plane data references (4 Hz sampling rate)

TABLE I
HUD SYMBOLOGY USER REQUIREMENTS AND DESIGN RATIONALE

HUD Symbology	User Requirement / Task-Relevance
Groundspeed Radial	• Real-time ground speed (digital and analogue clock format) and throttle (N1) setting. ○ Eyes out knowledge of ground speed and thrust setting when taxiing, without having to look down to relevant instrumentation. • Acceleration / deceleration vector arrows ○ If the aircraft is slowly accelerating or decelerating due to sloping surface condition, then action can be taken to adjust maintain the amount of throttle/brake being required.
Mini-Map	• Aircraft position indication relative to a scrolling 2D aircraft surface map. Depicts runway and taxiway features in the vicinity of the current aircraft track. • The cleared taxi route provided by ATC is shown as a bold dotted line overlaid on the 2D map together with the location of mandatory hold bars along the planned route.
Conformal Route Line	• The conformal route line is drawn on the HUD as a bold dotted line (together with mandatory hold bars) that overlays the natural view of the designated runway / taxiway features. • Aids identification and execution of turns of varying severity and provides particular benefit at complex junctions and under poor visual conditions.
Distance to Turn Indicator	• Comprises several features to provide turn awareness information in preparation for making a turn: ○ A digital / analog clock format countdown of distance to the next turn. ○ The integral runway / taxiway indicator presents the name of the current runway / taxiway together with an arrow indicating the direction and designation of the next turn.
Undercarriage Position Indicator	• The undercarriage position indicator provides a continuous indication of nose wheel and main gear position relative to the runway / taxiway centerline to aid judgement of whether wheels are too close to the edge. • Representation of both nose wheel and main gear aids the manoeuvring of large aircraft types which require offsetting of the nose wheel ahead of the centerline during tight turns to maintain the main gear on the centerline.

for performance analysis. Eye movement data were captured at 25 Hz using a Dikablis head mounted eye tracker (Ergoneers, GmbH). The eye tracker has both good gaze direction accuracy (0.25°) and precision (0.25 RMS).

D. Task and Procedure

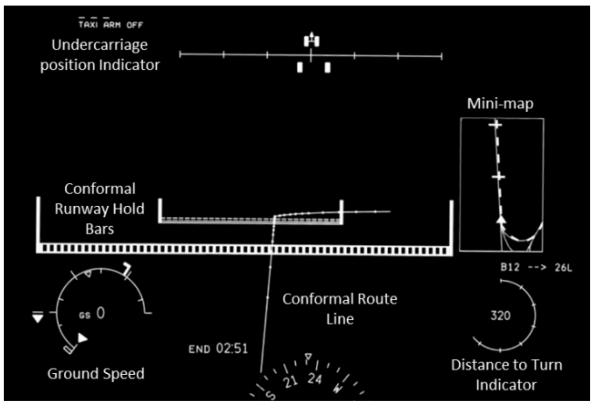
Munich airport (EDDM) was used for the experiment. Participants taxied in low-visibility conditions (CAT-III) along four different 5-min (approximate) routes that consisted of 120°, 90°, and S-Bend turns, and a series of straights. The different route segments are illustrated in Fig. 1(b). Participants received a video briefing to introduce them to the HUD interface features. This was followed by a practice session in the simulator to familiarize them with the layout, the aircraft's maneuvering capabilities, and the HUD feedback behaviors. Participants were equipped with the eye tracker, calibrated using a 4-point calibration array. Six trials were completed. HUD symbology was present in half of the trials, the order of which was counterbalanced. Eye tracker calibration was checked and amended (if necessary) between trials. At the end of the six trials, the participants took part in a 30-min structured debrief to obtain insights into the usability benefits of the HUD interface features. The experiment lasted approximately 2.5–3 h.

E. Outcome Measures

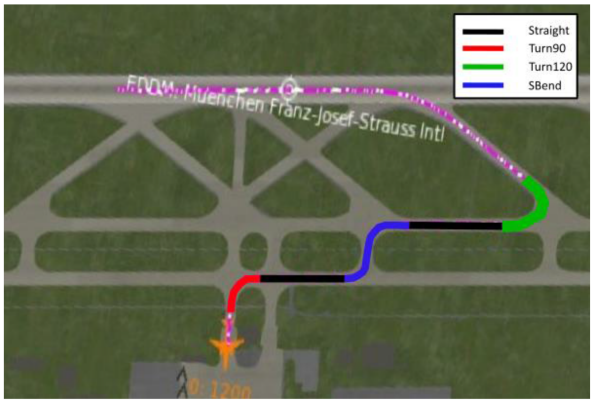
1) Performance: X-Plane data references generated the following variables: 1) participant main gear lateral deviation from the taxi centerline in meters (*MG* root-mean-squared-error (RMSE)), and; 2) ground speed in knots (*GS*).

2) Eye Tracking: Gaze point data mapped upon a 576 (height) x 768 (width) pixel forward facing field camera was analyzed offline using custom MATLAB Image Processing functions, including the removal of blinks and application of a velocity-based threshold to separate raw gaze point data into fixation and saccade eye movements. Fixation dwell times were calculated as the proportion of fixations allocated to ROIs created for the following head-up symbology elements (ROI sizes in VA): ground speed/throttle radial (*GS*—5° x 5°); undercarriage position indicator (*Wheel*—15.5° x 3°); airport mini-map (*Map*—4.25° x 5.5°); distance to turn radial (*Turn*—4.25° x 4.25°); conformal taxiway route (*Line*—8.75° x 8.75°).

3) Posttrial Interview: Qualitative feedback on the holistic benefits of the symbology, and the functional and physical qualities of the individual elements, was collected in interviews. Discussions were facilitated by requiring participants to provide quantitative usability feedback for each HUD symbology element using five custom-made scales (see supplementary material). Standardized usability scales were rejected as they



(a)



(b)

Fig. 1. (a) Labeled example of the taxi navigation head-up symbology. (b) Example of an EDDM route. The four different route type segments are highlighted.

provided an overly generalized measurement of system intuitiveness that offered less explicit, valuable, design insights. The current custom-made scales contained three functional properties scales (task relevance, safety benefit, intuitiveness) and two physical properties (size, location) scales. Participants rated their agreement with statements related to the elements (e.g., "The [Distance to Next Turn Indicator] contained useful task-related information") on a 7-point scale: 1 = Strongly Disagree; 4 = Neither Agree nor Disagree; 7 = Strongly Agree. They were asked to expand descriptively upon their rating. This allowed for comparison between symbology elements and encouraged a robust design-related discourse.

F. Data Analysis

1) Factorial GLMM: Preliminary factorial GLMM analyses were conducted separately on performance and gaze behavior data to inform the input parameters of the subsequent multivariate GLMM. GLMMs are a powerful and flexible variate of linear models that allow modeling of "fixed" and "random" effects. Fixed effects model *systematic* changes in experimental variance being manipulated in the experiment. Random effects enable a degree of structure to be assigned to a model's error variance—normally expressed as a generalized error term within traditional linear models. Critically, random effects for "participant" and "trial" are often defined, which can characterize/control for idiosyncratic variations that are due to individual differences (e.g., background) and changes in performance over the course of an experiment (e.g., fatigue). In turn, GLMM's power lie in their ability to accommodate the alternative correlation structures of repeated-measures research designs, making them preferable to traditional ANOVA where unavoidably small samples are involved [27]. A feature of research where professional populations (e.g., pilots) and costly laboratory procedures (e.g., high-fidelity simulation) are commonplace, as in the current study. GLMMs can also handle missing data, which would otherwise require listwise deletion or cumbersome data substitution solutions with traditional ANOVA. GLMM best-practice guidance by Meteyard & Davies [28] was followed. All GLMM analysis was conducted using the MATLAB (2021b) Statistical Toolbox.

The performance data, *MG* (log-transformed), and *GS*, from all six trials were fitted with a fixed effect for *HUD* (two levels: HUD ON/OFF). Performance comparisons between the taxi route segments [see Fig. 1(b)] were examined using a fixed effect named *Route* (4 levels: Straight, Turn90, Turn120, SBend).

The analysis of fixation dwell time included only data from the three trials, where HUD symbology was present. A fixed effect containing five levels representing the different HUD ROIs (*ROI*: Line, Map, Turn, Wheel, Speed) was used. The same *Route* fixed effect from the performance analysis was included.

Random participant and trial intercepts were used as random factors. In this way, participant random effects accounted for individual differences in pilot experience, which can contribute to changes in pilot gaze behavior [29]. Likewise, trial-based random effects controlled for practice related gaze behavior changes. Interpretation of interactions and main effects were checked using likelihood ratio tests that compared models with and without relevant terms, providing a Chi-Square (χ^2) means of model comparison [30]. Simple effect p-values were computed with a Satterwaite approximation to degrees of freedom. Size and confidence of significant fixed effects are described using model slope coefficients (β) alongside their respective standard errors (se).

2) Multivariate GLMM: The multivariate GLMM analysis that underpinned the vision-in-action paradigm, explored the relationship between fixation behavior and performance. Performance data was the dependent variable; fixation dwell time data of each ROI were predictor variables. Selection of ROI predictor variables was achieved by comparing candidate model Akaike Information Criterion (AIC). The model with the lowest AIC value was selected as the best model.

All GLMM analysis assumptions of reported models were met. Model residuals were normally distributed and exhibited homoscedasticity. All reported models converged.

3) Posttrial Interview: Interviews were recorded, transcribed, imported into NVivo (version 1.3), and analyzed using thematic analysis [31]. This involved the development of an initial template of higher order hierarchical thematic categories.

Three initial primary themes facilitated the UCD process: *Safety*, *Utility,* and *Design*. *Safety* and *Utility* themes concerned participant feedback describing how symbology enhanced safety

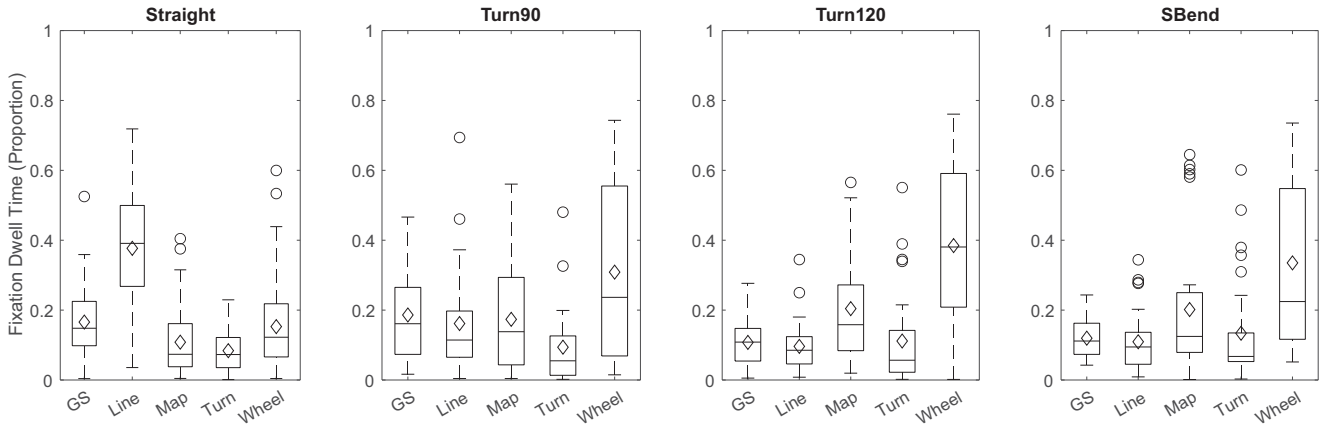


Fig. 2. Fixation dwell time proportion boxplots grouped by *ROI* and *route*. Means included as diamond symbols.

and how it was used during the taxiing task. The *Design* theme included evaluations from participants of the physical and functional properties of specific HUD symbology elements. Inductive coding for secondary themes described the structure of the three deductively coded primary themes in greater detail. No novel theoretical insights for secondary themes were generated after the third participant interview (the point that data saturation was reached). Coding consistency was assured by triangulation; a separate researcher reviewed and recoded the data, resulting in 86% agreement between original and recoded data. Subsequent dialog between coders tackled the discrepancies until agreement was met. Descriptive statistics from the interview scale data complemented the qualitative analysis.

TABLE II
FIXATION DWELL TIMES VERSUS MAIN GEAR DEVIATION CORRELATIONS

ROI	*Straight*	*Turn90*		*Turn120*		*SBend*	
GS	0.08	0.57	**	0.08		0.28	
Turn	-0.15	-0.12		0.33		<0.01	
Wheel	-0.14	-0.43	*	-0.46	**	-0.46	**
Map	0.20	0.34		0.18		0.16	
Line	-0.03	-0.23		0.20		0.39	*

Note: * $p < 0.05$ and ** $p < 0.01$

III. RESULTS

A. Factorial GLMM

1) Performance: Overall mean *MG* RMSE with and without the HUD taxi guidance was 2.72 and 3.79 m, respectively. A *HUD* by *Route* interaction was found, χ^2 (3) = 8.48, $p < 0.05$, such that there was a specific benefit of the HUD guidance in reducing *MG* RMSE on the 120° turn route segment ($\beta = 1.67$, se = 1.18, $t = -3.13$, $p < 0.01$). The *HUD* main effect was not significant, χ^2 (1) = 0.67, $p = 0.42$. *GS* results found neither a main effect for *HUD,* χ^2 (1) = 1.04, $p = 0.31$, nor an interaction including *HUD* and *Route*, χ^2 (3) = 1.39, $p = 0.71$.

2) Fixation Dwell Time: Fig. 2 presents the proportion of dwell time participants allocated to the different HUD ROIs across the route segments. Gaze behavior changes during straight route segments were characterized by a main effect of *ROI,* χ^2 (4) = 25.76, $p < 0.001$. Dwell time toward the *Line* ROI was significantly greater than other ROIs ($\beta = 0.22$, se = 0.02, $t = 9.48$, $p < 0.001$). A significant *ROI* by *Route* interaction, χ^2 (12) = 385.64, $p < 0.001$ revealed that *Line* ROI dwell times decreased during the turns ($\beta = -0.21$, se = 0.02, t = -9.44, $p < 0.001$). No difference ($p > 0.05$) was found in *Line* ROI dwell time between turn segments. Conversely, dwell time increased on the *Wheel* ROI on Turn90 segments ($\beta = 0.15$, se = 0.02, $t = 6.56$, $p < 0.001$), and increased equally toward the *Wheel* ROI during Turn120 and SBend segments ($\beta = 0.21$, se = 0.02, $t = 11.93$, $p < 0.001$). Compared to dwell times during the straight segments, fixations to the *Map* ROI did not increase during Turn90 segments ($p = 0.35$) but did on Turn120 and SBend segments ($\beta = 0.09$, se = 0.02, $t = 5.25$, $p < 0.01$). Dwell time for the *Turn* and *GS* ROI did not change across route segments.

3) Multivariate GLMM: The correlation matrix for taxi *MG* performance with ROI dwell times, grouped by route segment, is shown in Table II. Consistent negative correlations were found across the turn segments between *Wheel* ROI dwell time and *MG* RMSE ($p < 0.05$); RMSE lateral deviations decreased with a concomitant increase in percentage dwell time on the *Wheel* ROI.

Based on the factorial GLMM analysis, a multivariate GLMM analysis of *GS* was not undertaken. A multivariate analysis was conducted with *MG* RMSE, using turn route segment data only. Model selection using AIC comparisons included a maximal model containing predictor variables representing each of the five HUD ROIs. The top seven ranked AIC models are presented in Table III. The best model (ΔAIC = 0) included the *Wheel* and *Map* ROIs as predictors of *MG* RMSE during turn segments ($R^2_{\ Adjusted} = 0.53$). Model coefficients of the best model (see Table IV) reveal that the fit reflected the correlation findings (see Table II). Attention allocation during turn toward the *Wheel* ROI ($t = -2.46$, $p < 0.05$) was associated with reduced centerline deviation, while the opposite was found for attention toward the *Map* ROI ($t = 2.24$, $p < 0.05$). The gaze heat maps from two trials, where *MG* RMSE accuracy

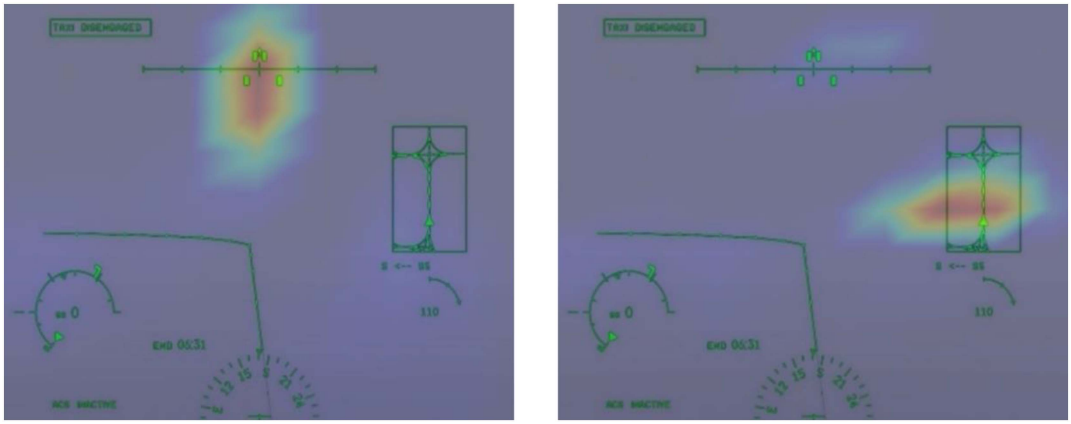


Fig. 3. Symbology gaze heat maps for two participants during an S-Bend turn. The left and right panels are from participants who had a *MG* RMSE of 1.4 (good) and 6.5 m (poor), respectively.

TABLE III
MULTIVARIATE MODEL SELECTION OF *MG* RMSE AND ROI DWELL TIME

Model	*K*	AIC_c	ΔAIC_c	*AICw*	*Rsq*
Generating Model: GS + Wheel + Line + Turn + Map Random Effects Structure : *(1|Participant) + (1|Trial)*					
Wheel + Map	3	167.3	0.00	0.16	0.53
Wheel + Line	3	167.4	0.11	0.15	0.53
GS + Map	3	167.6	0.34	0.13	0.54
Wheel	2	168.3	1.01	0.09	0.49
Map	2	168.9	1.57	0.07	0.52
Wheel + Map + GS	4	169.3	1.98	0.06	0.55
Wheel + Line + Turn	4	169.7	2.44	0.05	0.54

TABLE IV
MG RMSE BY HUD ROI DWELL MODEL SUMMARY

Equation: RMSE~Wheel + Map + (1|Participant)+(1|Trial)
Model fit: $R^2_{Adjusted}$= 0.53

Fixed Effects (*Transformed (exponentials) logarithmic model estimates are presented)*

	Beta	SE	95% CI	*t*	
Intercept	4.08	1.23	2.69 – 6.17	6.73	**
Wheel	-2.46	1.44	-5.08 – -1.19	-2.46	*
Map	2.42	1.48	1.12 – 5.32	2.24	*

Random Effects

	Variance	95% CI	Corr
Participant (Intercept)	1.46	1.26 – 1.88	-
Trial (Intercept)	1.16	1.05 – 1.60	-

Note: * $p < 0.05$ and ** $p < 0.01$

was either good or poor as shown in Fig. 3. The heat maps exemplify the model results from Table IV. Fig. 4 shows the plotted model estimates for *MG* RMSE during turn segments from the best model based on *Wheel* and *Map* ROI dwell time as predictors.

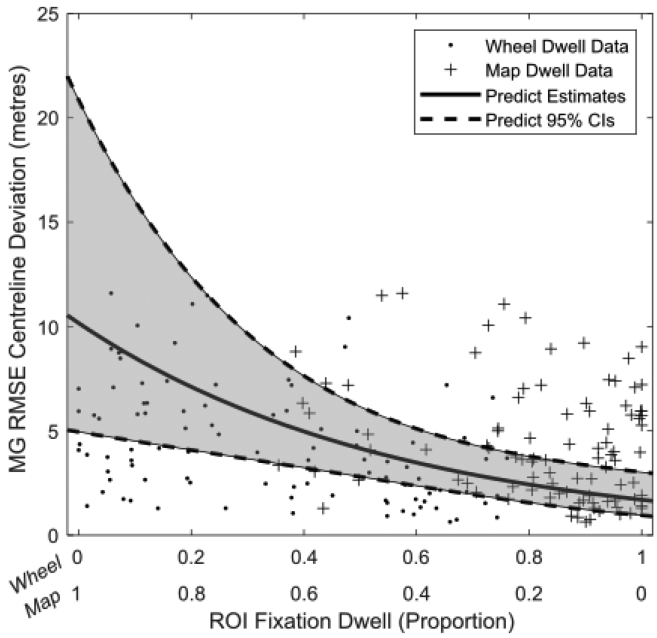


Fig. 4. *MG* RMSE centerline deviation model estimates and 95% CIs based on HUD Wheel and Map ROI fixations during turns. The *x*-axis presents dwell time proportions for both wheel and Map ROIs, with the scale of the latter being reversed.

B. Posttrial Interview

A selection of anonymized participant comments is provided which best exemplified the thematic analysis.

1) Safety: Three secondary themes were deductively identified within this primary theme: *Situation Awareness, Attention Capture,* and *Mechanism of Improvement.*

Situation Awareness related to participant reports on how symbology enhanced task and environmental awareness. For example, awareness enhancements during turns were attributed to the centerline deviation indicator: "*The main gear symbology massively aids SA as you need to know where your wheels are from a safety perspective.*" [participant 9] and "*Having the stop bars was useful, without them there would be no chance of me being able to see them*" [participant 1]. The mini-map

was praised for supporting task awareness: "*It's helpful to have the different taxiways presented on the mini-map, they help you count down to your turn.* [participant 8].

Negative opinions were voiced in the *Attention Capture* secondary theme that emphasized how undue attention could be directed to certain symbology elements. Notably, the conformal centerline element: "*So that was a bit of a concern for me, that I was just following the green line.*" [participant 1].

The final secondary theme, *Mechanism of Improvement,* represented remarks on how the symbology enhanced taxiing performance (e.g., managing speed). For example: "*The wheel indication allows you to more accurately maintain the centerline."* [participant 5] and "*the acceleration trend arrows were excellent and helped me manage my speed and acceleration"* [participant 2].

2) Utility: Scanning Behavior and *Task-Specificity* were the secondary themes identified. *Scanning Behavior* highlighted how pilots allocated their attention within the symbology space, supporting the eye tracking results: "*I was trying to scan nose wheel, speed, nose wheel, map, nose wheel, speed*." [participant 2]. Conversely, comments were also offered that indicated which parts of the display were ignored: "*I didn't find myself using the compass rose.*" [participant 3].

The *Task-Specificity* secondary theme concerned "when" symbology elements were attended to, reflecting the fixation dwell time findings (see Fig. 2). For example: "*In those turns, the nose wheel indicator becomes the center of your scan behavior.*" [participant 5]. Likewise: "*I found the speed indicator useful at all parts of the trial.*" [participant 4].

3) Design: This theme accompanied participant ratings of the different HUD symbology elements' functional and physical properties (see Fig. 5). Overall, functional properties of the symbology were rated highly across the three dimensions (mean/SD score: 6.02/1.35). A notable exception was a consensus among participants that the undercarriage indication was the least intuitive symbology element (mean/SD score = 4.90/1.79). In terms of the physical properties, participants expressed positive opinions regarding the size and position of the symbology within the display area (mean/SD score = 6.45/1.21). However, lower agreement scores were found for the size of the mini map element (mean/SD score = 5.00/2.45).

Positive comments on the functionality of the symbology were captured in a secondary theme *Intuitive*. Many participants commended the overall intuitiveness of the HUD symbology, praising the architecture and comparing it favorably to a PFD: "The *architecture made sense. It had some commonalities with the PFD, such as speed on the left for example which I think was a useful feature*." [participant 5] and "*The arrangement was very similar to what you would find on a PFD. For example, yaw on the top and a compass rose at the bottom. The layout is very intuitive and everything is where you would expect it to be"* [participant 8]. Specific elements considered, for instance, the ground speed: "*Intuitive to use. I mean this is just like a standard airspeed indicator, it is classic glass cockpit stuff*" [participant 11].

A secondary theme, *Training Requirements,* identified symbology that required additional familiarization time, for

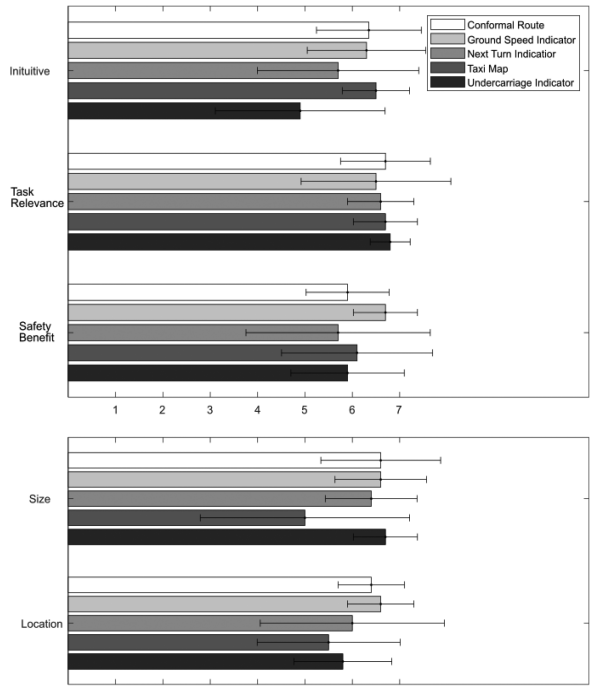


Fig. 5. Mean pilot ratings of HUD symbology functionality (intuitive, task-relevance, safety benefit) and physicality (size, location) for the five different HUD symbology elements. Rating standard deviations represented as error bars.

example, the undercarriage indicator: "*I found with the gear position symbol, which is something we are not used to using, took a little bit of time to figure out how it worked.*" [participant 6] and "*Once I had confidence using it, I can then saw how the lead and lag in the main gear worked in relation to the nose wheel.* [participant 7].

The final design-related secondary theme was *Suggested Changes* that arose from physical property discussions of the individual symbology elements. Comments were minor. For example, as per participant ratings (see Fig. 5), the size of the mini-map was an element participants would have preferred to be larger: "*I'd prefer if it was a bit bigger actually*" [participant 1]. Awareness that a simple size increase is problematic with HUD symbology design (due to HUD size restrictions) prompted some participants to offer novel suggestions in the form of including an adaptive zoom feature for the mini-map: "M*aybe you could do an auto zoom, like on a Garmin GPS*" [participant 8]. Suggested additions included improving spatial awareness during tight turns, when the conformal centerline was less visible and more difficult to track: "*It would be really useful to have almost flight director inputs that give you feedback on how much input you need to put in*" [participant 2].

IV. Discussion

This study presents the findings from a human-in-the-loop evaluation of a UCD HUD taxi symbology set. Alongside traditional usability debriefing procedures, the evaluation incorporated a novel multivariate analysis of gaze behavior and task performance to complement the review of the symbology design. Qualitative pilot feedback revealed that the HUD

interface was perceived as being functionally intuitive and would promote substantial safety and efficiency benefits during taxiing operations. These comments were reported with evidence of HUD-related performance improvements. The multivariate analysis of gaze behavior and task performance connected the above findings through a vision-for-action paradigm [15], [16], aiding the interpretation of how variability in user attention and interface utility translates into task performance. The findings highlight the benefits of implementing eye tracking techniques in the design and evaluation of complex systems/displays.

The architecture of the symbology was praised for its intuitive design, with participants expressing how it mirrored their mental model of the PFD. Safety benefits were mostly directed toward the presence of the conformal symbology, namely how the runway stop bars decreased the likelihood of incursions into active runways during low-visibility conditions. Participants reported the ground speed information, together with acceleration/deceleration trend information enabled speed to be managed more safely during turns. These findings reflected those from past studies that used similar UCD approaches as those in the current study [26], [32].

The subjective evaluation of the HUD taxi guidance was complimented by the objective measurement of taxiing performance. In low-visibility conditions, centerline deviation was reduced when taxiing with the taxi guidance (see Fig. 2). This corroborates the findings of previous studies, where taxi centerline deviations were greater while navigating with paper charts compared to a HUD-based taxi symbology set [8], [19], [33]. These results add to the accumulating evidence that AR symbology can support complex navigational tasks.

In contrast to the majority of aviation eye tracking research [8], [9], [11], the current results offer insight into the interaction between human gaze behavior and interface design at a finer spatial resolution. By employing ROIs that define specific elements within the HUD, it is possible to determine both: 1) the parts of the symbology attended to; and 2) whether allocation of visual attention is dependent upon changing task demands. The findings demonstrated that participants' gaze behavior conformed with this expectation. Visual attention toward the conformal centerline and undercarriage symbology was highest during straight and turn segments, respectively (see Fig. 3). More importantly, the multivariate analysis (see Table IV) confirmed that task-specific changes in attention corresponded to increases in performance. Greater allocation of attention to the undercarriage indicator showed reduced centerline deviations. This approach serves an important function in UCD to determine if design requirements have been achieved [26].

Pilot feedback further aided the holistic interpretation of the symbology set. Participants reported that the increased attention toward the undercarriage indicator was attributed to the increased aircraft position awareness it provided during tight turns, where the restricted cross-cockpit view inhibits judgement of the aircraft's undercarriage placement. Conversely, the multivariate analysis revealed that participants who attended more to the mini-map during turns exhibited larger centerline deviations (see Table IV). It is possible that greater visual attention to the mini-map versus the undercarriage indicator reflects different user strategies. It could be argued that the mini-map bestows greater strategic awareness of the aircraft's future navigational state while the undercarriage largely provides tactical information concerning immediate centerline deviation. Participants choosing to attend more to the mini-map might be willing to sacrifice centerline accuracy in favor of enhancing strategic awareness. This is supported by comments on the interpretation and utility of the mini-map as an aid that supported navigational planning.

Another explanation for increased attention toward the mini-map over the undercarriage indicator could be due to their intuitiveness. The *Design* theme emphasized that issues with the interface could be remedied with minimal additional training time. However, the undercarriage indicator was identified by many as an aspect of the symbology that required greater familiarization, potentially leading to some participants searching for other, more intuitive, symbology (such as the mini-map) during turns. Comments on training requirements echoed the usability feedback given by pilots on the HUD taxi guidance symbology developed by NASA [33]. The more focused analysis presented here is more explicit in the particular training requirements the symbology would demand.

Attention to the conformal centerline during S-bend turn segments was associated with greater centerline deviation (Table II), which could be interpreted as a detrimental effect of the conformal symbology, particularly if participants are not attending to more informative task related information (e.g., the undercarriage indicator). While conformal presentation can facilitate processing the symbology and the environment [23], [24], it can also lead to the filtering of task-related information that exists elsewhere in the visual scene [25]. Similar attentional effects have been reported when pilots fly with scene-linked flight symbology (i.e., HITS) during landing tasks [13]. The likelihood that attentional tunneling occurred is supported by comments made by participants suggesting an over reliance on task information communicated by the symbology. This includes comments that during turns participants often felt like they were "waiting" for the conformal route line to come back into view. This effect may have been exacerbated due to the limited field-of-view (FOV) issue that is inherent to HUDs. This meant important steering-relevant conformal information could not be drawn on the limited HUD FOV while in a turn.

V. Conclusion and Future Research

The findings underline the benefits of adopting a vision-for-action paradigm [15], [16] in the usability evaluation of design solutions developed within a UCD framework. The results demonstrate how detailed examination of gaze behaviors, defined by finer resolution ROI areas, alongside the implementation of multivariate analysis techniques, can provide robust evidence. Eye tracking measurements of pilot gaze behavior toward task-specific symbology were associated with enhanced task performance. The novel procedures demonstrated how the integration of posttrial structured qualitative interview data were implemented to enhance the explanatory power of the gaze behavior results.

This study showcases the utility of linear mixed effects (e.g., GLMM) analysis procedures in the context of human factors

research. This is important as the technique is highly relevant due to its suitability to small sample research (a challenge for costly simulation studies). The human factors field has lagged behind in the adoption of these more sophisticated statistical approaches compared to more fundamental fields of psychology (e.g., psycholinguistics [34], cognitive neuropsychology [35]). GLMM analysis methods have the capability to enhance the statistical robustness of human factors research, though future research is warranted that compares the application of GLMM and traditional ANOVA techniques on small sample size, repeated measures datasets across a range of human factors settings.

Future research endeavors in the domain of complex system and interface design will benefit from adopting similar mixed-method evaluation paradigms. In particular, the results have relevance for the burgeoning area of AR research for how eye tracking can be used in the evaluation of symbology design. Future studies would benefit from exploring the application of the current paradigm to evaluate AR applications intended for other safety critical domains, for instance, evaluating the use of AR to support human–robot interactions, or interactions with systems that possess varying levels of autonomy.

REFERENCES

[1] R. Wang, Y. Xu, and L. Chen, "GazeMotive : A gaze-based motivation-aware E-learning tool for students with learning difficulties," in *Proc. 17th IFIP Conf. Human-Comput. Interaction*, 2019, pp. 544–548.

[2] E. MacHado, I. Carrillo, M. Collado, and L. Chen, "Visual attention-based object detection in cluttered environments," in *Proc. IEEE SmartWorld, Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People Smart City Innov.*, 2019, pp. 133–139.

[3] R. Clark et al., "The potential and value of objective eye tracking in the ophthalmology clinic," *Eye*, vol. 33, pp. 1200–1202, 2019, doi: 10.1038/s41433-019-0417-z.

[4] R. Mao, G. Li, H. P. Hildre, and H. Zhang, "A survey of eye tracking in automobile and aviation studies: Implications for eye-tracking studies in marine operations," *IEEE Trans. Human-Mach. Syst.*, vol. 51, no. 2, pp. 87–98, Apr. 2021, doi: 10.1109/THMS.2021.3053196.

[5] G. Ziv, "Gaze behavior and visual attention: A review of eye tracking studies in aviation," *Int. J. Aviation Psychol.*, vol. 26, no. 3–4, pp. 75–104, 2016, doi: 10.1080/10508414.2017.1313096.

[6] K. Colvin, R. Dodhia, and R. Dismukes, "Is pilots' visual scanning adequate to avoid mid-air collisions?," in *Proc. Int. Symp. Aviation Psychol.*, 2005, pp. 1–6.

[7] C. D. Wickens, J. Goh, J. Helleberg, W. J. Horrey, and D. A. Talleur, "Attentional models of multitask pilot performance using advanced display technology," *Hum. Factors*, vol. 45, no. 3, pp. 360–380, Sep. 2003, doi: 10.1518/hfes.45.3.360.27250.

[8] J. Wilson, J. L. Hooey, and B. L. Foyle, "Head-up display symbology for surface operations: Eye tracking analysis of command-guidance vs. Situation-guidance formats," in *Proc. Int. Symp. Aviation Psychol.*, 2005, pp. 835–841.

[9] L. C. Thomas and C. D. Wickens, "Eye-tracking and individual differences in off-normal event detection when flying with a synthetic vision system display," *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, vol. 48, no. 1, pp. 223–227, 2004, doi: 10.1177/154193120404800148.

[10] N. B. Sarter, R. J. Mumaw, and C. D. Wickens, "Pilots' monitoring strategies and performance on automated flight decks: An empirical study combining behavioral and eye-tracking data," *Hum. Factors*, vol. 49, no. 3, pp. 347–357, Jun. 2007, doi: 10.1518/001872007X196685.

[11] P. K. Hughes and D. J. Creed, "Eye movement behaviour viewing colour-coded and monochrome avionic displays," *Ergonomics*, vol. 37, no. 11, pp. 1871–1884, 1994, doi: 10.1080/00140139408964955.

[12] E. M. Stelzer and C. D. Wickens, "Pilots strategically compensate for display enlargements in surveillance and flight control tasks," *Hum. Factors*, vol. 48, no. 1. pp. 166–181, 2006, doi: 10.1518/001872006776412225.

[13] C. D. Wickens and A. L. Alexander, "Attentional tunneling and task management in synthetic vision displays," *Int. J. Aerosp. Psychol.*, vol. 19, no. 2, pp. 182–199, 2009, doi: 10.1080/10508410902766549.

[14] D. B. Beringer and J. D. Ball, "A comparison of pilot navigation performance using conventional instrumentation, head-down, and head-up highway-in-the-sky primary flight displays," *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, vol. 45, no. 2, pp. 16–20, 2001.

[15] J. N. Vickers, "Advances in coupling perception and action: The quiet eye as a bidirectional link between gaze, attention, and action," *Prog. Brain Res.*, vol. 174, pp. 279–288, 2009, doi: 10.1016/S0079-6123(09)01322-3.

[16] D. Panchuk, S. Vine, and J. N. Vickers, "Eye tracking methods in sport expertise," in *Routledge Handbook of Sport Expertise*. England, U.K.: Routledge, pp. 176–187, Mar. 2015, doi: 10.4324/9781315776675-16.

[17] M. A. Goodale, A. D. Milner, L. S. Jakobson, and D. P. Carey, "A neurological dissociation between perceiving objects and grasping them," *Nature*, vol. 349, no. 6305, pp. 154–156, 1991, doi: 10.1038/349154a0.

[18] S. Wilke, A. Majumdar, and W. Y. Ochieng, "Airport surface operations: A holistic framework for operations modeling and risk management," *Saf. Sci.*, vol. 63, pp. 18–33, Mar. 2014, doi: 10.1016/J.SSCI.2013.10.015.

[19] J. Arthur et al., "Synthetic vision enhanced surface operations with head-worn display for commercial aircraft synthetic vision enhanced surface operations with head-worn display for commercial aircraft," *Int. J. Aviation*, vol. 9, no. 2, pp. 158–181, 2009, doi: 10.1080/10508410902766507.

[20] J. Blundell et al., "Low-visibility commercial ground operations: An objective and subjective evaluation of a multimodal display," *Aeronautical J.*, vol. 9, pp. 1–23, Feb. 2023, doi: 10.1017/AER.2022.81.

[21] A. Dey, M. Billinghurst, R. W. Lindeman, and J. E. Swan, "A systematic review of 10 years of augmented reality usability studies: 2005 to 2014," *Front. Robot. AI*, vol. 5, no. 4, 2018, Art. no. 37, doi: 10.3389/FROBT.2018.00037/BIBTEX.

[22] J. Blundell and D. Harris, "Designing augmented reality for future commercial aviation: A user-requirement analysis with commercial aviation pilots," *Virtual Real*, vol. 23, pp. 1–15, 2023, doi: 10.1007/s10055-023-00798-9.

[23] S. Fadden, P. M. Ververs, and C. D. Wickens, "Pathway HUDs: Are They viable?," *Hum. Factors*, vol. 43, no. 2, pp. 173–193, 2001, doi: 10.1518/001872001775900841.

[24] J. L. Levy, D. C. Foyle, and R. S. McCann, "Performance benefits with scene-linked hud symbology: An attentional phenomenon?," *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, vol. 42, no. 1, pp. 11–15, Oct. 1998, doi: 10.1177/154193129804200104.

[25] S. Fadden, C. D. Wickens, and P. Ververs, "Costs and benefits of head up displays: An attention perspective and a meta analysis," *J. Aerosp.*, vol. 109, pp. 1112–1117, 2000.

[26] D. Harris, *Human Performance on the Flight Deck*. New York, NY, USA, USA: Taylor & Francis Group, 2011.

[27] C. Muth, K. L. Bales, K. Hinde, N. Maninger, S. P. Mendoza, and E. Ferrer, "Alternative models for small samples in psychological research: Applying linear mixed effects models and generalized estimating equations to repeated measures data," *Educ. Psychol. Meas.*, vol. 76, no. 1, Feb. 2016, Art. no. 64, doi: 10.1177/0013164415580432.

[28] L. Meteyard and R. A. I. Davies, "Best practice guidance for linear mixed-effects models in psychological science," *J. Memory Lang.*, vol. 112, Jun. 2020, Art. no. 104092, doi: 10.1016/J.JML.2020.104092.

[29] D. J. Harris et al., "Assessing expertise using eye tracking in a virtual reality flight simulation," *Int. J. Aerosp. Psychol.*, vol. 150, no. 1, pp. 1–29, 2023, doi: 10.1080/24721840.2023.2195428.

[30] B. Winter, "A very basic tutorial for performing linear mixed effects analyses," 2013, *arXiv:1308.5499*.

[31] V. Braun and V. Clarke, "Qualitative research in psychology using thematic analysis in psychology using thematic analysis in psychology," *Qualitative Res. Psychol.*, vol. 3, no. 2, pp. 77–101, 2006.

[32] K. J. Parnell, R. A. Wynne, K. L. Plant, and N. A. Stanton, "User-centered design and evaluation of future flight deck technologies," in *Proc. Ergonom. Hum. Factors*, 2021, pp. 1–8.

[33] R. Bailey, J. Arthur, L. Prinzel, and L. Kramer, "Evaluation of head-worn display concepts for commercial aircraft taxi operations," *Head- Helmet-Mounted Displays XII: Des. Appl.*, vol. 6557, 2007, Art. no. 65570Y, doi: 10.1117/12.717221.

[34] B. Winter and S. Grawunder, "The phonetic profile of Korean formal and informal speech registers," *J. Phonetics*, vol. 40, pp. 808–815, 2012, doi: 10.1016/j.wocn.2010.11.010.

[35] J. Blundell et al., "Markers of cognitive function in individuals with metabolic disease : Morquio syndrome and tyrosinemia type III," *Int. J. Environ. Res. Public Health*, vol. 15, no. 6, May 2018, Art. no. 1113, doi: 10.1080/02643294.2018.1443913.

James Blundell received the Ph.D. degree in cognitive neuroscience from the University of Birmingham, Birmingham. U.K., in 2015.

He is currently an Assistant Professor with Coventry University, where he teaches human information processing and statistics.

Charlotte Collins received the M.Sc. degree in flight simulation of microair vehicles from Coventry University, Coventry, U.K., in 2010.

She is an Assistant Professor with the School of Mechanical Aerospace & Automotive at Coventry University.

Rod Sears received the M.Phil. degree in pilot performance in the Business Jet community from Cranfield University, Cranfield, U.K., in 2014.

He is an ex-military Test Pilot Instructor and Airline Examiner/Instructor.

Tassos Plioutsias received the Ph.D. degree in engineering and ergonomics from the National Technical University of Athens, Athens, Greece, in 2021.

He is a former fighter Pilot. He is an Assistant Professor with Coventry University, teaching safety management systems.

John Huddlestone received the Ph.D. in applied psychology from the College of Aeronautics, Cranfield University, Cranfield, U.K., in 2003.

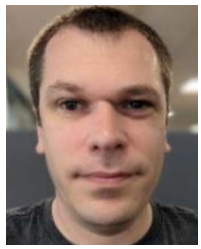
He is an Associate Professor of Human Factors at Coventry University

Don Harris received the Ph.D. degree in applied psychology from the College of Aeronautics, Cranfield University, Cranfield, U.K., in 1988.

He is a Professor of human factors at Coventry University, Coventry, U.K.

James Harrison received the B.Sc. degree in computer science with mathematics from the University of Leeds, Leeds, U.K., in 1998.

He works as a Principal Software Engineer for BAE Systems, where he specialises in human–machine interface systems.



Anthony Kershaw received the B.Sc. (Honors) degree in software engineering from the University of Kent, Kent, U.K., in 2001.

He works as a Principal Software Engineer for BAE Systems primarily in the aerospace industry.

Paul Harrison received the B.Sc. degree in computer science from Reading University, Reading, U.K., in 1994.

He qualified as a Pilot in 2002. He has 28 years of aerospace experience within BAE Systems working on flight simulation research and development programs.

Phil Lamb, photograph and biography not available at the time of publication.