Human Factors in Design, Engineering, and Computing, Vol. 238, 2026, 2636–2645
https://doi.org/10.54941/ahfe1008488

Eye-Tracking Factors During Spatial Ability Task Predict IRT-Based Item Difficulty and Naval Flight Training Performance

Alejandro Albizu[1], Laura K. Jamison[1], Kaylin D. Strong[1], Kayla M. Russell[2], Ciara M. Sibley[1], Shannon P. McGarry[1], Noelle L. Brown[1], Christopher D. Draheim[1], Cyrus K. Foroughi[1], Rebecca L. NeSmith[3], Sarah R. Melick[4], and Joseph T. Coyne[1]

[1]U.S. Naval Research Laboratory, Washington, DC 20375, USA
[2]Strategic Analysis Inc., Arlington, VA 22203, USA
[3]Naval Aerospace Medical Institute, Pensacola, FL 32508, USA
[4]Training Air Wing Six, Pensacola, FL 32508, USA

ABSTRACT

Since WWI, spatial reasoning abilities have demonstrated strong predictive validity for aviation training success. Eye-tracking offers millisecond-level indexing of underlying cognitive mechanisms: visual attention, visual search, and cognitive load. 228 U.S. Naval flight students performed the Terrain Orientation Task (TOT) while eye-tracking was recorded. Exploratory factor analysis reduced eight eye-tracking metrics into three factors: Visual Search, Visual Attention, and Cognitive Load. Objective difficulty was calibrated via a two-parameter logistic Item Response Theory model. Harder TOT items induced significantly larger peak pupil dilations ($b = 0.06$, $p < .001$) and greater pupil trace area under the curve ($b = 0.02$, $p < .001$), confirming pupillometry as an online cognitive load gauge. With escalating difficulty, cognitive load rose ($b = 0.04$, $p = .027$) while visual search narrowed ($b = –0.29$, $p < .001$), reflecting load-induced scan tightening. At the trial-level, broader visual search conferred a ~30% increase in the odds of a correct response ($p < .001$). At the participant level, the interaction between visual search and visual attention predicted Naval Introductory Flight Evaluation (NIFE) academic performance ($\beta = –0.189$, $p = .013$) whereas flight performance was predicted by interaction between attention and cognitive load ($\beta = –0.26$, $p = .006$). Adding eye-tracking factors to traditional selection metrics improved prediction of NIFE academics ($\Delta R^2 = +7\%$) and flight performance ($\Delta R^2 = +10\%$), while markedly enhancing attrition prediction ($\Delta AUC = +15\%$). These findings demonstrate that eye-tracking provides valuable cognitive workload and strategic signatures to enhance aviator selection, difficulty titration, and targeted coaching.

Keywords: Aviation, Eye-tracking, Training, Spatial ability

INTRODUCTION

Spatial ability is the capacity to encode, transform, and compare mental representations of objects in three-dimensional space. Spatial ability has long been recognized as a cornerstone of aviation performance.

Received October 30, 2026; Revised November 14, 2026; Accepted November 18, 2026; Available online December 1, 2026

© 2026 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

Military aviation selection research has consistently ranked spatial ability assessments among the strongest predictors of flight training success (Coyne et al., 2024). The Naval Aviation Test Selection Battery (ASTB) and Air Force Test Battery for Aviation Selection (TBAS) utilize several spatial ability assessments like the Direction Orientation Task (DOT; (Coyne et al., 2024, 2020) and Terrain Orientation Task (TOT; (Coyne et al., 2024). While these spatial ability assessment tools minimize the attrition risk, the attrition in Naval flight training remains around 10–15% (Prasad-Rao et al., 2023). Most attrition occurring during the Primary flight phase, where students first learn to fly military aircraft (Prasad-Rao et al., 2023). Nearly 24% of attrition in the primary phase is due to poor flight performance (Prasad-Rao et al., 2023). Each Naval flight student (NFS) attrition costs the Department of the Navy (DON) up to $2M in the advanced phase (OUSD Comptroller, 2023). Improving detection of students at risk of attrition and identifying coachable metrics remains an operational imperative.

Eye-tracking offers an understanding of how NFS allocate attention and cognitive effort with millisecond-level resolution. Task-evoked pupil dilation scales reliably with intrinsic cognitive load making it a potential biomarker of workload (Babu et al., 2019; Beatty, 1982; Bellenkes et al., 1997; Harris et al., 2023; Hess & Polt, 1964; Klingner et al., 2008; Lounis et al., 2021; Matton et al., 2022; Palinko et al., 2010; Robison et al., 2022). In both laboratory settings (Beatty, 1982; Hess and Polt, 1964; Klingner et al., 2008; Robison et al., 2022) and complex operational settings (Babu et al., 2019; Harris et al., 2023; Matton et al., 2022), pupil dilation has been used as a reliable gauge of cognitive workload. Further, gaze fixation behavior complements pupil dilation by revealing how attention is directed. In mental rotation tasks, high-performers showed longer gaze dwell times on target regions compared to low-performing peers (Bellenkes et al., 1997). Finally, visual scan patterns reveal the strategy with which visual information is sampled. A structured and systematic scan path is typically related to higher task performance (Bellenkes et al., 1997; Lounis et al., 2021). Crucially, all three of these metrics can be captured simultaneously and unobtrusively on commercial-off-the-shelf (COTS) eye-tracking devices. These COTS devices can be seamlessly integrated into virtual reality or flight simulators. COTS eye-tracking devices are currently being assess for integration into the cockpit.

Despite the promise of eye-tracking metrics, few studies have asked whether eye-tracking metrics can predict real flight training success. Recent work has demonstrated the predictive validity of the TOT (Coyne et al., 2024). However, the physiological dynamics underlying performance during the task remain opaque. Aviation workload experiments have demonstrated that pupil diameter forecasts moment-to-moment deviations from desired flight parameters (Babu et al., 2019), yet eye-tracking metrics have not been linked to longitudinal training outcomes. Broadening the understanding of

the cognitive processes underlying spatial ability can provide early objective indicators of trainees' capacity to allocate cognitive resources, something that current selection tools may miss.

The current investigation aimed to determine if eye-tracking metrics collected before flight training that can predict subsequent performance in Naval flight school. Three factors of eye-tracking were utilized for prediction: cognitive load, attention, and search strategy. Specifically, this study sought to: 1) confirm that eye-tracking metrics scale with task difficulty, establishing them as an online marker of cognitive load in NFS, 2) determine whether eye-tracking metrics predict online task performance (i.e., trial-level accuracy) beyond item difficulty, and 3) test the predictive utility of the eye-tracking metrics in forecasting NFS performance throughout flight school.

METHODS

Participants

Two hundred twenty-eight U.S. Naval flight students (208 male, 20 female; mean age = 25.1 ± 3.4 years) awaiting enrollment in Naval Introductory Flight Evaluation (NIFE) participated. All held Class-I military flight physicals, normal or corrected-to-normal vision, and no history of neurological disorders. The protocol was approved by the Naval Research Laboratory Institutional Review Board in accordance with the Declaration of Helsinki, with informed written consent obtained.

Materials and Equipment

Stimuli were presented on a 24-inch LCD monitor (1920 × 1440, 120 Hz). Binocular gaze and pupil diameter were sampled at 150 Hz using a Gazepoint 150HD eye-tracker.

Terrain Orientation Task (TOT)

The TOT presents 2D satellite reference scenes oriented North alongside a zoomed and rotated target scene (Figure 1A). Participants determine the cardinal camera heading of the target (e.g., Southeast) within a 30-second response window. Eighteen items were administered in randomized order.

Item Response Theory (IRT)

Item difficulty parameters (b) were estimated via a two-parameter logistic (2PL) IRT model in mirt (v1.38). The 2PL model demonstrated superior fit compared to a 3PL specification across nine test forms (nine form-unique and nine common linking items). Differential item and test functioning analyses confirmed invariance across sex and race. Forms were linked using the Mean/Mean method for unidimensional models to yield standardized difficulty parameters for all trials (Figure 1B & 1C).

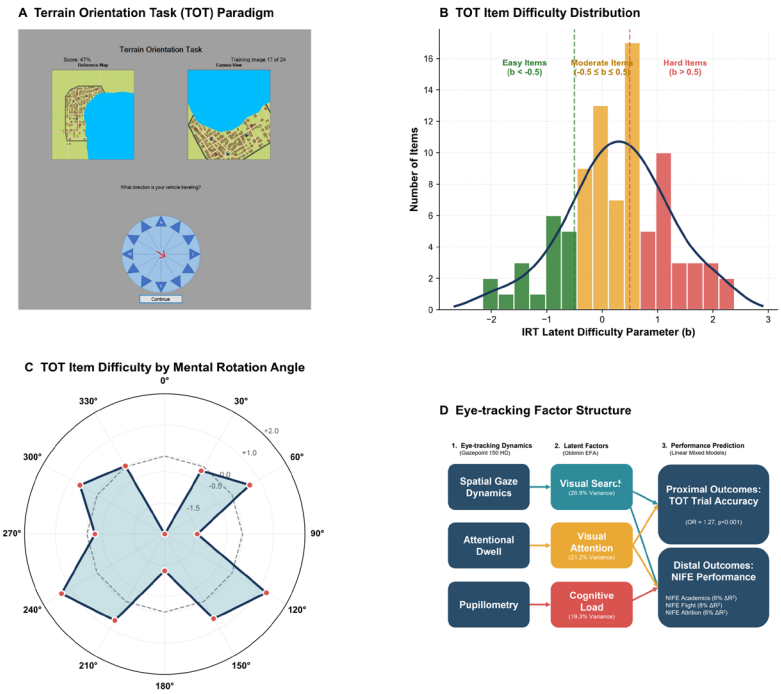


Figure 1: (A) A representative TOT trial. (B) Distribution of IRT difficulty parameters. (C) Polar distribution of item difficulty by angular offset. (D) Eye-tracking factor structure linked to trial accuracy and longitudinal Naval flight training outcomes.

Pre-Processing

Blinks were identified using first-derivative pupil velocity thresholds, padded by ±100 ms, and linearly interpolated for intervals ≤ 250 ms (Kret and Sjak-Shie, 2019). Pupil time-series were smoothed with a 5-sample running median and baseline-corrected to the 0–200 ms pre-stimulus interval. Two areas of interest (AOIs) were defined corresponding to the reference and target scenes. For each trial, three pupillometric indices (peak dilation, mean dilation, area under the curve [AUC]) and five gaze indices (total dwell time, fixation count, AOI transfer saccades, fixation entropy, and Earth Mover's Distance gaze heatmap similarity) were extracted. Metrics were normalized by response time and within-subject z-scored prior to factor analysis.

Statistical Analyses

Linear mixed-effects models (LMMs; lme4 in R 4.3.2) evaluated whether pupillary and gaze metrics scaled with IRT item difficulty. The eight normalized metrics were entered into an exploratory factor analysis (EFA) with oblimin rotation. Generalized linear mixed-effects models (GLMMs) predicted trial accuracy from factor scores. Participant-level factor averages were validated against external constructs using general linear models (GLMs). Downstream NIFE academic grades, flight performance (Navy Standard Score, NSS), and attrition were modeled using GLMs and binomial logistic regression. Finally, hierarchical regression evaluated incremental validity beyond ASTB Academic Qualification Rating (AQR) and TOT Rate of Correct Score (RCS).

RESULTS

Factor Structure of Eye-Tracking Metrics

Oblimin-rotated EFA of the eight normalized eye-tracking metrics revealed three interpretable latent factors explaining 67.4% of total variance (RMSEA = 0.10, TLI = 0.91, BIC = 227; Figure 2A). Factor 1 (Visual Search) loaded heavily on fixation entropy ($\lambda = 0.90$), AOI heatmap similarity ($\lambda = 0.89$), and pupil AUC ($\lambda = 0.66$), capturing systematic, distributed scanning. Factor 2 (Visual Attention) loaded on fixation count ($\lambda = 0.96$), dwell time ($\lambda = 0.72$), and AOI transfers ($\lambda = 0.49$), reflecting focal, high-density inspection. Factor 3 (Cognitive Load) aggregated pupil tonic change ($\lambda = 1.00$) and peak pupil dilation ($\lambda = 0.72$). Inter-factor correlations were small ($|r| \leq 0.25$; Figure 2B).

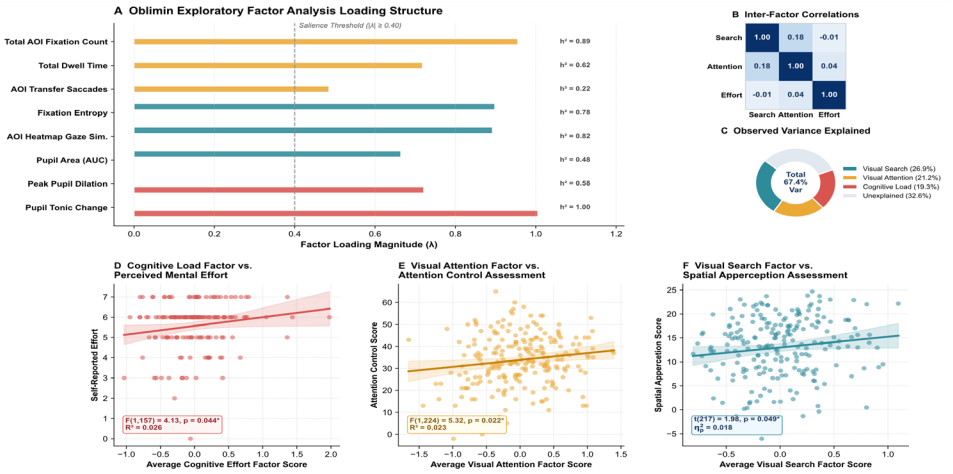


Figure 2: (A) Oblimin-rotated EFA loadings across eight eye-tracking metrics, identifying visual search (blue), visual attention (yellow), and cognitive load (red). (B) Inter-factor correlation matrix. (C) Percentage of variance explained per factor. (D) cognitive load factor scores correlate with perceived effort. (E) Visual attention scores correlate with attention control. (F) Visual search scores correlate with spatial apperception ability.

Trial-Level Pupil Response to IRT-Derived Difficulty

LMMs confirmed that pupillometric metrics scaled systematically with IRT difficulty. Harder items elicited significantly greater peak pupil dilation ($b = 0.06$, $p < .001$; Figure 3A) and larger pupil AUC ($b = 0.02$, $p < .001$; Figure 3C). Significant time-on-task reductions emerged for peak dilation ($b = -0.01$, $p = .002$; Figure 3B) and pupil AUC ($b = -0.001$, $p = .017$; Figure 3D), demonstrating cognitive workload habituation across the 18 trials. In contrast, baseline tonic pupil change was unrelated to item difficulty ($p = .941$) but declined over time ($b = -0.01$, $p < .001$). When composite factors were modeled, the Cognitive Load factor was significantly associated with IRT difficulty ($b = 0.04$, $p = .027$), particularly when visual search was constrained ($b = -0.06$, $p < .001$), consistent with load-induced scan tightening and perceptual tunneling.

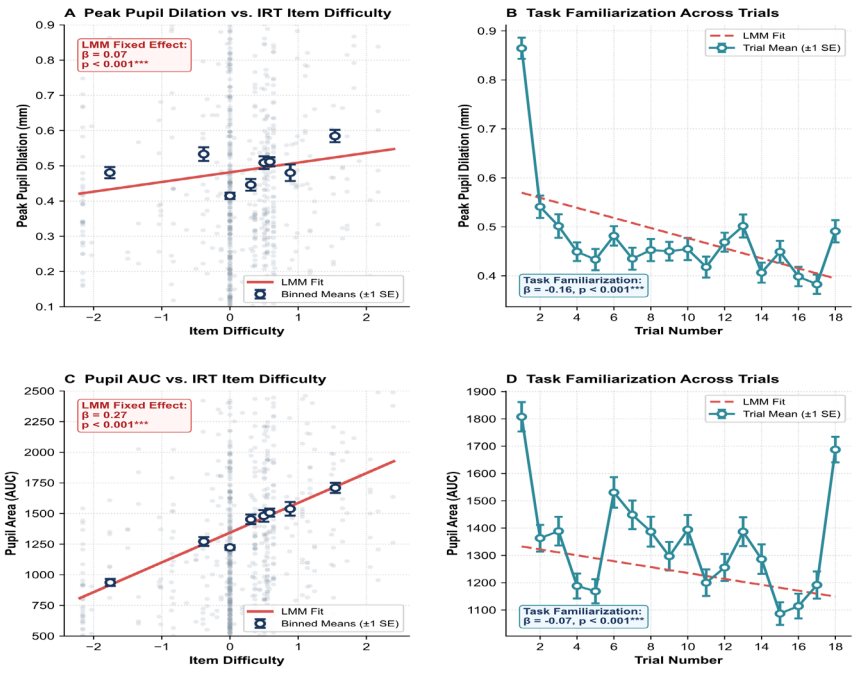


Figure 3: (A) LMM fit demonstrating peak pupil dilation scaling with IRT item difficulty. (B) Peak pupil dilation diminishing across the 18-trial sequence (b = –0.007 mm/trial). (C) LMM fit showing pupil AUC scaling with IRT difficulty. (D) Pupil AUC also significantly diminishes across trials.

Trial-Level Performance Prediction

In GLMM logistic regression, the three latent factors significantly predicted trial accuracy (Figure 4A). Higher Visual Search factor scores substantially increased the odds of a correct response (OR = 1.27, $p < .001$). A significant Search × Attention interaction (OR = 0.78, $p < .001$) indicated that broad visual search was most beneficial when visual attention demands were low, whereas elevated attention without distributed search impaired performance (OR = 0.65, $p < .001$; Figure 4B). Cognitive Load marginally reduced trial correctness odds ($p = .138$).

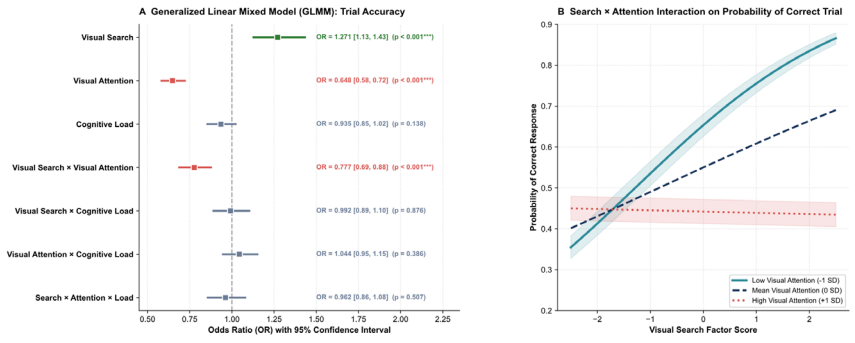


Figure 4: (A) Forest plot of odds ratios from GLMM predicting trial correctness. (B) Model-predicted trial accuracy across visual search stratified by visual attention recruitment (±1 SD).

Construct Validation

Validation analyses supported the construct validity of the latent factors. Mean Cognitive Load factor scores positively correlated with self-reported effort ($\beta = 0.16$, $p = .044$; Figure 2D). Visual Attention scores positively predicted assessed attention control ($\beta = 0.15$, $p = .022$; Figure 2E). Visual Search positively predicted with spatial apperception ($\beta = 0.13$, $p = .049$; Figure 2F).

Participant-Level Training Outcomes

Eye-tracking factors significantly predicted downstream training milestones. For NIFE academic performance, the eye-tracking model was significant ($F(7, 183) = 2.83$, $p = .008$, $R^2 = .098$), exhibiting a significant Visual Search main effect ($\beta = –0.20$, $p = .017$) and Search × Attention interaction ($\beta = –0.19$, $p = .013$; Figure 5A), demonstrating that high focal attention coupled with poor search predicts classroom learning.

For NIFE flight performance, the overall eye-tracking model was also significant ($F(7, 153) = 3.14$, $p = .004$, $R^2 = .13$). A marginal Cognitive Load effect ($\beta = –0.15$, $p = .085$) was moderated by Visual Attention ($\beta = –0.26$, $p = .006$) and a significant three-way Search × Attention × Effort interaction ($\beta = 0.12$, $p = .007$; Figure 5B). Elevated cognitive load degraded flight performance primarily when attention was high and search was restricted; broad search effectively buffered flight scores against high workload. Mapping students by these factors identified distinct operational profiles (Figure 5C): high-potential candidates who combine low physiological effort with broad search (Quadrant II), and vulnerable candidates exhibiting high load and restricted search (Quadrant IV).

Incremental Validity Analyses

In hierarchical regression, baseline ASTB AQR predicted NIFE academic performance ($R^2 = .189$). Adding TOT RCS yielded a minor gain ($R^2 = .194$, $\Delta R^2 = .005$). Incorporating eye-tracking factors contributed significant incremental variance ($R^2 = .262$, $\Delta R^2 = .060$), with Visual Search remaining a significant unique predictor ($\beta = –0.217$, $p = .018$). For NIFE flight performance, AQR accounted for 6.2% of variance ($p < .001$). TOT RCS added 1.9% ($R^2 = .081$, $\Delta R^2 = .019$). Crucially, adding eye-tracking factors more than doubled explained variance ($R^2 = .185$, $\Delta R^2 = .104$), driven by Attention × Effort ($\beta = –0.296$, $p = .005$) and Attention × Search × Effort ($\beta = 0.119$, $p = .013$) interactions. Finally, in binomial logistic regression forecasting NIFE training attrition, baseline AQR was significant (McFadden $R^2 = .074$, $p = .002$). Adding TOT RCS improved model fit (McFadden $R^2 = .162$, $\Delta R^2 = .088$, $p = .007$). Eye-tracking factors further enhanced classification (McFadden $R^2 = .228$, $\Delta R^2 = .066$, $p = .017$; Figure 5D), boosting the receiver operating characteristic area under the curve to AUC = 0.834 (vs. AQR AUC = 0.682; Figure 5E).

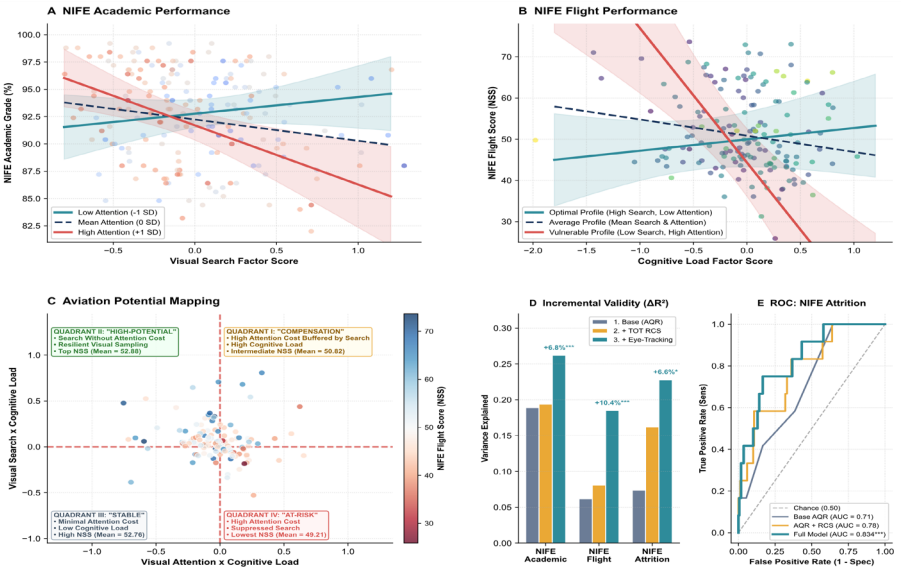


Figure 5: (A) NIFE academic score prediction by visual search and visual attention interaction. (B) NIFE NSS modeled across cognitive load for optimal (+1 SD search, −1 SD attention) vs. vulnerable (−1 SD search, +1 SD attention) profiles. (C) Aviation potential mapping: quadrant plot of attention × load vs. search × load colored by NIFE flight score, identifying high-potential and at-risk candidates. (D) Incremental validity (ΔR^2) across models for NIFE academic, flight, and attrition. (E) ROC curves for NIFE attrition.

CONCLUSION

This study establishes that eye-tracking metrics collected during spatial orientation testing provide objective biomarkers of cognitive load and powerful predictors of longitudinal flight training success. By decomposing high-dimensional oculomotor behavior into three latent dimensions—Visual Search, Visual Attention, and Cognitive Load—we demonstrate that real-time pupil and gaze dynamics scale with IRT task difficulty and forecast performance across both immediate spatial tasks and subsequent flight training phases.

Our findings confirm that task-evoked pupil dilation serves as a sensitive, continuous index of mental workload in Naval flight students, corroborating cognitive resource theory (Beatty, 1982; Robison et al., 2022). As difficulty increased, participants exhibited load-induced scan tightening and perceptual tunneling (Dirkin, 1983; Jarmasz et al., 2005), characterized by elevated focal attention and restricted spatial exploration. Importantly, trainees who maintained a broad, systematic search pattern achieved higher trial accuracy.

These results offer pathways for optimizing military aviation training. First, eye-tracking metrics can enable adaptive difficulty titration within flight simulators, dynamically tailoring scenario demands to keep students in an optimal learning zone without exceeding cognitive bandwidth. Second, objective oculomotor signatures provide instructors with actionable

diagnostic feedback. Finally, integrating eye-tracking into COTS virtual reality platforms provides a scalable, cost-effective screening framework to detect attrition risk prior to resource-intensive live-flight phases, directly addressing the substantial financial and operational costs of student flight failure.

ACKNOWLEDGMENT

This research was funded by Dr. Jeffery Morrison's Command Decision Making Program, LCDR Michael Natali's Manpower, Personnel, Training, and Education Program, and Natalie Steinhauser's Attention Control and Naval Strike Group Training and Assessment Program at the Office of Naval Research.

REFERENCES

Babu, M.D., JeevithaShree, D.V., Prabhakar, G., Saluja, K.P.S., Pashilkar, A., Biswas, P., 2019. Estimating pilots' cognitive load from ocular parameters through simulation and in-flight studies. J. Eye Mov. Res. 12. https://doi.org/10.16910/jemr.12.3.3

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting Linear Mixed-Effects Models Using lme4. J. Stat. Softw. 67. https://doi.org/10.18637/jss.v067.i01

Beatty, J., 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. Psychol. Bull. 91, 276–292. https://doi.org/10.1037/0033-2909.91.2.276

Bellenkes, A.H., Wickens, C.D., Kramer, A.F., 1997. Visual Scanning and Pilot Expertise: The Role of Attentional Flexibility and Mental Model Development. Aviat. Space Environ. Med. 68, 569–579.

Coyne, J., Draheim, C., Sibley, C., Foroughi, C., Melick, S., Armendariz, N., Burgoyne, A., Engle, R., 2024. Evaluation of new measures of spatial ability and attention control for selection of naval flight students, in: AHFE International. https://doi.org/10.54941/ahfe1005768

Coyne, J., Herdener, N., Sweet, J., Sibley, C., Brown, N., Foroughi, C., King, K., 2020. Toward Improving the Direction Orientation Task, in: Advances in Intelligent Systems and Computing. Springer, pp. 533–539. https://doi.org/10.1007/978-3-030-50791-6_68

Dirkin, G.R., 1983. Cognitive Tunneling: Use of Visual Information under Stress. Percept. Mot. Skills 56, 191–198. https://doi.org/10.2466/pms.1983.56.1.191

Embretson, S.E., Reise, S.P., 2013. Item Response Theory. Psychology Press. https://doi.org/10.4324/9781410605269

Harris, D.J., Arthur, T., De Burgh, T., Duxbury, M., Lockett-Kirk, R., McBarnett, W., Vine, S.J., 2023. Assessing Expertise Using Eye Tracking in a Virtual Reality Flight Simulation. Int. J. Aerosp. Psychol. 33, 153–173. https://doi.org/10.1080/24721840.2023.2195428

Hess, E.H., Polt, J.M., 1964. Pupil Size in Relation to Mental Activity during Simple Problem-Solving. Science 143, 1190–1192. https://doi.org/10.1126/science.143.3611.1190

Jarmasz, J., Herdman, C.M., Johannsdottir, K.R., 2005. Object-Based Attention and Cognitive Tunneling. J. Exp. Psychol. Appl. 11, 3–12. https://doi.org/10.1037/1076-898X.11.1.3

Klingner, J., Kumar, R., Hanrahan, P., 2008. Measuring the task-evoked pupillary response with a remote eye tracker, in: Proceedings of ETRA '08. ACM, p. 69. https://doi.org/10.1145/1344471.1344489

Kret, M.E., Sjak-Shie, E.E., 2019. Preprocessing pupil size data: Guidelines and code. Behav. Res. Methods 51, 1336–1342. https://doi.org/10.3758/s13428-018-1075-y

Lounis, C., Peysakhovich, V., Causse, M., 2021. Visual scanning strategies in the cockpit are modulated by pilots' expertise: A flight simulator study. PLOS ONE 16, e0247061. https://doi.org/10.1371/journal.pone.0247061

Matton, N., Paubel, P.-V., Puma, S., 2022. Toward the Use of Pupillary Responses for Pilot Selection. Hum. Factors 64, 555–567. https://doi.org/10.1177/0018720820945163

OUSD Comptroller, 2023. FY2024 DoD Fixed Wing and Helicopter Reimbursement Rates effective October 1, 2023.

Palinko, O., Kun, A.L., Shyrokov, A., Heeman, P., 2010. Estimating cognitive load using remote eye tracking in a driving simulator, in: Proceedings of ETRA '10. ACM, p. 141. https://doi.org/10.1145/1743666.1743701

Prasad-Rao, J., Pinon Fischer, O.J., Rowe, N.C., Williams, J.R., Puranik, T.G., Mavris, D.N., Natali, M.W., Tindall, M.J., Atkinson, B.W., 2023. Attrition Risk and Aircraft Suitability Prediction in U.S. Navy Pilot Training Using Machine Learning. Aerospace 10, 379. https://doi.org/10.3390/aerospace10040379

Robison, M.K., Coyne, J.T., Sibley, C., Brown, N.L., Neilson, B., Foroughi, C., 2022. An examination of relations between baseline pupil measures and cognitive abilities. Psychophysiology 59, e14124. https://doi.org/10.1111/psyp.14124