PSYCHOPHYSIOLOGICAL RESEARCH METHODS TO ASSESS AIRLINE FLIGHT CREW RESILIENT PERFORMANCE IN HIGH-FIDELITY FLIGHT SIMULATION SCENARIOS

Chad L. Stephens, Tyler D. Fettrow, Lawrence J. Prinzel III, Jon B. Holbrook,
& Kathryn M. Ballard
NASA Langley Research Center
Hampton, VA

Daniel J. Kiggins
San Jose State Research Foundation
San Jose, California

New concepts in aviation system safety thinking have emerged to consider not only what may go wrong, but also what can be learned when things go right. This approach forms a more comprehensive approach to system safety thinking. A need exists for methods to enable a better understanding of human contributions to aviation safety and how they may inform Safety Management Systems (SMS). A high-fidelity 737-800 simulation study was conducted to study how current type-rated commercial airline flight crews anticipate, monitor, respond to, and learn from expected and unexpected disturbances during line operations. A number of dependent measures were collected that included traditional SMS data types, but also non-traditional safety data to include multiple psychophysiological metrics. This paper describes the psychophysiological measures results that evinced the capability of measures to help identify resilient flight crews. Implications for future research and design of future In-time Aviation Safety Management Systems are discussed.

The NASA System-Wide Safety (SWS) Project is focused on developing new technologies and operational concepts for the aviation industry to meet the increasing global demand while maintaining the current ultra-safe level of system safety. To achieve this, the project is studying safety producing behaviors (e.g., Hollnagel, 2016) and developing research priorities, including In-time System-wide Safety Assurance (ISSA) and In-time Aviation Safety Management System (IASMS; Ellis et al., 2019). Challenges currently being addressed include identifying data sources, analyzing data to detect and prioritize risks, and optimizing safety awareness and decision support. The project is focused on developing domain-specific safety monitoring and alerting tools, integrated predictive technologies, and adaptive in-time safety threat management to expand the knowledge base of resilience engineering and inform ISSA and IASMS for traditional and emerging operational concepts. One test case for this effort concerns non-adherence of area navigation standard terminal arrival route (RNAV STAR) procedures used at major airports.

Stewart, Matthews, Janakiraman, and Avrekh (2018) conducted a study on aircraft flight track data for over 10 million flights into 32 domestic airports and revealed that only 12.4% of flights fully complied with the published arrivals' vertical and lateral profiles. Based on that study, Holbrook et al. (2020) collected data from pilots, air traffic controllers, and airlines to examine safety behaviors during RNAV STAR arrivals at Charlotte Douglas International

Airport (KCLT). The takeaway was that the majority of non-adherences were to sustain operations under dynamic real-world conditions. These findings suggest that traditional approaches to risk and safety management may not be sufficient to address the misalignment between published procedures and routine safe operations, and a complementary approach that includes ensuring that "things go right" is necessary. The study by Holbrook et al. highlights that to maintain safety, humans will likely need to continuously adjust their work to match their operating conditions (Hollnagel, 2014).

Historically, resilience engineering research has centered on the theoretical aspects of productive safety. To address the gap in guidance on measuring resilient performance, we designed and conducted a human-in-the-loop (HITL) flight simulation study to gather empirical data to be used to understand productive safety (Stephens et al. 2021). Neuroergonomics research examining human operators in the context of safety-critical behavior has incorporated traditional human factors methods, including psychophysiological methods, to study human error (Dehais et al., 2020). We are extending this research by developing psychophysiological measures of resilient performance of pilots in simulated flight scenarios. Additionally, exploration of the data generated will determine how to analyze this data to prioritize risks and optimize decision-making support for safety awareness.

The main research objective for this study was to create a data testbed our team and the research community could explore to determine how commercial airline pilots manage routine contingencies and safety during RNAV arrivals. Studying actual operational events in airline operations is challenging because there is a limited amount of data that can be collected and analyzed for productive safety research due to pragmatic, logistical, procedural, or regulatory constraints. This research study involved gathering a comprehensive dataset of candidate measures to facilitate future data science efforts and to gain a better understanding of the phenomena of productive safety. To this end, traditional human factors data collection methods were employed including operator-generated data (e.g., self-report measures of workload, situation awareness, and resilient performance), observer-generated data (e.g., psychophysiological measures: electroencephalography, electrocardiography, galvanic skin response, and eye tracking) and system-generated data (e.g., simulated flight track data) were captured during the flight simulation. However, for the current analysis, we are focused specifically on the eye tracking data.

Methods

Data presented herein were collected during the SWS Operations and Technologies for Enabling Resilient In-Time Assurance (SOTERIA) flight simulation study conducted at NASA Langley Research Center in Hampton, VA USA during May-June 2022. Details of the full data collection plan and flight simulation scenarios are described in Stephens et al. (2021). Twenty-four (24) healthy airline transport pilots (9 women, M = 49.2 years) from a major US airline volunteered for the study. Subjects provided informed verbal and written consent to participate. The experiment was conducted under approval from NASA's Institutional Review Board.

After explaining the experiment and obtaining consent from each crew, each pilot was outfitted with a combined electroencephalography (EEG) and electrocardiography device (ABM X10, CA, USA), and a smart watch that measures galvanic skin response, skin temperature, and heart rate (Empatica, MA, USA). The impedance of each EEG electrode was verified to be less than 10 megaohms. Following the checkout of the outfitted systems, each pilot proceeded to the

simulator flight deck and performed an eye tracking (Smarteye, MA, USA) calibration procedure.

All psychophysiological devices were time synced and triggered for recording through eyesDX Multi-modal Analysis of Psychophysiological and Performance Signals (MAPPs; IA, USA). The data were exported from MAPPs for processing with custom python (Python3) scripts. At this time, eye tracking data analysis is ongoing; therefore only data processing details are discussed. Several metrics of interest were derived from the eye tracking data. These metrics were derived from different raw data generated by the eye tracking system, and had different methods of filtering, calculation, etc. For each variable, we averaged over time epochs of 10 seconds. We use the following definitions for each eye tracking metric:

- Head Heading Velocity: The rate (degrees/second) of the head turning left or right. We only retained indices where the reported % quality was greater than 60%.
- Pupil Diameter: The diameter of the pupils (mm). Because this variable is the most difficult to acquire, in order to keep sufficient indices, we retained indices where the reported percent quality was greater than 40%.
- Gaze Velocity: The velocity of the gaze vector (degrees/second). We retained indices where % quality was greater than 60%, and the gaze velocity of a particular frame did not exceed 700 degrees/second (Wilson et al. 1992).
- Gaze Variance: The variance (spread) score of the gaze vector. We converted the unit vector to a plane using standard stereographic mapping (Marcus, 1966). We retained indices where the % quality was greater than 60%, and the velocity of the raw gaze vector of respective indices did not exceed 700 degrees/second.

In addition to the psychophysiological sensors, we administered an array of traditional human factors measures including self-reported workload and situation awareness. We also created a custom resilience questionnaire, “Resilient Performance Self-Assessment” (RPSA). The RPSA consists of 16 questions that were modeled on American Airlines Learning Improvement Team (LIT) Proficiencies (American Airlines, 2020). The participants were required to specify whether they made use of a particular behavior, and if so, rate their perceived success of implementing that behavior. The choices consisted of a discrete scale from 1 (very unsuccessful) to 5 (very successful). Here, we are only focused on the RPSA scores, and not the other questionnaire data.

We investigated whether pilots exhibit behaviors that can be captured via eye tracking sensors (Smarteye system) that have a relation to their perceived resilience scores. We ran statistics for two questions. 1) Do resilience scores differ by crew? 2) Do the same crews that exhibit different resilience scores, exhibit differing psychophysiological behaviors, specifically in eye tracking measures?

To test our hypotheses, we used *lme4* (Bates, 2015) within R (version 4.1.2; R2021) to perform linear mixed effects analyses. We fit multiple linear mixed models and ran a single model for each variable of interest, including RPSA, Head Heading Rate, Pupil Diameter, Gaze Velocity, and Gaze Variance. For RPSA, we treated each of the 16 questions as repeated measures, assuming equal weighting, used fixed effects of Crew and Seat (left vs right), and subject as a random effect. The psychophysiological data consisted of varying total repeated measures per crew and scenario since we used the average across the 10 second epochs for each dependent variable. The models contained the same factors as the model for RPSA. We

performed post hoc pairwise analyses for each model by calculating the least squares means and estimating the 95% confidence intervals, using a Kenward-Roger approximation implemented in the R-package *emmeans* (Lenth, 2016).

Results

All participants volunteered for all aspects of the experimental protocol. In general, all participants completed every scenario successfully, without any mishaps. Figure 1 shows the results by crew for the reported resilience scores (combined across questions). Crews 6, 8, and 10 showed the lowest RPSA scores, and were significantly different from 1, 2, 11, and 13 (95% confidence intervals did not overlap). Our primary goal here, is to identify psychophysiological measures that exhibit similar crew differences, and therefore indicate resilient or non-resilient behavior.

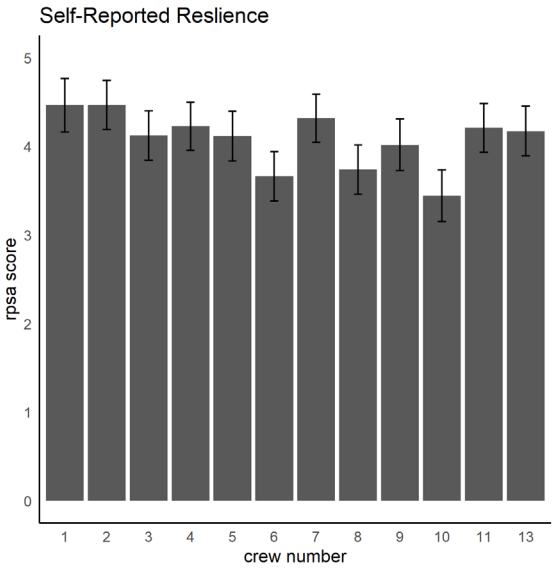

Figure 1: Mixed model results for RPSA scores. The bar graphs depict the estimated marginal mean (bar), and 95% confidence interval (error bar) for each crew's self-reported resilience scores. Lack of overlap between any crews' confidence interval indicates statistical significance between those crews.

Here we are interested in identifying whether the same crews that had statistically significant RPSA scores, also showed differences in metrics we derived from the eye tracking data. Figure 2 shows the statistical results of the metrics derived from the eye tracking data. Crew 11 had a statistically significant difference in Gaze Variance. Crew 11 showed significantly higher variance scores compared to all other crews, which suggests that this crew was looking at more of the cockpit than the rest of the crews throughout the scenarios. The significant findings for Crew 11's Variance score did not transfer to any other metric. Crew 8 exhibited the lowest Gaze Velocity out of all crews. Low gaze velocity indicates less shifting of attention over time. In addition, Crew 8 exhibited the largest Pupil Diameter out of all crews. Crew 8 was one of the crews that showed relatively lower resilient scores, therefore Gaze Velocity and Pupil Diameter appear to be likely candidates for predicting resilient behavior (or lack thereof).

Discussion

In the current preliminary analysis of a subset of the psychophysiological data captured during the study, we were interested in identifying metrics that can predict resilient (safe) behavior. In general, we showed significant differences between some crews in self-reported resilience scores and the psychophysiological measures.

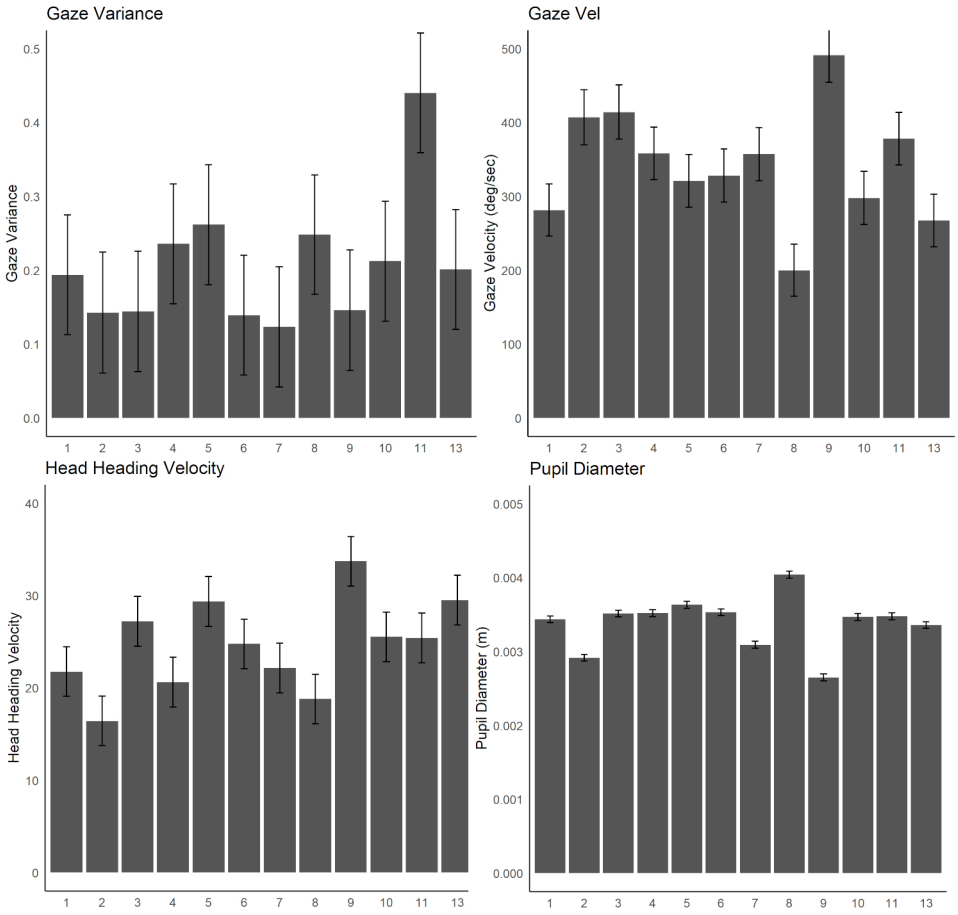

Figure 2: Mixed model results for eye tracking metrics. The bar graphs indicate the estimated marginal mean (bar), and 95% confidence interval (error bar) for each crew's self-reported resilience scores. Lack of overlap between any crews' confidence interval indicates statistical significance between those crews.

The self-reported resilience scores showed significant differences between crews, where Crews 6, 8, and 10 had the lowest scores. Despite these crews being significantly lower than the other crews, the average reported resilience score was well over 3, indicating a self-reported resilience of more than successful.

Psychophysiological measures also showed significant differences between crews, however, the same crews did not exhibit the same differences across all the psychophysiological measures. For example, Crew 11 showed the highest Gaze Variance, but was medial for all other metrics. Crew 8, which was one of the crews that reported lower resilient scores, showed the highest Pupil Diameter and the lowest Gaze Velocity. This finding might suggest that these two metrics could be used to predict resilient behavior. Future work will include direct analysis between resilience scores and the psychophysiological values.

There are several considerations that should be noted while interpreting this work. First, the psychophysiological analyses were performed without consideration of whether the data fell within a certain window or when an "event" occurred. Specifically, the reported results include

data from the entirety of the scenarios, which may actually hide more significant effects if we focus the analyses on specific event timings. Second, we intentionally did not want to perform a direct analysis between the RPSA and psychophysiological measures. The RPSA was created for use in this study, but it is not a psychometrically validated measure. We assumed equal weighting of the individual questions towards the overall "resilience score", but it is possible that some participants showed resilience in one category (i.e., adapt) and not another (i.e., learn). There were also several missing responses which is reasonable if the participant was not able to exhibit a specific resilient quality, they were not able to rate themselves on the scale. Furthermore, we are still experimenting with ways to analyze both the RPSA scores and the psychophysiological scores. A direct comparison did not seem fair given all these considerations.

Future work will address the issues discussed in the Considerations section, but also expand on the current work. There are several other psychophysiological sensors that were used to collect data including electroencephalography and electrocardiography that we plan to analyze in similar format. Furthermore, we also plan to extract more detailed resilience scores for each crew. Each scenario had video and audio recording that we plan to have observations completed by The LOSA Collaborative and American Airlines LIT that will provide resilience metrics for each scenario and crew. This will improve our resolution and expand the types of analyses we could perform with the dataset.

Acknowledgements

This work was funded by NASA's System-Wide Safety Project, part of the Aeronautics Research Mission Directorate's Aviation Operations and Safety Program.

References

American Airlines' Department of Flight Safety. (2020). Trailblazers into Safety-II: American Airlines' Learning and Improvement Team, A White Paper Outlining AA's Beginnings of a Safety-II Journey.

Bates, D., Machler, M., Bolker, B., & Walker, S., 2009. Fitting linear mixed-effects models using lme4. Science 325, 883–885.

Dehais, F., Lafont, A., Roy, R., & Fairclough, S (2020) A Neuroergonomics Approach to Mental Workload, Engagement and Human Performance. Frontiers in Neuroscience 14, 268, 1-17.

Gray, D.E. (2013). Ethnography and participant observation (Chapter 17). Doing research in the real world. London: Sage.

Holbrook, J. Prinzel, L., Stewart M., & Kiggins, D. (2020). How do pilots and controllers manage routine contingencies during RNAV arrivals? In Proceedings of the 11th International Conference on Applied Human Factors & Ergonomics.

Hollnagel, E. (2014). Safety-I and Saftey-II: The Past and Future of Saftey Management. Farnham, UK: Ashagte.

Lenth, R.V. (2016). Least-Squares Means: The R Package lsmeans. Journal of Statistical Software 69.

Marcus, C.F. (1966). The stereographic projection in vector notation. Mathematics Magazine, 39, 2, 100-102.

Stephens, C., Prinzel, L., Kiggins, D., Ballard, K., & Holbrook, J. (2021). Evaluating the use of high-fidelity simulator research methods to study airline flight crew resilience. 21st International Symposium on Aviation Psychology, 140-145.

Stewart, M., Mathews, B., Janakiraman, V., & Avrekh, I., (2018). Variables influencing RNAV STAR adherence. IEEE/AIAA. Proceedings of the 37th Digital Avionics Systems Conference (DASC). London, UK.

Wilson, S.J., Glue, P., Ball, D., Nutt, D.J. (1992). Saccadic eye movement parameters in normal subjects. Electroencephalograph Clinical Neurophysiology. 86 (69-74).