AI-based pilot training tools and their implications for instructors and pilot training outcomes

[1]Ayala, N., [2]Bresee, J., & [2]McKenna, C.

[1]Biometrics & Performance Consulting Inc., Ontario, Canada
[2]Vocavio, Dublin, Ireland

November 2023

Executive Summary

As we advance into an era of Artificial Intelligence (AI) in pilot training, it is important to recognize the role of the subject matter expert- the flight instructor. As more pilot training tools leverage the use of AI to enhance the accuracy and efficiency of data and provide insights into crew performance, there is a greater need to understand the degree to which we can rely on AI data output and how to integrate it into current industry practices.

A sprint project between Biometrics & Performance (B&P) Consulting Inc. and Vocavio demonstrated how much of the variance in instructor grading could be accounted for by Vocavio voice analytics, as well as how the integration of AI-based software could reduce instructor workload. High yield instructor data sessions revealed how Vocavio Communication Performance Scores reliably accounted for a significant portion of instructor rated Crew Workload Scores.

Report findings also demonstrated that Vocavio software generated workload events (or indicators) are closely associated with instructor ratings of Crew Workload and Crew Stress Levels. More importantly, it was found that Vocavio software is finding more workload events than instructors are observing. For instance, for every 10 workload events that were identified by voice analytics, only 4 were observed and annotated by instructors. The findings and recommendations discussed contribute to the development and validation of AI-based pilot training tools, support instructor rating standardization, and have potential implications in the way human-AI training outcomes can be optimized.

Introduction and Research Objective

As the aviation industry navigates the shift of pilot training from an hours-based training system to competency based training and assessment solutions (CBTA), there is an increasing need to understand how competencies are identified and assessed. Under current practices, qualified instructors rate the performance of pilot observable behaviours (OBs) that correspond to the various IATA competencies (International Air Transport Association, 2023). Evaluating the communication competency has consistently posed a challenge to instructor standardization as it relies on subjective judgment calls about the presence or absence of specific aspects of communication, collectively known as OBs. Vocavio software analyses speech dialog and automatically output a range of metrics and events relating to communication and team dynamics during a training scenario. These events provide insight into communication quality, teamwork effectiveness, and crew workload, which relate to observable communication behaviours and are key factors that can impact pilot performance.

Methodology

The goal of the collaborative effort between B&P Consulting Inc. and Vocavio was to broadly examine the relationship between instructor OB data and Vocavio signal output. Six recordings and datasets from training scenarios were examined that involved crew members at a simulator training facility. These sessions were analyzed by qualified instructors in parallel with Vocavio voice analytics. The relationship between instructor ratings and Vocavio signal output

was specifically investigated through linear regression models and basic descriptive statistics. These linear regression models were particularly focused on understanding the relationship between instructor rated Crew Workload/Stress Level scores and Vocavio signal output (i.e., Communication Performance Score, and CRM/Workload Flags) at the level of the individual crew pairing. Crew Workload/Stress Level instructor ratings were correlated with their time matched Communication Performance Score. Additionally, Vocavio-generated CRM events and Workload Flags were redefined as a Workload scale (i.e., 1=low workload, no CRM event, or Workload Flag; 2= medium workload, Vocavio-generated CRM event; 3= high workload, Vocavio-generated Workload Flag) to allow for the correlation of Vocavio workload signal output and instructor ratings of Crew Workload/Stress Level.

Findings and Results

Model performance was shown to be heavily dependent on the quality of data provided. Namely, the volume of Vocavio signal data provided, and more importantly, instructor data quantity and quality within sessions was shown to be an influential factor in how informative and accurate the prediction models were. In other words, high yield data sessions produced significant relationships between instructor rated Crew Workload Scores and Vocavio Communication Performance Scores. For instance, the preliminary analysis found that for every grade increase in Vocavio Communication Performance Score, there was an associated 8% decrease in Crew Workload instructor ratings (**Figure 1**). This weak correlation (r=-0.20) was

shown to account for 18% (r^2=0.18) of instructor rated Crew Workload. Although 18% seems low, there is a widely accepted a priori threshold of 10% that is typically applied in the context of behaviour analytics where noise poses a significant challenge to signal processing and analysis. What this signifies is that Vocavio Communication Performance Scores can reliably account for a significant portion of instructor rated Crew Workload Scores.

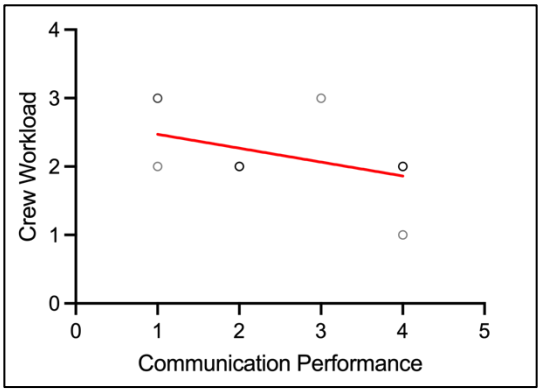

Figure 1. Scatter plots demonstrate the relationship between Communication Performance Score and instructor rated Crew Workload. Black circles indicate data points for an exemplar crew pairing with high-yield Vocavio and Instructor data output. Red line indicates the associated trend (i.e., line of best fit).

Data modeling also revealed significant trends between Vocavio Workload (e.g., Workload Level 1= No Vocavio generated events/flags; Workload Level 2= Vocavio generated

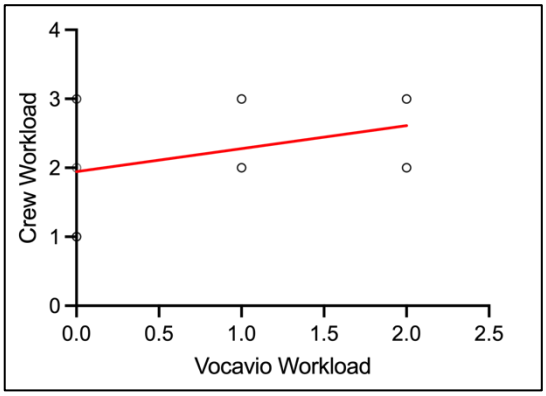

Figure 2. Scatter plots demonstrate the relationship between Vocavio Workload Level and instructor rated Crew Workload. Black circles indicate data points for an exemplar crew pairing with high-yield Vocavio and Instructor data output. Red line indicates the associated trend (i.e., line of best fit).

CRM events; Workload Level 3= Vocavio generated workload flags) and instructor rated Crew Workload Scores, as well as instructor rated Crew Stress Level. With respect to Vocavio

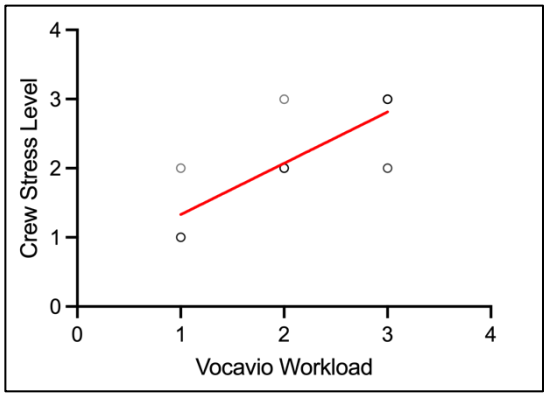

Figure 3. Scatter plots demonstrate the relationship between Vocavio Workload Level and instructor rated Crew Stress level. Black circles indicate data points for an exemplar crew pairing with high-yield Vocavio and Instructor data output. Red line indicates the associated trend (i.e., line of best fit).

Workload Level and instructor rated Crew Workload Scores, preliminary analysis found that for every level increase in Vocavio Workload, there was an 11% increase in Crew Workload instructor ratings (**Figure 2**). This weak-moderate correlation (r=0.33) was shown to account for 17% (r^2=0.17) of instructor rated Crew Workload. Additionally, Vocavio Workload Level and instructor rated Crew Stress Level analysis found that for every level increase in Vocavio Workload, there was a 14% increase in instructor rated Crew Stress Level (**Figure 3**). This strong correlation (r=0.74) was shown to account for 13% (r^2=0.13) of instructor rated Crew Stress Level. In summary, the data show significant connections between the Vocavio generated workload events/flags that are closely associated with instructor ratings of Crew Workload and Crew Stress Levels.

A second line of questions related to data capture emerged as a result of some 'red flags' that became apparent in some data sets (*Note: these problematic low yield data sets were not used to generate the modeling trends discussed above)*. Namely, a number of sessions had

minimal instructor data annotations (i.e., low data yield), which brought on several issues when trying to compare the data with high yield Vocavio signal output- automatically generating data points every 30 seconds[1]. We found that, on average, instructors observed and annotated 40% of what Vocavio voice analytics observed in relation to workload during a 20-minute LOFT exercise. This means that for every 10 workload events/flags that were identified by voice analytics, only 4 were observed and annotated by instructors. Top data yield instructors observed up to 82% of what Vocavio voice analytics observed, whereas lower data yield instructors observed as little as 16%. Indeed, higher yield instructor data helped improve model performance (i.e., validity and reliability) and demonstrated the significance of the connections between Vocavio voice analytics and instructor ratings related to Crew Workload and Stress Level. In contrast, low yield instructor data failed to reach significance, and were also more likely to violate modeling assumptions. More importantly, the ability to identify low yield instructor data highlights a critical aspect about improving instructor standardization. In particular, Vocavio software can support instructors with a stream of objective insights (evidence), where an OB or competency may have been overlooked during a simulator training session.

Another important factor to consider is the lag between instructors observing and inputting of annotations on e-grading systems. Naturally, this is to be expected, and the data

[1] This made modeling very tricky as low data yield quickly became a source of modeling assumption violations (e.g., lacked homoscedasticity, constant parameters, linear model, normality, etc.), which are typically required when data modeling is applied appropriately. Although one could not possibly expect to have some form of instructor rating generated every 30 seconds, it is still useful to have instructors provide as much annotated data as possible for two main reasons: 1) to reduce the risk of violating modeling assumptions, 2) to improve the data quantity and quality that the regression models could then use to identify and generate relationships between objective Vocavio signal output and instructor ratings.

shows there was an average lag of 46 seconds. However, more extreme lag examples were shown to extend up to 160 seconds between behaviours observed and data annotated into the e-grading system. This, of course, is less of a concern than low-yield instructor data input. However, lag may still impact model fit by reducing the validity of the way the instructor ratings are temporally lined up with more ambiguous (i.e., less neatly defined) CRM events and Workload flags reported through Vocavio signal output.

Recommendations for Application and Further Research

In light of all the data modeling outcomes and descriptive statistics reported here, it is important to note that highly populated (e.g., high yield) data sets produced models that accounted for up to 20% of the instructor rated crew workload/stress level scores; suggesting that these specific Vocavio metrics are capturing a significant portion of the behaviour that is being observed and rated by instructors with respect to Crew Workload and Stress Levels. In contrast, sparsely populated (e.g., low yield) data sessions failed to produce relationships to this effect (e.g., weak models accounting for only 4% of instructor rated Crew Workload/Stress Scores) and are limited by too few data points that often violate modeling conventions. More importantly, the data shows clear benefits to implementing AI-based assessment software like Vocavio as it captures more events than instructors are identifying. As a result, the data suggests that Vocavio software has the potential to reduce instructor workload and improve the accuracy of identifying communication and team related events and their associated workload and stress levels.

Conclusion

Taken together, these findings demonstrate the potential Vocavio metrics have in providing behavioural insights that align with and augment existing instructor evaluation methods. These results suggest that further research with large data sets will increase the precision of the identification of crew communication, teamwork, workload and stress events, enabling instructors to provide a data-rich debrief and enhanced training experience for crew.

References

International Air Transport Association (2023). *Competency Assessment and Evaluation for Pilots, Instructors and Evaluators* (2nd ed). Guidance Material.

About the Authors

Naila Ayala
Naila received her Master's of Neuroscience degree at the University of Western Ontario in 2019. She is currently a PhD candidate of Neuroscience with a focus on eye-tracking and visuomotor control. Her research explores how gaze behaviour can provide an objective measure of information processing, decision-making and complex skill performance. In collaboration with WISA, she is working to apply these questions to aviation education with the broader goal of enhancing our understanding of the complex human factors involved in piloting an aircraft and guiding the development of novel training paradigms and industry practices.

Jerome Bresee, FRAeS
VP Human Systems at Vocavio, Jerry has over 40 years in aviation training systems design ranging from the 767 to the F-35. He led AQP development for five airlines and supported the development of the U.S. Flight Operations Quality Assurance (FOQA) program.

Conor McKenna
CEO & Cofounder at Vocavio, Conor has over 25 years experience leading teams to develop novel web based technologies and experiences. He qualified as a business analyst and has led teams in corporates, start ups and academic institutions. Vocavio was inspired by his commercialisation project work at Trinity College Dublin and his exposure to the data-rich 'STRAVA' cycling app – the ultimate debriefing tool for all lycra loving cyclists who obsess about their strava metrics and data.