5-1-2021

Training Airline Pilots for Improved Flight Path Monitoring: The Sensemaking Model Framework

Dorrit Billman

Peter Zaal

Randall Mumaw

Thomas Lombaerts

Isabel Torron

See next page for additional authors

Authors
Dorrit Billman, Peter Zaal, Randall Mumaw, Thomas Lombaerts, Isabel Torron, Saad Jamal, and Michael Feary

TRAINING AIRLINE PILOTS FOR IMPROVED FLIGHT PATH MONITORING: THE SENSEMAKING MODEL FRAMEWORK

Dorrit Billman,[1,2] Peter Zaal,[3,2] Randall Mumaw,[4,2] Thomas Lombaerts,[5,2] Isabel Torron,[4,2] Saad Jamal,[4,2] Michael Feary[1,2]

NASA Ames Research Center[1]
Metis Technology Solutions[3]
San José State University Research Foundation[4]
KBR Wyle Services[5]
Moffett Field, California[2]

The importance and benefit of improved monitoring is increasingly recognized. Improved training may be a valuable intervention. Our study (conducted 2019) assessed and trained airline First Officers on flight path monitoring skills. The exploratory study assessed monitoring pre-training in a simulator session that included monitoring challenges (8 or 7 events). A 1-hour interactive training followed, based on the Sensemaking Model of Monitoring; it presented concepts and examples using a slide deck, discussion, and simple activities. Post-training assessment used scenarios with analogous monitoring challenges (7 or 8 events) but a different setting. Performance showed significant and relatively consistent improvement. Training monitoring as sensemaking merits further investigation.

The importance of monitoring performance is gaining increased attention. Monitoring gaps are a pervasive contributor to accidents & incidents (e.g., CAST, 2014) and is found at high rates in both line (Dismukes & Berman, 2010) and simulated flight (Mumaw, et al, 2010). Designation of the non-flying pilot as Pilot Monitoring (PM) and increased prominence in NOTECH (NonTechnical) and CRM (Crew/Cockpit Resource Management) training are other indicators of its importance. Despite the recognized need for improved monitoring training, there is not a standard approach for how best to achieve this.

We have proposed the Sensemaking Model of effective monitoring, which emphasizes that monitoring is far more than pointing one's eyes and detecting stimuli (Billman, Mumaw, & Feary, 2020; Mumaw, Billman, & Feary, 2020). Rather it is an active process of building and maintaining a relevant, accurate model of the dynamically unfolding situation. This depends both on activating an accurate mental models of "how things work" from long term knowledge and on using that understanding to guide collecting and assessing relevant information. The process of updating the situation model may frequently contain important gaps or errors (e.g., about modes, Sarter & Woods, 1995) and may contain errors of understanding, not just a failure to notice, that lead to accidents (AAIASB, 2006, cited in Dismukes & Berman, 2010). Thus, improving the overall process from noticing to understanding the situation may be a critical target of training. Further, the active, structured nature of this type of monitoring may also help the pilot stay engaged and interested. Monitoring for flight path management maybe particularly critical for safety and is the focus of our work. We characterize the monitoring process as initiating a situation model by drawing on relevant models in memory, followed by a three-phase cycle of updating this model. The pilot 1) identifies a key, relevant question that needs to be answered in the current situation, such as a gap or inconsistency in the situation model; 2) gathers relevant evidence, which requires identifying the sources of information, analyzing the information, and comparing current versus expected values; and 3) identifies what actions are needed. The central role of the Situation Model and the 3-phase cycle are illustrated in Figure 1. The monitoring cycle also provides anchor points for when and what to communicate with the other pilot. This model simplifies monitoring, highlighting certain aspects over others.

We conducted an exploratory training study that suggests the sensemaking model provides a useful framework for training monitoring. A collaborating US carrier provided us use of a training flight simulator and the pilot union supported recruitment of First Officer pilots to participate. The primary goal was to assess whether providing airline pilots with foundational training based on the Sensemaking Model of Monitoring would produce measurable improvement in flight path monitoring. To do this, we needed to develop sensitive measures of flight path monitoring that are capable of detecting change and a training intervention to produce change.

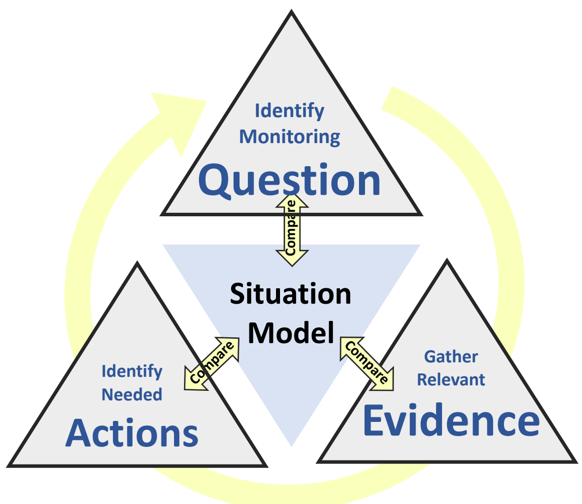

Figure 1. The Situation Model and the processes for drawing on and updating it.

Study Method

Design

We used a 2 Training (pre- vs post-training, within subject) x 2 Scenario Order (Scenarios 1&2 first vs Scenarios 3&4 first, between subject) x 2 Display Configuration (with vs without Flight Director, within subject) design. The pre- versus post-training variable assessed the impact of training, while Scenario Order was a counterbalancing factor. In Scenario Order 1, participants flew Scenarios 1 & 2 before training; in Scenario Order 2, they flew Scenarios 3&4 before training. Performance of each individual and on each item was assessed pre- and post-training. Pragmatic factors of scheduling and simulator availability constrained feasible designs, as discussed in Limitations. The study also asked whether presence versus absence of the Flight Director affected pilot monitoring. Eye tracking data was also collected. See Zaal et al, 2021.

Participants

The participants were 19 First Officers (FOs) who were active and current on the 737 NG. Flight hours ranged from 4100 to 14000 with a median of 7000. Participants were recruited through the union and were offered $100 and NASA stickers as an honorarium. We sought pilots in their first five years at the airline, as they might benefit more from additional training than more experienced pilots; 17 met this criterion.

Procedures and Equipment

The study took 3.5 hours and had five phases: the orientation and demographics interview; simulator session one training tutorial; simulator session two; and the study feedback interview. The simulator was a CAE 737-700 full-flight simulator used in the airline’s standard configuration, equipped with Seeing Machine eye-tracking. Participants always served as FO and PM; unknown to the FO, the Captain was a confederate who introduced scripted errors. Sessions were also staffed with an Instructor Pilot introducing ATC clearances, and two experimenters.

Training was conducted as a tutorial structured by a slide deck that standardized coverage. The slide deck included reading information based on the sensemaking model of monitoring, answering both comprehension and examples-from-experience questions, debriefing exercises both of their own simulator flight and of a video snippet, and follow-up discussions of the participant's response.

Key Materials

Approach Scenarios 1&2 (Airport A) versus 3&4 (Airport B) were counterbalanced between Pre- and Post-training sessions. To measure monitoring performance 15 challenging events (see Table 1) were designed so that noticing and understanding the event would lead to specific, identifiable behaviors, and enable objective scoring. Behaviors were typically talking to the Pilot Flying (PF), but some were control actions. Integrating so many issues for the PM to catch while maintaining plausibility relied critically on collaboration with senior pilots, drawing on reported safety events and their own line experience. The challenging events in Scenarios 1&2 versus 3&4 were designed in pairs to pose challenges with similar difficulty, but in different airports and conditions (see Table 1). Matched pairs proved possible for 14 of the 15 events.

Table 1*. Challenge Event Descriptions, by Matched Pair (where possible)*

Challenge Type	Scenario 1	Scenario 3
High on Path (ATC)	**#1** Slowed by ATC	**#9** Held high by ATC
Inappropriate mode	**#2** PF remains in VNAV	**#10** PF selects HDG SEL
Instrument issues	**#3** Given wrong altimeter setting	**#11** False glideslope
Did not enter value	**#4** Field elevation not set on MCP	
	Scenario 2	Scenario 4
Inappropriate mode	**#5** Auto-flight/PF interaction VS	**#12** PF engages LVL CHG
Shortened lateral path	**#6** ATC gives direct-to	**#13** ATC gives direct-to
Inappropriate mode	**#7** PF selects LNAV	**#15** PF fails to arm APP
Airspeed error	**#8** PF calls flaps 25 when too fast	**#14** PF fails to call for flaps 5

The training materials introduced the key concepts of monitoring as active inquiry and understanding as characterized by the Sensemaking Model and included activities to support understanding and integration. The training session was guided by an experimenter and was intended to maximize participant understanding of the key concepts and their application. The importance and centrality of a situation model were explained and illustrated. Training addressed: 1) identifying what question about the situation is the priority to answer, 2) how to gather evidence and assess it against expected values to answer the priority question, and 3) how to identify whether and what actions need to be taken by the pilots. The importance of talking to share information and align the situation models of each pilot was addressed. Pilots applied these concepts in creating a short self-debriefing of an event they had just flown and, later, of a video of another pilot's monitoring the same situation. Examples of monitoring situations were used to illustrate the concepts, and pilots were asked to produce examples from their experience as well. The training materials were focused on information content and were not designed to resemble official training materials; delivery was supported by flexible interaction with an experimenter. Individualized feedback was provided as our goal was to maximize learning, not to control or identify specific training components.

Results

We scored performance on each monitoring event based on the crew videos and simulator data streams.The intended scoring was to code 4 operationally distinct performance levels, two passing, two unsuccessful. A 3- level scale was used for 6 events, because behavior provided only 3 operationally consequential levels of performance .

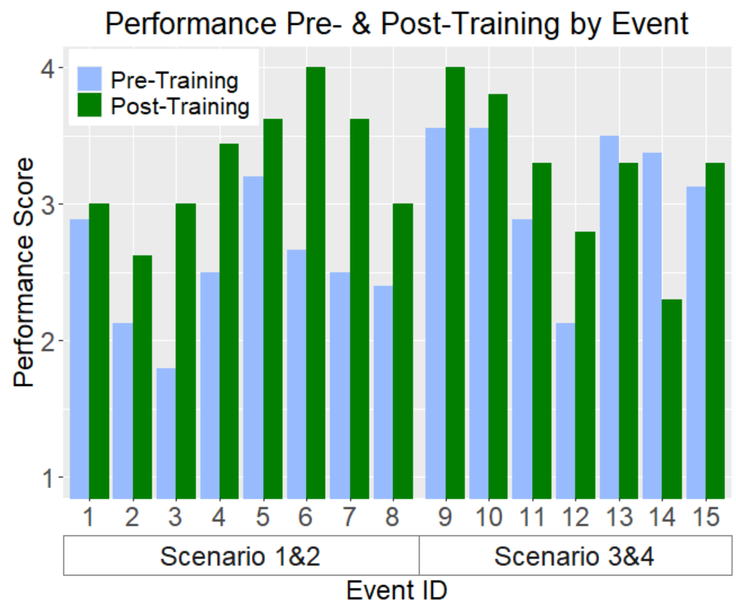

Figure 2. Pilot performance per event.

Overall mean performance score pre- training was 2.8 and post-training was 3.23, showing better performance after training. Figure 2 shows that performance on 13 of the 15 events was higher after training than before. Considered by individual, 13 of the 19 participants improved, 1 had identical pre- and post-training scores, while 5 scores declined. Table 2 shows pre- and post- training performance broken down by the counterbalancing factor of presentation order. To test the significance of training, we modeled Performance Score (ordinal) as predicted by Training (fixed factor), by Scenario Order (fixed factor), by the Training X Scenario Order interaction, by Participant (categorical random factor), and by Event (categorical random factor) using the Gamma distribution family in cumulative link mixed models (clmm routines in R). We compared this to a model that did not include the Training factor. Comparison using a likelihood ratio test showed that including the Training factor significantly improved fit ($\chi2(2)=10.868$, $p=.00437$). While an interaction between presentation order and training is suggested in Table 2, neither the effect of Scenario Order, nor the Scenario Order X Training interaction was significant (for Scenario Order, $\chi2(2)= 2.066$, $p=.356$, or for Scenario Order X Training, $\chi2(1)=.624$, $p=.430$).

Table 2. *Mean Performance Score Before and After Training*

Subject Group	Pre-training	Post-training
Grp1:Scenarios 1&2 first	2.51 *Items 1-8*	3.19 *Items 9-15*
Grp 2:Scenarios 3&4 first	3.17 *Items 9-15*	3.28 *Items 1-8*
Overall	2.80	3.23

An analogous assessment of the pass/fail score using binomial tests also found a significant effect of training ($\chi2(2)=10.341$, $p=.00568$) but not of Scenario Order or of the

Scenario Order X Training, though these became marginally significant, Scenario Order, $\chi2(2)= 4.999$, $p=.0821$, or for Scenario Order*Training, $\chi2(1)=3.066$, $p=.07999$. Improvement on the pass/fail score as well as the multi-point scale suggests that training did not simply shift subtle levels of performance within pass or fail categories, but was associated with shifts from unacceptable to satisfactory performance.

Interpretation of time-of-successful-completion is complicated because different events contribute to this measure before and after training and the role of time differed across events. However, for all events that were successfully performed, there was no difference in time to successfully complete the events before (mean = 103 s) versus after (mean = 97 s) training.

Conclusions

Summary

The study found significant improvement in monitoring simulator events designed to require active monitoring, after a brief training and a pre-test experience simulating approaches on a different airport. Our two before- and two after-training scenarios in an airline simulator, assessed 15 challenging events in total. Our design allowed the variation from participants and from events to be modeled, thus increasing statistical sensitivity. The design of the scenarios was challenging, and the fact that they were sensitive to changes after training suggests they can provide a useful resource for future research. The tutorial introducing the sensemaking concepts of monitoring was brief and had not been intensively developed. Nevertheless, this exploratory study suggests that monitoring can be improved with a modest intervention.

Limitations

Our design has two particular limitations, concerning diagnosticity and transfer. First, our study lacked a control group who flew both our pre- and post- training scenarios but had no training. We used performance in the two simulator sessions as our *measurement* of training, but experience in a simulator is itself a powerful learning opportunity. Without a no-training control, we cannot tell whether changes after training are due to the tutorial or due to learning from the first simulator session. Of course, both learning from the simulator sessions and from the tutorial would be valuable if that learning was retained and transferred. However, turning to the second limitation, we did not measure performance after a delay or in a new context. At short delay, memory for the specific cues, such as the confederate pilot or memory of the pilot's own self-debriefing may have produced a brief, context-specific benefit. Our study cannot diagnose whether change in performance is due to the tutorial, the initial simulator session, or the combination. Further, we do not know if performance would transfer to different or delayed simulator sessions, let alone to line operations. Our study uses a limited number of participants and events, and it lacked sensitive measures of pilot skill or experience and of item difficulty.

Future Directions

Monitoring continues to gain importance as system supervision becomes an increasing part of pilots' roles .Our study contributes a valuable initial finding about the ability to improve monitoring performance. This suggests that further efforts to use the Sensemaking Model of

Monitoring to guide training may be worthwhile. A basic replication with a new set of participants would be useful. Very recently a partial replication using a more polished training session but without performance measurement was undertaken (Baron, 2021). Even more informative would be the ability to test performance over a delay and in a setting which did not as strongly cue the training experiences. Rather than aiming to separate the effects of direct training about monitoring from the effects of additional time in the simulator, it may be more valuable to address how simulator time and targeted, out-of-simulator training can be integrated. In particular, linking concepts to simulator exercises through prebriefs and debriefs, and delayed assessment may provide a powerful integrative structure. This strategy for improved monitoring training is ambitious and may require incremental development, but we are excited about the prospects.

References

Baron, B. (April 2021) The Sensemaking Model for Airline Pilot Training: Building and Maintaining Expert Flight Path Management. accepted at Teaching, Colleges & Community. tcchawaii.org.

Billman, D., Mumaw, R.J., & Feary, M. (2020). A model of monitoring as sense-making: Application to flight path management and pilot training. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 64.

CAST (2014). Airplane state awareness joint safety analysis team. https://www.skybrary.aero/bookshelf/books/2999.pdf

Dismukes, R. K., & Berman, B. A. (2010). Checklists and monitoring in the cockpit: Why crucial defenses sometimes fail, NASA Technical Memorandum (NASA TM-2010–216396). MoffettField, CA: NASA Ames Research Center.

Mumaw, R.J, Billman, D. & Feary, M. (2020) Analysis of Pilot Monitoring Skills and a Review of Training Effectiveness. NASA Technical Memorandum (NASA/TM-20210000047). MoffettField, CA: NASA Ames Research Center.

Mumaw, R.J.,Roth,E.,Vicente, K.J.,Burns, C.M. (2000) There Is More to Monitoring a Nuclear Power Plant than Meets the Eye. *Human Factors, 42*, 36–55.

Sarter, N. B., & Woods, D. D. (1995). How in the World Did We Ever Get into That Mode? Mode Error and Awareness in Supervisory Control. *Human Factors*, *37*(1), 5–19.

Zaal, P., Lamberts, T. et al (2021). An Eye-Tracking Study into the Effect of Training on Pilot Monitoring, to be presented at the AIAA 2021 AVIATION Forum.

Acknowledgements

The research was supported in part by the FAA, “Training Methods for Improved Pilot Monitoring Performance.” We thank the airline’s pilots and simulator engineers who helped develop the experiment, instructor pilot who recorded two training snippets, the pilots who served as confederates and instructors during the simulator sessions, and the pilots who participated in the experiment. Thanks to Megan Shyer for help with graphs, to Kathryn Ballard for statistical discussions, and to Immanuel Barshi for comments on the paper.