Pilot problem detection during manual and automated flight

Van Den Eijkel, S. J.; Landman, A.; Van Paassen, M. M.; Stroosma, O.; Mulder, M.

DOI
10.1016/j.trpro.2025.05.014

Publication date
2025

Document Version
Final published version

Published in
Transportation Research Procedia

Citation (APA)
Van Den Eijkel, S. J., Landman, A., Van Paassen, M. M., Stroosma, O., & Mulder, M. (2025). Pilot problem detection during manual and automated flight. *Transportation Research Procedia*, *88*, 112-118. https://doi.org/10.1016/j.trpro.2025.05.014

Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 88 (2025) 112–118

Transportation Research Procedia

www.elsevier.com/locate/procedia

European Association for Aviation Psychology Conference EAAP 35

Pilot problem detection during manual and automated flight

S. J. van den Eijkel[a], A. Landman[a,b]*, M. M. (René) van Paassen[a], O. Stroosma[a], M. Mulder[a]

[a]*Delft University of Technology, Kluijverweg 1, Delft, 2629HS, The Netherlands*
[b]*TNO, Kampweg 55, Soesterberg, 3769DE, The Netherlands*

Abstract

We tested whether pilots would detect low-salient controllability problems more quickly during manual compared to automated flight. Using a moving-base simulator and a Piper Seneca aerodynamic model, airline pilots (n = 20) performed scenarios in which either a gradually ensuing single-engine failure or an icing accumulation occurred. Both scenarios were performed once during manual flight and once during automated flight, and were alternated with distraction scenarios. The icing accumulation was detected marginally significantly more quickly during manual flight, while there was no significant difference for the engine failure. Problems in manual flight were, as expected, most likely discovered from aircraft motions or control forces. Interestingly, there were several late detections during manual flight which appeared to be caused by subconscious manual corrections. In automated flight, the engine failure was discovered most often from the engine manifold pressure indication, while the icing accumulation was most often discovered from control column movement. The results therefore underline the importance of using back-driven controls, and further indicate that manual flight does not necessarily improve detection of problems that occur without display indications.

Peer-review under responsibility of the scientific committee of the European Association for Aviation Psychology Conference EAAP 35

Keywords: Situation awareness; Vigilance; Aviation; Automation

* Corresponding author.
E-mail address: h.m.landman@tudelft.nl

Peer-review under responsibility of the scientific committee of the European Association for Aviation Psychology Conference EAAP 35
10.1016/j.trpro.2025.05.014

1. Introduction

Technological advances in automated flight have successfully reduced pilot workload in the cockpit (Ephrath & Young, 1981; Masalonis, Duley & Parasuraman, 1999) and allowed for improved monitoring (Hancock & Williams, 1993). The autopilot has increased efficiency and flexibility in flight operations, and improved safety compared to the previous generation of aircraft (Boeing Commercial Airplane Group, 1997). These innovations also led to a reconsidering of the role of humans on the flight deck (Draper, Young & Whitaker, 1964). Concerns about the interaction between the crew and the automation are growing, as a growing proportion of accidents are currently caused by pilot interactions with automation (IATA, 2014; Snow, 2015). Humans are thought to be not well-suited for monitoring processes without actively controlling these processes (Bainbridge, 1983). Several studies have shown that pilots sometimes have a poor understanding of autopilot modes, leading to confusion about the active mode or selection of incorrect modes (Active Pilot Monitoring Working Group, 2014; Sarter & Woods, 1995), or to over-reliance on the autopilot (Young and Stanton, 2002; Parasuraman, Sheridan & Wickens, 2000; Masalonis et al., 1999). Without an effective feedback loop, pilots may also become unaware that their monitoring habits are degrading (Active Pilot Monitoring Working Group, 2014).

The optimal degree of automation may depend on phase of flight. Earlier studies have shown that adaptive task allocation (i.e., switching between manual and automatic control) is a way to increase vigilance (Davies & Parasuraman, 1982). However, experimental studies comparing detection of problems between manual and automated flight show mixed results. Participants who actively controlled a single-axis compensatory tracking task were faster to notice a change in the dynamics of the controlled element, compared to passive observers (Ephrath & Young, 1981; Kessel & Wickens, 1982; Wickens & Kessel, 1979). Similar results were found for an automated driving task (Desmond, Hancock & Monette, 1998). In contrast, heading and pitch deviations from a desired flight path in a fixed-base simulator were detected more quickly by passive controllers than by active controllers of large jet aircraft types (Ephrath & Young, 1981; Ephrath & Curry, 1977), possibly due to increased complexity and workload of the task. In real operations, the autopilot may counteract deviations from the flight path, mask failure cues or cause cues that differ from those which a pilot expects to result from a failure (Billman, Mumaw & Feary, 2020). This may lead to incorrect interpretations of such failure situations. If the autopilot counteracts a failure and is disengaged, this can in some cases cause a sudden exacerbation of controllability issues, leading to startle, confusion and inappropriate responses.

In the current study, we compared pilot detection and sense-making of low-salient (i.e., gradually increasing) controllability problems between automated flight and manual flight. Two well-known problems were selected, namely a single-engine failure and icing accumulation. The autopilot can mask the consequences of such controllability problems in respectively the roll/yaw axes and the pitch axis. Nevertheless, the problems will still be detectable in automated flight as the control column will move in line with increased autopilot inputs due to the use of back-driven controls (i.e., controls that move in line with autopilot inputs). Most aircraft currently feature such back-driven controls, but newer models are transitioning to fly-by-wire controls that are non-back-driven. The current study thus aims to provide insight into how effectively pilots can detect problems based either on cues that present themselves in manual flight, or on cues that present themselves in automated flight. We hypothesize that gradually-occurring problems are more quickly detected in manual flight, as this requires pilots to close the control loop themselves.

2. Method

2.1. Participants

Licensed pilots (n = 20, mean age = 39.6 years, SD = 10.6 years) participated, who had either an Airline Transport Pilot License (ATPL) or a Commercial Pilot License (CPL). All but two participants had experience with flying twin-propeller aircraft. One participant was actively flying twin-propeller aircraft and had more than 3000 hours of experience. Nine participants were currently Captain, ten were First Officer and one was Second Officer (i.e., the third in command on long-haul flights). One participant had an active type-rating on an aircraft without back-driven controls, five others had a mixed type-rating that included aircraft without back-driven control. The distribution of the

pilots' total flight hours is shown in Figure 1 (left).

2.2. Apparatus

The SIMONA Research Simulator (Figure 1, right) at Delft University of Technology was used to conduct the study. The SIMONA is a full-motion research simulator with a six-degrees-of-freedom hexapod motion system. The outside visuals are rendered with FlightGear and displayed on a collimated, 180 degrees horizontal by 40 degrees vertical field of view. A 5.1 surround sound system provided realistic 3D sound in the simulation. Participants were able to communicate with the experiment leader using a headset. The aerodynamic model used in this study was a Piper Seneca III, a popular twin-engine propeller aircraft that is widely used in general aviation. It was outfitted with a control column. For this study, the model has been enhanced with back-driven controls so that the input from the autopilot is observable to the pilot.

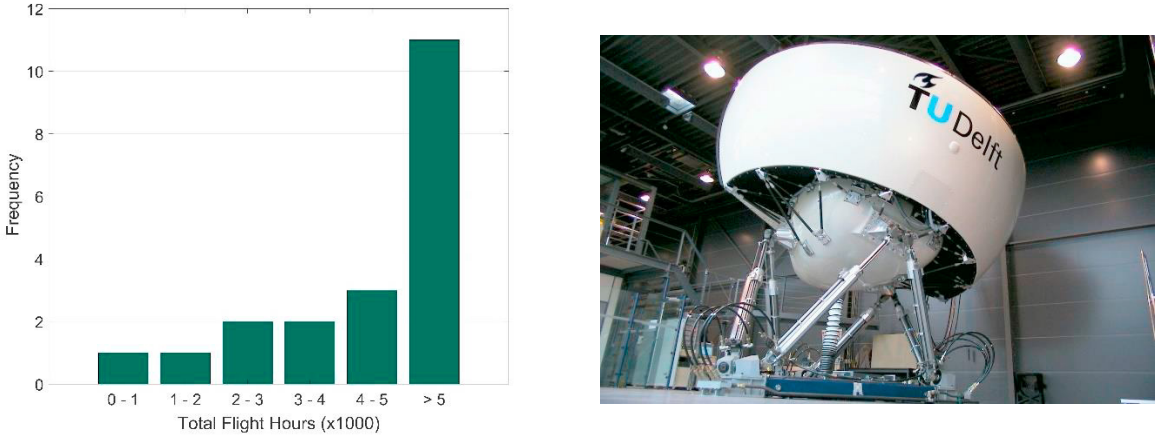

Fig. 1. Left: the participants' flight hours distribution. Right: the SIMONA research simulator.

The Piper Seneca III was equipped with instruments similar to the Garmin G1000 avionics and an autopilot inspired by the Garmin GFC700 with several autopilot modes. The autopilot had elevator and roll control but did not have auto throttles or rudder control. The G1000 avionics consisted of a Primary Flight Display (PFD) with speed tape, altitude tape, Horizontal Situation Indicator (HSI), Flight Mode Annunciator (FMA), Flight Path Vector (FPV) and Flight Director (FD), as well as controls for the autopilot system and an Multi-function Display (MFD) which displayed engine parameters (Manifold pressures, RPM, Cylinder Head Temperatures (CHT) etc.) as well as Outside Air Temperature (OAT) and Local Time.

2.3. Experimental design

The experiment had a within-subjects design to test whether the mode of flight control (i.e., either fully manual or using the autopilot) had an effect on the time required to detect problems. Each participant performed two scenarios in both a manual and autopilot condition, totalling four test scenarios (see, section 2.3.1). Pilots were instructed to speak out loud as if a co-pilot was present, and to call out any anomalies they spotted. The dependent measures were the detection time and the cue which led to the detection. Four distraction scenarios were added to mitigate recognition of the test scenarios. The order of the scenarios was counterbalanced between participants to prevent order effects.

2.3.1. Procedure

The total testing session lasted a maximum of three hours. Participants received a briefing in which they were told that the goal of the experiment is to evaluate the realism of the Piper Seneca III model and simulated events. Participants then performed three familiarization flights (traffic patterns) to practice with manual control, automated

flight, and calling out failures. Next, there were four test scenarios intermixed with four distraction scenarios. Each scenario lasted 8-12 minutes. Each test scenario was flown both manually and on autopilot and is described below. The distraction scenarios were: one automated flight and one manual flight scenario without problems, one autopilot flight scenario with a blocked pitot tube, and one manually flown scenario with a blocked static tube.

Each scenario began with a briefing on the aircraft state and context. Participants were tasked to use the automation modes that the experiment leader set at the start of each scenario, unless it was required for safety to disengage them. For the manual control condition, pilots flew manually without FD (i.e., a visual indication of the required flight path to follow selected headings, speeds or altitudes). For the automated flight condition, pilots used the FD and either the heading (HDG) or altitude (ALT) mode. When using the autopilot, the participants had to control the throttle setting themselves, as there was no autothrottle. Each test and distraction scenario concluded with a brief set of questions on what happened during the scenario (see, Dependent measures). This included a distraction question on the simulator cueing. After the third scenario, there was a 20-minute break.

Test scenarios were developed to feature realistic and relevant failures that have a gradual onset but be detectable by all pilots eventually. These failures would necessarily express themselves differently in the autopilot and manual conditions. The result was a gradual engine failure scenario and an icing accumulation scenario.

2.3.1.1. Engine failure scenario

Participants were instructed to fly various headings and flight level changes in the first 400 seconds of the flight until they were commanded an altitude change to 7,000 feet at 115 knots. At around 30 seconds after leveling off, a failure was triggered to reduce the left engine power to zero in 60 seconds. The manual condition scenario was the same, except that altitude change was to 5000 instead of 7000 feet. The indicated manifold pressure on the MFD will immediately start to drop from about 12 to 10 inHg in about 14 seconds. The RPM starts dropping about 30 seconds into the failure. The failing engine causes a disturbance in the roll and yaw axes. In the autopilot condition, the autopilot will compensate the roll angle to maintain the selected heading, but yaw and side slip will still occur. The autopilot will cause the control column to move from zero to 30 degrees roll right in 60 seconds. If the participant took no corrective action, the aircraft would enter a left-hand spin 72 seconds into the failure due to loss of speed. In the manual condition, no correction from the participant would lead to gradually increasing left-hand bank to 45 degrees and gradually pitching down to 10 degrees at 22 seconds.

2.3.1.2. Icing scenario

Participants were instructed to climb to 7,500 feet after about 450 seconds of various heading and flight level changes. At 40 seconds after leveling off, icing started to accumulate. The outside air temperature would be between -3 and 0 degrees Celsius. The manual condition scenario was the same, except that the altitude change was to 8,000 feet. The icing would decrease elevator effectiveness, causing the aircraft to slowly pitch up. In the autopilot condition, the pitch-up movement would be counteracted, visible by the control column moving forward. At 92 seconds into the failure, the most forward position of the column would be reached. Without corrective action, the aircraft would stall. There was no trim indicator or control surface deflector indicator available.

2.4. Dependent measures

The following dependent measures were obtained in the test scenarios: *Problem detection Time:* The time between the start of the engine failure or icing accumulation and the moment the pilot started mentioned that something was wrong. This was analysed using the audio recording. *Primary Detection Cue:* In the post-scenario interview, the pilots were asked which cue was the first trigger to cause them to detect that something was wrong.

2.5. Statistical analysis

Detection times were checked for normality. If normally distributed, they were compared between conditions for each scenario separately using paired-samples t tests. Else, Wilcoxon signed-rank tests were used. Alpha was set to 0.05. Effects with $p < 0.06$ were considered as marginal significant.

3. Results

3.1. Problem detection time

The comparisons of detection times between automated and manual flight are shown in Table 1. All problems were detected eventually. There was no significant difference between automated and manual flight for the engine failure detection, $Z = -0.141$, $p = 0.888$ (see, Figure 2, left). There was a marginally significant difference in the icing scenario, $Z = -1.89$, $p = 0.058$, suggesting faster detection in manual flight (see, Figure 2, right). Participants were on average about eight seconds quicker to detect the icing accumulation in manual flight. Participants who noticed this failure after 90 seconds, noticed it due to the aircraft starting to stall. This occurred eight times during automated flight and two times during manual flight.

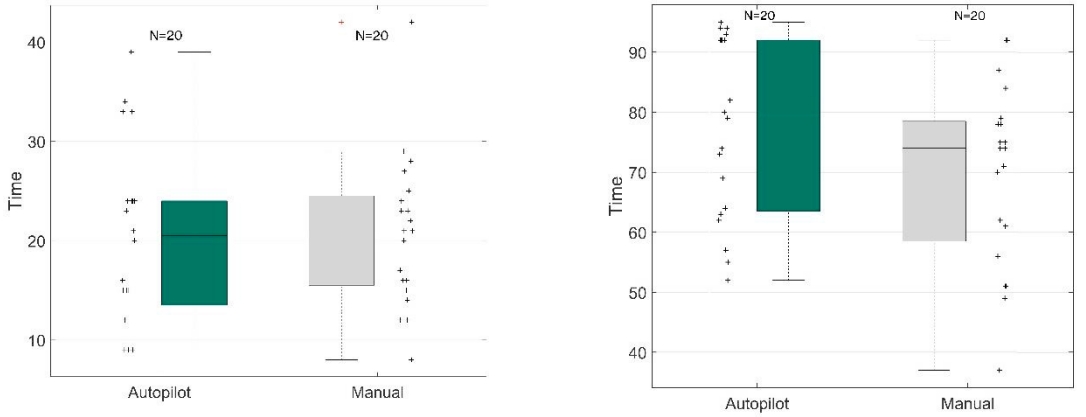

Fig. 2. Left: Tukey boxplot of problem detection times of the engine failure. Right: Tukey boxplots of problem detection times of the icing accumulation.

Table 1. Results of problem detection time.

Problem	Mode	Mean (s)	SD	Minimum (s)	Maximum (s)
Engine failure	Autopilot	20.40	20.5	9	39
Engine failure	Manual	20.75	21.0	8	42
Icing accumulation	Autopilot	77.70	79.5	52	95
Icing accumulation	Manual	69.80	74.0	37	92

3.2. Cues leading to detection

The primary cues that led to detection are listed in Table 2. The engine failure in automated flight was discovered in 13/20 cases by a drop in the indicated engine manifold pressure, in 5/20 cases by a drop in the indicated airspeed, and in 2/20 cases by aircraft pitch and yaw movement. Thirteen out of twenty participants had their hands on the control column in this scenario, but no pilots mentioned movement of the control column as primary detection cue. In manual flight, the aircraft's tendency to roll and yaw, as well as the need to correct this, was named as primary cue in 13/20 cases and a drop in engine manifold pressure was named in 7/20 cases.

The icing accumulation in automated flight was detected by 12/20 pilots from the movement of the control column. The other eight pilots noticed it from the aircraft pitching up as the aircraft lost lift and began to stall. Thirteen out of twenty pilots had their hands on the control column while monitoring the autopilot. Only two out of these thirteen discovered the problem after stalling. Those who had their hands on the control column detected the icing on average 23 seconds earlier than those who did not. This was a significant difference as confirmed with a Mann-Whitney U test, $Z = -2.546$, $p = 0.008$.

The icing accumulation in manual control was detected in all cases by the need to trim down repeatedly. Some pilots reported checking if something was wrong by not trimming down and applying no force on the control column, to check what the aircraft would do.

Table 2. Results of primary cue leading to problem detection

Problem	Mode	Primary cue	N cases
Engine failure	Autopilot	Engine Manifold Pressure indication	13/20
Engine failure	Manual	Aircraft motions	13/20
Icing accumulation	Autopilot	Control column motion	12/20
Icing accumulation	Manual	Pitch up tendency	20/20

3.3. Validation check of scenarios

None of the participants required reminding during the experiment to talk as if there was a co-pilot flying with them. The engine failure was correctly diagnosed by all participants before the scenario was stopped. One participants commented that the absence of display or aural warnings was unexpected, another that loss of power should be accompanied with a more clearly distinguishable decrease in engine sound. The icing accumulation was often diagnosed as a trim runaway, jammed elevator, or a cargo or passenger shift. Six pilots named icing as the plausible cause. Participants named the absence of a trim indicator and trim cut-out switches as a hindrance for detecting and correctly diagnosing the problem. After the second time participants experienced the same failure, only two participants mentioned that they recognized cues of the engine failure, and one of the icing accumulation, from having experienced the failure before in the experiment.

4. Discussion

In this experiment, we tested whether two low-salient problems that are well-known to pilots would be detected more quickly under manually-controlled flight than under autopilot-controlled flight. Two problems were simulated, which expressed themselves through a gradual deviation in controllability either in the pitch axis (icing accumulation) or in the roll and yaw axes (gradual single engine failure). There was a marginally significant effect indicating that detections were quicker for the icing accumulation event. This problem only expressed itself though an increase in required pitch up input (manual flight) or by the control column moving back (automated flight), but not through any obvious display indications. Thus, it seems that abnormal manual inputs are noticed more easily than autopilot-induced movement of the control column. Interestingly, there were still some late detections of this problem in manual flight, as 2/20 pilots detected the problem only when the aircraft began to stall. It seems that these pilots subconsciously counteracted the problem by pitching forward or trimming down. This indicates that using manual control is not a catch-all for the detection of gradually-occurring controllability issues.

In automated flight most pilots (12/20) detected the icing accumulation due to control column movement, whereas the rest (8/20) detected it after the stall. This underlines the importance of using back-driven controls, i.e. controls that will move in response to autopilot actions, although it does not seem to be a catch-all solution for problem detection. The number of pilots detecting the problem only while stalling is surprisingly high, as the control column would have moved full forward before the stall occurred. Six of the eight pilots who detected the problem after the stall did not have their hands on the control column, and holding the column significantly decreased detection times. Pilots commented that their employer recommends keeping hands near the controls below 10,000 feet, and the results of the icing accumulation scenario support the effectiveness of this policy.

For detection times in the engine failure scenario, no significant difference was found between manual and automated control. Most pilots detected this issue in automated flight based on the indicated manifold pressure change, supporting the notion that using automated flight freed attentional resources so that the secondary flight instruments were scanned better. Compared to the icing accumulation scenario, which featured no display indications, this would have given them an advantage to detect problems in the automated flight condition. The engine failure scenario in

automated flight also required manual input on the throttle to correct the decreasing speed, which possibly also helped pilots detect the problem, as five mentioned a decrease in airspeed as the primary detection cue.

A limitation of the current experiment is the use of an aircraft type with a generic cockpit interface, which pilots do not use in their daily work. This made it likely more difficult to detect abnormal aircraft behavior. A second limitation for applying results to operational practice is that pilots were likely more vigilant due to the simulated setting. For the gradually-occurring problems, scenarios were used that would induce visual or aural alerts (e.g., approach to stall, engine failure) in jet transport aircraft. This was done to avoid the use of problems which depend on inadequate flight path monitoring or autopilot monitoring, but this also limits generalizability of the findings to jet transport operations.

In conclusion, our results support the notion that manual flight increases detection of low-salient controllability issues. With display indications of a problem, the advantage of manual flight seems to disappear. Additionally, the use of hand-on monitoring of the control column, as well as the use of back-driven controls, appeared crucial in detecting low salient controllability issues in this experiment.

References

Active Pilot Monitoring Working Group. (2014). A Practical Guide for Improving Flight Path Monitoring. Technical Report. Flight Safety Foundation.

Bainbridge, L. (1983). Ironies of automation. Automatica 19, 775–779.

Billman, D., Mumaw, R., Feary, M. (2020). A model of monitoring as sensemaking: Application to flight path management and pilot training. Proceedings of the Human Factors and Ergonomics Society Annual Meeting 64, 244–248.

Boeing Commercial Airplane Group. (1997). Statistical summary of commercial jet aircraft accidents: Worldwide operations. Technical Report.

Davies, D., Parasuraman, R. (1982). The psychology of vigilance. Academic Press London: New York.

Desmond, P. A., Hancock, P. A., Monette, J. L. (1998). Fatigue and automation-induced impairments in simulated driving performance. Transportation Research Record 1628, 8–14.

Draper, C. S., Young, L. R., Whitaker, H. P. (1964). The Roles of Man and Instruments in Control and Guidance Systems for Aircraft. 15th International Astronautical Congress.

Ephrath, A. R., Curry, R. E. (1977). Detection by pilots of system failures during instrument landings. IEEE Transactions on Systems, Man, and Cybernetics 7, 841–848.

Ephrath, A. R., Young, L. R. (1981). Monitoring vs. Man-in-the-Loop Detection of Aircraft Control Failures. Springer US, Boston, MA. pp. 143–154.

Hancock, P., Williams, G. (1993). Effect of task load and task load increment on performance and workload, in: Seventh International Symposium on Aviation Psychology, Citeseer. pp. 328–334.

IATA (2014). Loss of Control In-Flight Accident Analysis Report. Technical Report.

Kessel, C. J., Wickens, C. D. (1982). The transfer of failure-detection skills between monitoring and controlling dynamic systems. Human Factors 24, 49–60.

Masalonis, A.J., Duley, J.A., Parasuraman, R. (1999). Effects of manual and autopilot control on mental workload and vigilance during simulated general aviation flight. Transportation Human Factors 1, 187–200.

Parasuraman, R., Sheridan, T.B., Wickens, C.D. (2000). A model for types and levels of human interaction with automation. IEEE transactions on systems, man, and cybernetics. Part A, Systems and humans : a publication of the IEEE Systems, Man, and Cybernetics Society 30, 286–97.

Sarter, N., Woods, D. (1995). How in the world did we ever get into that mode? mode error and awareness in supervisory control. Human Factors, 37, 5–19.

Snow, M. A. (2015). Preventing loss of control in flight. Boeing Aero Magazine, No 57, 9–12.

Wickens, C. D., Kessel, C. (1979). The effects of participatory mode and task workload on the detection of dynamic system failures. IEEE Transactions on Systems, Man, and Cybernetics 9, 24–34.

Young, M., Stanton, N. (2002). Malleable attentional resources theory: A new explanation for the effects of mental underload on performance. Human factors 44, 365–75.