DOT/FAA/TC-22/10

TNO 2022 R10427

Federal Aviation Administration
William J. Hughes Technical Center
Aviation Research Division
Atlantic City International Airport
New Jersey 08405

Negative transfer of training: Simulator study into effects of going beyond alarms during stall recovery training

May 2022

Final report

U.S. Department of Transportation
Federal Aviation Administration

NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The U.S. Government assumes no liability for the contents or use thereof. The U.S. Government does not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the objective of this report. The findings and conclusions in this report are those of the author(s) and do not necessarily represent the views of the funding agency. This document does not constitute FAA policy. Consult the FAA sponsoring organization listed on the Technical Documentation page as to its use.

This report was originally published by the Netherlands Organization for Applied Scientific Research (TNO), Human Performance; Soesterberg, The Netherlands as TNO 2022 R10427. The FAA has not edited or modified any content of this report.

This report is available at the Federal Aviation Administration William J. Hughes Technical Center's Full-Text Technical Reports page: actlibrary.tc.faa.gov in Adobe Acrobat portable document format (PDF).

Technical Report Documentation Page

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

1 Report No DOT/FAA/TC-22/10	2 Government Accession No	3 Recipient's Catalog No
4 Title and Subtitle Negative transfer of training: Simulator study into effects of going beyond alarms during stall recovery training		5 Report Date April 2022
		6 Performing Organization Code
7 Author(s) Landman, A. Mol, D., Emmerik, M. L. van, Groen, E. L.		8 Performing Organization Report No
9 Performing Organization Name and Address Netherlands Organization for Applied Scientific Research (TNO), Human Performance; Soesterberg, The Netherlands.		10 Work Unit No (TRAIS)
		11 Contract or Grant No
12 Sponsoring Agency Name and Address FAA Policy and Innovation Division Aircraft Certification Service Attn: Jeff Schroeder Bldg. N243 Moffett Field, CA 94035		13 Type of Report and Period Covered
		14 Sponsoring Agency Code

15 Supplementary Notes

This report was originally published by the Netherlands Organization for Applied Scientific Research (TNO), Human Performance; Soesterberg, The Netherlands as TNO 2022 R10427

16 Abstract

An often-used practice in pilot simulator training is to let pilots respond to a situation that has gone wrong. This may require the trainee to postpone their intervention, so that they can experience the cues and sensations signaling the critical situation, and have an opportunity to practice the appropriate response. However, some aviation-training experts have raised concerns about this practice. They argue that it might cause negative transfer of training when pilots suppress responses to alerts. These experts suggested that it is better to only train situations in which pilots immediately respond to situations going wrong.
This study tested whether these two approaches to stall recovery training differently affect the learning and performance of pilots. Commercial airline pilots (N = 40) practiced stall recoveries for 30 minutes in the Desdemona flight simulator, which has a representative aerodynamic model and an extended motion envelope suitable for stall cueing. One group of pilots recovered from training scenarios that started in pause at the moment of the required response (Freeze group), while another group manually flew the aircraft into the stalls, going beyond alarms (Dynamic group). Before and after the training, pilots performed a stall recovery test that was not surprising. Other actively flown post-tests included a surprising ground proximity warning, a surprising stall, and a false stall alarm. In an additional post-test, the pilots' ability to recognize stall cues was tested in six passive situations of true and false stalls, and pilots had to indicate whether the presented situation involved a stall or not. An extensive number of stall recovery performance parameters and other behavioral parameters was tested.

The training affected stall recovery positively for the whole group on nearly all parameters. When inspecting group differences, we unexpectedly found that the Dynamic group experienced more time pressure than the Freeze group when confronted with the surprising ground proximity warning and surprising stall. They also showed no significant decrease in experienced time pressure from Non-surprise pre-test to post-test while the Freeze group did. Trends suggest more aggressive pitch down responses to a surprising stall in the Freeze group, and better recognition of stall and non-stall situations by the Dynamic group. No other significant differences or trends in performance were found.

The increase in experienced time pressure may point to a potential hazard of too much contrast between self-paced stall events in the simulator and externally-paced surprising events in operational practice. The results further suggest that Dynamic training increased the pilots' ability to recognize stall cues.

17 Key Words Training, Transfer, Simulation, Pilot behavior, Alarms	18 Distribution Statement This document is available to the U.S. public through the National Technical Information Service (NTIS), Springfield, Virginia 22161. This document is also available from the Federal Aviation Administration William J. Hughes Technical Center at actlibrary.tc faa.gov.

19 Security Classif (of this report) Unclassified	20 Security Classif (of this page) Unclassified	21 No of Pages 49	19 Security Classif (of this report) Unclassified

Contents

1 Introduction 1

2 Method 3

2.1 Design 3

2.2 Participants 3

2.3 Apparatus 4

2.3.1 Stall warnings 5

2.3.2 Motion cueing 5

2.4 Procedure 5

2.5 Training scenarios and manipulation 6

2.6 Pre-test and post-tests 7

2.6.1 Non-surprise stall pre-test and post-test 7

2.6.2 Surprise ground proximity warning post-test 7

2.6.3 Surprise stall post-test 8

2.6.4 False stall post-test 8

2.6.5 Stall recognition test 8

2.7 Dependent measures 9

2.7.1 Definitions used for several dependent measures 9

2.7.2 Response time 10

2.7.3 Stall recovery parameters 10

2.7.4 Parameters specific to the EGPWS post-test 12

2.7.5 Parameters specific to the False stall post-test 12

2.7.6 Subjective responses 12

2.7.7 Stall recognition test 12

2.7.8 Manipulation checks 13

2.8 Data analysis 13

2.9 Hypotheses 14

3 Results 15

3.1 Performance example 15

3.2 Missing data 17

3.3 General learning effect 17

3.4 Response time 19

3.4.1 Pitch input response time 19

3.4.2 AP/AT disconnect response time 19

3.5 Stall recovery parameters 20

3.5.1 Summary 20

3.5.2 Input variability 20

3.5.3 Input aggressiveness 21

3.5.4 Order of control inputs 24

3.5.5 Safety margin during the recovery 24

3.5.6 Duration of the first stick shaker activation 26

3.5.7 Binary parameters 26

3.6 Parameters specific to the EGPWS post-test 27

3.7 Parameters specific to the False stall post-test 28

3.8 Subjective responses 28

3.9 Stall recognition test 29

3.10 Evaluation of the training 30

3.10.1 Subjective self-confidence in stall recovery 30

3.10.2 Interest and enjoyment during the training 31

3.10.3 Preference for training type 31

4 Discussion 31

4.1 General training effect 31

4.2 Group differences 32

4.2.1 Response time 32

4.2.2 Occurrences of loss of control 32

4.2.3 Stall recovery performance 33

4.2.4 Subjective responses 33

4.2.5 Stall recognition test 34

4.2.6 Pilots' evaluation of the training 34

4.2.7 Limitations 35

5 Conclusions 36

6 References 37

Figures

Figure 1. The PFD of situation 3 (left) and 6 (right) of the stall recognition test. 9
Figure 2. The alarms and events and the control inputs (lower plot) of participant 5 in the surprise stall post-test. 16
Figure 3. The V-N diagram of participant 5 in the surprise stall post-test. 17
Figure 4. Pitch input response time. 19
Figure 5. AP and AT disconnect response time. 20
Figure 6. Pitch input variability 21
Figure 7. Roll input variability. 21
Figure 8. Maximum pitch down input. 22
Figure 9. Minimum N_z reached. 23
Figure 10. Maximum roll input. 23
Figure 11. Time between the first pitch input and first roll input 24
Figure 12. Altitude loss. 25
Figure 13. Recovery duration. 25
Figure 14. Duration of the first stick shaker activation 26
Figure 15. Parameters specific to the EGPWS post-test 28
Figure 16. The subjective measures obtained after each scenario. 29
Figure 17. Difference between results. 30
Figure 18. Subjective self-confidence rating for stall recovery. 31

Tables

Table 1. Characteristics of the participants 4
Table 2. Outcomes of the pre-test-post-test comparisons. 18
Table 3. Binary parameters of stall recovery 27

Acronyms

Acronym	Definition
AP	Autopilot
AT	Autothrottle
CAS	Calibrated airspeed
EGPWS	Enhanced Ground Proximity Warning System
FAA	Federal Aviation Administration
FO	First officer
IE	Interest and Enjoyment subscale of the Intrinsic Motivation Inventory
IFR	Instrument flight rules
MSA	Minimum safe altitude
PLI	Pitch limit indicator
SO	Second officer
TNO	Netherlands Organization for Applied Scientific Research
TRI	Type rating instructor
TRE	Type rating examiner
UPRT	Upset prevention and recovery training

Executive summary

In a previous study, aviation experts raised a concern that deliberately letting events go wrong in the simulator may lead to incorrect responses in real situations. Others stated that this is sometimes necessary to let pilots observe the cues of undesirable or hazardous aircraft states. In the current study, we experimentally tested the hypothesis that going beyond alarms is detrimental to performance in the context of stall recovery training.

Commercial airline pilots (*N* = 40) practiced stall recovery for 30 minutes in a simulator designed for stall cueing. One group of pilots recovered from situations, which started in pause, at the moment of intervening (Freeze group). Another group manually flew the aircraft into the stalls, and subsequently recovered (Dynamic group). Before and after the training, pilots performed an non-surprising stall recovery test. Post-test responses to a surprising ground proximity alert, a surprising stall, and a false stall alarm were also examined, as well as pilots' recognition of stall and non-stall situations. We hypothesized that the Dynamic group would show signs of negative training in that they would respond more slowly to the alarms. However, the Freeze group was hypothesized to respond too hasty to false alerts, and to perform worse at stall recognition due to having had less opportunity to observe the stall cues.

The training affected stall recovery positively for the whole group. When inspecting group differences, we unexpectedly found that the Dynamic group experienced more time pressure in post-tests compared to the Freeze group. Trends further suggest more aggressive pitch down responses in the Freeze group, and better recognition of presented stall and non-stall situations by the Dynamic group.

The finding concerning time pressure suggests that training stall scenarios only in a self-paced manner is unadvisable, as it may create a contrast between the self-paced events in training and externally paced events in reality. While the Freeze situations were themselves not realistic, the majority of our participants found the required behavior in this training more realistic, and they indicated that such training may be useful especially if combined with surprise. The results further confirm that the Dynamic training improved the recognition of stall cues, which would specially be important in surprising situations. We therefore advise that training consist of both self-induced stalls for practicing cue recognition, and recovering from unknown and surprising situations for practicing quick sensemaking and automatic responses. For both types of training, dynamic instead of static situations should be used, as these are more realistic.

1 Introduction

Negative transfer of training occurs when performance in a new context or task becomes worse for those who received a particular training compared to those who did not (Alexander, Brunyé, Sidman, & Weil, 2005; Borgvall, Castor, Nählinder, Oskarsson, & Svensson, 2007; Burke, 1997; Woltz, Gardner, & Bell, 2000). In a previous study performed by TNO, training experts from aviation and other domains were asked about their concerns about training approaches that theoretically result in negative transfer of training (Pennings, Oprins, Schoevers, & Groen, 2019). One concern or question, voiced by various experts, was as follows: how to let situations deliberately go wrong, so that trainees can practice responses to hazardous situations. Because they say prevention is better than cure, some experts see a risk of negative transfer of training when trainees are instructed to deliberately go beyond signals or alarms, rather than taking precautionary measures to prevent the imminent dangerous situation from occurring. We use the word alarms for short, but these may also include interface alerts, interface warnings, or alarming cues outside of the interface. The main concern of these experts is that when trainees are instructed to suppress preventive actions, this may have repercussions for their mental task model. Their assumption is that the execution of normally improper actions in the simulator, even if this is knowingly done for educational reasons, increases the chance that pilots will respond similarly in operational practice. However, according to other experts, pilots can easily distinguish between actions performed for training purposes and actions that should be performed in reality. They feel that going beyond alarms is a valuable training exercise, as it allows the trainee to experience the cues and sensations that signal an unsafe situation.

In the aviation domain, this debate particularly applies to upset prevention and recovery training (UPRT). More specifically, the debate focuses on how the recovery from an aerodynamic stall should be practiced during simulator training. Following prominent accidents, such as Colgan Air flight 3407 and Air France flight 447, aviation authorities have implemented special requirements for UPRT, including hands-on application of the stall recovery procedure in the simulator (FAA, 2015). However, experts have different views on how to bring the simulated aircraft into the stall (Pennings, Oprins, Schoevers, & Groen, 2019). One approach is to let the trainee pilots fly the aircraft into the stall themselves, so that they learn to recognize pre-stall and stall cues and alerts. However, this approach has been criticized, as it requires the pilots to suppress, or delay their immediate response to the alarms. Alternatively, the instructor can fly the aircraft into the stall, while the trainee pilot is watching. The disadvantage of this approach is that it does not actively engage the trainee pilot. Another, more active approach would be to start the simulation at a more progressed state of stall, and have the pilot respond immediately to the alarms that become active as soon as the exercise starts. .This is more or less similar to the approach where the instructor (or confederate pilot) sets up an exercise, while the trainee pilot closes their eyes until he or she suddenly receives controls and has to recover. However, having a trainee pilot close their eyes before “handing-over” the aircraft may limit the development of situational awareness, and introduce confounding by the factor of surprise.

The objective of the current study is to investigate the effects of two different ways of bringing a simulated aircraft into aerodynamic stall in training: 1) pilots fly the aircraft into the stall themselves, and subsequently perform the recovery, and 2) the simulation starts from a freeze, with the aircraft already stalled when the pilots perform the recovery. To eliminate the confounding of surprise in the latter, the pilot is allowed to observe the frozen situation before starting. To our knowledge, no prior studies have investigated this issue. Although the current study focuses on stall recovery training for pilots, the results may have implications for other aspects of aviation training, or for training in other domains that involves exercises that deliberately bring the trainees beyond restrictions or alarms.

2 Method

2.1 Design

The experiment had a between-subject design. One group of pilots received an experimental training in which they were instructed to go beyond alerts and alarms in most of the scenarios, and let the situation develop dynamically (i.e., the Dynamic group). A second group received training with scenarios that started from a pause, or freeze, either in a progressed state of stall or with a negative speed trend, so that the alarm or alarming cues would occur immediately after the unfreeze. They were instructed to immediately respond when the scenario was unfrozen and the alarm sounded (i.e., the Freeze group). All pilots of both groups performed a pre-test before the experimental training, providing a baseline measure of some performance parameters. By administering the same test at the end of the experiment, a measure of the overall training effect for both groups was obtained.

2.2 Participants

Pilots (38 men and 2 women) with a current commercial pilot license were invited through a message on an online news bulletin of their company and through word-of-mouth. The exclusion criteria used were military flying experience, having an aerobatics rating, and having had a glider flying rating in the last 20 years. Pilots were divided into the Dynamic and Freeze group (both N = 20), which would receive different training manipulations (see 2.5). The groups were balanced on the characteristics listed in Table 1. Several factors possibly influencing stall recovery performance were considered: flying experience, type of aircraft flown, rank, and working or having worked as Type rating instructor or Type rating examiner (TRI/TRE). The number of flight hours in large and small aircraft were not normally distributed, with several pilots in both groups having much more experience than the mean. All pilots were employed at the same airline company. Of the pilots who were not currently flying Boeing B737 (i.e., most similar to the simulated cockpit and aircraft model) in the Dynamic group, six were flying B777 and four were flying Embraer E175/190. In the Freeze group, there were also eight pilots flying B777, two flying Embraer E175/190, and one flying Airbus A330. When asked to rate how rested they felt before the experiment started, both groups reported a median of 4 on a scale ranging from 1 (‘not at all’) to 5 (‘very well’) and a Mann-Whitney U test did not suggest a significant difference, *p[[* = 0.543.

Table 1. Characteristics of the participants

	Freeze	Dynamic
Age (mean years ± SD)	40.7 ± 9.5	40.2 ± 9.5
Work experience as pilot (mean years ± SD)	17.3 ± 2.1	16.2 ± 2.1
Flying experience medium/large twin-jet (mean hours ± SD)	8449 ± 4917	8406 ± 5068
Flying experience in smaller aircraft (mean hours ± SD)	153 ± 263	180 ± 450
Rank (Captains/FOs/SOs*)	8/10/2	8/11/1
Currently flying B737	9	10
Previously flown B737	13	12
TRI/TRE past or present	5	6
Gender (M/F)	19/1	19/1

* SO: Second Officer. The third in line of command, a rank sometimes used on international or long haul flights.

2.3 Apparatus

The experiment was performed in the Desdemona flight simulator (AMST Systemtechnik), located at TNO Soesterberg. Desdemona features a gimbaled system that allows for continuous rotations around three axes. This system can be moved within a stroke of two meters vertically on a heave axis, and eight meters laterally on a horizontal track. The track itself can rotate around a planetary axis to induce g-loads when the simulator cockpit is positioned off-center (analogous to a human centrifuge). The cockpit mockup was styled after the Boeing 737NG, and included the left-side seat, primary flight display with pitch limit indicator (PLI), navigation display (not used), engine indications, crew-alerting system (not used), and a partial mode control with autopilot mode controls. There was no overhead panel or flight management system. Controls consisted of a yoke (pitch and roll), rudder pedals with rudder limiter, throttles and a stabilizer with electric trim (tabs), and silent trim wheels. The yoke had control loading on pitch only. Flaps and speed brakes were not used. The aerodynamic model used in the experiment featured an extended aerodynamic envelope of medium-sized modern transport category aircraft (e.g., Boeing 737NG, Airbus A321, Tu-204) into high angles of attack (Groen, et al., 2012). The model includes aerodynamic phenomena like buffeting, longitudinal and lateral instabilities, dynamic hysteresis, and degradation of control response (Goman & Khrabrov, 1994).

2.3.1 Stall warnings

The alerts, alarms and other cues of (approaching to) stall were, in order of time to stall if speed is decreasing at 1 g:

- Continuous decrease of indicated speed;
- Continuous increase in angle of attack (alpha, i.e., the angle between the relative wind and the wing);
- Auditory speed low warning and blinking of the speed box (both at 70% of the amber band);
- Appearance of the PLI;
- Stick shaker and audio stall alarm. A stall audio alarm is normally not featured in a Boeing 737NG, after which the systems were largely modeled. However, it was still included in the simulation to make it a more generic model for pilots who did not fly Boeing;
- Stall buffet (motion cues and audio);
- Lateral instability (roll off) and sloppy controls.

2.3.2 Motion cueing

The motion platform was used in an extended hexapod mode, without the centrifuge capability (i.e., no g-loads). As described in detail by (Nooij, Pretto, Oberfeld, Hecht, & Bulthoff, 2017), the extended hexapod mode is based on the classical washout scheme used in training simulators, rendering onset cues by high-pass filtering of aircraft accelerations, and sustained accelerations by tilt-coordination. However, employing Desdemona’s larger motion envelope, some cues are amplified for the purpose of stall recovery exercises. For example, aerodynamic buffeting is reproduced by the heave axis, and a sensation of (un)loading is amplified by vertical prepositioning of the simulator cabin by almost 1 m. The sensation of (un)loading is further improved by pitching the cabin to a maximum of 30 degrees (nose-up for loading and 30 degree nose down for unloading), similar to tilt coordination. The effective motion limits in the extended hexapod mode are: 155 deg/s (45 deg/s^2) for the central yaw drive; 2.2 m/s (0.5g) for heave; 90 deg/s (120 deg/s^2) for pitch, roll, and yaw rotations.

2.4 Procedure

The participants arrived in couples and received the briefing together. In the briefing, they first received a 15-min general instruction about the experiment. It was explained that the experiment aimed to investigate and compare different methods of stall recovery training. They were explicitly warned that the training they would receive was experimental, that it could therefore be suboptimal, and that they should normally rely on their own company's training. Their memory about the stall recovery template was refreshed, and the aspects from this template, which would be measured, were briefed. Furthermore, the pilots were told to prioritize the safety of the recovery, i.e., intervening in time and unloading fluently without risking secondary stalls, or overspeed. Overspeed was said to be less important than excessive g (i.e., -1 to 2.5 g). Pilots were also instructed that they should respond to any situation in the scenarios as if it actually happened, except when the experimenter would ask them explicitly to act in a special manner. The pilots were encouraged to give callouts to remain close to the normal procedure.

Because the exercises were performed with conventional motion cueing (and no centrifuge), it was not expected that the pilots would experience any simulator sickness. Still, in case of any discomfort they were allowed to take a break. Finally, the pilots filled in the pre-experiment questionnaires, and provided their informed consent before receiving a 15-minutes briefing on the aerodynamic model, cockpit mockup and the display indications and sounds that were used in the experiment.

After the briefing, the pilots each performed the simulator session (ca. 60 min) individually. Instructions in the simulator were given by an experienced UPRT instructor from the Desdemona BV. The simulator session consisted of the following elements:

- Familiarization with the controls at low (5,000 ft.) and high altitude (38,000 ft.);
- Pre-test (Non-surprise stall);
- Experimental training scenarios;
- Three surprising post-test scenarios;
- Stall recognition test;
- Post-test (same as pre-test).

In total, the experiment lasted between 1.5 and 2 hours. All data were collected between June 23rd and July 6th, 2021.

2.5 Training scenarios and manipulation

The training was designed to give pilots a refresher training in stall recovery. All pilots recovered (approach to) level-flight stall situations from three levels of (approach to) stall severity:

1. The aural speed low alert with coinciding flickering of the speed box (70 % into the amber band). Recovering from this situation did not require executing every step of the stall recovery template, as it could be solved by adding thrust, or solved quickly by also slightly reducing pitch.

2. The moment of stick shaker activation.
3. The moment of roll off due to increased instability, with the roll angle exceeding 45 degrees.

Recovering from these stadia was always repeated once, with the order: 1-1-2-2-3-3. This whole set of practice trials was first performed at low altitude (i.e., 10,000 ft), and then again at high altitude (i.e., 38,000 ft). Stall recovery at high altitude is more difficult, since the controls are more sensitive due to the higher speed and the smaller margin between under- and overspeed.

The manner of how the stall was introduced to the pilots in these training scenarios was manipulated. For the Dynamic group the training scenarios always started with the autopilot connected and holding the current altitude (Altitude Hold mode). Pilots were then instructed to set the throttle to idle to slow down the aircraft until they were confronted with one of the cues mentioned above, at which they were instructed to recover.

For the Freeze group the same scenarios were used, but these were presented in a paused setting, so that the alarm would present itself as soon as the scenario was “unfrozen”. This ensured that pilots of this group were not exposed to the progression of the alarms. Prior to the scenario, pilots were informed about the critical situation at the scenario start, they were given time to check the settings, and were instructed to react to the alarms as soon as the scenario started.

2.6 Pre-test and post-tests

2.6.1 Non-surprise stall pre-test and post-test

Before the training session, pilots performed a pre-test to obtain a measure of their general stall recovery performance, as well as a baseline measure of the subjective measures. This scenario was performed at 38,000 ft, and consisted of a level stall that was announced by the instructor. Pilots were flying level with autopilot and autothrottle connected until a sudden extreme tailwind brought the aircraft quickly into a stall. They recovered as soon as the stick shaker (automatically) activated, which activated four seconds after tailwind onset, and three seconds after the speed low alert. The same scenario was used in the post-test at the end of the training and testing.

2.6.2 Surprise ground proximity warning post-test

Following the training session, the first test scenario required pilots to respond to the Enhanced Ground Proximity Warning Systems (EGPWS), indicating that terrain was closing (i.e., aural: “Terrain, Pull up” and visual alert). This was achieved by letting pilots fly in instrument flight rules (IFR) on autopilot heading east at 6,000 ft. near a mountain range in northern Italy. They were then instructed to select heading north, causing the aircraft to fly closely over a mountain range, which triggered the EGPWS. The appropriate response would be to disconnect autopilot and autothrottle, pitch up to 20 degrees, and apply maximum thrust. If terrain remained a threat, pitch should be increased to the PLI, stick shaker, or initial buffet. Speed brakes should be retracted. After the pilot's initial response, the instructor informed that the minimal safe altitude (MSA) in this area was 10,000 ft, which gave the pilot a target to climb towards.

2.6.3 Surprise stall post-test

The second test scenario was a surprising stall at low altitude. The scenario started in an autopilot-flown climb at 2500 ft, with the autopilot initiating a turn. When in the turn, instructions were given to change the vertical speed for the autopilot so that the margin to under-speed would increase. When pilots were making the change, and looked away from the instruments, a wind gust brought the aircraft quickly into a stall.

2.6.4 False stall post-test

The third test scenario had the same setup as the Surprise stall post-test. At the same moment in the turn, when the adjustment of settings was required, instead of a wind gust, now the stick shaker and auditory stall alarm were triggered due to a simulated technical malfunction. These remained activated until the scenario ended, which was when participants maintained level flight, or a climb. This scenario was included to test whether the Freeze group would tend to follow the alarms longer due to taught behavior, or less familiarity with the aerodynamic stall cues than the other group.

2.6.5 Stall recognition test

After the post-test, pilots performed a stall recognition test. This was meant to test whether the groups differed in their ability to recognize stall cues. We showed them six situations, which were announced to be either static or dynamic, and we required them to indicate as quickly as possible within 20 seconds whether they thought the situation was a stall or not. They pressed the autopilot disconnect button to indicate stall, and a comm button to indicate no stall. They were explicitly instructed that the task was not to indicate whether they would initiate a recovery just in case or not. All situations were in IFR. Pilots were told that the speed tape was covered, which was achieved by disabling it. The five situations are listed below:

1. No stall: a false stick shaker at level flight. This situation was dynamic;
2. No stall: a high-speed buffet at high altitude. This situation was dynamic;

3. Stall: An overbank situation with stall (see Figure 1, left). This situation was static;
4. No stall: a false stick shaker in climb. This situation was dynamic;
5. Stall: level flight stall with extremely high angle of attack, so that the flight path vector was not visible. This situation was static;
6. No stall: an overbank situation with no stall (see Figure 1, right). This situation was static.

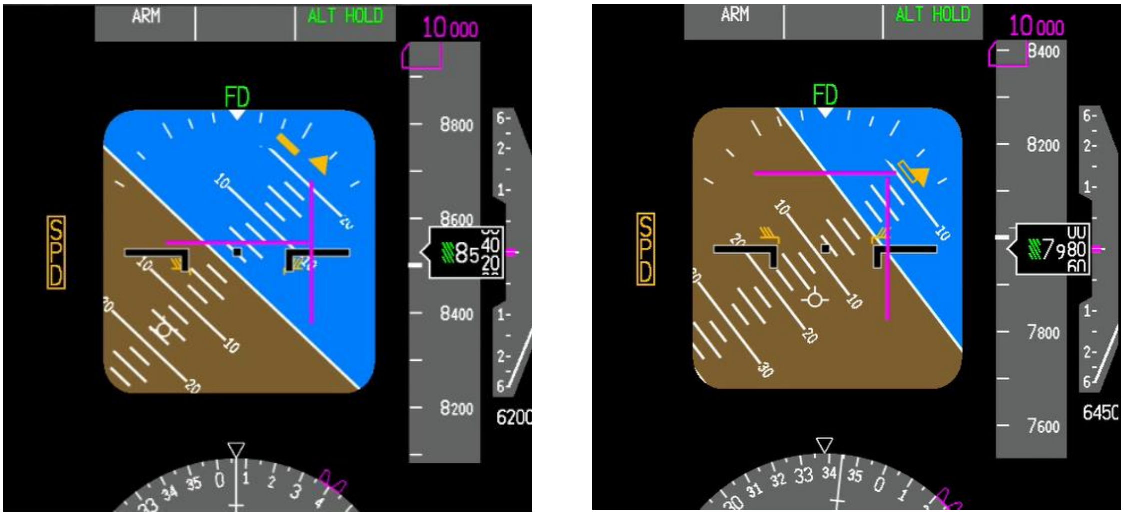

Figure 1. The PFD of situation 3 (left) and 6 (right) of the stall recognition test.

2.7 Dependent measures

2.7.1 Definitions used for several dependent measures

For the computation of some of the dependent measures, we defined a trigger cue, to which the pilot could respond. These were:

- The activation of the stick shaker (Non-surprise pre-test and post-test, False stall post-test).
- The start of EGPWS alert (EGPWS post-test).
- The start of the wind gust (Surprise stall post-test).

The end of the recovery was defined as the moment the descent stopped, and the end of the response to the EGPWS was defined as the moment the minimum save altitude was reached.

2.7.2 Response time

2.7.2.1 Pitch input response time

Pitch input response time is the time between the trigger and the start of the first pitch input in the required direction. This was pitch down for the Non-surprise stall pre-test and post-test and Surprise stall post-test, and pitch up for the EGPWS post-test. The measure was not obtained in the False stall post-test.

The start of the first pitch input was defined as the first time following the trigger at which the cumulative sum of the pitch input diverged more than 10 standard deviations (*SD*s) from the mean pitch input. The mean and *SD* were obtained from the five seconds before the trigger.

2.7.2.2 Autopilot and autothrottle disconnect

Time to autopilot (AP) and autothrottle (AT) disconnect was the time from the trigger until the AP and AT were both disconnected. This was obtained in all scenarios except the False stall post-test.

2.7.3 Stall recovery parameters

In the Non-surprise stall pre-test and post-test, and in the Surprise stall post-test, parameters characterizing stall recovery behavior were obtained.

2.7.3.1 Input variability

Input variability was obtained for pitch and roll inputs during the time between the trigger and the end of the recovery, which was defined as the end of the descent. We operationalized the variability of pitch inputs by taking the root mean square (RMS) of the pitch input rate. The rate was used, because a consistent non-zero pitch input is required during the recovery. We operationalized the variability of roll inputs by taking the root mean square (RMS) of the roll inputs. Low variability indicates a fluent recovery.

2.7.3.2 Input aggressiveness

As measures of input aggressiveness, the maximum pitch input, minimum N_z (resulting from pitch down input magnitude and duration), and the maximum roll input were obtained between the trigger and the end of the recovery.

2.7.3.3 Order of control inputs

The pilot’s performance in correctly executing pitch-down control inputs before roll control inputs was defined as the time between the first moment where pitch-down inputs exceeded

25%, and the first moment where roll inputs exceeded 25%. The time will be negative if this roll input occurs before the pitch-down input, which would indicate inadequate performance. In the Non-surprise pre-test and post-test, there may be no roll to correct, meaning that inputs are likely much smaller.

2.7.3.4 Safety margin

To test how much safety margin was used by the pilots, the altitude loss and the time between the trigger and the end of the recovery were obtained. Pilots experiencing multiple stick shaker events were excluded for these measures. Therefore, less altitude loss and shorter recovery duration indicates the use of a smaller safety margin. Performances with multiple stick shaker events were excluded from these parameters. These were reported separately as a categorical measure.

2.7.3.5 Duration of first stick shaker activation

We regard the duration of the first stick shaker activation as an indicator of timely unloading. The occurrence of multiple stick shaker activations was reported as a categorical measure.

1. We measured the time between the trigger and the disconnection of both the autopilot and AT as an indicator of pilots forgetting to disconnect these at the start of the recovery.

2. Pilots were instructed to call out their speed brake check during the recovery. Actual manual speed brake check could not be logged, but callouts were used as an indication of speed brake checking.

3. We checked if participants used rudder during the recovery as a binary measure. A significant input was defined in the same manner as a significant pitch input, namely when the cumulative sum diverged more than 10 *SD*s from the mean input.

2.7.3.6 Binary parameters

The binary parameters that were obtained were all coded in the direction so that higher values suggest less optimal performance. All parameters were collected between the trigger and the end of the recovery. The parameters were:

- Not disconnecting the AT;
- Not using trim;
- No verbal speed brake check;
- Using rudder;
- Experiencing multiple stick shaker activations;
- Keeping the AP disconnect aural on by not pressing the AP disconnect button twice.

2.7.4 Parameters specific to the EGPWS post-test

In the EGPWS post-test, we first measured the time until the MSA was reached. After the first reaction, the instructor told the pilots that the MSA was 10,000 ft. To exclude the time that was used for leveling off, the time needed to reach an altitude of 9,000 ft was used for this measure. The starting altitude was 6,000 ft.

To measure the pilots’ aggressiveness in pitch-up behavior, we recorded the maximum pitch angle pilots reached during the scenario, as well as the minimum calibrated airspeed (CAS).

2.7.5 Parameters specific to the False stall post-test

In the False stall post-test, we tested whether or not pilots responded to the false stall alarm by pitching down. For those who responded, we also measured the amount of time and altitude loss before the pilots stopped attempting to recover, and leveled off.

2.7.6 Subjective responses

Subjective ratings were obtained immediately after the scenario had ended. The pilots gave their rating verbally, referring to a scale that was presented on a sheet of paper attached to the simulator cabin next to their seat. The following ratings were obtained:

- Subjective anxiety, measured with the anxiety scale (Houtman & Bakker, 1989). This is a 10 cm horizontal scale, which we made to range from 0 (no anxiety) to 100 (maximum anxiety) to allow for verbal indications;
- Subjective time pressure, measured with the Temporal demand subscale of the NASA-TLX (Hart & Staveland, 1988). Pilots indicated on a scale from 0-100 how much “time pressure they felt due to the pace at which the tasks or task elements occurred.” Since our goal was not to measure overall task load, we did not use other subscales of the NASA-TLX;
- Subjective time taken for diagnosis. Using a custom non-validated scale with the same format as the NASA-TLX, we asked pilots to indicate to what extent they felt they acted immediately (0) or took time for diagnosis (100).

2.7.7 Stall recognition test

For the situations presented in the stall recognition test, the proportion of correctly recognized situations (percentage) and the average response time were calculated. We took their final decision as the definite answer. Finally, pilots rated their certainty in their decision after each decision on a scale from 0-100%. We report the average of this certainty rating.

2.7.8 Manipulation checks

To test whether the training had an effect on pilots' stall recovery, performance on the Non-surprise stall pre-test and post-test was compared.

To check whether participants experienced the events in the test scenarios as surprising, we measured subjective surprise on a scale ranging from zero (minimum surprise) to 100 (maximum surprise). The format is the same as the anxiety scale, but it is not validated for surprise. There exists no such validated scale yet. Nevertheless, we can compare scores with those of previous experiments to obtain a rough estimate of participants' surprise. Surprise scores above, or around the midpoint of the scale should indicate that a large proportion of participants did not see the event coming.

A second manipulation check was the Interest and Enjoyment (IE) subscale of the Intrinsic Motivation Inventory (Ryan, 1982), which consists of nine questions regarding how interesting or boring pilots found the training session. This was to control for potential group differences in perception during the training.

2.8 Data analysis

Data analysis will start with an analysis of a general learning effect for the whole group from Non-surprise stall pre-test to post-test. This is done to validate our chosen parameters, which are expected to show an increase in performance and risk taking. However, if performance on certain parameters does not change for the whole group, the groups can still respond differently to the training, so these parameters are not excluded from further analysis. The pre-test-post-test comparison is also used as a check whether our training functioned as expected, in that it increased pilot performance with the presented aerodynamic model.

When testing the two groups' responses to the training, outcomes will be clustered for different scenarios as much as possible to make optimal use of test-to-test correlations in responses, and to reduce Type-I errors that can be caused by too large a number of outcomes.

Only Group × Test interaction effects are reported, as main effects of Group could be caused by unbalanced groups, and main effects of Test are to be expected due to differences between scenarios. The Non-surprise stall pre-test is always used as a baseline for comparisons when assessing effects of training. If a significant interaction effect is found, post-hoc comparisons are performed in which each scenario is compared with the pre-test only.

- Response time data were analyzed for all scenarios which required a response, with 2 × 4, Group (Dynamic, Freeze) × Test (pre-test, post-test, EGPWS, Surprise stall) mixed-model ANOVAs;
- The subjective responses are analyzed for all scenarios, with 2 × 5, Group (Dynamic, Freeze) × Test (pre-test, post-test, EGPWS, Surprise stall, False stall) mixed-model ANOVAs;
- Stall recovery performance is analyzed for the stall recovery scenarios, with a 2 × 3, Group (Dynamic, Freeze) × Test (pre-test, post-test, Surprise stall) mixed-model ANOVA.

Other scenario-specific outcomes are compared between groups using independent-samples *t* tests or Mann-Whitney *U* tests for non-normally distributed or ordinal data.

Binary measures were compared separately per group between each post-tests and the pre-test using McNemar's test for paired-samples of binary data.

2.9 Hypotheses

For response time (pitch, AP and AT disconnect, and first stick shaker duration), we hypothesize slower reactions in the Dynamic group in all post-test scenarios. We expect that this group has learned in the training to observe before they react, whereas the Freeze group has learned to react immediately. Slower reactions may affect scenario outcomes positively and negatively, depending on the presence of time pressure and the need to react quickly. The latter is the case in the EGPWS post-test and Surprise stall post-test. In the False stall scenario, slower reactions may lead to more positive scenario outcomes.

For the stall recovery parameters, we expect that the pilots in the Freeze group will demonstrate higher input variability, higher input aggressiveness, use of a larger safety margin (more altitude loss and longer recovery duration), and more binary indicators of suboptimal performance. This is due to either hasty responses (see above), having had less opportunity during the training to experience the stall aerodynamics of the aircraft, and possibly being more stressed or experiencing more time pressure during the test scenarios because of this.

It follows that we expect the subjective parameters to indicate that the Freeze group takes less time for diagnosis, and experiences more time pressure and anxiety during the post-test scenarios.

For the stall recognition test, we expect that the Dynamic group will be more successful and faster, or more certain in recognizing the presented situations, or both, as they have had the opportunity to focus more on stall cues and signals during the training.

3 Results

3.1 Performance example

Figure 2 shows an example of the alarms and events (top), and a pilot's control inputs (bottom) in the Surprise stall post-test. Black circle markers in the plots were used to visually inspect whether the script adequately determined, for example, changes in control inputs. In this example, the pilot responded to the stall alarms by unloading slightly by pitching down, however these pitch down inputs did not yet reach the predetermined threshold of 20% of the pitch input range to be registered as the true moment of unloading. The first registered response were roll inputs, as the pilot rolled wings level using nearly 100% roll inputs while setting thrust to maximum. Pitch down inputs crossed 20% somewhat later, which led to a time of -1.9 seconds on the variable of time between roll inputs and pitch down inputs. The sign indicates a suboptimal ordering of these actions. The roll inputs show a high variability resulting in a relatively high RMS of roll inputs of 35% (mean of all pilots = 25%). The stick shaker was active for 5.2 seconds, which was the longest time of all participants in this scenario. This indicates that sufficient unloading was late and slow.

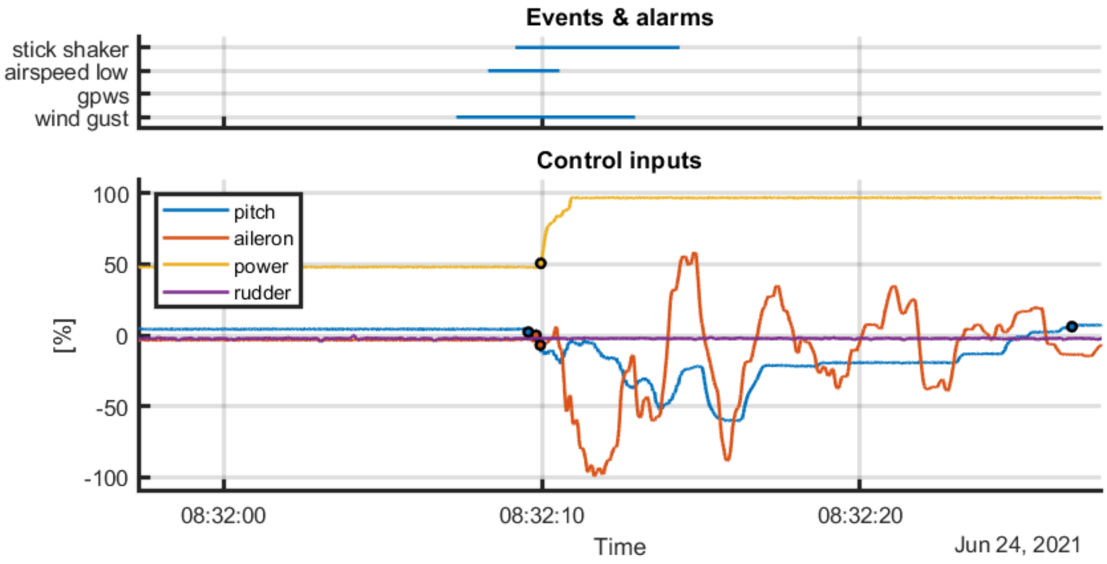

Figure 2. The alarms and events and the control inputs (lower plot) of participant 5 in the surprise stall post-test.

The V-n diagram corresponding to the same example is shown in Figure 3. The wind gust is given while the pilot flies around 220 knots, after which CAS is quickly reduced due to the wind gust. The stick shaker activation causes the pilot to unload, but there is a brief moment of reloading too early around 1 *g*, causing prolonged stick shaker activation. N_z can be seen to vary after the stick shaker activation. This variability is captured by our measure of RMS of the pitch rate. The lowest N_z reached was around 0.8 g, which is captured by our dependent measure of minimum N_z.

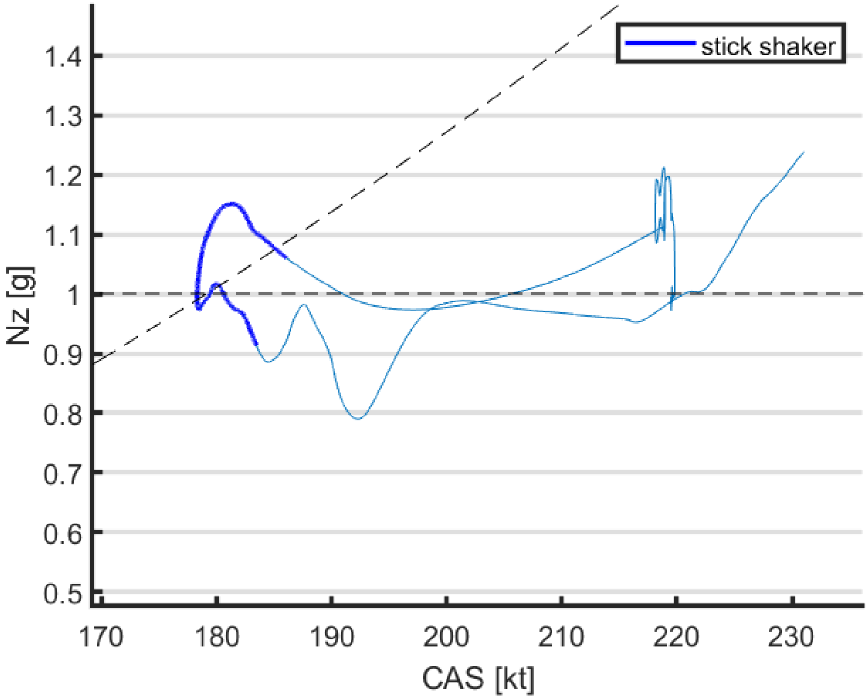

Figure 3. The V-N diagram of participant 5 in the surprise stall post-test.

3.2 Missing data

Data of one participant in the Freeze group were lost for the Surprise stall test due to an error in setting up the scenario. Three other participants (two Dynamic, one Freeze) responded too early, before stick shaker activation in the Non-surprise stall post-test, leading to loss of the reaction time and duration of stick shaker activation measures in this test.

The continuous parameters of one participant in the Surprise stall test, and the response time and stick shaker activation in the post-test, were replaced with the group means of these scenarios. This was done so that not all data in the ANOVAs would be list-wise excluded from the analyses. This means that in the 2 × 3 ANOVAs of these parameters, 3.3 % (4/120 values) of the data was imputed, and 2.5% (4/160) in the 2 × 4 ANOVAs.

3.3 General learning effect

Table 2 shows results on all measures of the Non-surprise stall pre-test and post-test. Data were used to conduct paired-samples *t* tests. The shaded cells contain the measures that were significantly affected by training, in a direction that indicated improved performance. Parameters for which a significant improvement was not detected (non-shaded cells) were as follows: pitch-down reaction time, minimum N_z, stick shaker duration, and time to AP/AT disconnect.

Correlations between pre-test and post-test performance were very low for pitch-down reaction time, recovery duration and AP/AT disconnect, indicating that performance on these measures was not consistent within the same participant. An effect size of 0.2 is considered small, 0.5 is considered medium, and 0.8 is considered large (Cohen, 1992).

Table 2. Outcomes of the pre-test-post-test comparisons.

	Pre-test Mean (*SD*)	Post-test Mean (*SD*)	*t*	*p*	*df*	*r*	Effect size (Cronbach's alpha)
Pitch response time (s)	1.3 (0.7)	1.1 (0.9)	0.93	0.357	39	0.046	0.15
Time to AP & AT disconnect (s)	1.8 (2.4)	1.9 (1.4)	0.33	0.743	39	-0.036	0.32
RMS of pitch input rate (%/s)	0.0283 (0.0054)	0.0256 (0.0035)	4.34	< 0.001	39	0.689	0.69
RMS of roll inputs (%)	22.7 (7.8)	15.2 (4.5)	6.07	< 0.001	39	0.275	0.96
Maximum pitch down input (%)	44.2 (15.7)	37.2 (15.8)	2.72	0.01	39	0.474	0.43
Minimum N_z (g)	0.42 (0.18)	0.46 (0.15)	1.22	0.229	39	0.414	0.19
Maximum roll input (%)	62.3 (16.6)	43.1 (12.9)	7.00	< 0.001	39	0.325	1.11
Time between pitch and roll (s)	0.4 (1.3)	2.5 (4.2)	3.27	0.002	39	0.070	0.49
Altitude loss (ft)	3232 (893)	2842 (716)	3.10	0.004	39	0.531	0.49
Recovery duration (s)	40.0 (5.5)	35.6 (3.2)	4.61	< 0.001	39	0.099	0.68
Stick shaker duration (s)	2.8 (0.8)	2.5 (0.5)	2.05	0.051	39	0.147	0.28

Of the measured binary parameters (see section Binary parameters for an overview), McNemar’s test showed only a trend towards less secondary stick shaker activations in the post-test, *p* = 0.063.

3.4 Response time

Summary: we did not find a significant effect of the training manipulation (Dynamic versus Freeze) on response time parameters.

3.4.1 Pitch input response time

The 2 × 4 mixed-model ANOVA revealed no significant interaction effect, *F*(3,114) = 0.56, *p* = 0.645 (Figure 4). One outlier in the Freeze group did not respond to the first trigger of the EGPWS alert. This participant responded after 36 seconds, when a second mountain ridge was reached and the EGPWS alert activated a second time. When removing this outlier, there was still no significant interaction effect on reaction time, *F*(3,111) = 0.41, *p* = 0.747.

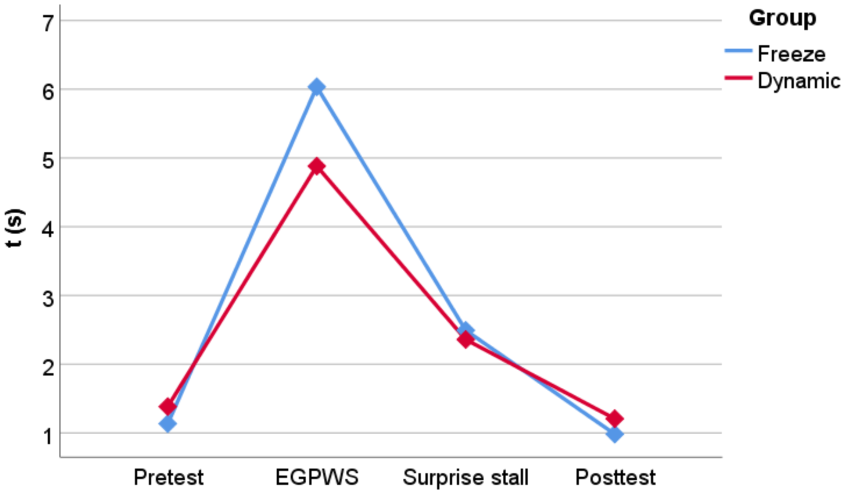

Figure 4. Pitch input response time.

3.4.2 AP/AT disconnect response time

For the duration until the AT and AP were both disconnected, the 2 × 4 mixed-model ANOVA showed no significant 2 x 4 interaction effect, *F*(3,114) = 0.57, *p* = 0.636 (Figure 5).

Two participants (Freeze group) did not turn off the AT in the Non-surprise stall pre-test, and three participants (Freeze group) did not do so in the Surprise stall post-test. For these participants, only the duration until AP was disconnected was used.

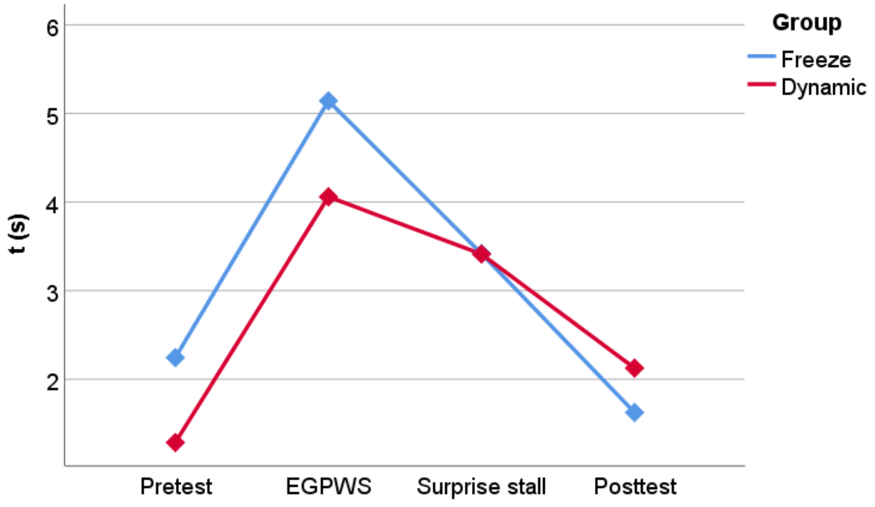

Figure 5. AP and AT disconnect response time.

3.5 Stall recovery parameters

3.5.1 Summary

There were some marginally significant effects, which suggested the Freeze training led to more aggressive pitch down inputs.

3.5.2 Input variability

For the pitch input variability (RMS of the differentiated pitch inputs), the 2 × 3 mixed-model ANOVA showed no significant interaction effect $F(2,76) = 0.11$, $p = 0.900$ (Figure 6).

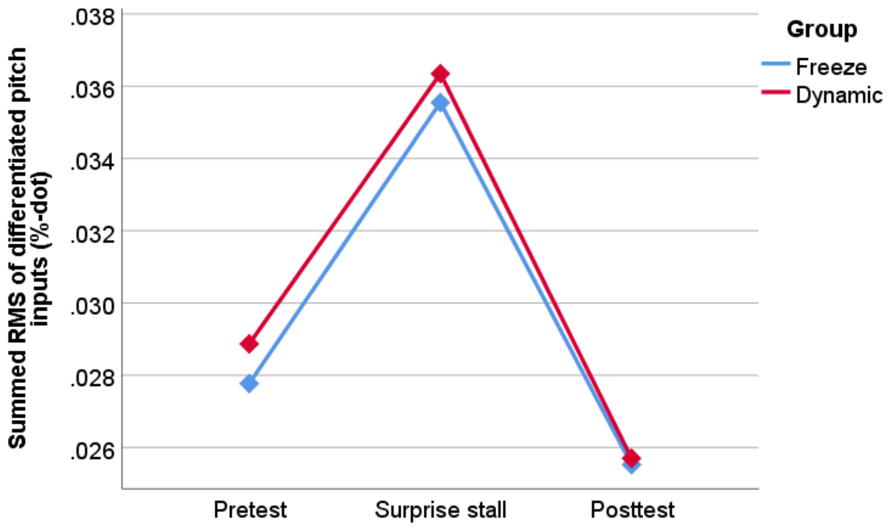

Figure 6. Pitch input variability

For the RMS of roll inputs, the 2 × 3 mixed-model ANOVA showed no significant interaction effect *F*(2,76) = 0.84, *p* = 0.436 (Figure 7).

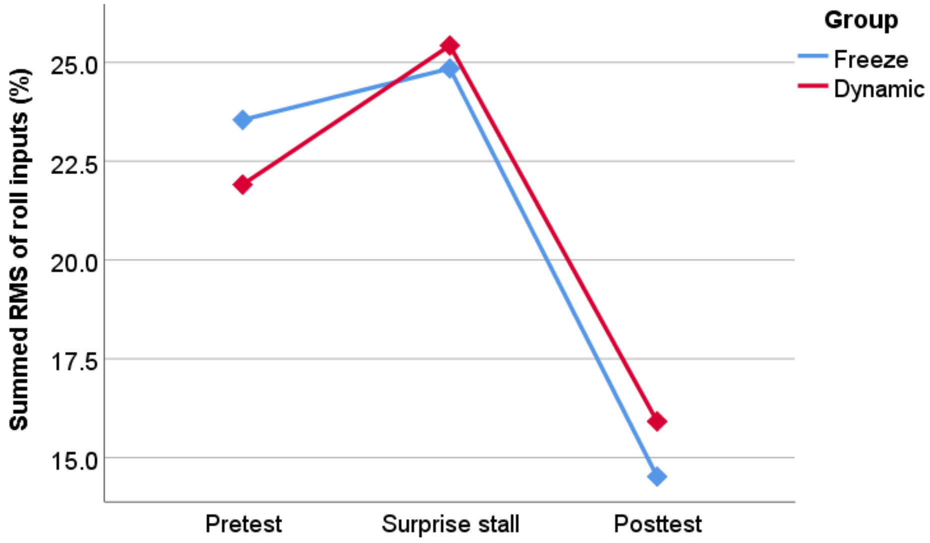

Figure 7. Roll input variability.

3.5.3 Input aggressiveness

For the maximum pitch down input, the 2 × 3 mixed-model ANOVA showed a marginally significant interaction effect indicating differences between the groups' response to the training, *F*(2,76) = 2.64, *p* = 0.078 (Figure 8). Post-hoc comparisons showed that both groups pitched down significantly stronger in the Surprise stall post-test compared to the pre-test. However the Freeze group pitched down significantly stronger than the Dynamic group in the Surprise stall post-test, Δ = 12%, $p = 0.017$. The Dynamic group also pitched down significantly less in the post-test compared to the pre-test, Δ = 9.4%, $p = 0.014$, whereas the Freeze group showed no significant difference, $p = 0.216$.

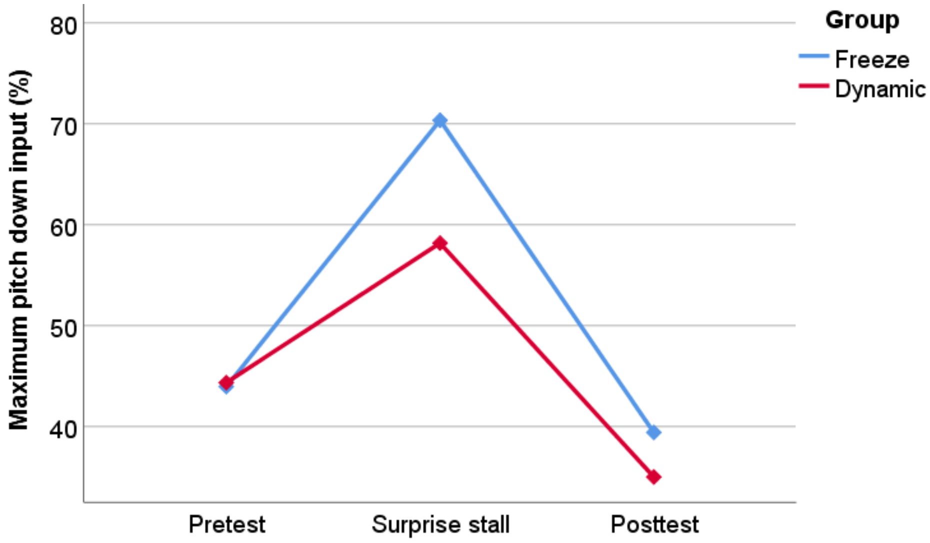

Figure 8. Maximum pitch down input.

For the minimum N_z reached, the 2 × 3 mixed-model ANOVA showed a marginally significant interaction effect indicating differences between the groups' response to the training, $F(2,76) = 2.96$, $p = 0.058$ (Figure 9). Post-hoc comparisons showed that the Freeze group reached a lower N_z in the post-test than in the pre-test, Δ = 0.093 g, $p = 0.049$, while the Dynamic group did not, $p = 0.261$. The Freeze group also reached a marginally significant lower N_z in the Surprise stall post-test than the Dynamic group, Δ = $p = 0.070$, which was not present in the pre-test, $p = 0.647$.

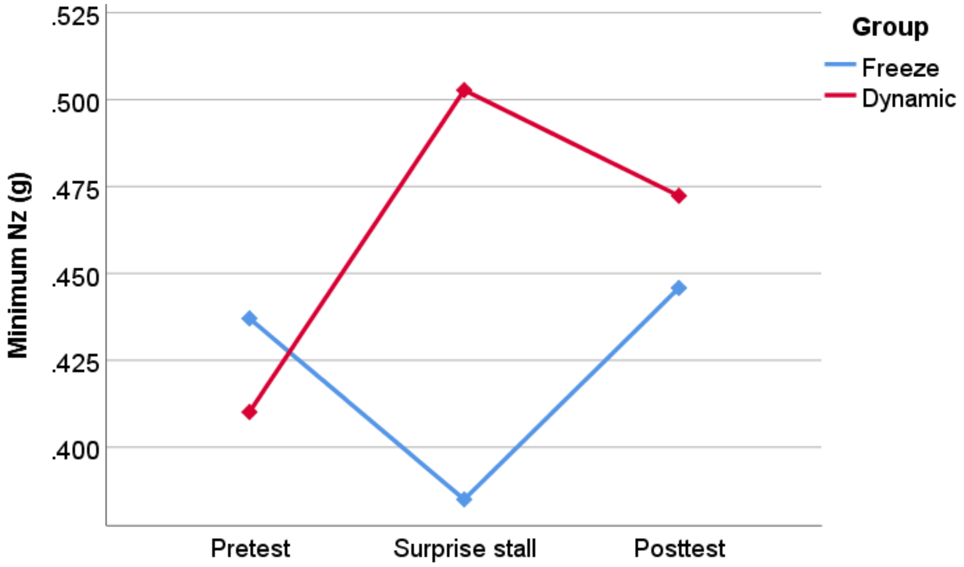

Figure 9. Minimum N_z reached.

For the maximum roll input, the 2 × 3 mixed-model ANOVA showed no significant interaction effect $F(2,76) = 0.20$, $p = 0.816$ (Figure 10).

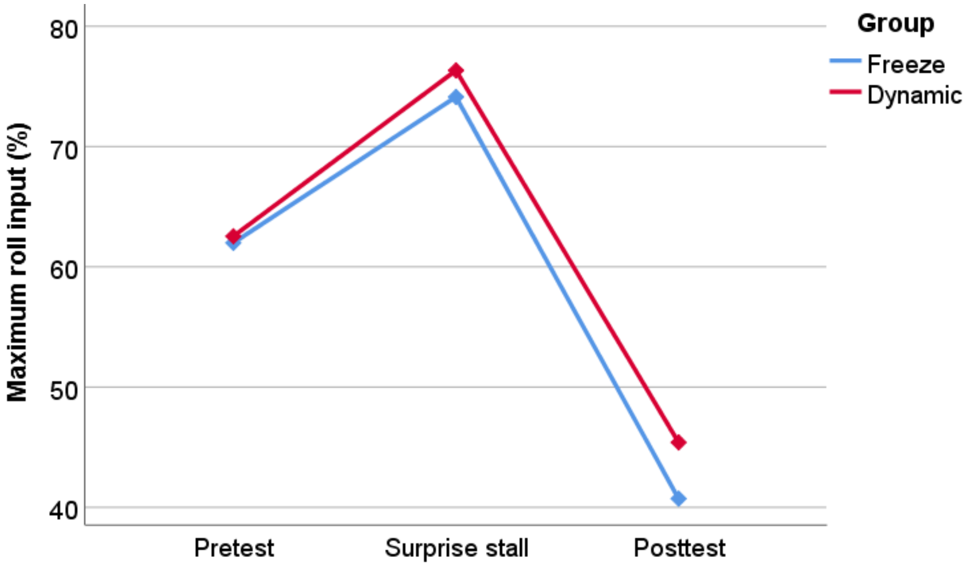

Figure 10. Maximum roll input.

3.5.4 Order of control inputs

Two pilots in the Non-surprise pre-test did not reach 25% pitch input. We scanned the data for lower present pitch inputs with decrements of 5% and substituted missing values with the first present value. The same was done for the pitch inputs of nine pilots and roll inputs of three pilots in the Non-surprise post-test, and for the roll inputs of one pilot in the Surprise stall post-test.

For the time between pitch and roll inputs, the 2 × 3 mixed-model ANOVA showed no significant interaction effect *F*(2,76) = 0.44, *p* = 0.645 (Figure 11).

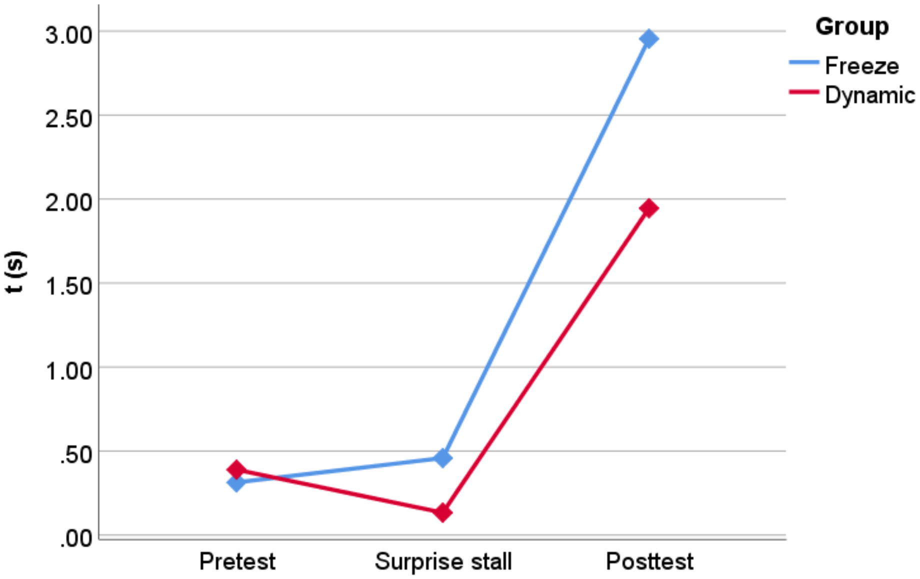

Figure 11. Time between the first pitch input and first roll input

3.5.5 Safety margin during the recovery

The secondary stick-shaker activation events were about equally distributed over the groups in each stall recovery test (Table 3). No significant differences were detected between the pre-test and any of the post-tests.

For altitude loss of those who experienced no secondary stick shaker events, the 2 × 3 mixed-model ANOVA showed no significant interaction effect *F*(2,62) = 1.06, *p* = 0.351. (Figure 12)

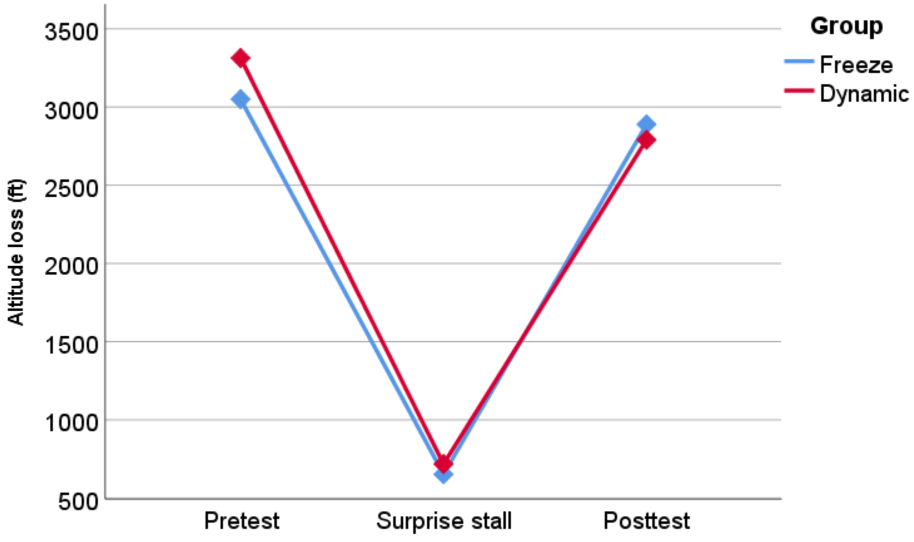

Figure 12. Altitude loss.

For recovery duration of those not experiencing secondary stick shaker events, the 2 × 3 mixed-model ANOVA showed no significant interaction effect $F(2,62) = 0.91$, $p = 0.409$. (Figure 13)

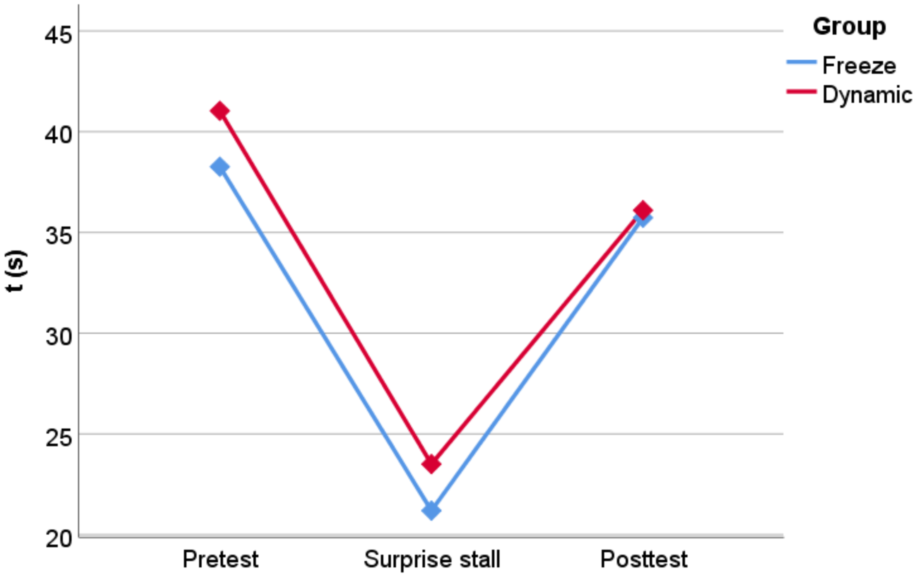

Figure 13. Recovery duration.

3.5.6 Duration of the first stick shaker activation

For the duration of the first stick shaker activation, the 2 × 3 mixed-model ANOVA showed no significant interaction effect *F*(2,76) = 0.23, *p* = 0.793. The numbers of participants experiencing secondary stick shaker events are listed in the section: Binary parameters. (Figure 14)

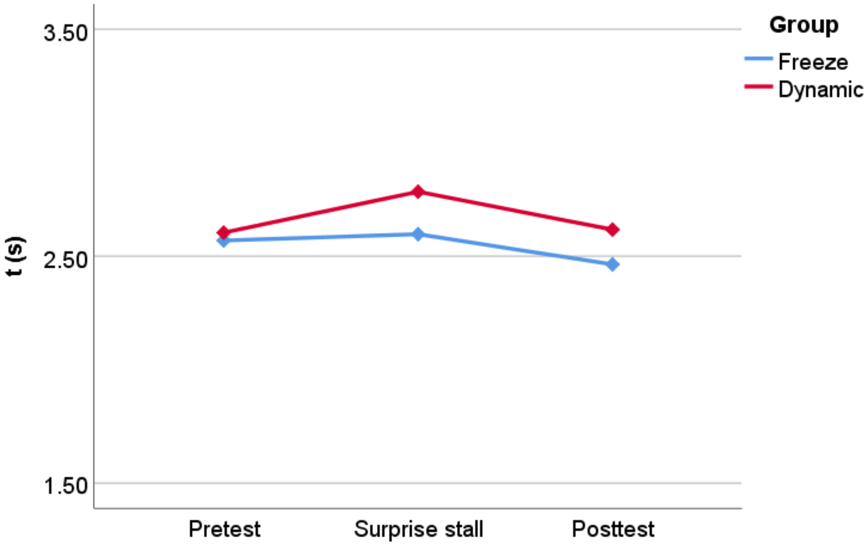

Figure 14. Duration of the first stick shaker activation

3.5.7 Binary parameters

The binary parameters obtained in the experiment are shown in Table 3. The measures are coded so that higher numbers always indicate less desirable behavior or effects. There were no significant differences in responses compared with the pre-test, except for trim use in the Surprise stall post-test. McNemar's test showed no significant differences between post-test scenarios and the pre-test for both groups. The binary measures can be seen to slightly change from pre-test to post-test but these differences were non-significant.

Table 3. Binary parameters of stall recovery

	Pre-test		EGPWS		Surprise stall		Post-test	
	Dynamic N (%)	Freeze N (%)	Dynamic N (%)	Freeze N (%)	Dynamic N (%)	Freeze N (%)	Dynamic N (%)	Freeze N (%)
No disconnect of AT	0/20 (0%)	2/20 (10%)	1/20 (5%)	1/20 (5%)	0/20 (0%)	3/20 (15%)	0/20 (0%)	0/20 (0%)
No use of trim	16/20 (80%)	15/20 (75%)	4/20 (20%)	2/20 (10%)	9/20 (45%)	8/19 (42%)	13/20 (65%)	14/20 (70%)
No verbal speedbrake check	2/19 (11%)	6/19 (32%)	3/14 (21%)	4/15 (27%)	6/20 (30%)	5/16 (31%)	2/18 (11%)	4/20 (20%)
Used rudder	1/20 (5%)	5/20 (25%)	-	-	3/20 (15%)	5/20 (25%)	1/20 (5%)	4/20 (20%)
Secondary stick shaker	2/20 (10%)	3/20 (15%)	-	-	1/20 (5%)	1/19 (5%)	0/20 (0%)	0/20 (0%)
Kept AP disconnect aural on	0/20 (0%)	3/20 (15%)	0/20 (0%)	0/20 (0%)	2/20 (10%)	2/20 (10%)	0/20 (0%)	1/20 (5%)

3.6 Parameters specific to the EGPWS post-test

Summary: there were no significant effects indicating that the training affected pilot’s behavior in the EGPWS post-test.

The time to reach 9,000 ft, the maximum pitch angle, and the minimum CAS, are shown in Figure 15. One participant in the Freeze group did not respond to the EGPWS alert, and responded instead to the second time the alarm was triggered by a second mountain ridge.

There were no significant differences between the groups in time to reach 9,000 ft, $p = 0.894$, maximum pitch angle, $p = 0.458$, and minimum CAS, $p = 0.882$. There were also no significant differences when the outlier, who responded exceptionally late, was removed.

Thirteen pilots in the Dynamic group and ten pilots in the Freeze group triggered the speed low warning. Two pilots in each group also triggered the stick shaker. This led to loss of altitude in two pilots from the Freeze group, and one from the Dynamic group. This further led to flight into terrain for one pilot of each group.

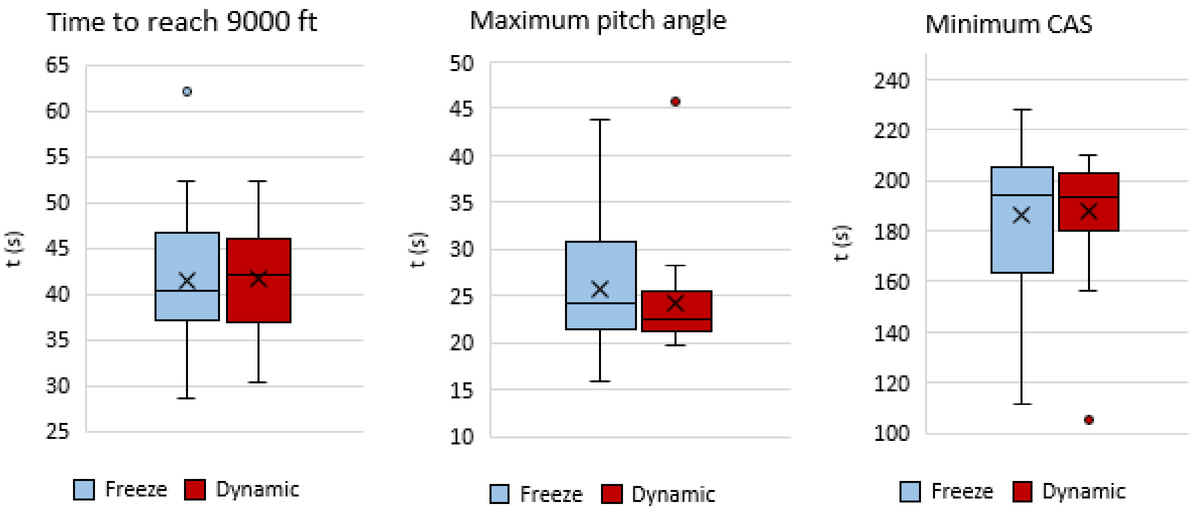

Figure 15. Parameters specific to the EGPWS post-test

3.7 Parameters specific to the False stall post-test

Summary: there were no significant effects indicating that the training affected pilots' behavior in the False stall post-test.

When comparing any unloading response (i.e., a significant pitch-down input), there was no significant difference between the groups, $X^2 = 0.902$, $p = 0.342$. When comparing the altitude loss, there was no significant difference between the groups, $U(38) = 184.0$, $p = 0.650$. When comparing the duration of the descent, there was no significant difference between the groups, $U(38) = 200.0$, $p > 0.999$.

3.8 Subjective responses

The subjective time pressure, subjective anxiety, and subjective time for diagnosis are displayed for all scenarios in Figure 16. There was only a significant effect of the training on subjective time pressure.

The 2 × 5 Group (Dynamic, Freeze) × Test (pre-test, EGPWS, Surprise stall, False stall, post-test) mixed-model ANOVA revealed a significant interaction effect for subjective time pressure, $F(4,152) = 3.13$, $p = 0.017$, $\eta_p^2 = 0.076$. Post-hoc analysis showed that there was a significant increase from pre-test to the Surprise stall post-test $\Delta = 19.8$, $p < 0.001$, and EGPWS post-test, Δ = 17.3, *p* = 0.015, for the Dynamic group, which was not the case for the Freeze group, *p* = 0.770 and .807, respectively. There was also a significant decrease from pre-test to post-test in the Freeze group, Δ = 22.0, *p* < 0.001 whereas this was not the case for the Dynamic group, *p* = 0.210.

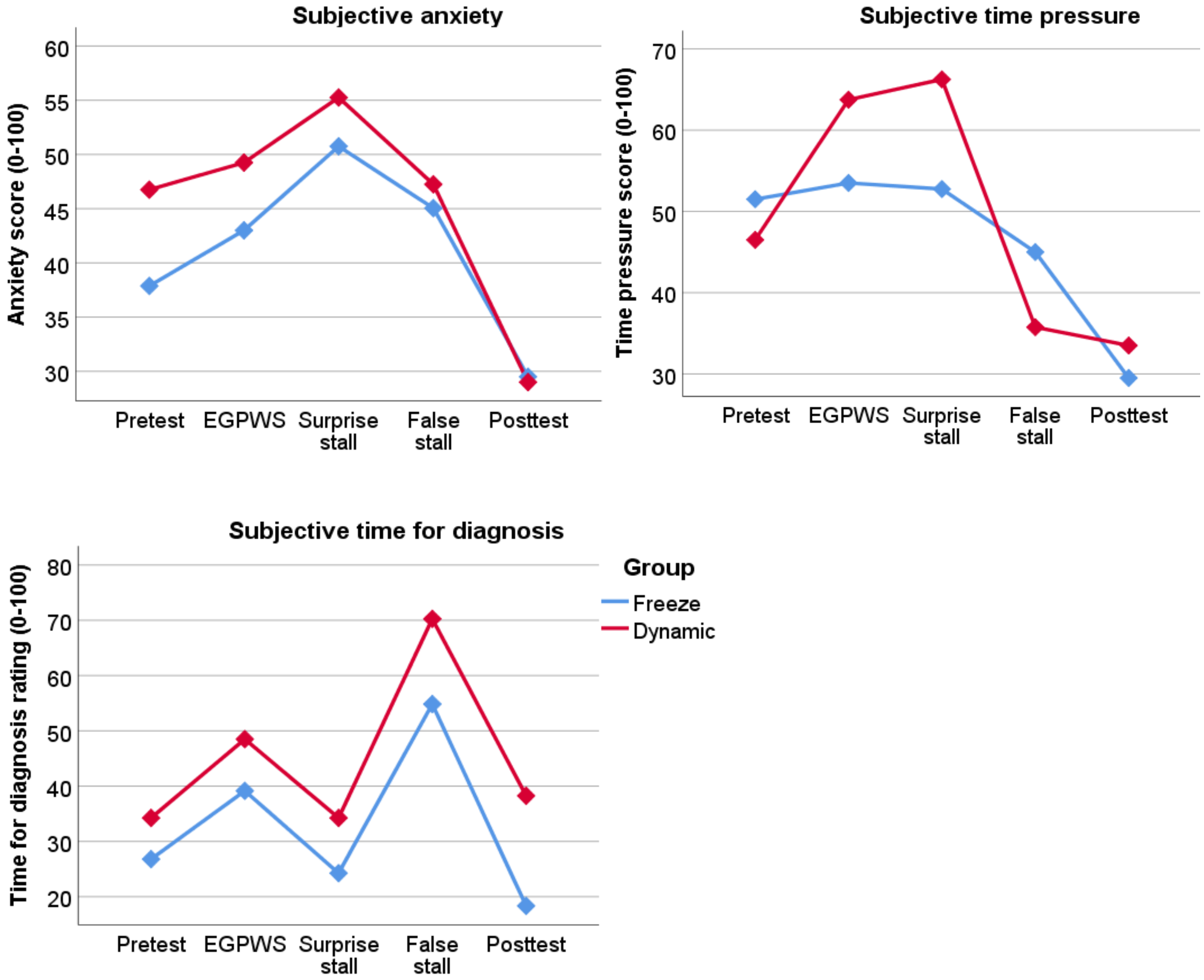

Figure 16. The subjective measures obtained after each scenario.

3.9 Stall recognition test

When comparing the proportion of correct diagnoses of the six test situations between the groups, there was a marginal trend towards better performance in the Dynamic group, *U*(38) = 140.0, *p* = 0.072. When looking at the situations separately, this difference was most prominent in the Overspeed buffet situation, where four pilots in the Freeze group and none in the Dynamic group made an incorrect diagnosis, $X^2 = 4.44$, $p = 0.035$. The number of correct diagnoses was not significantly different for other situations separately.

Over the whole group, the lowest accuracy was observed in situation 6: an overbank situation with no stall (78% correct), and in situation 1: a false stick shaker at level flight (87% correct).

There was no significant difference between the groups in response speed, $t(38) = 0.91$, $p = 0.367$.

There was no difference in reported certainty of the answers between the groups, $t(38) = 0.25$, $p = 0.801$, nor when only correct answers were included, $t(38) = 0.72$, $p = 0.474$, nor when only the incorrect answers were included, $t(18) = 0.18$, $p = 0.858$. The lowest certainty ratings were observed for situation 6: an overbank situation with no stall ($M = 76\%$), for situation 4: a false stick shaker in climb ($M = 80\%$) and for situation 1: a false stick shaker at level flight ($M = 81\%$). (Figure 17)

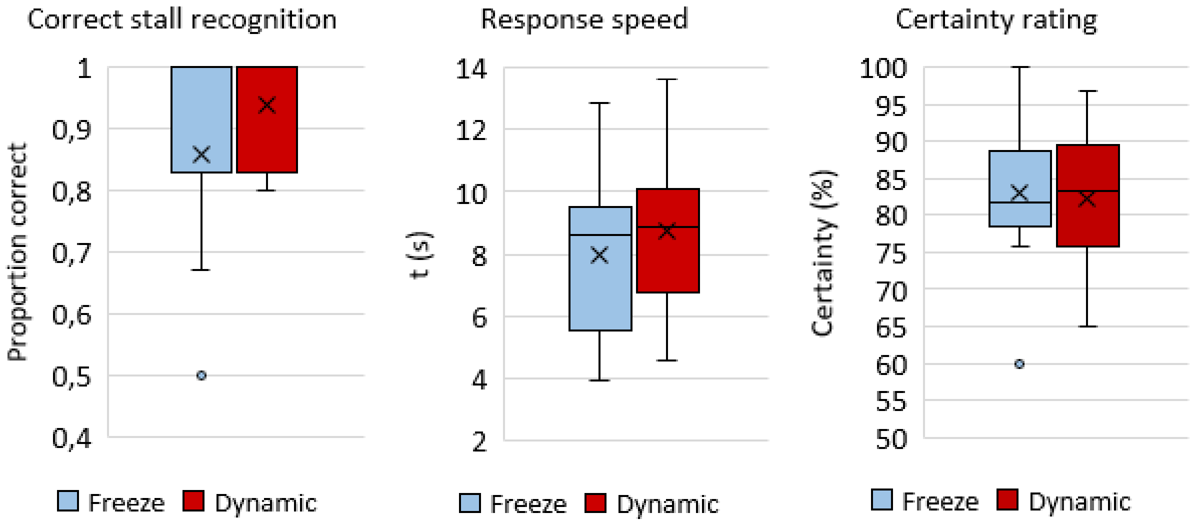

Figure 17. Difference between results.

3.10 Evaluation of the training

3.10.1 Subjective self-confidence in stall recovery

For subjective self-confidence to recover stalls in the aircraft model, there was no significant interaction effect between group and time of measurement (i.e., before and after the training), $B(1,38) = 1.14$, $CI = -0.33 – 2.62$, $p = 0.128$. A Wilcoxon signed ranks test indicated that experienced self-confidence significantly increased for the whole group, $Z = 2.45$, $p = 0.014$. (Figure 18)

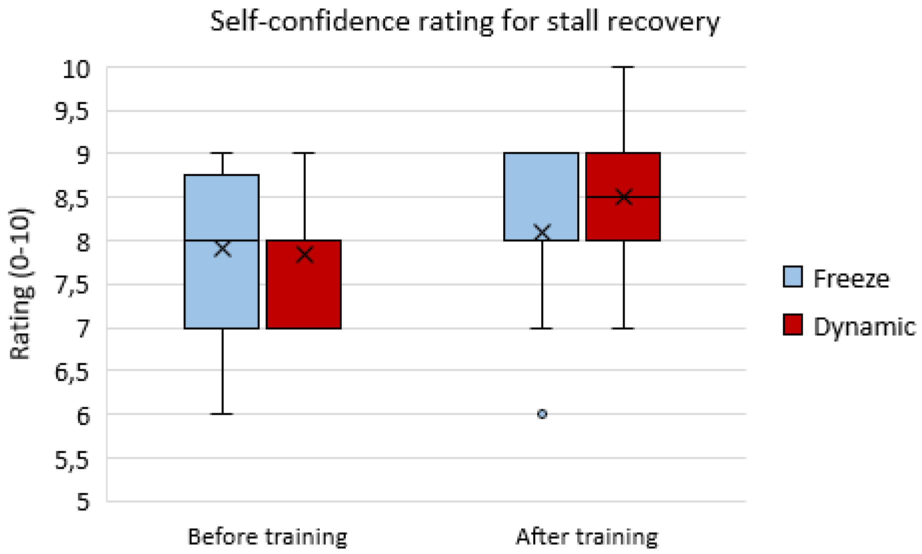

Figure 18. Subjective self-confidence rating for stall recovery.

3.10.2 Interest and enjoyment during the training

For the pilots' Interest and Enjoyment ratings of the training, there was no significant difference between the groups, p = 0.529. The mean rating was 43.2 for the Dynamic training, and 44.2 for the Freeze training, which is near the maximum of the scale (i.e., 49).

3.10.3 Preference for training type

When explaining the group differences to the pilots, thirteen pilots in the Freeze group and seven pilots in the Dynamic group preferred the training, which they received. One pilot in the Freeze group and four pilots in the Dynamic group preferred the other group's training. Five pilots in the Freeze group and nine pilots in the Dynamic group preferred something else, namely a combination of both dynamic scenarios and freeze scenarios. These proportions were significantly different, $X^2 = 10.39$, $p = 0.006$.

4 Discussion

4.1 General training effect

The results show that most of our selected stall recovery parameters responded significantly in the expected direction for both groups of pilots. Exceptions were measures of response speed, and minimum N_z.. This is in line with the parameters being valid indicators of performance or, in the case of altitude loss and recovery duration, possibly the use of a smaller safety margin. It also underlines that our training improved stall recovery in the used aerodynamic model, and the simulator in general, for the pilots in both groups.

4.2 Group differences

4.2.1 Response time

There was no significant difference in how fast pilots responded to stall alarms or the EGPWS alert, whereas we expected faster responses in the Freeze group due to having practiced with immediate responses in the training. In the Dynamic group, we expected signs of skepticism, or ignoring an acute threat indicated by alarms, as this would coincide with the concept of defensive avoidance, a potentially counterproductive coping mechanism against stress (Janis & Mann, 1977).

It is interesting that one pilot in the Freeze group seemed confused about the EGPWS alert, and did not respond to it the first time it activated. This pilot commented: *"When the EGPWS activated, my speed and pitch seemed okay.*" This could indicate that the pilot was confused and checked for stall. "*[I experienced] startle and surprise effect. [I was] unsure whether I should act."* This confusion exacerbated in a self-induced stall when the pilot pulled up after a second activation of the alert. Since this is only person out of 40 participants, we cannot draw conclusions about group differences based on this finding. Furthermore, the stick shaker was triggered by an equal number of participants in both groups, and one pilot in the Dynamic group impacted with terrain as a result.

4.2.2 Occurrences of loss of control

Although the two incidents with terrain impact in the EGPWS scenario do not indicate a group difference, these findings do raise an alarming issue concerning pilots' response to a ground proximity warning and their handling of the subsequent stall, at least in the current sample group. Whether this result can be generalized to airline pilots is unclear, since all participants came from the same company. It seems likely, though, that pilots who are willing to participate in a performance-related study, are relatively more confident about their skills than the general population of pilots, meaning that the observed performance level is probably higher than that of the general population.

No statistical comparison can be made between those who lost control and those who did not, however some characteristics are interesting to note as they may explain why these incidents occurred. The involved pilots were both FOs, had less working experience compared to the rest (12 years versus 17 years), and reported feeling less rested on the day of testing (3 versus 4 on a scale from 1-5). They both had flown most of their flight hours in aircraft types that have envelope protection which prevents manually exceeding the angle of attack limit, and one pilot was currently still flying such a type. It could be that startle and surprise confused them about the envelope protection, but their comments were insufficient to confirm this. These findings, although possibly coincidental, may be relevant to investigate further in future research.

4.2.3 Stall recovery performance

When analyzing the stall recovery behavior, there was a trend towards more aggressive pitch inputs and minimum N_z in the Freeze group in Surprising scenarios compared to the Dynamic group. Furthermore, the Dynamic group showed a decrease in pitch aggressiveness in the Non-surprise post-test compared to the pre-test, which was not observed in the Freeze group. This suggests that practicing only with immediate responses to alerts may lead to slightly more aggressive inputs, but the outcomes were not significant.

No signs of hasty and incorrect responses, such as extensively long recoveries, were found in the Freeze group in the False stall scenario. We had expected these, based on the hypothesis that under high stress, people involuntarily tend to take more simple and well-practiced courses of action, even when these are perhaps erratic and irrational in hindsight (Ozel, 2001). Worst-case, repetitive practice of delaying preventive actions may become ingrained, so that trainees may incorrectly fall back to this behavior when startled or surprised in reality.

4.2.4 Subjective responses

The subjective responses showed a significant increase in experienced time pressure from pre-test to the Surprise stall and EGPWS post-tests in the Dynamic group, but not in the Freeze group. There was also a significant decrease in experienced time pressure in the Freeze group from Non-surprise stall pre-test to post-test, which was not present in the Dynamic group. This is in contrast to our expectation, because we expected the Dynamic group to respond slower and more controlled than the Freeze group in the post-test scenarios, leading to less experienced time pressure. If the Dynamic group's training increases experienced time pressure in surprising situations, this may point to a potential hazard of practicing stall scenarios only in a self-paced manner. Such training may create a contrast between the experienced pacing of events in training, and the experienced pacing of events in operational practice, since real stall situations are likely to be surprising and not self-paced. This contrast may overwhelm pilots when experiencing a real stall, causing stress and confusion.

A potential explanation for this unexpected finding can be found in the pilot's comments in the debrief. Pilots were significantly more likely to prefer the Freeze training, for which they found their responses to be more comparable to those in an actual, necessarily surprising, stall situation. Even though pilots received time to observe the frozen situation during the Freeze training to eliminate the potential advantage of training with surprise, the sudden development of the situation once unfrozen may still have allowed them to practice quick sensemaking better than the other group. This practice may have given them an advantage in the surprising EGPWS and Surprise stall post-tests, which also required quick sensemaking. The contrast between the required sensemaking speed in the Dynamic training scenarios and the post-test scenarios may have been larger, causing an increase in experienced time pressure in the Dynamic group. The question remains why the performance parameters do not confirm this by indicating signs of quicker sensemaking in the Freeze group.

4.2.5 Stall recognition test

The stall recognition test indicated a trend towards better recognition in the Dynamic group, which is in line with our hypothesis. The Dynamic group had more opportunity to observe the aircraft’s behavior and responses to control inputs during stall, which may have allowed them to better analyze the stall and non-stall situations. The difference was most prominent in the high-speed buffet situation, which the Dynamic group was perhaps better able to distinguish from stall buffet due to more focused and longer observation of the stall buffet during the training. Experiencing the dynamic sequence of cues and alarms in different contexts, together with commentary from the instructor, may create a better frame of the situation, prompting better recognition of these situations when startled (Landman, 2017)

4.2.6 Pilots' evaluation of the training

There was no difference in the increase in self-confidence to recover from a stall in the aerodynamic model and simulator. Self-confidence scores increased significantly for the whole group, and the median score was eight on a range from 0-10 before and after the training.

There was also no difference in the Interest and Enjoyment ratings for the training between the groups, suggesting that one training was not experienced as more boring or interesting than the other training.

As mentioned before, most pilots preferred the Freeze training when the training manipulation was explained during the debrief, however most preferred this training type due to the potential of introducing surprise, which was not used in the experiment, to eliminate surprise in training as a possible confounder. Starting the training scenarios in pause was done for experimental reasons, but we advise to only use this setup in training practice when the simulator allows a start from an out-of-trim situation. The advantage of the Desdemona simulator is that we had full access to the flight model and the avionics simulation, and thus could set up the stimulated aircraft in the various stages of (approach to) aerodynamic stall. However, limitations of conventional training simulators make it difficult to start from a paused aerodynamic stall in training practice.

4.2.7 Limitations

One limitation of the study is the employment of well-trained professional pilots as participants. Although this ensures representative stall recovery skills and behavior, it also means that our practice session may have had little effect on their, previously extensively learned, stall recovery behavior. For less-trained or non-trained participants, the different training types may cause larger differences, but the responses of a sample of non-pilots may not be representative of pilots. In addition, if the stall recovery behavior of private pilots would be strongly negatively influenced by our training, the experiment would be unethical.

One technical limitation is that the initial settings of the frozen situations cannot exactly match those in dynamic situations if the situation needs to start with initial motion. This was a problem in the training scenario starting with a roll-off, as there is a roll rate present in the Dynamic scenario but not in the Freeze scenario. We attempted to solve this by letting the Dynamic group intervene at a smaller roll angle than the Freeze group, but it may still have made the training scenario somewhat less challenging for the latter.

The number of participants (i.e., 20 per group) causes the tests to have limited statistical power when testing for small effects. Our *a priori* power analysis of the tests shows that the power to find medium-size interaction effects (i.e., $\eta^2 = 0.06$) at a significance level of 0.05, for our 2×3 mixed-model ANOVAs is 0.679, for our 2×4 ANOVAs 0.755, and for our 2×5 ANOVAs 0.816. Statistical power above 0.8 is considered strong, but this value still means that there is a chance of 20% of Type II errors, i.e., incorrect non-detection of the effects.

In addition, the high number of examined variables increases the chance of Type I errors, for which the *p* values were not corrected due to the explorative nature of the study. The few effects observed on performance (more aggressive inputs and less trim use in one scenario by the Freeze group) did not coincide clearly with other parameters, so it seems assumable that these effects were caused by chance.

Finally, although the Desdemona simulator offers currently optimal motion cueing for ground-based simulation of stall, the centrifuge function could not be used due to unpredictable pilot control inputs. This of course limits the comparability of pilot behavior in the simulator with reality. The stress level in the simulator is not comparable to that in a real stall situation, although pilots still rated their anxiety around the midpoint of the scale, which is a satisfactory level for save, simulated situations without consequences.

5 Conclusions

The main conclusions of this study are as follows:

- The training improved the stall recovery performance in both groups of pilots, indicative of an overall learning effect;
- We did not find statistically significant differences between the Dynamic and Freeze group in the response time to surprising test scenarios, indicating that practicing an immediate response to an unfreeze did not result in more hasty responses;
- We did not find statistically significant differences between the Dynamic and Freeze group in objective recovery parameters to a surprise stall scenario, indicating that both training approaches equally improved the pilots' recovery performance;
- Pilots in the Dynamic group reported significantly higher time pressure in surprise post-test scenarios than did pilots in the Freeze group, suggesting that only self-paced practicing, and actively delaying the response to alarms, made pilots more sensitive to time pressure in an unanticipated event;
- Pilots in the Dynamic group were slightly better in distinguishing stall from non-stall situations compared to pilots in the Freeze group, indicating that experiencing the progression of alarms improves one's ability to recognize an aerodynamic stall.

Based on these conclusions, we advise that stall recovery training should include dynamic, self-induced stalls to practice cue recognition, as well as more sudden "handover" scenarios to practice quick sensemaking. As it may be difficult to pre-set the flight model as well as the moving base of a commercial flight simulator, and reproduce our "frozen" aerodynamic stall conditions, the handover scenarios can be achieved by letting pilots close their eyes while the instructor brings the aircraft into a stall.

6 References

Alexander, A. L., Brunyé, T., Sidman, J., & Weil, S. A. (2005). From gaming to training: a review of studies on fidelity, immersion, presence, and buy-in and their effects on transfer in pc-based simulations and games. *DARWARS Training Impact Group, 5*, 1-14.

Borgvall, J., Castor, M., Nählinder, S., Oskarsson, P. A., & Svensson, E. (2007). *Transfer of training in military aviation.* Swedish defence research agency (FOI).

Burke, L. A. (1997). Improving positive transfer: a test of relapse precention training on transfer outcomes. *Human Resource Development Quarterly, 8*(2), 115-128.

Chandrasekaran, R., Payan,, A., Collins,, K., & Mavris, D. (2019). *A survey of wire strike prevention and protection technologies for helicopters.* Technical Report, U.S. Department of Transportation, Federal Aviation Administration. Retrieved from http://actlibrary.tc.faa.gov

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155.

FAA. (2015). *Stall prevention and recovery training (Advisory Circular No. 120-109A).* Retrieved from www.faa.gov/documentlibrary/media/Advisory_Circular/AC_120-109A.pdf

Goman, M., & Khrabrov, A. (1994). State-space representation of aerodynamic characteristics of an aircraft at high angles of attack. *Journal of Aircraft, 31*(5), 1109-1115.

Groen, E., Ledegang, W., Field, J., Smali, H., Roza, M., Fucke, L., . . . Grigoryev, M. (2012). SUPRA-enhanced upset recovery simulation. *AIAA modeling and simulation technologies conference*, (p. 4630).

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. *Advances in psychology, 52*, 139-183.

Houtman, I. L., & Bakker, F. C. (1989). The anxiety thermometer: a validation study. *Journal of personality assessment, 53*(3), 575-582.

Houtman, I. L., & Bakker, F. C. (1989). The anxiety thermometer: a validation study. *Journal of personality assessment, 53*(3), 575-582.

Janis, I. L., & Mann, L. (1977). *Decision making: a psychological analysis of conflict, choice and commitment.* New York: Free Press.

Kaplan, A. D., Cruit, J., Endsley, M., Beers, S. M., Sawyer, B. D., & A, H. P. (2021). The effects of virtual reality, augmented reality, and mixed reality as training enhancement methods: a meta-analysis. *Human factors, 63*(4), 706-726.

Kochhar, S., & Friedell, M. (1990). User control in cooperative computer-aided design. *UIST '90: Proceedings of the 3rd annual ACM SIGGRAPH symposium on user interface software and technology* (pp. 143-151). ACM. doi:https://doi.org/110.11445/97924.9794

Landman, A. G. (2017). Dealing with unexpected events on the flight deck: a conceptual model of startle and surprise. *Human factors, 59*(8), 1161-1172.

Nooij, S. A., Pretto, P., Oberfeld, D., Hecht, H., & Bulthoff, H. (2017). *Vection is the main contributor to motion sickness induced by visual yaw rotation: implications for conflict and eye movement theories.* Retrieved from https://doi.org/10.1371/journal.pone.0175305

Ozel, F. (2001). Time pressure and stress as a factor during emergency egress. *Safety science, 38*(2), 95-107.

Pennings, H. J., Oprins, E. A., Schoevers, E., & Groen, E. L. (2019). *Current insights in negative transfer of training (R11747).* Soesterberg: TNO.

Ryan, R. M. (1982). Control and information in the intrapersonal sphere: an extension of cognitive evaluation theory. *Journal of Personality and Social Psychology, 43,* 450-461.

Scruton, R. (n.d.). The eclipse of listening. *The New Criterion, 15*(3), 5-13.

Woltz, D. J., Gardner, M. K., & Bell, B. G. (2000). Negative transfer errors in sequential cognitive skills: strong-but-wrong sequence application. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(3), 601-25.