Original Article

Does deliberately going beyond alarms in stall recovery exercises lead to negative training?

Annemarie Landman[1,2]

Douwe Mol[1]

Martijn van Emmerik[1]

Eric Groen[1,3]

[1]TNO: Department of Human Performance, Soesterberg, The Netherlands

[2]Delft University of Technology, Control and Operations Department, Delft, The Netherlands

[2]Cranfield University: Safety and Accident Investigation Centre, Cranfield, United Kingdom

Abstract:

We investigated if deliberately going beyond alarms during aerodynamic stall recovery exercises may result in negative training. Two groups of 20 airline pilots received stall recovery training in a moving-base simulator. The "Delayed-response" group induced the stall themselves. The "Immediate-response" group was presented with paused situations and had to recover immediately after unpausing. In a surprising transfer test, the pilots in the Delayed-response group showed less aggressive unloading, and experienced significantly more time pressure compared to the pilots in the Immediate-response group. In a stall cue recognition test, the Delayed-response group performed nearly significantly better. We conclude that

experiencing the progression of an aerodynamic stall during training has positive effects on pilot performance, even if this requires unprocedural behavior.

Keywords: upset recovery, flight simulation, transfer of training, aviation, human factors

Corresponding author

Annemarie Landman
Human Performance
TNO
Kampweg 55
3769 DE, Soesterberg
The Netherlands

annemarie.landman@tno.nl

Acknowledgments

An earlier version of this paper was presented at the 34th European Association for Aviation Psychology (EAAP) Conference (Gibraltar, United Kingdom, September 2022).

The authors wish to thank all participants for their valuable contributions to this study as well as Mr. Pjotrek Bellers, and Mr. Robert-Jan Burgers for their expert advice on the study design.

Publication Ethics

Informed consent was obtained from all participants included in the study. All procedures in the study involving human participants were performed in accordance with the ethical standards of the institution's Ethics Committee.

Authorship

Annemarie Landman, methodology, writing – original draft; Douwe Mol, data curation; Martijn van Emmerik, project administration, writing – review & editing; Eric Groen, conceptualization, supervision, writing – review & editing.

All authors approved the final version of the article.

Open Data

The data (without participant characteristics to ensure anonymization) that support the findings of this study are available on request from the corresponding author, Annemarie Landman. The data are not publicly available due to privacy restrictions.

Funding

This research was supported by the Dutch ministry of Infrastructure and Water management (assignment 31165538.0002)

INTRODUCTION

Following prominent aircraft accidents, such as Colgan Air flight 3407 (Buffalo, New York, February 2009) and Air France flight 447 (Atlantic Ocean, June 2009), aviation authorities have issued special requirements for Upset Prevention and Recovery Training (UPRT), including simulator exercises to practice recovery from an aerodynamic stall (FAA, 2015). This seems to introduce a training paradox: the prevention part requires pilots to promptly respond to stall indications, or "alarms", while the recovery part may require them to deliberately go beyond these alarms. The latter is illustrated by the following quote from EASA Safety Information Bulletin no. 2013-02: *"During training, the pilot may be asked, for demonstration purposes, to ignore some aural and visual indications of impending stall in order to practice the more difficult control movements needed to recover from the stick shaker."* (European Aviation Safety Agency, 2013), section 1.4.5). We previously interviewed several training experts from aviation industry on this topic, and some expressed their concern that instructing pilots to deliberately go beyond alarms may engrain improper responses to these situations, leading to negative training (Pennings, Oprins, Schoevers, & Groen, 2019). Negative training in this context is training that "*unintentionally introduces incorrect information or invalid concepts, which could actually decrease rather than increase safety.*" (European Aviation Safety Agency, 2015). For example, a head of a flight training organization who was interviewed in (Pennings, Oprins, Schoevers, & Groen, 2019) stated that allowing pilots, or even instructors, to fly into a stall sends an inconsistent message if it takes place after teaching how to prevent a stall. Another flight instructor remarked that he sometimes put trainees in a difficult situation, but would not let them fly into that situation themselves. In contrast, other interviewees found it particularly valuable to have the trainees fly into a stall themselves, as it allows them to experience the progression of the stall, including various visual, aural and motion indications of the stall.

These opposing opinions inspired us to compare the effects of two different approaches for stall recovery training on pilot performance in a quasi-transfer test. One approach was to let the pilots fly into the aerodynamic stall themselves, and go beyond the successive alarms, before recovering. In the other approach the pilots were confronted with a pre-set simulation of a stall. After unpausing the simulator, they had to recover immediately upon noticing that something was going wrong. This is similar to a hand-over scenario where the instructor hands over the control of the aircraft to the trainee. Although the latter approach conforms with operational procedures that dictate recovery at the first indication of a stall, it does not

allow pilots to experience the development of the stall. We therefore hypothesized that, on the one hand, the "non-procedural" approach of going beyond alarms could lead to delayed responses to alarms, while on the other hand, it could have a positive effect on the pilots' recognition of stall indications.

METHODS

Design

Stall recovery performance was tested using a mixed (Test × Group) design with two levels for each factor. Two pilot groups received a stall recovery refresher training, but with a different approach (see, Training manipulation). One group always induced the aerodynamic stall themselves before recovering (Delayed-response group), whereas the other group always had to respond immediately upon unpausing a pre-set stall situation (Immediate-response group). Second, there were two tests. Before the training session a baseline test was performed, consisting of a non-surprising stall event to obtain a baseline measure of the pilots' stall recovery performance. Immediately after the training the pilots were exposed to a transfer test, which included a surprising stall event to test for group differences in the pilots' response to an unexpected stall. Although strictly speaking this should be designated "quasi-transfer" test as it took place in the (same) simulator (Taylor, Lintern, & Koonce, 1993), we will simply refer to it as transfer test. At the end of the simulator session, a passive stall cue recognition test was included to test for group differences in the pilots' ability to identify various stall indications.

Participants

A total of 40 commercial airline pilots (38 men and 2 women, all from one company) participated in the experiment. Nineteen pilots currently flew the Boeing B737 (i.e. the most similar to the experimental generic aerodynamic model and cockpit), fourteen the Boeing B777, six the Embraer E175/190, and one the Airbus A330. Exclusion criteria were: military flying experience, having an aerobatics rating, and having had a glider flying rating in the last 20 years. Two balanced groups were created based on the characteristics listed in Table 1. No significant differences between the groups were detected, indicating that balancing was successful.

Table 1. Characteristics of the groups

	Immediate-response	Delayed-response
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
Type rating instruction or examining experience	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.

Simulator

The experiment was performed in the moving-base Desdemona flight simulator (manufactured by AMST Systemtechnik), located at TNO in Soesterberg (The Netherlands). The simulator's motion platform consists of a gimbaled system for unlimited rotations in three axes; two linear motion drives which allow for 8 meter surge and 2 meter heave motion, respectively; and a planetary yaw drive that allows for centrifugation up to 3g. The aerodynamic model used in this study comprised a generic model of a medium-sized modern transport category aircraft (e.g., Boeing 737NG, Airbus A321, Tu-204; see, Nooij et al., 2017). The model includes aerodynamic phenomena like buffeting, longitudinal and lateral instabilities, dynamic hysteresis, and degradation of control response (Goman & Khrabrov, 1994). The sensation of (un)loading is amplified by vertical prepositioning of the simulator cabin by nearly 1 meter.

The cockpit mockup was styled after the Boeing 737NG, and included the left-side seat, primary flight display with pitch limit indicator, navigation display (not used), engine indications, crew-alerting system (not used), and a partial mode control panel. 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 in pitch. Flaps and speed brakes were not used.

The following alerts, alarms and other cues of (approaching to) an aerodynamic stall would occur in successive order as a stall developed due to decreasing speed at level flight: 1) Decrease of indicated air speed on the speed tape; 2) Increase of pitch angle difference between the flight path vector and aircraft attitude; 3) Speed low aural and blinking of the speed box (both at 70% of the amber band); 4) Appearance of the pitch limit indicator (PLI); 5) Stick shaker and (Airbus) aural warning of stall. The aural cue is in reality not featured in a Boeing 737NG, but it was included to make the cockpit more generic and suitable for different pilots; 6) Stall buffet with motion and audio cues; 7) Lateral instability and "sloppy" controllability due to detachment of the airflow.

General procedure

The participating pilots arrived in pairs and received the experimental briefing together. All pilots provided informed consent and filled in the pre-experiment questionnaires before receiving a briefing on the aerodynamic model, cockpit mockup and the display indications and sounds that were used in the experiment.

They were told that the experiment was aimed to investigate different methods of stall recovery training. They were explicitly told that this training was experimental, and that in operational practice they should always rely on their own company's training. The stall recovery template that was briefed was taken from the Federal Aviation Administration (FAA, 2015). Furthermore, the pilots were instructed to prioritize safety of the recovery over recovery time or minimum altitude loss. Overspeed was said to be less important than excessive g-loads (i.e., -1 to 2.5 g). Pilots were also instructed that they should respond to any situation in the test scenarios as if it was real, except when the experimenter would explicitly tell them otherwise.

After the briefing, each pilot individually performed the training and testing session, lasting for about 60 minutes in total. Instructions in the simulator were given by an experienced UPRT instructor. The simulator session consisted of the following elements: 1) Familiarization with the controls at 5,000 feet and 38,000 feet altitude; 2) Baseline test; 3) Experimental training; 4) Transfer test; 5) Stall recognition test. All instructions given by the instructor

were written out in a script. Pilots were always clearly briefed in advance whether upcoming scenarios were for training or for testing.

Training manipulation

All pilots recovered or corrected (approach to) stall situations from three stages of severity: 1) Recovery at the speed low aural with coinciding flickering of the speed box. This situation did not require executing the stall recovery template, as it could be solved by only adding thrust. 2) Recovery at the stick shaker activation. 3) Recovery at 45 degrees angle of bank, caused by stall-induced lateral instability (i.e., a roll-off or wing drop).

Recovering from these stages was always practiced twice in succession. This set of six exercises was first performed at low altitude (i.e., 10,000 feet), and then again at high altitude (i.e., 38,000 feet). Stall recovery at high altitude is more difficult, because the controls are more sensitive due to the higher speed and the smaller margin between under- and overspeed.

For the Delayed-response group, the training exercises started with the autopilot connected and holding the current altitude (Altitude Hold mode). They were instructed to set the throttle to idle to slow down the aircraft until one of the three trigger moments, described above, at which they had to intervene. The Immediate-response group performed the same exercises, but starting from a paused setting. The alarm(s) would present itself as soon as the scenario was unpaused. This approach ensured that the pilots were not exposed to the progression of successive alarms, and thus always responded to the first appearance of one, or more, stall indications. Prior to each exercise, the pilots in this group were always informed at which stage of the stall the scenario would start, and they were given time to check the settings in the paused situation. This was done to prevent their responses from being influenced by surprise.

Baseline test

Before the training session, a baseline test was performed at 38,000 feet, consisting of an aerodynamic stall at level flight that was announced beforehand by the instructor. The pilot was flying level with autopilot and autothrottle connected until an extreme tailwind brought the aircraft quickly into a stall. The pilots were instructed to recover upon stick shaker activation, which occurred almost four seconds after onset of the tailwind, and three seconds after speed-low alert.

Transfer test

The transfer test after the training session included an actively flown scenario in which the pilots had to respond to an unanticipated aerodynamic stall. To temporarily take the pilots' focus away from stall recovery, this scenario was always preceded by a scenario with an unexpected activation of the enhanced ground proximity warning system (EGPWS). During the latter scenario certain performance measures were collected (e.g., response time, maximum climb rate). However, because post-hoc analysis of these measures did not reveal differences between the two groups, and also because the focus of the current paper is on stall recovery, the results of the EGPWS scenario will not be described in detail here. The transfer test scenario started in climb (500 feet/minute) at 2500 ft. after take-off from Aviano Air Base in Italy, which is a military airport not used by commercial airlines. The autopilot was connected, and initiated a turn. When in the turn, pilots were instructed to change the vertical speed for the autopilot to 1000 feet/minute. This change caused the margin to underspeed to decrease. To set up a false expectation, pilots were also instructed to climb to 6000 feet, and then change their heading to fly north. At 3000 ft., when pilots were setting the vertical speed and looked away from the PFD, a strong wind gust was induced bringing the aircraft quickly into a stall.

Stall cue recognition test

After the transfer test, all pilots performed a passive stall recognition test while they were still in the simulator. The pilots were presented with six different situations, and were asked to identify as soon as possible, or at least within 20 seconds, whether or not the situation was an aerodynamic stall. They could respond by pressing the autopilot disconnect button to indicate "stall", and pressing a 'comm' button to indicate "no stall". All situations were without outside visuals (hence, Instrument Meteorological Conditions, or IMC), and the speed tape was covered. The six situations were always shown in the same order: 1) False stick shaker event at 10,000 feet level flight; 2) High speed buffet (no stall) at 38000 feet level flight; 3) Stall: 45 degrees bank, 3 degrees pitch up, stalled situation at 8500 ft. (shown paused); 4) False stick shaker event while climbing 10 degrees pitch up and wings level at 5000 feet; 5) Level flight stall at 38000 feet with an extremely high angle of attack, so that the flight path vector was not visible (shown paused); 6) 45 degrees bank, 8 degrees pitch-down situation with no stall (shown paused).

Dependent measures

The following dependent measures were obtained in the baseline and transfer test.

Reaction time. The time between the trigger (i.e., stick shaker activation in the baseline test or wind gust onset in the transfer test) and the start of the first subsequent pitch down input was obtained. The start of this input was the first time at which the cumulative sum of pitch input diverged more than ten standard deviations (SDs) from the mean pitch input in the five preceding seconds.

Control input variability. The root mean square (RMS) was obtained for either pitch input changes or for roll inputs between the trigger and the end of the recovery. The end of recovery was defined as the first moment level flight was achieved. Pitch input changes instead of pitch inputs were used to account for pitch trim. Higher control variability indicates that pilots needed more adjustments for the recovery, suggesting that they needed more effort to handle the situation.

Unloading aggressiveness. The minimum N_z (vertical *g*s) was obtained during the recovery to indicate how strongly the pilot pitched down and unloaded. Lower N_z values correspond to more aggressive unloading. More aggressive unloading may help maintain a higher safety margin in terms of altitude loss, but it can also be a sign of startle.

Perceived stress. Pilots rated their stress during the stall event on the anxiety scale (Houtman & Bakker, 1989). This is a 10 cm horizontal scale ranging from 0 (no anxiety) to 100 (maximum anxiety).

Perceived time pressure. Pilots rated time pressure experienced during their response on the Temporal Demand subscale of the NASA-TLX (Hart & Staveland, 1988), ranging from 0-100. The question asked was: “How much time pressure did you feel due to the pace at which the tasks or task elements occurred as you responded to the stall?” Since our goal was not to measure overall task load, we did not use other subscales of the NASA-TLX.

Stall cue recognition. In the stall cue recognition test, the proportion of correct answers was obtained as an ordinal variable. The average speed of answering was obtained as their response time.

Perceived surprise. This was only measured in the transfer test to obtain an indication of whether this scenario was truly surprising. Pilots rated this on a custom scale made similar to the NASA TLX. It ranged from 0 (minimum surprise) to 100 (maximum surprise).

Interest and enjoyment. A second manipulation check was the Interest and Enjoyment (IE) subscale of the Intrinsic Motivation Inventory (Ryan, 1982), which consists of nine questions with regards to how interesting or boring the training exercises were. This was scored after the training session to check for potential group differences in boredom during the training.

Data analysis

Performance variables and subjective responses of the groups during the baseline and transfer tests were analyzed in a 2×2 (Approach × Group) mixed-model ANOVA. Significant interaction effects were followed-up by independent-samples *t*-tests between groups.

The proportion of correctly recognized situations in the stall recognition test was compared between groups using a Mann Whitney *U* test.

Interest and enjoyment during training, and surprise in the transfer test, was compared between the groups with an independent-samples *t*-test.

RESULTS

Performance example

Figure 1 shows an example of relatively poor performance in the transfer test. The pilot responded to the stall alarms by pitching down quite forcefully, which led to a minimum N_z of 0.11 g. The pitch input variability was relatively high (i.e., RMS = 0.048 %/s, mean = 0.036 %/s), while the roll input variability was close to the mean (i.e., RMS = 24.5 %, mean = 25%). At around t = 19 s, the pilot gives a pitch up input, which leads to loading too soon as seen from the secondary stick shaker activation.

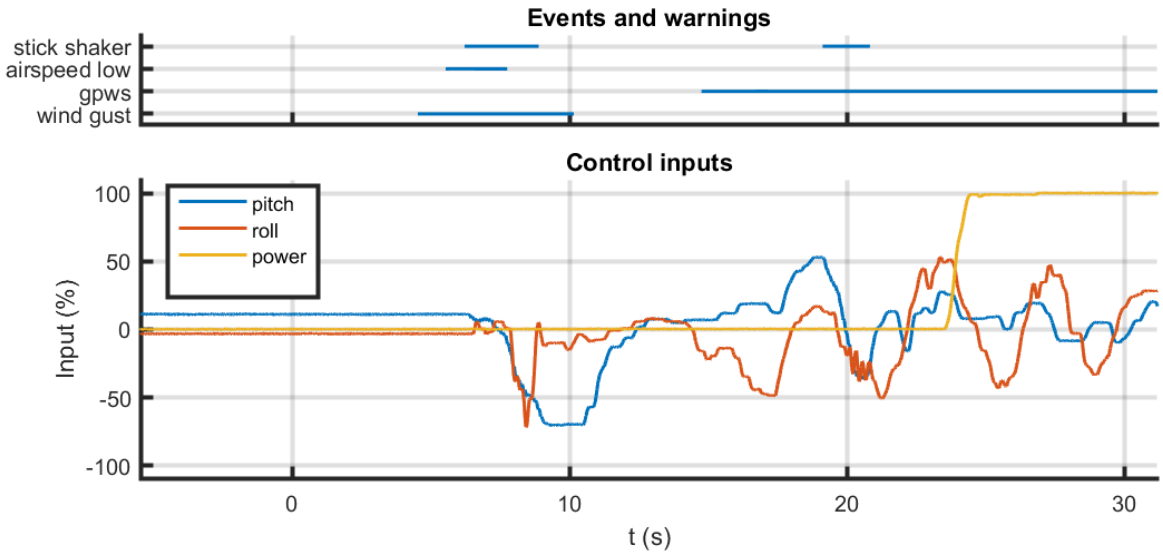

Figure 1. Example of events and alarms (top) and pilot response (bottom) during the transfer test.

Performance of stall recovery

There was a significant 2×2 interaction effect regarding the unloading aggressiveness, i.e., the minimum N_z reached during stall recovery (see, Table 2). Post-hoc comparisons of this interaction showed that the Delayed-response group unloaded significantly less aggressively in the transfer test compared to the baseline test, $\Delta = 0.093$ g, $p = 0.049$, whereas there was no significant difference in the Immediate-response group, $p = 0.261$. This led to a nearly significant difference between the groups in the transfer test, $p = 0.070$, and no significant difference in the baseline test, $p = 0.647$ (see, Figure 2).

Table 2. Analysis of the performance variables in the stall recovery tests.

Variable	Immediate-response Mean (SE)		Delayed-response Mean (SE)		Test × Group $F(1,38)$	p
	Baseline	Transfer	Baseline	Transfer		
Reaction time (s)	1.13 (0.16)	2.49 (0.12)	1.38 (0.16)	2.36 (0.12)	1.62	0.212

RMS pitch input changes (%/s)	0.028 (0.001)	0.026 (0.001)	0.029 (0.001)	0.026 (0.001)	0.54	0.466
RMS roll inputs (%)	23.5 (1.75)	24.8 (1.42)	21.9 (1.75)	25.4 (1.42)	0.57	0.455
Minimum N_z (g)	0.44 (0.041)	0.39 (0.045)	0.41 (0.041)	0.50 (0.045)	5.03*	0.031

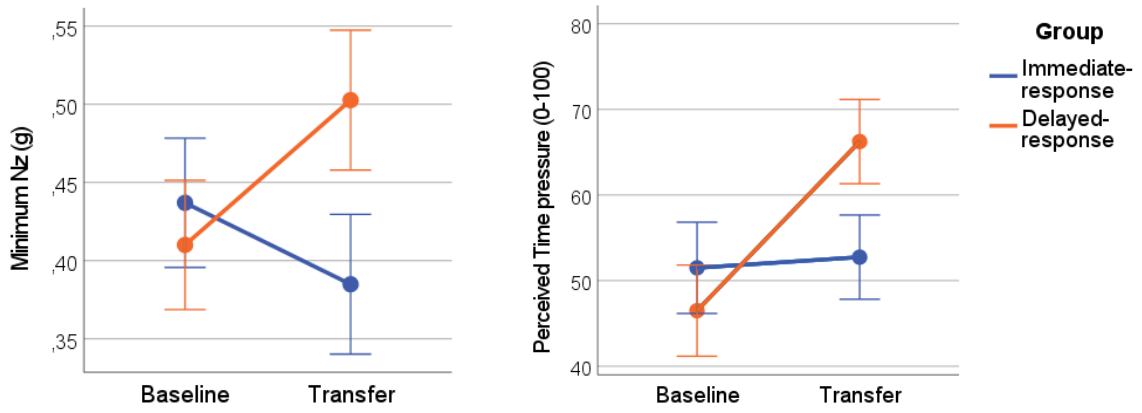

Figure 2. Minimum Nz (left) and perceived time pressure (right) in the baseline and transfer test. Error bars indicate Standard Errors.

Subjective responses during stall recovery

There was a significant interaction effect for perceived time pressure, but not for perceived stress (see, Table 3). Post-hoc comparisons revealed that the Delayed-response group perceived significantly higher time pressure in the transfer test than in the baseline test, $\Delta = 19.8$ points, $p < 0.001$, whereas time pressure perceived by the pilots in the Immediate-response group did not differ between the tests, $p = 0.807$ (see, Figure 2).

Table 3. Pilot subjective responses during the stall recovery tests.

Variable	Immediate-response Mean (SE)		Delayed-response Mean (SE)		Test × Group $F(1,38)$	p
	Baseline	Transfer	Baseline	Transfer		
Stress (0-100)	37.9 (5.0)	50.8 (4.2)	46.8 (5.0)	55.3 (4.2)	0.714	0.527
Time pressure (0-100)	51.5 (5.3)	52.8 (4.9)	46.5 (5.3)	66.3 (4.9)	6.63	0.014

Stall cue recognition test

In the stall cue recognition test, the proportion of correct answers was nearly significantly higher in the Delayed-response group, median = 1.0, than in the Immediate-response group, median = 0.83, $U(38) = 140.0$, $p = 0.072$. When inspecting the answers separately, the difference was most prominent in the overspeed situation, which was incorrectly judged as a stall by four pilots in the Immediate-response group, and none in the Delayed-response group, $X^2 = 4.44$, $p = 0.035$. There were no significant differences on other answers. There was also no significant difference between the groups in response speed, $t(38) = 0.91$, $p = 0.367$.

Over the whole group, the lowest accuracy in answering was observed in the 45 degrees bank with no stall (78% correct), and the false stick shaker at level flight (87% correct).

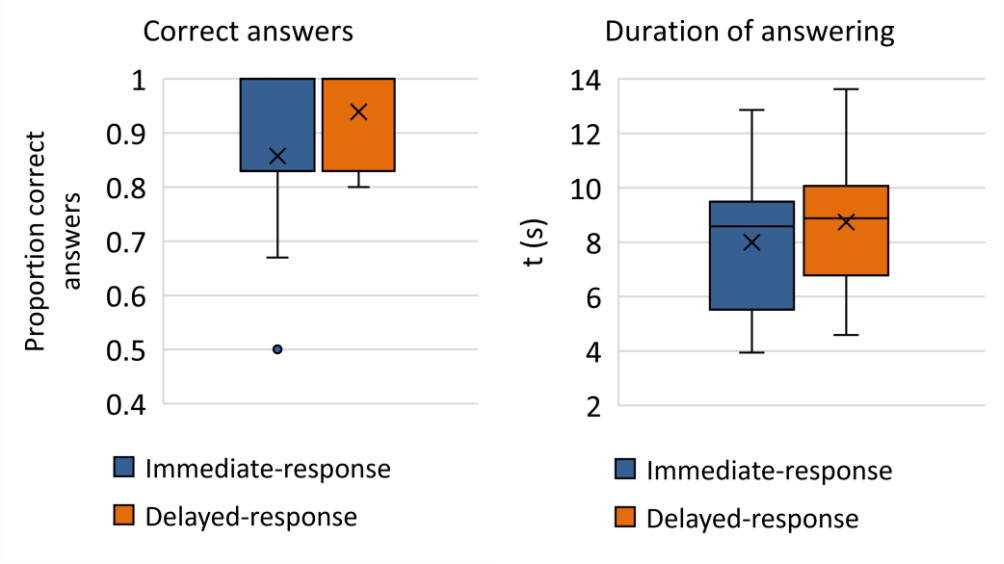

Figure 3. Correct answers and duration of answering in the stall cue recognition test.

Manipulation checks

An analysis of the perceived surprise in the transfer test indicated no significant difference between the groups, $t(38) = -0.63$, $p = 0.531$. The mean score was 61 (median = 70), indicating that the transfer test was generally successful in inducing surprise. The raw scores revealed that there were two pilots with very low scores (i.e. < 20 on the scale). Because this concerned one pilot of each group, we decided not to exclude them from analysis.

The interest and enjoyment ratings during the training did not differ significantly between the groups, $t(38) = 0.64$, $p = 0.529$. The mean score was 43.2 for the Delayed-response group, and 44.2 for the Immediate-response group, which is near the maximum of the scale (i.e., 49).

DISCUSSION

Our first hypothesis was that the Delayed-response group, compared to the Immediate-response group, would show delayed responses to the stall alarms in the surprising transfer test. However, we did not find a significant difference between the groups in reaction time to support this hypothesis. The RMS on control inputs indicated that there was also no difference between the groups in the number of adjustments made in the pitch and roll axes, nor was there a difference in perceived stress, indicating that the groups maintained a similar level of control over the situation when recovering the surprising stall.

Interestingly, we did find that the Delayed response group unloaded marginally significantly less aggressively in the transfer test than the Immediate response group. Aggressive unloading cannot be classified as either positive or negative behavior. This result could indicate positive transfer of training for the Delayed-response group, as quicker obtaining situational awareness may lead to more control over the situations, allowing for smoother unloading. Having seen the situations develop during their training may have helped the Delayed-response group to attain situation awareness in the transfer test. However, the result could also indicate that the Delayed-response group was more hesitant to unload aggressively due to not having practiced such responses in the training session in contrast to the Immediate-response group. If this is the case, this does not indicate a risk of *going beyond alarms* in training, but instead of *lack of practice with hand-over scenarios* which require immediate responses. Supporting this explanation is the fact that pilots are known to be uncomfortable with aggressive control inputs and have to learn to overcome this when recovering stalls (ICAO, 2017). Additionally, the self-reported perception of time-pressure in the transfer test was significantly higher in the Delayed-response group. Hesitation to unload aggressively could lead the Delayed-response group to perceive that more time is needed to recovery smoothly, leading to higher perceived time-pressure.

Our second hypothesis was that the Delayed-response group would show better performance in recognizing stall cues because they were exposed to the progression of various stall alarms during the training exercises. Indeed, the results of the stall recognition test indicated a trend towards better recognition in the Delayed-response group. They appeared especially better recognizing high-speed buffet, possibly because they had been able to pay more attention to the stall buffet cues during their training than the pilots in the Immediate-response group..

When we explained both experimental training approaches during the debrief, most pilots preferred the Immediate-response approach as they saw it as being more comparable to real situations. However, many pilots mentioned the potential of introducing surprise, which was deliberately excluded as much as possible from the current experiment, as the Immediate-response group was asked to observe the paused training scenarios before these were unpaused.

When extrapolating the results of this study to operational practice, several limitations should be acknowledged. First, even though the Immediate-response group received time to observe the paused situations during the training, the sudden development of the situation once unpaused may still have induced some surprise, meaning that the advantage of training with surprise may not have been eliminated completely. Second, it is likely that the surprise and stress

induced by the transfer test is an underestimation of surprise and stress in real situations. Third, the number of variables examined is high, which increases the chance of Type-I errors. The *p* values were not corrected for this, thus the study needs to be seen as explorative. Fourth, only well-trained professional pilots volunteered in this study. Although this ensures adequate stall recovery skills, this limits the potential effects of our training session on the pilots' stall recovery behavior.

In summary, the findings indicate that both the Delayed-response training and the Immediate-response training had different elements of positive transfer of training in a surprising transfer test. Letting pilots fly into stall situations themselves, going beyond various alarms, had a positive effect on their ability to recognize stall cues. Experiencing the progression of alarms may thus improve one's ability to recognize an aerodynamic stall. Letting pilots respond immediately, for instance by using "hand-over" training scenarios, may help pilots with overcoming their hesitation of giving aggressive control inputs.

Recommendations

Since we found evidence of benefits of responding immediately to upset situations in simulator training, as well as of attentively experiencing the cues of (developing) upsets, we advise that UPRT includes both scenario-based training exercises with surprising hand-over situations, as well as maneuver-based training exercises with self-induced upsets.