

The Effectiveness of a Mnemonic-Type Startle and Surprise Management Procedure for Pilots

Annemarie Landman^{a,b}, Sophie H. van Middelaar^a, Eric L. Groen^b, M. M. (René) van Paassen^a, Adelbert W. Bronkhorst^b, and Max Mulder^a

^aControl and Operations Department, Delft University of Technology, Delft, The Netherlands; ^bTNO, Human Factors, Soesterberg, The Netherlands

ABSTRACT

Background: Mnemonic-type startle and surprise procedures were previously proposed to help pilots cope with startle and surprise in-flight, but effects on performance after procedure execution have not yet been investigated.

Objective: Thus, we tested the effectiveness a new mnemonic-type procedure in a moving-base simulator with a non-linear model of a small twin-propeller aircraft flown single-pilot.

Method: An experimental group of twelve line pilots was trained to use a four-item procedure: 1. *Calm down*: take a deep breath, sit up straight and relax shoulders and hands. 2. *Observe*: call out the basic flight parameters. 3. *Outline*: formulate a hypothesis about the problem. 4. *Lead*: formulate and execute a plan of action. A control group of twelve line pilots received a control training. Next, all pilots performed four scenarios with startling and surprising events. Data were obtained on pilot performance, stress, procedure application and evaluation.


Results: Application of the procedure in the test scenarios was high (90.0% full, 100.0% partly), and pilots evaluated the procedure positively (median: 4 on a 1–5 point scale). There was significantly superior decision-making in the experimental group, but immediate responses were significantly less optimal. Pilots sometimes applied the procedure at inappropriate moments.

Conclusion: The results of the tested mnemonic-type procedure were promising. The procedure may benefit, however, from modifications to reduce complexity and to stimulate application at the appropriate moment.

Introduction

Aviation safety organizations have issued new regulations or recommended that pilots receive targeted training to manage startle and surprise (European Aviation Safety Authority, 2015; Federal Aviation Administration, 2015). A startle consists of a rapid stress response in reaction to a sudden or threatening event, whereas a surprise occurs when one observes information that mismatches with one's expectations (Rivera et al., 2014). A surprise prompts an analysis of the situation and, possibly, an adjustment of one's mental model, or frame. This so-called "reframing" is particularly difficult to do under high stress (Klein et al., 2007; Landman et al., 2017a), as stress occupies working memory and impairs top-down or goal-directed attentional processes (Eysenck et al., 2007). Without an appropriate frame for the situation, the perspective on the relevance and meaning of the present information may be lost, leading to confusion. It was shown that in-flight situations are indeed significantly more difficult to handle when they are surprising instead of predictable (Casner et al., 2013; Landman et al., 2017b; Martin et al., 2016).

CONTACT Annemarie Landman  h.m.landman@tudelft.nl  Section of Control and Operations, Faculty of Aerospace Engineering, Delft University of Technology, Kluyverweg 1, Delft 2629 HS, The Netherlands

 Supplemental data for this article can be accessed on the [publisher's website](#).

© 2020 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

It is still unclear how airline companies should approach this problem and train pilots effectively to manage startle and surprise. Recent research suggests that introducing unpredictability and variability in pilot training is important (Landman et al., 2018). A second approach is to teach pilots a specific startle and surprise management procedure. Decision-making procedures already exist to systemically deal with emergencies, for instance: FOR-DEC (Hörmann, 1996), DESIDE (Murray, 1997) or DODAR (Walters, 2002). However, these procedures all start from a diagnosis of the problem, whereas startle and surprise may severely deteriorate a pilot's ability to understand what is going on (Landman et al., 2017a). Therefore, new procedures have been proposed which aim to "de-startle" pilots before they engage in a problem-solving routine. Examples are Breathe-Analyze-Decide (BAD) by Martin (2017), and Unload-Roll-Power (URP), by Field et al. (2018), which later evolved into Reset-Observe-Confirm (ROC; Boland, 2016). Until now, there exist no peer-reviewed publications about the effectiveness of these procedures. One report (Field et al., 2018) indicated high pilot appreciation of the URP procedure (on average 8.3 on a 1–10 point scale; page 87), and an increase in the calling out of observations (page 74) in a simulator scenario. However, the experimental task was not to make optimal decisions, but to focus on proper URP procedure execution. Thus, no data exists yet to indicate whether these procedures achieve their intended results, that is, lead to better decision-making and performance in surprising and startling situations.

The current experiment aims to change this by testing the effects of a mnemonic-type startle and surprise management procedure on pilot decision-making and other outcomes. We developed a new procedure to have complete control over its design and presentation. Elements of the procedure are similar to those in the other proposed procedures. Based on the above-mentioned theoretical framework of stress and surprise, the primary aims our procedure are to manage the effects of stress and to aid pilots in reframing, so that they can come up with appropriate responses. The procedure is purposefully kept concise for use under high stress and can only be applied when there are no immediate threats present. The first step is to manage stress through breathing and muscle relaxation. Similar techniques are applied in the military (e.g., U.S. Marine Corps, 2010), competitive sports (Pelka et al., 2016), and education (e.g., Paul et al., 2007). This might help pilots to start with the troubleshooting process from a calmer state. The next step is to systematically observe the overall situation. The rationale behind this is that it (re) establishes an overview, which might be impaired due to a long period of automated flight, or due to the surprise involving a sudden change in the situation. Establishing an overview may further reduce stress and prevent tunnel-vision or rushed responses. After this, the pilot analyzes the issue and its implications (reframes) and formulates a plan of action. Pilots would likely perform these last two steps regardless of a mnemonic procedure. However, the procedure aims to improve the execution of these steps by having the first two steps precede these. A more detailed description of the procedure is presented in the Materials and methods section.

Materials and Methods

Participants

Twenty-four Dutch, currently employed, line pilots participated in the experiment. Pilots with military flying experience were excluded as they are likely to have had extensive training on dealing with startle and surprise. Each pilot was randomly assigned to an experimental ($n = 12$) or control group ($n = 12$), unless balance, in terms of the characteristics listed in Tables 1 and 2, tended to be distorted. Interventions into the random assignment occurred four times. All pilots had basic experience (< 100 hours) in flying multi-engine piston (MEP) aircraft, similar to the aircraft model used in this study. Most pilots came from one company, with eight in the experimental and six in the control group. Other companies featured one or two pilots each. Pilots' trait anxiety was evaluated using the State-Trait Anxiety Inventory (STAI) test (Spielberger et al., 1970). There were no significant differences between the groups. This research complied with the tenets of the Declaration of Helsinki and informed consent was obtained from each participant.

Table 1. Characteristics of the participants.

	Experimental group Mean (SD)	Control group Mean (SD)
Age (years)	37.4 (12.7)	39.6 (11.7)
Hours large aircraft	7172 (5549)	7544 (5851)
Hours small* aircraft	265 (107)	393 (431)
Employed as pilot (years)	13.5 (10.8)	14.7 (10.9)
STAI (20–80)	28.9 (12.3)	24.9 (4.3)

* CS-23/FAR part 23.

Table 2. Characteristics of the participants (continued).

	Experimental group n	Control group n
Aerobatics experience	2	4
Glider rating	4	3
Instructor (large jet)	4	3
Rank: Captain	4	6
Rank: First officer	6	5
Rank: Second officer	2	1
Gender: male	12	11

Apparatus

The experiment was conducted in the SIMONA research simulator at the Delft University of Technology (Stroosma et al., 2003). This is a full-motion simulator with a six-degrees-of-freedom hydraulic hexapod motion system. The simulator has a collimated 180 degrees horizontal by 40 degrees vertical field of view for outside vision rendered with FlightGear. A 5.1 surround sound system was installed for realistic 3d sound effects of startling events, alarms, flaps, gear, aerodynamic noise, ground rumble and engines.

A generic model of the Piper PA-34 Seneca III, a light MEP aircraft, served as the aircraft model throughout the experiment. The model is suitable for testing the pilots' general flying skills, instead of their application of type-specific standard operating procedures. None of the participating pilots had the advantage of having more than basic flight experience (< 100 hours) on this or similar types. The corresponding software model was a non-linear, six-degrees-of-freedom model developed by De Muynck and Hesse (1990), which has been adapted to simulate failures by Koolstra (e.g., Koolstra et al., 2015). This model was expanded to include the failures simulated in this experiment. The flight deck of the research simulator was modeled after a generic multi-crew cockpit. The flight controls and instruments included a control column and pedals with force feedback, pitch trim on the column, throttle, gear, and flap lever with three flap settings: 0°, 25° and 40°. The (digital) instruments were based on a Cessna Citation II and included a Primary Flight Display, a gear- and flap indicator, Exhaust Gas Temperature display, RPM and torque indicators, fuel quantity and oil temperature/pressure displays. A stickshaker functionality, which the real aircraft does not feature, was added to the model for the goal of this experiment.

Experimental Design and Tasks

The experimental and control group followed the same protocol, except that the experimental group received instructions and practice regarding the COOL procedure (see, Figure 1).

Pilots performed the tasks as single-pilot crews. In the familiarization, and most training and posttest scenarios, pilots were to take off at EHAM (Schiphol, Amsterdam) runway 18 C, make two left turns, join a left-handed traffic pattern at 1000 ft, and land again on 18 C (see, Figure 2). This will hereafter be referred to as “standard pattern”. Pilots had the required settings, as shown in

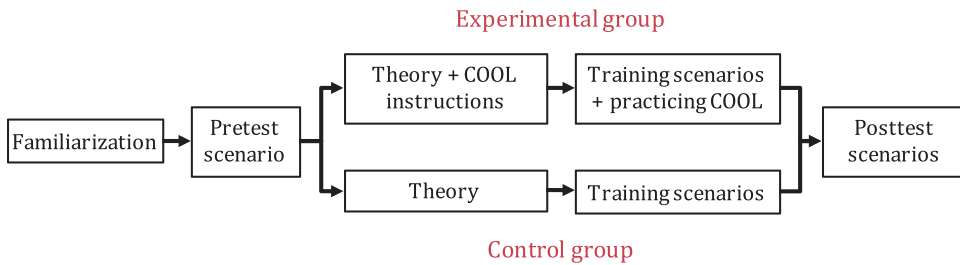


Figure 1. Experimental design.

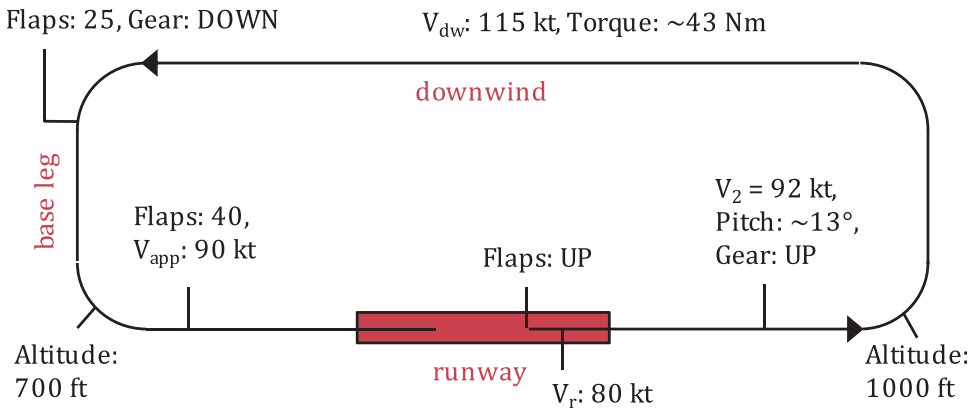


Figure 2. The standard traffic pattern with the target settings.

Figure 2, available on a checklist in the cockpit. The single-engine minimum control speed air ($V_{mca} = 80$ kt) was also listed on the checklist.

Pre-Flight Briefing and Familiarization

Each pilot first received a pre-flight briefing about the experiment, the aircraft model and the required flight patterns. Pilots were instructed to refrain from making go-arounds, leaving the pattern, or landing on different runways. These limits were set to keep performance comparable and to create time pressure. Within these limits, they were free to adjust speed, altitude or configuration as they felt necessary. They were then seated in the simulator and practiced two takeoffs and three standard patterns. The second and third pattern featured crosswind (10 kt pure crosswind from the left). The third pattern was used to demonstrate the aircraft model's stall behavior, the stall alarms (audio and stick shaker) and the gear-up alarm, by letting the pilot trigger a stall and reduce throttle on downwind. At the end of the familiarization session, all pilots confirmed that they could handle the aircraft model satisfactorily, and none required help in determining the turn points of the pattern.

Pretest Scenario

Following the familiarization, a pretest scenario was performed to compare the two groups in terms of performance, surprise or stress responses. Pilots were to perform a precision landing in crosswind conditions (10 kt pure crosswind from the left). An unannounced left engine failure occurred at circa 600 ft altitude, 1.5 minute before touchdown.

Theory

Pilots came out of the simulator to receive a briefing about the next parts of the experiment. Both groups received a 10-minutes briefing about the concepts of startle and surprise (see, Introduction), and about the current relevance of research into pilot reactions to startling and surprising events. The reason for this briefing was to ensure that both groups had similar expectations with regard to the startling and surprising nature of the upcoming experimental scenarios. Next, only the experimental group received a second 10-minutes briefing in which they learned about the startle and surprise management procedure and the reasoning behind it (see, Introduction). As a memory aid, the procedure used the mnemonic COOL:

C – Calm down. Take a deep breath, sit upright, relax shoulders and hands, and be aware of applied control forces.

O – Observe. Instead of immediately attempting to analyze the problem, take a step back and observe the situation. Call out basic instrument readings: pitch, speed, bank angle, altitude and vertical speed. Call out what the aircraft is doing (e.g., “continuously yawing to the right”) as well as other unusual perceptions. Check secondary instruments and configuration if possibly related to the observed issue.

O – Outline. Consider what does and does not make sense and formulate a diagnosis. This can be a technical cause (e.g., damaged elevator) or, if not understood, the aircraft’s behavior (e.g., controllability issue in pitch).

L – Lead. Formulate a plan for action (e.g., “thus, I’m going to ...”) and follow through. This can also involve testing out the effect of certain inputs to analyze the problem further.

The experimental group was told that the purpose of the experiment was to test the usefulness of the procedure for dealing with startling and surprising events, and they were asked to apply the procedure whenever an unusual event occurred. However, it was emphasized that immediate actions required to fly the aircraft (e.g., recovering an upset, maintaining altitude) always took precedence over the COOL procedure. All pilots agreed to this and indicated that they had learned this principle in their own training as well. Going back into the simulator, the experimental group now had a note with the COOL procedure steps attached to the control column. The control group was told that the experiment was about measuring pilot responses to startling and surprising events.

Training Scenarios

Pilots went back into the simulator to practice the COOL procedure with feedback on the execution (experimental group) or to simply respond to the presented issues (control group) in five training scenarios. They were told that their performance in these training scenarios was not monitored yet. In the first scenario, with no unusual events, the experimental group was asked to execute the COOL procedure at several phases in the pattern. The second scenario consisted of an approach and landing with strong crosswind (19 kt pure from the left), while the rudder malfunctioned and remained centered at ca. 300 ft altitude, two minutes before touchdown. The third scenario consisted of a standard pattern with an RPM indicator failure on the left engine when turning into downwind. The fourth scenario consisted of an approach and centerline flyby. Shortly before reaching the runway, the rudder deflected and remained stuck at 15 degrees to the right. The fifth scenario involved a right engine failure occurring shortly after rotation.

Posttest Scenarios

The pilots were informed that four posttest scenarios would follow, in which performance would be monitored. The experimental group was told that resolving the situation safely should take precedence over applying COOL. Distractions were included to increase workload and stress: reduced visibility, flying in a different area, crosswind, and instructions to make a precision landing (in all scenarios). The scenarios were developed so that they did not require type-specific knowledge, and ATC communication was not included. Three out of four scenarios (all except the mass shift)

involved malfunctions that also feature in pilot training. The order in which the scenarios were presented was counterbalanced using the Latin square method.

The “flap asymmetry” scenario (FLAP, see Figure 3) consisted of a standard pattern with low visibility. The runway was just visible when turning toward base leg. When selecting flaps 25, the left flap remained up, which was simulated by adding an aileron and rudder offset depending on the flap angle. This caused a roll as well as a yaw moment, which required correcting. If selecting flaps 40, the asymmetry increases and landing would become very difficult. Thus, appropriate decisions would be to land with flaps up, or to leave flaps at 25 degrees.

The “mass shift” scenario (MASS, see Figure 4) consisted of a standard pattern. Upon rotation, it was simulated that a piece of cargo broke loose and shifted toward the tail, with a loud scraping noise that sounded from the back. The center of gravity (CoG) shifted backwards, producing a violent pitch-up moment. If this was not counteracted quickly with the elevator, this would cause the airspeed to decrease to such an extent that giving maximum elevator could not make the aircraft pitch down. In that case, the aircraft could stall, and the only ways to recover would be to reduce thrust or roll to the side. Since selecting flaps 25 causes a pitch up moment and increased drag (which pilots had experienced in the familiarization flights), appropriate decisions included early configuration at a higher altitude, or landing with flaps up.

For the “false stall warning” scenario (STALL, see Figure 5) pilots were asked to fly a right-handed pattern at 2000 ft. Visibility was moderate. When reaching 1500 ft, a bird struck the angle of attack vane. This created a loud impact noise coming from the front and triggered a continuous (false)

Flap asymmetry when selecting flaps 25

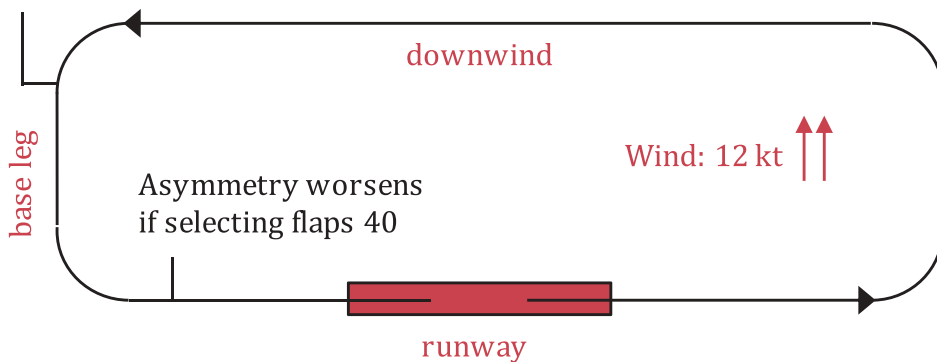


Figure 3. An overview of the pattern and events in FLAP.

Selecting flaps 25 causes excessive pitching up again

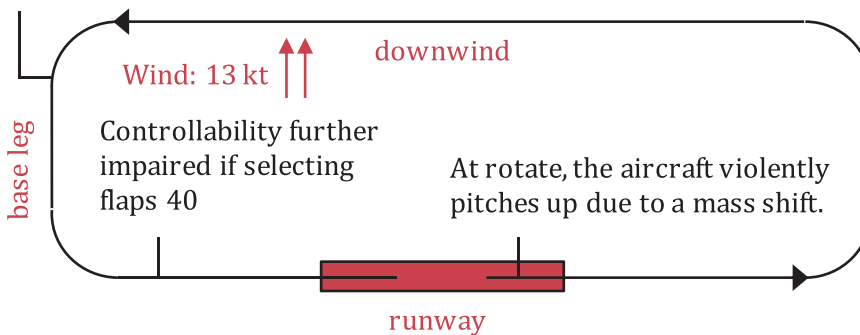


Figure 4. An overview of the pattern and events in MASS.

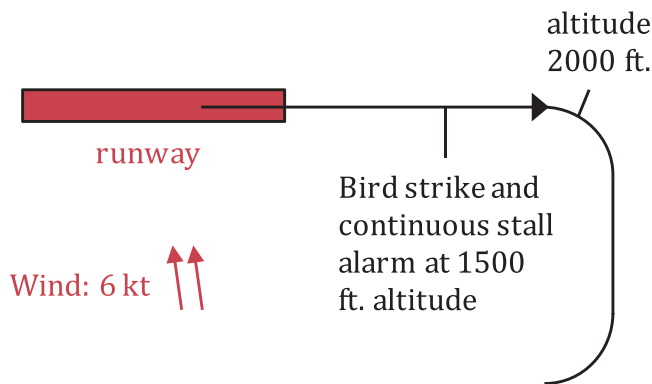


Figure 5. An overview of the pattern and events in STALL.

stickshaker and stall audio alarm. Pilots were familiar with these alarms from the familiarization session, and were expected to respond first by unloading (i.e., decreasing their rate of climb or descending), then figure out that the alarm was false and resume the pattern. The scenario was stopped in downwind.

The “airspeed unreliable” scenario (ASU, see Figure 6) featured a standard pattern at a different airport (EHLE 05). Upon rotation, the indicated airspeed decreased by 1 kt every second from the actual airspeed. This would prevent the pilots from achieving their target speed during climb at the required pitch angle (see Figure 2), even while maintaining maximum throttle. After realizing that the airspeed was unreliable, standard operating procedures (see e.g., Bureau d’Enquêtes et d’Analyses pour la sécurité de l’aviation civile, 2012, Appendix 6) dictate reverting to the known pitch and power settings for the required speed. These settings were provided on a checklist in the cockpit.

Dependent Measures

Flight parameters and pilot inputs were logged at 100 Hz. Questionnaires were filled in following the practice session and following each posttest scenario. Pilots were informed about the nature of the issue in each posttest scenario after they had filled in the questionnaire.

Application and Usefulness of the Procedure

Pilots in the experimental group reported which steps of the COOL procedure they had applied. This was confirmed by checking the audio recording for the application of *Observe*, which was to be

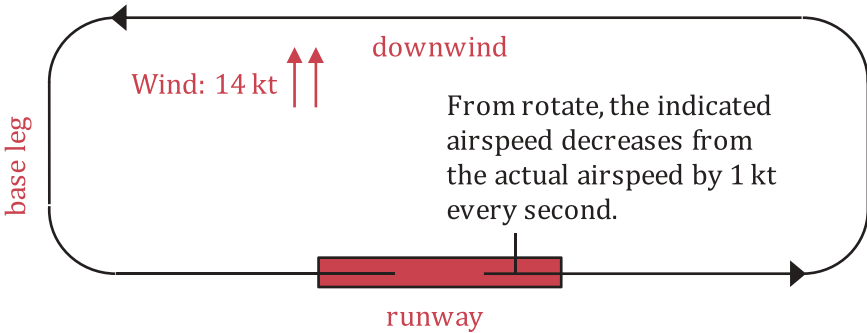


Figure 6. An overview of the pattern and events in ASU.

performed out loud. If applied, pilots rated the usefulness of the procedure on a 1–5 Likert-type scale ranging from “very little” to “very much”.

Performance

Pilot performance in the pretest was measured by checking whether the speed fell below V_{mca} (i.e., 80 kt), the duration at which it remained below V_{mca} , whether loss of aileron control occurred (increasing roll angle while giving maximum inputs to the opposite direction), and whether the pilot successfully landed on the runway.

In the posttest scenarios a number of binary performance criteria were defined for which adherence indicates appropriate or optimal responses. The criteria were selected based on being clearly identifiable in the logged flight parameters, and based on advice by an expert (i.e., a SEP flight instructor and small twin-jet test pilot). The criteria were clustered into three main performance aspects: *Aviate*, *Diagnosis* and *Decision-making*. The scores on these aspects indicate the percentage of criteria adhered to. This clustering was done to avoid having too many separate outcomes, and to decrease the variance in outcomes due to chance.

The first aspect, *Aviate* (Table 3), referred to the pilot’s immediate responses to ensure a safe flightpath. This is in line with the first item of the common phrase “Aviate, Navigate, Communicate” (see e.g., FAA Safety Team, 2018), outlining the order of importance of piloting actions to ensure safety. For *Aviate*, we intended to test whether the COOL procedure caused a performance impairment, for instance, due to additional workload or distraction. The limits selected for excessive bank angle (A1) and pitch angle (A3) were both 5 degrees below the FAA Safety Team’s (2018) definition of an aircraft upset. This was done because pilots are manually flying in the experiment and can therefore be expected to intervene before an upset is reached.

Immediate responses were in particular necessary in FLAP (i.e., stopping the roll motion; A1 in Table 3), in MASS (i.e., stopping and recovering the pitch-up motion; A3 in Table 3), and, to a lesser extent, in STALL (i.e., unloading; A5 in Table 3). To obtain more insight into potential interference by the COOL procedure, we also checked in the audio recordings if pilots started to execute the *Observe* step before recovering the roll or pitch-up angle (i.e., before bringing it back under 40 or 20 degrees respectively), or before unloading (as defined in A5 in Table 3).

Diagnosis referred to the pilots’ ability to identify the cause of the problem (see, Table 4), which was measured by inquiring the pilot about the cause subsequent to each scenario. Our hypothesis was that the *Observe* and *Outline* steps of the COOL procedure could improve *Diagnosis*. In ASU, the gradually increasing error in the instrument readings can be expected to eventually be identified by all pilots (see also, Landman et al., 2018). Therefore, the time it took pilots to identify this problem was obtained instead of their ability to identify the problem. An earlier realization that the ASU is incorrect would lead to an earlier reduction of the throttle setting, as this would cause the indicated airspeed to drop below minimum (D4 in Table 4).

Table 3. Criteria defined for the performance aspect *Aviate* (A).

Criterion	Scenario	Action	Description
A1	FLAP	Prevented excessive bank angle	Following the selection of flaps 25, the pilot responds quickly enough to prevent the bank angle from exceeding 40 degrees.
A2	FLAP	Maintained speed	After selecting flaps 25, on base leg (i.e., heading 060 to 110), the pilot is vigilant enough to not let the speed drop below V_{mca} (80 kt).
A3	MASS	Prevented excessive pitch angle	When the mass shift occurs, the pilot responds quickly enough to not let the pitch angle exceed 20 degrees.
A4	MASS	Recovered quickly	When an excessive pitch angle occurs (i.e., A3 is not met), the pilot responds quickly enough to bring the pitch angle back to below 20 degrees, within 10 seconds after the mass shift.
A5	STALL	Unloaded	Following the bird strike (< 10 seconds), the pilot responds to the stall alarm by unloading during the climb (decreasing pitch > 4 × the SD from the mean pitch angle, both taken over the 20 seconds before the bird strike).

Table 4. Criteria defined for the performance aspect Diagnosis (D).

Criterion	Scenario	Action	Description
D1	FLAP	Identified problem	Identified a flap asymmetry or malfunctioning flaps.
D2	MASS	Identified problem	Identified a cargo or mass shift.
D3	STALL	Identified problem	Identified a false stall alarm. Identifying a bird strike or a malfunctioning angle of attack vane was not necessary.
D4	ASU	Identified problem quickly	Reverting to a lower than full throttle setting within two minutes. This was the average identification time for the same scenario in Landman et al., 2018.

Decision-making referred to the actions pilots performed to ensure safety following the startling and surprising events (see, Table 5). We expected that the *COOL* procedure would improve situation awareness through *Calm down* and *Observe*, which should lead to better comprehension of the situation (*Outline*), and better projection of how the situation would or could evolve and what could be done (*Lead*; Endsley, 1995). Thus, the experimental group was expected to display more decisions which indicate awareness of the issues and risks. Increasing the distance for final (DM5) would require planning in downwind, which was only applicable in MASS and ASU. However, in MASS, execution of the procedure after selecting flaps 25 at the end of downwind can cause pilots to unintentionally increase the length of final. Therefore, only ASU was analyzed.

Subjective Responses to the Scenarios

To test if the scenarios were challenging, pilots rated startle and surprise on a 0–10 Likert-type scale ranging from “not at all” to “extremely”. No validated rating scales of startle and surprise exist. Therefore, these scales were based on the anxiety scale of Houtman and Bakker (1989). A horizontal analogue version of the anxiety scale was also used to measure perceived anxiety during the scenarios. We did not expect that the procedure would affect the intensity of the ‘startle and surprise responses, however we did expect that better management of these responses and the events would lead to lower anxiety experienced during the scenario in the experimental group. Mental effort was scored on the Rating Scale Mental Effort (RSME; Zijlstra & Van Doorn, 1985), to check if the procedure increased mental workload in the experimental group.

Data Analysis

For the performance aspects: Aviate, Diagnosis and Decision-making, a percentage of adherence to the criteria was obtained. These were compared using Mann-Whitney *U* tests. Other ordinal variables, such as non-combined Likert scale data, or non-normally distributed data, was also

Table 5. Criteria defined for the performance aspect Decision-making (DM).

Criterion	Scenario	Action	Description
DM1	FLAP	Refrained from selecting flaps 40	The pilot does not exacerbate the asymmetry and refrains from selecting flaps 40.
DM2	MASS	Configured early	Recognizing that configuration changes may exacerbate the controllability issues, the pilot configures flaps and/or gear earlier at higher altitude (before turning to base leg, heading 030), or keeps the flaps up.
DM3	MASS	Increased altitude	To increase the safety margin, the pilot flies the pattern at a higher altitude than the given target altitude of 1000 ft (average > 1200 ft in downwind, i.e. between heading 330 to 030). To exclude inadvertent altitude increases, those who selected flaps in downwind (which causes altitude increase) are excluded.
DM4	MASS	Selected flaps carefully	Recognizing that the ballooning effect may again cause excessive pitch up, the pilot takes measures to prevent pitch from exceeding 20 degrees when selecting flaps. Those keeping flaps up are not included.
DM5	ASU	Increased final	To increase the safety margin, the pilot increases the distance and time at final, by turning to final (heading 080) at least 1500 m further from the runway compared to the last familiarization pattern.

compared with Mann-Whitney U tests between the groups. T -tests were used for comparing continuous data. The binary data in the pretest were compared using Chi squared tests. To reduce false positive findings, the performance aspects for which we predicted a positive effect of the *COOL* procedure (i.e., *Diagnosis* and *Decision-making*) were corrected for multiple (two) comparisons using Holmes-Bonferroni correction.

Results

Application and Perceived Usefulness of the COOL Procedure

The application of the *COOL* procedure was high according to self-report (Table 6). The full procedure was executed by 89.6% of pilots on average over the scenarios. *Observe* was executed most consistently, with 100.0% of pilots reporting application. This could be confirmed for all but three pilots in the audio recordings. For one of these three, *Observe* could not be heard, and the recordings of the remaining two pilots were lost. Three out of the nine pilots for whom *Observe* was confirmed did not strictly follow the instructions for *Observe*, as they called out the parameters' meaning (e.g., "*Speed is low*", "*Speed makes sense*", "*Speed is as I'd like it to be*") instead of the required value (e.g., "*Speed is 100*"). Pilots found the procedure generally useful, mostly in FLAP and STALL, and least in MASS (Table 6).

Interestingly, 60.0% of pilots in the control group called out instrument readings or aircraft behavior, similar to *Observe*, on their own initiative. However, unlike the experimental group, they made very few callouts, which were highly specific to the failure. On average over the scenarios, 25.0% pilots in the control group remained (nearly) entirely silent during the scenarios.

Examples of Application of the COOL Procedure

Tables 7 and 8 present two examples of pilots applying the *COOL* procedure. The transcripts are partly translated from Dutch. The first example (Table 7) is in ASU. *Observe* was performed by calling out the parameters' meaning instead of the values. The pilot appropriately interrupted the procedure multiple times to aviate or navigate. The second example in MASS (Table 8) shows a pilot starting to execute *Calm down* and *Observe* while the aircraft was still stalling. The pilot did not anticipate what would happen when selecting flaps, but the procedure seemed to increase attention to the possibility to control pitch with thrust. Afterward, the pilot indicated not knowing what the issue was.

A third example of *COOL* procedure application can be viewed in the video that is attached in the supplementary files (Appendix A). This example occurred when the pilot selected flaps 25 in MASS.

Performance in the Pretest

In the pretest, no significant performance differences between the groups were found. Eight pilots in both groups let the airspeed drop below V_{mca} (80 kt). We could not detect a significant difference in the time flown below V_{mca} , $U = 64.0$, $p = .638$. Three (experimental) versus four (control) experienced loss of aileron control, and one (experimental) versus two (control) did not land on the runway (p 's

Table 6. Self-reported application of the *COOL* procedure items by the experimental group, and perceived usefulness of the procedure.

	FLAP	MASS	STALL	ASU
Calm down (n)	11	10	12	12
Observe (n)	12	12	12	12
Outline (n)	12	10	12	12
Lead (n)	12	11	11	12
Full procedure (n)	11	9	11	12
Perceived usefulness median (1–5)	4	3	4	3–4

Table 7. An audio script showing an example of the COOL procedure being applied in ASU. Author comments are in [brackets].

Category	Pilot comments
(Aviate/navigate)	<i>The speed is not really increasing. Pitch is lower. Ok. So. Now the speed is increasing. I'm still going straight ahead. So, change our feet [on the rudder pedals]. So, we're going to climb.</i>
Calm down	[Pilot breathes deeply.]
Observe	<i>So, for now straight ahead. Thrust. Climbing 1000 ft [per minute]. Less pitch. Full thrust. Secondary: flaps are up, gear is up. Indications look normal.</i>
(Aviate/navigate)	<i>1000 ft [altitude]. Initiating slight turn.</i>
Outline	<i>It ... It could also be an airspeed indicator failure, but I'll check that later.</i>
(Aviate/navigate)	<i>Let's see, [heading] 230, there's the runway. Let's not climb too much.</i>
Outline	<i>Airspeed is ... Might be indication failure. This is difficult, it's in the stall [region] but it still doesn't stall.</i>
Lead	<i>Let's try to stall then. Or approach stall.</i>
Outline	<i>Yeah this isn't possible at 50 knots. ASI [air speed indicator] problem. So then ...</i>
(Aviate/navigate)	<i>Ho! We must keep it at 1000 ft.</i>
Lead	<i>Then I'm just going to fly pitch-power.</i>

Table 8. An audio script showing an example of the COOL procedure being applied in MASS. Author comments are in [brackets].

Category	Pilot comments
Calm down	[Mass shift occurs. Pilot breathes deeply.]
Observe	<i>Ok, pitch keeps increasing over 30 degrees, bank is still zero, I hear a stall warning. No, just 980 ft [altitude] now, climbing at 800 [ft per minute].</i>
Outline/Lead	<i>I want to recover after the stall warning, but my aileron seems unresponsive. Maybe if I decrease power a bit. [Pilot recovers.]</i>
Lead	<i>My plan is really to just keep flying and use my throttle as much as possible to control pitch. That seems to be working. With full power I cannot keep the nose down anyway.</i>
(Aviate/navigate)	<i>I'm going to go back to 1000 ft.</i> [Pilot selects flaps 25 in base leg.]
Observe	<i>Ok, we're still flying west, 1000 ft, speed is a little low, 60 degrees [heading], pitch just below the horizon at 2.5 degrees.</i>
Outline	<i>Again, I don't seem to be able to keep the nose down, or at the right position, so I have limited authority in my pitch axis.</i>
Lead	<i>I do have enough to turn to final later, and to land. So that is my plan. I notice that when I increase throttle, I cannot hold my pitch down. So, I'd like to keep the configuration as it is and see if I can keep enough speed to land, so I'll do that with flaps in the approach configuration, and maybe even flaps up if needed. Yeah, I'm going to try with flaps up. [Pilot selects flaps up.]</i>

> .500). There were also no significant differences in surprise, $U(22) = 69.0$, $p = .860$, startle, $U(22) = 54.0$, $p = .289$, and anxiety, $U(22) = 64.0$, $p = .644$.

Performance in the Posttest

The performance outcomes of the groups per performance aspect are shown in Figure 7.

The median adherence to the criteria defined for the aspect *Aviate* was significantly lower for the experimental group than for the control group, $U(22) = 33.5$, $p = .023$ (see, Table 9), indicating that the procedure caused less optimal immediate responses.

Of the four pilots in the experimental group who experienced an excessive bank angle in FLAP, two audio recordings were missing. Of the remaining two pilots, none executed *Observe* before recovering. Of the ten pilots in the experimental group who experienced an upset in MASS, also two audio recordings were missing. Five pilots out of the eight remaining cases executed *Observe* before recovering (e.g., the example in Table 8). Of the ten pilots in the experimental group who unloaded in STALL, also two audio recordings were missing. One pilot of the remaining eight cases executed

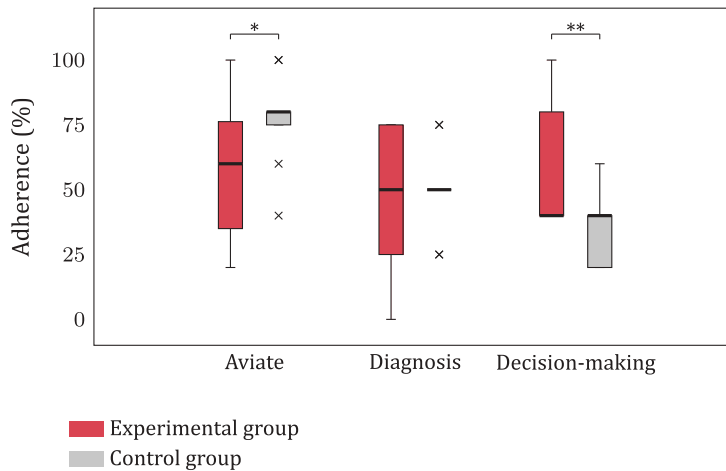


Figure 7. Tukey boxplots of adherence to the criteria defined for the three performance aspects. * $p < .05$; ** $p < .01$.

Table 9. Adherence to the criteria defined for Aviate. Adherence to A3 resulted in exclusion from A4.

	Group			
	Experimental		Control	
	N	%	N	%
A1. FLAP: prevented excessive bank angle	8	66.7	12	100.0
A2. FLAP: maintained speed	7	58.3	11	91.7
A3. MASS: prevented excessive pitch angle	2	16.7	4	33.3
A4. MASS: recovered quickly	4 (/10)	40.0	6 (/8)	75.0
A5. STALL: unloaded	10	83.3	11	91.7
Overall adherence, mean (<i>SD</i>)	12	54.6 (26.8)	12	77.5 (15.9)
Overall adherence, median	12	60.0	12	80.0

Observe before unloading. Thus, only in MASS, suboptimal immediate responses were perhaps caused by the pilots applying the procedure too soon.

Diagnosis

There was no significant difference between the groups in adhering to the criteria defined for *Diagnosis*, $U(22) = 67.0$, $p = .750$ (see, Table 10). In FLAP, those who did not identify a flap issue guessed that it was an aileron issue. In MASS, all but three pilots thought the issue was with the elevator or pitch trim. Three pilots in the experimental group stated that they did not know what the issue was, due to its reoccurrence when selecting flaps. In STALL, the two pilots who did not recognize the issue did not recognize the stall alarms and either linked the vibrations to aerodynamics or had no idea.

Decision-Making

The experimental group scored significantly higher than the control group on the performance aspect *Decision-making*, $U(22) = 28.0$, $p = .005$ (see, Table 11). In MASS, one pilot in the control group selected flaps 40 for landing, which led to loss of control in-flight and the scenario had to be ended prematurely to avoid a crash.

Table 10. Adherence to the criteria defined for performance aspect: Diagnosis.

	Group			
	Experimental		Control	
	N	%	N	%
D1. FLAP: identified issue	7	58.3	7	58.3
D2. MASS: identified issue	0	0.0	0	0.0
D3. STALL: identified issue	11	91.7	11	91.7
D4. ASU: decreased throttle in under two minutes	6	50.0	6	50.0
Overall adherence, mean (<i>SD</i>)	12	50.0 (26.1)	12	50.0 (15.1)
Overall adherence, median	12	60.0	12	60.0

Table 11. Adherence to the criteria defined for Decision-making. Selecting flaps 25 in downwind resulted in exclusion from DM3, whereas not selecting flaps 25 during the flight resulted in exclusion from DM4.

	Group			
	Experimental		Control	
	N	%	N	%
DM1. FLAP: did not select flaps 40	8	66.7	5	41.7
DM2. MASS: configured early	4	33.3	2	16.7
DM3. MASS: increased altitude	5 (/11)	45.5	1 (/11)	9.1
DM4. MASS: prevented 2nd upset	9 (/11)	81.8	7 (/11)	63.6
DM5. ASU: increased distance final	7	58.3	4	33.3
Overall adherence, mean (<i>SD</i>)	12	59.6 (25.3)	12	34.4 (12.3)
Overall adherence, median	12	40.0	12	40.0

Manipulation Checks and Stress Response

Table 12 shows the subjective ratings averaged over the four posttest scenarios. Pilots generally scored around the midpoint of the scales, indicating that the scenarios induced moderate startle, surprise and mental effort. Although not significant, the trends in surprise, perceived anxiety and mental effort were in the direction of higher scores in the experimental group. For surprise, no significant difference was expected. For anxiety, the trend was in the opposite direction of the hypothesis. For mental effort, the trend may indicate increased workload due to the *COOL* procedure.

Discussion

The startle and surprise management procedure tested in the current experiment had significant positive effects on pilot decision-making under startle and surprise. The elements of the procedure, which were to take a moment to relax oneself (if time allows), observe the whole situation, analyze the problem and then select and execute a course of action, thus appear to improve decision-making in startling and surprising events in-flight. It is not certain though which of these elements were the most important underlying factors of this performance difference.

In contrast to decision-making performance, the pilots' "aviate" actions (e.g., stopping an upset motion, upset recovery and paying attention to speed), were significantly less optimal. This indicates

Table 12. The means and standard deviations of the subjective measures.

	Group		<i>t</i>	<i>p</i>
	Experimental Mean (SD)	Control Mean (SD)		
Startle (0–10)	5.83 (1.60)	5.85 (2.13)	.03	.979
Surprise (0–10)	7.23 (1.06)	5.81 (2.55)	1.78	.089
RSME (0–150)	63.6 (15.9)	53.0 (14.7)	1.70	.104
Anxiety (0–10)	5.03 (1.90)	3.88 (1.51)	1.64	.115

that the procedure may have had a distraction effect. This distraction, as well as the monitoring of the experimental group's verbal analyses by the experimenters, were perhaps reasons why the experimental group did not report significantly lower anxiety, and why there was a trend to higher mental effort. In line with this potential distraction effect, we found that a large proportion of pilots (62.5%) started the *Observe* step in MASS before the upset was recovered. This happened despite explicit instructions given beforehand. It could be that inappropriate prioritizing of the *COOL* procedure was an artifact caused by the experiment, as pilots were perhaps unnaturally focused on trying out the procedure. Still, it seems advisable that the recognition of when to execute a startle and surprise management procedure is sufficiently practiced and tested. Adding first step of "Aviate" or "Fly" may also be beneficial.

The application of the procedure in the experiment was high and pilots rated the procedure on average as very useful. Some pilots remarked that the procedure may be more applicable in operational practice, since both time pressure and situation awareness are generally much lower if an issue were to occur in cruise. The *Calm down* and *Observe* steps were regarded as being the "core" of the procedure. A criticism was that the procedure was a bit too elaborate to execute when startled in a real situation. Some improvements to reduce workload suggested by the pilots were to: "Call out instruments' meaning instead of the absolute values in the *Observe* step, for a more natural feel", "Reduce the number of parameters to call out", "Remove *Outline* and *Lead*", and: "Let only the pilot monitoring perform *Observe* in a two-pilot crew". The *Observe* step can also be combined with reverting to a more static situation and known settings, as suggested by Gillen (2016), which may further decrease workload and stress.

The single-pilot nature of the tasks, the simulated environment, as well as using a small MEP aircraft model with limited functionalities for the experiment, limit the ability to generalize the results to operational practice of airline piloting. Receiving the tests immediately upon the training session is likely to have enhanced the positive, as well as the negative effects of the procedure on performance. Pilots indicated that the experimental scenarios were believable and generally challenging. Although the pilots had little experience in small MEP aircraft, almost all were able to respond sufficiently to the presented scenarios. Not all issues were easy to identify (e.g., MASS), however all performance criteria were selected to reflect responses that are appropriate even if the exact issue is not identifiable.

It is important to note that the manner in which the procedure was trained in this experiment was designed to ensure a comparable practice session of the groups, and it therefore does not reflect an optimal training session. In operational practice, training can be optimized by practicing the procedure at a higher frequency within a training session, and by presenting scenarios of an increasing level of stress and difficulty.

In conclusion, it appears that the tested mnemonic-type procedure positively influences pilot decision-making in situations of startle and surprise. The most useful elements of the tested procedure, pilots remarked, was a step to manage stress and one to observe the overall situation before analyzing the problem. It is advisable to improve on the tested procedure by reducing complexity, sharing workload, and to ensure a proper prioritization of immediate issues over procedure execution.

Disclosure Statement

No potential conflict of interest was reported by the authors.

References

- Boland, E. (2016). Managing startle & surprise. Presented at PACDEFF, Adelaide. <http://pacdeff.com/wp-content/uploads/2016/11/PACDEFF-Startle-Surprise-Management.pdf>
- Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile. (2012). *Final report on the accident on 1st June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France, flight AF 447 Rio de Janeiro – Paris*. <https://www.bea.aero/docspa/2009/f-cp090601.en/pdf/f-cp090601.en.pdf>
- Casner, S. M., Geven, R. W., & Williams, K. T. (2013). The effectiveness of airline pilot training for abnormal events. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 55(3), 477–485. <https://doi.org/10.1177/0018720812466893>

- De Muynck, R., & Hesse, M. V. (1990, June). *The a priori simulator software package of the Piper PA34 Seneca III*. (Technical report). TU Delft.
- Endsley, M. R. (1995). Measurement of situation awareness in dynamic systems. *Human Factors*, 37(1), 65–84. <https://doi.org/10.1518/001872095779049499>
- European Aviation Safety Authority. (2015). *Loss of control prevention and recovery training (Notice of proposed amendment 2015-13)*. <https://easa.europa.eu/system/files/dfu/NPA%202015-13.pdf>
- Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion*, 7(2), 336–353. <https://doi.org/10.1037/1528-3542.7.2.336>
- FAA Safety Team. (2018). Fly the aircraft first. *GA Safety Enhancement Topic Fact Sheets*. Federal Aviation Administration. https://www.faa.gov/news/safety_briefing/2018/media/SE_Topic_18-07.pdf
- Federal Aviation Administration. (2015). *Upset prevention and recovery training (Advisory circular No. 120-111)*. https://www.faa.gov/documentLibrary/media/Advisory_Circular/AC_120-111_CHG_1.pdf
- Field, J. N., Boland, E. J., Van Rooij, J. M., Mohrmann, J. F. W., & Smeltink, J. W. (2018). *Startle Effect Management*. (report nr. NLR-CR-2018-242). European Aviation Safety Agency.
- Gillen, M. W. (2016). *A study evaluating if targeted training for startle effect can improve pilot reactions in handling unexpected situations in a flight simulator*. The University of North Dakota.
- Hörmann, H.-J. (1996). Training of aircrew decision making. In *AGARD Conference proceedings*, (CP-588, 19). North Atlantic Treaty Organization.
- Houtman, I. L. D., & Bakker, F. C. (1989). The anxiety thermometer: A validation study. *Journal of Personality Assessment*, 53(3), 575–582. https://doi.org/10.1207/s15327752jpa5303_14
- Klein, G., Phillips, J. K., Rall, E. L., & Peluso, D. A. (2007). A data-frame theory of sensemaking. In *Expertise out of context: Proceedings of the sixth international conference on naturalistic decision making* (pp. 113–155). Lawrence Erlbaum.
- Koolstra, D. V. C. C., Herman, J., & Mulder, J. A. (2015). Effective model size for the prediction of the lateral control envelope of damaged aircraft. In *AIAA Modeling and Simulation Technologies Conference*. American Institute of Aeronautics and Astronautics. <https://doi.org/10.2514/6.2015-2036>
- Landman, A., Groen, E. L., Van Paassen, M. M., Bronkhorst, A. W., & Mulder, M. (2017a). Dealing with unexpected events on the flight deck: A conceptual model of startle and surprise. *Human Factors*, 59(8), 1161–1172. <https://doi.org/10.1177/0018720817723428>
- Landman, A., Groen, E. L., Van Paassen, M. M., Bronkhorst, A. W., & Mulder, M. (2017b). The influence of surprise on upset recovery performance in airline pilots. *The International Journal of Aerospace Psychology*, 27(1–2), 2–14. <https://doi.org/10.1080/10508414.2017.1365610>
- Landman, A., van Oorschot, P., van Paassen, M. M., Groen, E. L., Bronkhorst, A. W., & Mulder, M. (2018). Training pilots for unexpected events: A simulator study on the advantage of unpredictable and variable scenarios. *Human Factors*, 60(6), 793–805. <https://doi.org/10.1177/0018720818779928>
- Marine Corps, U. S. (2010). *Combat and operational stress control*. (nr. 144 000083 00). Marine Corps Reference Publication.
- Martin, W. L. (2017). *Developing startle and surprise training interventions for airline training programs*. Presented at PACDEFF 2016, Melbourne. <http://pacdeff.com/wp-content/uploads/2017/08/PACDEFF-FC-Forum-Presentation-on-Startle.pdf>
- Martin, W. L., Murray, P. S., Bates, P. R., & Lee, P. S. (2016). A flight simulator study of the impairment effects of startle on pilots during unexpected critical events. *Aviation Psychology and Applied Human Factors*, 6(1), 24–32. <https://doi.org/10.1027/2192-0923/a000092>
- Murray, S. R. (1997). Deliberate decision making by aircraft pilots: A simple reminder to avoid decision making under panic. *The International Journal of Aviation Psychology*, 7(1), 83–100. https://doi.org/10.1207/s15327108ijap0701_5
- Paul, G., Elam, B., & Verhulst, S. J. (2007). A longitudinal study of students' perceptions of using deep breathing meditation to reduce testing stresses. *Teaching and Learning in Medicine*, 19(3), 287–292. <https://doi.org/10.1080/10401330701366754>
- Pelka, M., Heidari, J., Ferrauti, A., Meyer, T., Pfeiffer, M., & Kellmann, M. (2016). Relaxation techniques in sports: A systematic review on acute effects on performance. *Performance Enhancement & Health*, 5(2), 47–59. <https://doi.org/10.1016/j.peh.2016.05.003>
- Rivera, J., Talone, A. B., Boesser, C. T., Jentsch, F., & Yeh, M. (2014). Startle and surprise on the flight deck: Similarities, differences, and prevalence. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 58, pp. 1047–1051). <https://doi.org/10.1177/1541931214581219>
- Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1970). *Manual for the state-Trait anxiety inventory*. Consulting Psychologists Press.
- Stroosma, O., Van Paassen, M. M., & Mulder, M. (2003). Using the SIMONA research simulator for human-machine interaction research. In *AIAA modeling and simulation technologies conference and exhibit* (p. 5525).
- Walters, A. (2002). *Crew resource management is no accident*. Aries.
- Zijlstra, F. R. H., & Van Doorn, L. (1985). *The Construction of a Scale to Measure Perceived Effort*. Technical Report. Delft University of Technology.