TUDelft

Delft University of Technology

Managing startle and surprise in the cockpit

Landman, Annemarie

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Managing Startle and Surprise
in the Cockpit

Managing Startle and Surprise
in the Cockpit

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. dr. ir. T. H. J. J. van der Hagen,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op woensdag 4 september 2019 om 12:30 uur

door

Annemarie LANDMAN

Master of Science in Kinesiology,
VU University, Amsterdam, The Netherlands,

geboren te Bunschoten-Spakenburg, Nederland

Dit proefschrift is goedgekeurd door de

promotor: dr. ir. M. M. van Paassen
promotor: prof. dr. E. L. Groen

Samenstelling promotiecommissie:

Rector Magnificus,	voorzitter
Dr. ir. M. M. van Paassen,	Technische Universiteit Delft, promotor
Prof. dr. E. L. Groen,	TNO Soesterberg, promotor

Onafhankelijke leden:

Prof. dr. D. B. Kaber,	University of Florida at Gainesville, USA
Prof. dr. J. M. C. Schraagen,	Universiteit van Twente
Dr. H.-J. Hörmann,	DLR, Germany
Prof. dr. J. Dankelman,	Technische Universiteit Delft
Prof. dr. ir. H. A. P. Blom	Technische Universiteit Delft

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

To my mother.
You will always be missed and loved.

Contents

Summary

After several recent flight safety events, such as the accident of Air France flight 447 in 2009, investigators determined that surprise and startle can severely disrupt pilot responses. They concluded that pilots need to be better prepared for unexpected and potentially startling situations. In response, aviation safety authorities have recommended and mandated that startle and surprise should receive more attention in pilot training. However, there is insufficient scientific data available on pilots' behavior in startling and surprising situations, and on how they can best be trained for these situations. This thesis addresses this problem, by studying startle and surprise in pilots, and by investigating which training interventions can strengthen the pilots' response to unexpected situations.

One of the tools developed in this thesis is a conceptual model of the effects of startle and surprise on pilot performance and sensemaking. The model uses the concept of "frames", which are knowledge structures, similar to mental models, with regard to, for instance, situations or systems. An unexpected situation requires an adaptation or change of one's frame to recognize, understand and explain the events. Frames are based on previous experiences and are stored in long-term memory. Frames include knowledge of how situations evolve or how things work, what can be expected and which behavior is appropriate. Frames aide a person in directing attention to relevant information and responding automatically, thereby allowing for the conservation of mental energy. When an inconsistency is detected between an observation and the activated frame, one experiences a surprise. One then has to adapt, or even completely switch ("reframe"), the activated frame so that the observation and its implications can be understood.

If an unexpected situation is quickly appraised as posing a threat, it is likely to induce a startle response. A startle is a quick, defensive response, increasing stress and readying the body for fight or flight. Events with a more slowly evolving threat can cause stress and surprise, but do not necessarily involve startle. Stress hypothetically interferes with the reframing process by increasing attentional focus on stimuli in the environment (bottom-up) and preoccupying working memory. This thesis attributes the confusion associated with unexpected events to a frame mismatch, and not necessarily to being overwhelmed by acute stress. Reframing may also be impaired by other factors, like fatigue, decreased situation awareness, insufficient system knowledge or suboptimal interface design.

One prediction of the conceptual model is that the degree in which an event is unexpected will affect the level of confusion and the impact on pilot performance. To test this, a simulator experiment was performed in which the expectation of an upset event was manipulated. Even though the pilots received a refresher training on the specific upset event just before the test, their adherence to the recovery procedure was significantly worse when the test event occurred unexpectedly compared to expectedly.

The pilots also indicated increased surprise and confusion in the unexpected condition, showing that the manipulation was successful. Interestingly, none of the twenty participating pilots refrained from rolling the wings level to give precedence to unloading, whether the event was expected or unexpected. This indicates that intuitive responses are difficult to suppress when startled and surprised.

Another prediction of the model is that the active frame influences how information is being perceived and interpreted. This was confirmed both in a simulator experiment and in an in-flight experiment, in which participants were misled with regards to the bank angle of the aircraft. This induced confusion about the bank angle, also known as "the Leans", a form of spatial disorientation that is prevalent in aviation. The results show that one's expectation, or frame, about the bank angle elicit misinterpretations of the artificial horizon and incorrect control inputs when asked to level the aircraft. The incorrect responses were too quick to be caused by being overwhelmed by startle and surprise, and originated instead from a frame mismatch. In some cases, the correction of the response took a considerable amount of time, indicating that frame mismatch sometimes leads to reframing issues.

Besides investigating the mechanisms of startle and surprise, several experiments were executed to investigate potentially effective training interventions. In the first of these experiments, we looked into the potential advantage of variable and unpredictable training scenarios. One group of ten pilots practiced responses to events under more varying circumstances, in a mixed order of exercises, and with little information being given beforehand. A control group of ten pilots practiced the same responses in scenarios featuring constant circumstances, in a grouped order of exercises, and while always being told beforehand what would happen and how to respond. The results showed that the experimental group was more likely to apply the practiced skills in a novel situation, which led to more successful landings compared to the control group. The outcomes warn against pilot training designs which only feature scenarios that are highly scripted and predictable. It is in unpredictable circumstances that pilot reframing skills are truly practiced.

A second training intervention that was tested was a checklist that can be applied when a surprising and startling event happens. Compared to variable and unpredictable training, this intervention is applicable to a broader range of events, even events that are completely new and untrained. A simulator experiment was set up to test the effectiveness of such a checklist. A checklist was developed, consisting of four steps. It started with a moment of active relaxation, followed by calling out basic flight parameters, outlining the issue and formulating a plan. An experimental group of twelve pilots received training that included the checklist, while a control group of another twelve pilots received training without the checklist. Both groups then performed four startling and surprising test scenarios. A manipulation check showed that the test scenarios were surprising, startling and mentally demanding. The outcomes showed that immediate responses in the experimental group were impaired, which could be explained by the finding that pilots had a tendency to apply the checklist too early. However, long-term planning and proactive decision-making were significantly better in the experimental group. This suggests that these type of checklists are helpful, but that it is important to keep it simple and to practice applying it at the correct moment.

Whereas both aforementioned training interventions originate from research, a third intervention was evaluated in collaboration with an airline and proposed by the aviation industry. This intervention method consisted of a slow, non-verbal and goal-directed action to be applied by pilots when feeling overwhelmed by startle and surprise. The action was to turn one's head from the side window, over the instruments, and ending with facing one's fellow pilot. This was intended to help with obtaining an overview, preventing immediate actions, reattaining goal-directed functioning and checking with one's fellow pilot. The intervention was introduced to pilots in a recurrent training session, but very few pilots applied it in the subsequent simulator scenario. Their feedback indicated low appreciation of the method. This could indicate that the non-verbal method was less effective than the previously tested checklist-based intervention. However, there were several important differences between the manner in which the startle management methods were taught and between the experimental designs. There was, in particular, a difference in how extensively the method was explained and practiced. These and other differences were analyzed to identify potential pitfalls when implementing a training intervention in practice.

In conclusion, performance issues due to startle and surprise stem from the need to reframe under pressure. It is possible to induce startle and surprise with simulator scenarios, and elicit the performance issues. This indicates that simulator training can be used to practice responses under startle and surprise. It follows that a training intervention should focus on facilitating reframing under pressure. This can be done with variable and unpredictable training, and/or by providing pilots with a checklist that aides them with stress management and reframing. To ensure pilot appreciation and application of such a method, it is important to explain the purpose of a method thoroughly, and to practice the method in the simulator.

The next step would be to test the training interventions, that were shown to be effective in research environments, in pilot training practice. The effects of the training interventions on performance should then first be checked in training simulators and second with retrospective research in operational practice.

Samenvatting

Na verschillende recente veiligheidsincidenten in the luchtvaart, zoals het ongeval van Air France vlucht 447 in 2009, stelden onderzoekers vast dat schrik en verrassing de reacties van een bemanning ernstig kunnen belemmeren. Zij concludeerden dat vliegers beter moeten worden voorbereid op onverwachte en mogelijk schokkende situaties. In reactie daarop hebben luchtvaartautoriteiten aanbevolen en bepaald dat schrik en verrassing meer aan bod moeten komen in de training. Desalniettemin zijn er onvoldoende wetenschappelijke gegevens beschikbaar over het gedrag van vliegers in schokkende en verrassende situaties, en over hoe men het beste voor deze situaties zou kunnen trainen. Deze thesis adresseert dit probleem door schrik en verrassing bij vliegers te bestuderen, en door te onderzoeken welke trainingsmethoden de reacties van vliegers kunnen verbeteren in onverwachte situaties.

Één van de middelen die in dit onderzoek ontwikkeld is, is een conceptueel model van de effecten van schrik en verrassing op het presteren en betekenisgeven door vliegers. Het model gebruikt het concept van "frames"(kaders). Dit zijn kennisstructuren, lijkend op mentale modellen, met betrekking tot bijvoorbeeld situaties of systemen. Frames zijn gebaseerd op eerdere ervaringen, en zijn opgeslagen in het lange-termijn geheugen. Frames omvatten kennis over hoe bepaalde situaties zich ontwikkelen of hoe dingen werken, wat men kan verwachten, en welk gedrag gepast is. Dit helpt men om de aandacht te richten op de dingen die belangrijk zijn en om automatisch te kunnen reageren, om op die manier mentale inspanning te minimaliseren. Wanneer er een inconsistentie is tussen een observatie en de verwachtingen gebaseerd op het frame, ervaart men een verrassing. Deze alarmeert de persoon dat er wellicht een aanpassing of wisseling van het frame nodig is ("reframen"). Deze aanpassing kan oppervlakkig zijn (situationeel) of dieper (lange termijn kennis). Als de aanpassing slaagt, kunnen de observatie en haar implicaties verklaard worden.

Als de verrassende observatie direct wordt ingeschat als een dreiging, dan zal het waarschijnlijk een schrikreactie veroorzaken. Schrik is een snelle, defensieve reactie die stress opwekt en het lichaam klaarmaakt om te vechten of te vluchten. Gebeurtenissen waarin een dreiging zich langzaam ontwikkelt zullen ook stress en verrassing veroorzaken, maar niet per se schrik. Stress kan verondersteld worden te interfereren met het reframing proces, doordat het een focus van aandacht bevordert op stimuli in omgeving (bottom-up) en het werkgeheugen bezet. Dus, in de thesis wordt verwarring in noodsituaties verklaard als veroorzaakt door een discrepantie tussen het frame en de situatie, en niet door het per se overweldigd raken door stress. Het reframen kan ook belemmerd worden door andere factoren, zoals vermoeidheid, verminderd situationeel bewustzijn, onvoldoende (systeem)kennis, of suboptimaal ontwerp van interfaces.

Een voorspelling van het model is dat de mate waarin een gebeurtenis niet geanticipeerd is, bepaalt hoezeer er verwarring en prestatieverslechtering plaats vindt. Om

dit te testen werd een simulatorexperiment uitgevoerd, waarin de anticipatie van een upset gebeurtenis gemanipuleerd werd. Ondanks dat de vliegers net een opfristraining gekregen hadden, hielden ze zich significant minder aan een procedure, wanneer ze reageerden op een gebeurtenis die niet, in plaats van wel, geanticipeerd was. Ze gaven ook meer verrassing en verwarring aan in de niet-geanticipeerde conditie, wat aangeeft dat de manipulatie succesvol was. Interessant was dat alle twintig vliegers de vleugels direct probeerden recht te rollen, of de situatie nu geanticipeerd of niet geanticipeerd was, terwijl dit niet volgens de procedure is. Dit geeft aan dat intuïtieve reacties onder schrik en verrassing lastig te voorkomen zijn.

Een andere voorspelling van het model is dat het actieve frame beïnvloedt hoe informatie wordt geïnterpreteerd. Dit werd bevestigd in een simulator experiment en in een vliegexperiment, waar deelnemers werden misleid met betrekking tot de rolhoek van het vliegtuig. Deze verwarring, genaamd de "Leans" is een vorm van ruimtelijke desoriëntatie die veel voorkomt in de luchtvaart. De resultaten lieten zien dat de verwachting over de rolhoek, gebaseerd op het frame, misinterpretaties veroorzaakte van de kunstmatige horizon en stuurfouten bij het recht rollen van het vliegtuig. In een paar gevallen kostte het deelnemers zeer veel tijd om de eerste foutieve reactie te corrigeren, wat aangeeft dat de verrassing soms leidde tot verwarring.

Naast het onderzoeken van schrik en verrassing, werden er ook verschillende experimenten uitgevoerd om mogelijk effectieve training interventies te testen. Met het eerste experiment hiervan, onderzochten we het mogelijke voordeel van variabiliteit en onvoorspelbaarheid in simulator trainingsscenarios. Één groep van tien vliegers oefende reacties op storingen in meer variërende omstandigheden, in een gemengde volgorde van oefeningen en zonder voorkennis. Een controlegroep van tien vliegers oefende dezelfde reacties in scenarios met eenzijdige omstandigheden, in een gegroepeerde volgorde van oefeningen, en met de gebeurtenissen en vereiste reacties van tevoren aangekondigd. De resultaten toonden aan dat de experimentele groep de getrainde vaardigheden meer toepaste in een nieuwe situatie, om daardoor succesvol te landen. Deze uitkomsten waarschuwen ons tegen het aanbieden van trainingsscenarios die erg eenzijdig en voorspelbaar zijn. Alleen in onvoorspelbare omstandigheden kunnen vliegers reframing vaardigheden echt oefenen.

Een tweede trainingsinterventie die getest werd, was een checklist die toegepast kan worden wanneer men schrikt of verrast is. Vergeleken met variabele en onvoorspelbare training, is deze interventie breder toepasbaar, ook bij gebeurtenissen die totaal nieuw en ongetraind zijn. Een simulatorexperiment werd uitgevoerd om te testen of zo'n checklist daadwerkelijk effectief is. De geteste checklist bestond uit vier stappen. Het startte met een moment van actieve relaxatie, gevolgd door een call-out van de algemene vliegparameters, het vaststellen van het probleem en het formuleren van een plan. Een experimentele groep van twaalf vliegers ontving training met de checklist, terwijl een controlegroep van nog eens twaalf vliegers dezelfde training ontving zonder de checklist. Beide groepen voerden vervolgens vier schrikwekkende en verrassende scenarios uit. De uitkomsten suggereerden dat de checklist de eerste reacties van vliegers belemmerde. Dit kwam overeen met een andere uitkomst, dat vliegers de neiging hadden om de checklist te vroeg toe te passen. Desalniettemin waren er significante verbeteringen in de experimentele groep in lange-termijn plannen en proactief

beslissingen nemen. Dit suggereert dat dit soort checklists nuttig zijn, maar dat het belangrijk is om het eenvoudig te houden en te oefenen met toepassingen op het juiste moment.

Waar de eerdere twee trainingsinterventies vanuit een wetenschappelijke benadering ontwikkeld werden, werd een derde interventie geëvalueerd in samenwerking met een vliegmaatschappij, en was deze voorgesteld door de luchtvaartindustrie. Deze interventie bestond uit een langzame, non-verbale, doelgerichte actie die men kon uitvoeren wanneer men zich overweldigd voelde door schrik en verrassing. De actie bestond uit een langzame draaiing van het hoofd, vanaf kijkend uit het zijraam, over de eigen instrumenten, over de instrumenten van de medevlieger, naar de medevlieger. De interventie werd geïntroduceerd in een periodieke trainingssessie, maar zeer weinig vliegers pasten het toe in een verrassend simulatorscenario dat volgde. De feedback van vliegers gaf aan dat ze de interventie over het algemeen niet waardeerden. Dit suggereert dat de methode minder effectief was dan de eerder geteste checklist. Maar er waren ook belangrijke verschillen in de manier waarop de methode werd aangeboden die het verschil kunnen verklaren. Er was met name een verschil in de uitgebreidheid van de uitleg en van het oefenen met de methode. Deze en andere verschillen werden geanalyseerd om mogelijke valkuilen te identificeren bij het implementeren van een trainingsinterventie in de praktijk.

In conclusie, prestatieproblemen bij schrik en verrassing komen voort uit de noodzaak om te moeten reframen onder druk. Het is mogelijk om de verrassing en prestatieproblemen op te wekken met simulatorscenarios, wat suggereert dat simulatortraining gebruikt kan worden om reacties onder schrik en verrassing te oefenen. Een trainingsinterventie zou dus ook gericht moeten zijn op het faciliteren van reframen onder druk. Dit kan gedaan worden door middel van variabele en onvoorspelbare training, en/of door vliegers een checklist te geven die hen helpt om stress te managen en te reframen. Om waardering en toepassing van zulk een methode te waarborgen is het belangrijk om het doel ervan goed uit te leggen, en om de methode in de simulator te oefenen.

De volgende stap zou zijn dat de trainingsinterventies, die effectief bleken in de experimenten, getest worden in de trainingspraktijk. De effecten van de interventies op het presteren zouden dan eerst gecheckt moeten worden in trainingssimulatoren en vervolgens in retrospectief onderzoek in de praktijk.

1

Introduction

1.1. Background

Technical advances in aviation have greatly improved safety over the years. Since 1970, the ratio of fatalities per (passenger × distance) has decreased by a factor of 54 [1]. Unfortunately, however, accidents are still occurring. As the pilot's role has shifted from actively controlling the aircraft towards monitoring automation, the causal factors in accidents have shifted as well. New issues have emerged, involving cooperation between the pilot and the automation. Efforts to ensure resilience of the human-automation interactive system are therefore highly relevant at the present time.

One issue here is that the situations which cannot be handled by automation, and which thus require human intervention, are typically unforeseen and complex, demanding quick judgment and decision making [2]. Such situations may arise after long periods of automated flight, which can be difficult as pilots suddenly need to switch from a passive to an active role [3, 4]. At the same time, automation may decrease the transparency of the flying process to the flight crew. If the system is malfunctioning, it may not be immediately clear which information the system is using, how it is using this information, and why it is taking certain actions. This can lead to automation surprises [5, 6], in which the automation does something which the crew does not expect or understand. Furthermore, pilots may be hesitant to intervene and take manual control due to having become complacent with the automation [7]. If intervening, pilots' manual flying skills may have eroded due to extensive use of automation [8]. Thus, the rarity of unsafe events can actually make it more difficult for pilots to intervene and solve the unsafe events that do occur.

As can be seen in Figure 1.1, loss of control in-flight currently forms the largest category of fatal accidents. In most cases, these loss of control in-flight situations involve some time and opportunity to respond to the problem after pilots become aware of it. It has been recognized for some time that such responses require specific crisis management skills. An important development in the targeted training of such skills was the implementation of crew resource management (CRM) in the 1980s and 1990s [9].

1

1

CRM is a set of training procedures that focuses on interpersonal skills, self awareness, problem solving and decision making, which are aimed at preventing and reacting to unsafe situations.

Fatalities by CICTT Aviation Occurrence Categories
Fatal Accidents | Worldwide Commercial Jet Fleet | 2008 through 2017
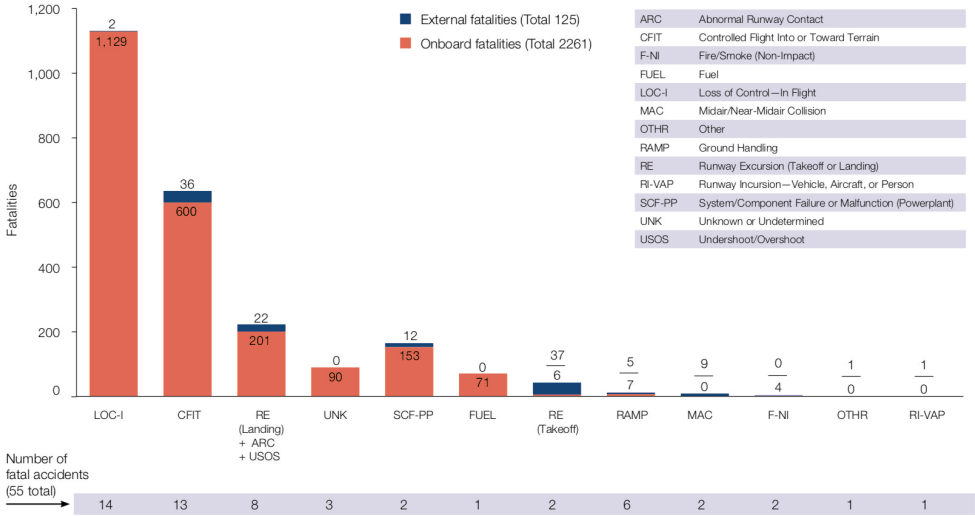
Figure 1.1: Fatalities of worldwide commercial jet fleet between 2008 and 2017, by categories defined by the Commercial Aviation Safety Team / ICAO Common Taxonomy Team. Source: [10].

Since then, however, several events have occurred which have revealed that the unexpectedness of unsafe situations can induce a "startle factor", which can severely complicate the crew's troubleshooting [11–15]. A key accident in this regard was that of Air France flight 477 in 2009. The final report, published in 2012, reads [16]:

> "The startle effect played a major role in the destabilisation of the flight path and in the two pilots understanding the situation. Initial and recurrent training as delivered today do not promote and test the capacity to react to the unexpected. Indeed the exercises are repetitive, well known to crews and do not enable skills in resource management to be tested outside of this context. All of the effort invested in anticipation and predetermination of procedural responses does not exclude the possibility of situations with a "fundamental surprise" for which the current system does not generate the indispensable capacity to react. The rapid increase in crew workload in an unusual and unexpected situation led to the degradation of the quality of communication and coordination between the pilots." (page 209).

The accident investigators recommended the European Union Aviation Safety A-gency (EASA) to review the requirements for initial and recurrent training, and to is-sue new guidelines of integrating targeted training for startle and surprise. They also recommended high simulator fidelity for reproducing abnormal upset situations, and the introduction of surprise in training scenarios so that pilots are exposed to it. This has led to changes in EASA's requirements for CRM in 2016 [17], and for upset preven-tion and recovery training (UPRT) in 2018-2019 [18]. Meanwhile, the Federal Aviation Administration (FAA) issued an Advisory Circular in which they encourage UPRT in-structors to: *"be inventive and introduce various ploys to achieve a startle or surprise response in simulation"* (page 14) [19].

Despite these regulatory changes and recommendations, it is still unclear in which way startle and surprise should be integrated in the training. Exposing pilots to startle and surprise in the simulator is one possible way, but there is no data about the effec-tiveness (or counter-effectiveness) of such interventions. The same is true for prac-ticing specific startle and surprise management techniques, or for providing merely theoretical training on the subject.

Airline pilot recurrent training involves approximately 16 hours in the simulator and 8 hours of theory per year. Half of this time is spent on true training (i.e., learning and practice), and the other half is spent on testing or checking. If startle and surprise are to be induced in the simulator, this would be appropriate in the training section instead of in the checking section, since the latter is necessarily highly standardized (which makes the events well known to pilots). Complicating the matter is that pi-lot performance data are also collected in the training section, meaning that there are standardization requirements for this section as well. This limits the possibility to of-fer different scenarios to different pilots, which makes the scenarios very predictable as pilots often share information among each other.

In 2015, EASA assigned a tender on "startle effect management" to a different con-sortium than our own [20], showing that the issue is receiving attention. During that same year, the work on the current thesis started. It was then entitled: "Inducing star-tle response in flight crew". This title already reveals a certain mindset, which is that the main issue is startle, and that exposing pilots to it in the simulator may be a solu-tion. However, during the course of this project, new insights caused the focus to shift from startle to surprise, and startle exposure as a solution was let go. To explain the difference between startle and surprise, and to outline the scope of this thesis, the two concepts will first be defined.

1.2. Startle and surprise

The terms startle and surprise are often used interchangeably in everyday life, as well as in aviation operational practice [21]. However, it is for this thesis important to dif-ferentiate between the responses, as they have different causes and effects. A "startle" or "startle response" is a brief, fast, and highly physiological reaction to a sudden, in-tense, or threatening stimulus, such as the sound of a pistol shot [22, 23]. Aspects of startle include eye blinks, contraction of facial and neck muscles, arrest of ongoing be-haviors, increased physiological arousal, and emotions of fear or anger. The reaction is extremely fast, with the first measurable responses starting at 10-20 ms post-stimulus.

Following the first reflex, people tend to inspect the startling stimulus to determine if a threat is really present. If the situation is determined to be safe, then the startle was a false alarm and our physiological arousal and stress will subside. In contrast, if the stimulus is perceived to indicate a threat, the stress response will remain and potentially increase in intensity.

The stress response starts in the amygdala, which is involved in emotional processing. The amygdala communicates with the hypothalamus, which activates the sympathetic nervous system by controlling the release of stress hormones: adrenaline/epinephrine, noradrenaline/norepinephrine and later cortisol. This prepares our body to respond to the threat by fighting or fleeing. The heart rate and rate of breathing increase (see, Figure 1.2a), blood sugar is released into the blood, pupils dilate, alertness increases and hand palms become more sweaty. Cognitively, this stress response is thought to inhibit the functioning of the goal-directed (top-down) attentional system, and to facilitate that of the stimulus-driven (bottom-up) attentional system [24]. As a result, it becomes more difficult to focus on plans, goals, problem-analysis and task-relevant stimuli, and to prevent distraction by task-irrelevant stimuli.
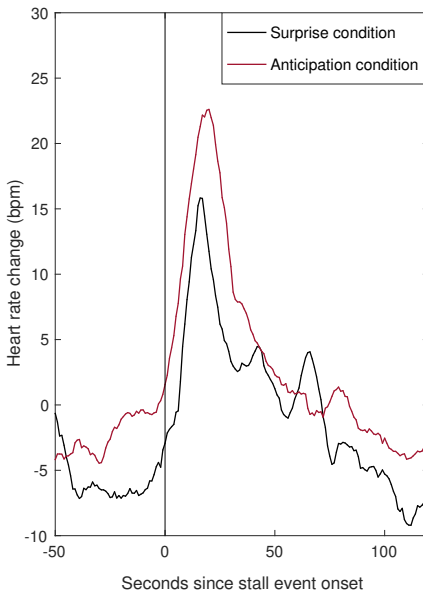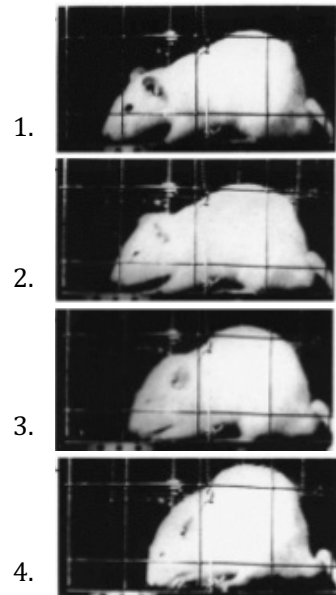
(a)   (b)

Figure 1.2: (a) The filtered heart rate response (relative to the preceding two minutes) of pilots during a simulated stall event (see, Chapter 3). Although the surprising stall event was rated as significantly more startling than the anticipated event, both events were stressful and there was no significant difference in the heart rate response. (b) A rat displaying a startle response (from: [25]).

Startle research stretches back to the start of the 20th century [26], when it was performed mainly within a framework of mechanistic behaviorism and involved stimu-

lus-response experiments. By far, most studies have been done in rats. For example, Figure 1.2b depicts an experimental setup with a rat being startled by a loud noise. By measuring the intensity of the response, researchers have, for instance, investigated brain functions or medications. A startling stimulus can also be used to measure a subject's pre-stimulus state such as fearfulness, which intensifies the startle response (fear-potentiated startle). A repetitive exposure to a startling stimulus in short succession will cause the response to diminish (habituation), but the response will recover in strength after some time of non-exposure.

In human beings, the sudden realization that there is a serious problem may also elicit a "startle" without an external stimulus. However, whether such a reaction is truly a startle response or simply a quick increase of stress, is debatable. Although unexpectedness increases the intensity of the startle response, it is not a prerequisite [22, 27–29]. As an example of the contrary, startles or so-called "jump scares" in movies work best when they are preceded by a build-up in tension (fear-potentiation), compared to when they appear out of nowhere. A typical example of an event in aviation that is startling but not unexpected would be a lightning strike when flying in a thunderstorm.

Whereas startle is reflex-like and related to the intensity and threat of an event, surprise is a slower emotional and cognitive response to unexpected events that are (momentarily) difficult to explain [30–33]. Surprise stimulates investigation and possibly a change in understanding of the situation. Research into this field is relatively young, and surprise is somewhat more difficult to instill reliably compared to startle. Meyer et al. [31] describe surprise in the framework of schema theory, as being the result of a mismatch between what is perceived and the activated schema. According to schema theory, perception, action, emotions and thoughts are controlled by structures of implicit knowledge (or theories) about situations.

The evolutionary function of surprise is to enable the brain (through action interruption) and to provide the motivational drive (through curiosity) to analyze the event and update the schema. Surprise may occur in the absence of startle, for instance when an event is simply odd and curious. Although surprise, like startle, increases arousal and draws attention to its cause, it does so in a more orienting manner (i.e., the orienting response) and less in a defensive or "flinching" manner [34]. Examples of highly surprising events in aviation include technical failures or automation actions that are "baffling" and difficult to explain.

Similar to the general psychological literature, research focused on startle in the domain of aviation is older than research focused on surprise. In the 1960s and 1970s, Thackray showed that pilot tracking performance was disrupted for at least 30 seconds following a startling stimulus [23]. Since the 2000s, there has been a renewed interest in startle as well as surprise in aviation literature (e.g., [13, 35–37]). Similar renewed interest can be found in the domain of health care, which involves comparable challenges requiring crisis management (e.g., [38, 39]). Evidence of the effectiveness of pilot training interventions for startle and surprise is still lacking, although there are promising indications with regards to discussing hypothetical events among pilots [40], and managing startle and surprise through a brief checklist [41].

The differences between startle and surprise have also been discussed in aviation

literature [21]. In light of the recent regulatory changes, these differences have become more relevant, because it has implications for the effectiveness of training interventions. If startle is the main cause of performance impairments, then training interventions may focus more on measures like relaxation or startle exposure to perhaps desensitize the pilots. On the other hand, if surprise is the main problem, then training interventions could focus more on measures to help pilots to make sense of situations that seem inexplicable.

1.3. Scope

The first limitation that stands out when investigating startle and surprise in a simulated setting, is that it is impossible to capture the levels of startle and surprise that are present in real emergency situations. There are two reasons for this. First, participating pilots will not assume that they enter the simulator to perform an uneventful flight. By knowing that they are in an experiment, there will always be a certain preparedness for "something" to happen. A second reason is due to ethical reasons. Participants cannot be mislead to feel unsafe, and they must be informed about possibly unpleasant experiences like startle before agreeing to participate.

The fact that pilots know that they are participating in an experiment does not mean that they cannot be surprised at all. They can, for instance, be misled to expect a different upcoming event. Events can also be chosen with which pilots are unfamiliar. Furthermore, even when pilots are not surprised that an event occurs, they can still experience surprise by the type of the event, and the timing. Periods of uneventful flight are deliberately included in the scenarios to decrease alertness. To check if our manipulation of surprise is successful, pilots rate their level of surprise on Likert-type scales. These scales are not validated, but they still provide insight into whether scenarios were surprising or not.

To somewhat compensate for the unrealistic high level of alertness in the simulator compared to reality, measures are taken to increase task difficulty, stress and workload. For instance, pilots are distracted with distraction tasks, different failures occur simultaneously, or decision options are limited to ensure time-pressure. These measures may in some cases make the scenario events seem somewhat extreme, coincidental, or unrealistic. The reader should keep in mind that this is done to compensate for the experimental setting. Nevertheless, the aim is not to make the scenarios as difficult as possible. If all pilots fail in a scenario, this will not provide us with many useful data. Control groups or control conditions are included in the experiments to test if the developed tasks are in principle "doable".

A second limitation is that it is not possible to surprise or startle pilots using the same event multiple times within a reasonable timespan. They will likely learn to expect what is going to happen or desensitize to the startling stimulus. However, if we present a surprise event only once in an experimental session, our ability to eliminate unwanted variance, for instance due to luck, is highly limited. Other measures are therefore used to eliminate variance in performance, like using within-subject comparisons of a straight-forward procedure (Chapter 3, limiting decision-making opportunities (Chapters 5 and 6), and combining performance outcomes of several different test scenarios (Chapter 6).

As the current thesis aims to provide recommendations for training design for the commercial air transport industry, the research is focused on airline pilots. As much as possible, we attempt to include airline pilots as participants in the experiments, and use research simulators or certified training simulators. The aircraft models we employ, in combination with the experimental tasks, are selected to be manageable by the participants, considering their experience. Pilots with a military background are likely to have had extensive training time with regard to managing startle and surprise, so they are excluded from our experiments.

The effects of interventions and manipulations on pilot responses are tested using within- or between-subject designs. This means that we are interested in performance differences between conditions or groups, instead of in pilots' absolute level of performance. Measuring pilots' absolute performance in a valid manner would require type-specific simulators and evaluation by certified type-rating instructors, which was outside the scope of this thesis.

The selection of the investigated training interventions is based on literature, insights obtained from our experiments, and/or opinions from the industry. The selected training interventions are not meant to be a complete list. Other effective interventions can be imagined that are not included in this thesis. Also, other intervention methods besides training, like pilot selection, reduction of fatigue, or improvements in interface design, can be imagined, but these fall outside of the scope of the current thesis.

1.4. Research objectives and key questions

An overview of the research objectives and key questions is shown in Table 1.1.

The thesis consists of two general parts, each one focusing on a research objective. Our first research objective is to obtain more insight into the mechanisms that cause pilot performance issues in startling and surprising situations. Increased understanding of these mechanisms is relevant for the development of intervention measures. One category of such intervention measures, on which our second research objective is focused, is training interventions. Our second research objective is to identify effective simulator training interventions for startle and surprise.

For the first research objective, the first key question we attempt to answer is: How do startle and surprise cause pilot performance issues in unexpected situations, according to literature? This question is investigated by reviewing the literature and by creating a conceptual model of startle and surprise. These insights are used to design the simulator- and in-flight experiments in the rest of this thesis.

The second key question within the first research objective is: Can we induce surprise and ensuing performance issues in the simulator? A simulator experiment is set up to test if we can surprise pilots in the simulator, and if this surprise causes relevant performance impairments in a critical situation. To induce a surprise in a controlled manner, pilots are confronted with an upset situation (an aerodynamic stall). Pilot performance in recovering this situation is measured by checking their adherence to a memorized recovery procedure. As this required response is relatively straightforward, the stall recovery task allows us to measure practically relevant performance in a highly controlled manner. An effect of surprise on performance would underline

Table 1.1: An overview of the thesis

Research objective 1. Obtain more insight into the mechanisms that cause pilot performance issues in startling and surprising situations	
Key questions	**Chapters**
1. How do startle and surprise cause pilot performance issues in unexpected situations, according to the literature?	Chapter 2
2. Can we induce surprise and ensuing performance issues in the simulator?	Chapter 3
3. Can we induce interpretation and response errors by inducing an inappropriate expectation through spatial disorientation?	Chapter 4
Research objective 2. Identify effective simulator training interventions for startle and surprise	
Key questions	**Chapters**
4. Does variable and unpredictable simulator training help pilots to solve a startling and surprising situation?	Chapter 5
5. Does a startle and surprise management checklist help pilots to solve startling and surprising simulator scenarios?	Chapter 6
6. What are potential pitfalls when implementing a startle and surprise training intervention in practice?	Chapter 7

that targeted training for surprise is important. Besides an effect on performance, we also test if surprise causes increased startle, mental workload and stress. This would indicate a (perceived) lack of resources when being mentally unprepared for a task.

The third key question for the first research objective focuses on spatial disorientation as a case of surprise, or, at least, mismatching expectations: Can we induce interpretation and response errors by inducing an inappropriate expectation through spatial disorientation? This key question was chosen based on our literature review (Chapter 2), from which it followed that an expectation that mismatches with the situation may cause interpretation and response errors. This is investigated in two experiments (simulator and in-flight) employing non-pilots, whom we attempt to mislead about the state of the aircraft which they controlled. Such a mismatching expectation can occur in aviation practice when a pilot is spatially disoriented. It is therefore a relevant case to test our hypothesis.

The second research objective of the thesis is to identify effective pilot simulator training interventions for startle and surprise. For this, we use the insights obtained through answering the first three key questions, as well as literature with regards to training for stressful situations in other domains. Several potentially relevant training interventions are selected for investigation in simulator experiments. This is done while keeping the limitations and challenges of current pilot training practice in mind.

One challenge that currently exists is the low variety and high predictability in training scenarios. In the final report of the accident of Air France 447, the investigators noted that: *"Indeed the exercises are repetitive, well known to crews and do not enable skills in resource management to be tested outside of this context."* (page 209) [16] This inspired us to perform a simulator experiment in order to answer key question 4: Does variable and unpredictable training help pilots to solve startling and surprising events in a simulator scenario?

A second challenge that currently exists in operational practice, is that limited training resources and training time are available, while there is a high variety of potential safety events that can occur in-flight. Also, new unsafe events can occur that have never occurred before. It is impossible to train for every possible issue. Therefore, training a general method or procedure that can be applied to a wide range of issues seems useful. One of these methods, that has been proposed by several others, is the use of a startle and surprise management checklist [41, 42]. With key question 5, we aim to investigate such a checklist: Does a startle and surprise management checklist help pilots to solve startling and surprising simulator scenarios? This intervention does not exclude the use of variability and unpredictability. Variability and unpredictably would be more suitable for later stages of training, whereas a checklist can be taught in initial training.

Our second research objective also comprises the identification of potential pitfalls that may impede the effectiveness of a newly designed training intervention method. Is it enough to merely attend pilots to the problems of startle and surprise? Or are the specifics of the training intervention, and the way it is presented to pilots, of influence? Key question 6 is thus: What are potential pitfalls when implementing a startle and surprise training intervention in practice? To investigate this, we collect data on pilot application of several training interventions, as well as pilot feedback. One of the interventions is tested in a highly practical setting and with a representative sample of airline pilots.

1.5. Structure of the thesis

In the first part of the thesis, (Chapters 2, 3 and 4), different aspects of the problem of startle and surprise are investigated. The second part, (Chapters 5, 6 and 7) focuses on the effectiveness of selected training interventions, and on practical issues in implementing such interventions.

In **Chapter 2**, we review the literature on startle and surprise. Using this literature, a conceptual model of the effects of surprise and startle on pilot performance is created. Potential training interventions and other factors that may positively or negatively influence pilot performance in surprising situations are linked to the model. Finally, the model is used to describe and explain the events that occurred in several in-flight incidents and accidents.

Chapter 3 explores the issues caused by surprise on pilot responses to a critical situation in the simulator. To create a critical situation in a controlled manner, an aerodynamic stall event is induced. Airline pilot responses to an anticipated and an unanticipated stall event are analyzed using a motion-base simulator that is outfitted with an advanced stall model. Data are collected on pilot adherence to the stall recov-

ery procedure, subjective experience of surprise, startle, stress, confusion and mental workload, as well as heart rate and galvanic skin response. Since the unexpected stall event poses a sudden increase in demands, and the situation is likely not immediately understood, it is expected to cause lower adherence to the recovery template, as well as higher stress, workload and confusion.

In **Chapter 4**, the conceptual model is applied to the issue of spatial disorientation and errors in reading the instruments. In two experiments, non-pilots are given the expectation that they are flying with a certain bank angle, which mismatched with the actual bank angle. In one experiment, this is done by letting participants perform a flying task in a fixed-base simulator. In a second experiment, misleading vestibular cues in-flight are presented while participants have their eyes closed. Following the induction of the expectation, the participants are tasked with rolling the plane level using the artificial horizon. When performing this leveling task, previous experiments have found that pilots sometimes roll towards the incorrect direction. However, the expectation was not manipulated in these experiments. In the experiments we perform, the effect of expectations on the occurrence of such errors is examined by letting the artificial horizon sometimes match, and sometimes mismatch with the manipulated expectation. Performance is further analyzed to test the extent to which such errors were caused by misperceiving the artificial horizon or by neglecting it.

Chapter 5 details a simulator study on the effect of variability and unpredictability in simulator training scenarios, as a means to improve performance when surprised. Pilots first practice managing asymmetric thrust in a research simulator featuring an aircraft model that is largely unfamiliar to them. For one group, this training session involves a variety of scenario settings, and the events are presented in a more unpredictable manner. A control group practices the same scenarios in a more one-sided and predictable manner. Both groups are then confronted with a novel and demanding problem in a test scenario, in which they can apply the trained principles. The experimentally trained group is expected to have a better understanding of managing asymmetric thrust, as they were required to actively make sense of the training scenarios. This better understanding should allow them to better generalize the learned knowledge and apply it to the novel situation. Thus, the experimental group should have less difficulty with understanding the problems solving them.

In **Chapter 6**, we test if a startle and surprise management checklist may help pilots to respond to unsafe events. The checklist, which is kept very brief, consists of several steps to stimulate active relaxation, observation and formulation of plans or actions. By taking a moment to manage stress before responding to the problem, the negative effects of stress on the rest of the problem-solving process may be reduced. Analyzing the problem in a structured manner may help pilots to (re)assume goal-directed focus. This could facilitate slow appraisal and the development of a strategy to proceed. The method is trained and tested in several scenarios in a research simulator, using an aircraft type largely unfamiliar to the pilots. Performance outcomes are analyzed, as well as pilot subjective experience of the test scenarios and feedback on the checklist. This information is used to evaluate the checklist, identify potential pitfalls, and to provide suggestions for improvement.

In **Chapter 7**, we investigate a training intervention developed by an airline com-

pany and introduced during their type-rating recurrent simulator training. This approach to the problem provides a more practical perspective. Pilot application and feedback on the intervention method is tested in a relatively complex emergency situation presented in a training simulator, which featured an aircraft model on which the pilots were type-rated. This, as well as testing the method in two-pilot crews, creates an environment that more accurately reflects the pilots' real work environment compared to the other experiments in this thesis. The sample group in this experiment is also more representative of the airline pilot population, as participating cost them little effort. Following the scenario, pilots were asked to fill in a questionnaire regarding the method's applicability in the training scenario and in operational practice.

Finally, in **Chapter 8** we synthesize the different chapters to answer the key questions and to give recommendations on training methods for managing startle and surprise in aviation.

References

[1] J. I. Mediavilla, *Aviation safety evolution (2018 update),* https://theblogbyjavier.com/2019/01/02/aviation-safety-evolution-2018-update/ (2019), accessed: 2019-03-01.

[2] L. G. Militello and R. J. Hutton, *Applied cognitive task analysis (ACTA): a practitioner's toolkit for understanding cognitive task demands,* Ergonomics **41**, 1618 (1998).

[3] M. R. Endsley, *Automation and situation awareness,* in *Human factors in transportation. Automation and human performance: Theory and applications*, edited by R. Parasuraman and M. Mouloua (Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, US, 1996) pp. 163–181.

[4] M. S. Young and N. A. Stanton, *Malleable attentional resources theory: a new explanation for the effects of mental underload on performance,* Human factors **44**, 365 (2002).

[5] R. J. de Boer and K. Hurts, *Automation surprise,* Aviation Psychology and Applied Human Factors **7**, 28 (2017).

[6] N. B. Sarter, D. D. Woods, and C. E. Billings, *Automation surprises,* Handbook of human factors and ergonomics **2**, 1926 (1997).

[7] R. Parasuraman and V. Riley, *Humans and automation: Use, misuse, disuse, abuse,* Human Factors **39**, 230 (1997).

[8] A. Haslbeck and H.-J. Hoermann, *Flying the needles: flight deck automation erodes fine-motor flying skills among airline pilots,* Human factors **58**, 533 (2016).

[9] Federal Aviation Administration, *Crew resource management training: Advisory circular 120-51a,* (1993).

1

[10] B. C. A. Aviation Safety, *Statistical summary of commercial jet airplane accidents, worldwide operations, 1959-2017,* (2018).

[11] C. M. Belcastro and J. V. Foster, *Aircraft loss-of-control accident analysis,* in *Proceedings of AIAA Guidance, Navigation and Control Conference, Toronto, Canada, Paper No. AIAA-2010-8004* (2010).

[12] J. Bürki-Cohen, *Technical challenges of upset recovery training: Simulating the element of surprise,* in *Proceedings of the AIAA Guidance, Navigation, and Control Conference* (2010).

[13] J. A. Kochan, E. G. Breiter, and F. Jentsch, *Surprise and unexpectedness in flying: Database reviews and analyses,* in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 48(3) (SAGE Publications, 2004) pp. 335–339.

[14] W. L. Martin, P. S. Murray, P. R. Bates, and P. S. Lee, *A flight simulator study of the impairment effects of startle on pilots during unexpected critical events,* Aviation Psychology and Applied Human Factors (2016), 10.1027/2192-0923/a000092.

[15] S. Shappell, C. Detwiler, K. Holcomb, C. Hackworth, A. Boquet, and D. A. Wiegmann, *Human error and commercial aviation accidents: an analysis using the human factors analysis and classification system,* Human Factors **49**, 227 (2007).

[16] Bureau d'Enquêtes et d'Analyses pour la Sécurité de l'Aviation Civile, *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,* (2012).

[17] European Aviation Safety Agency, *Decision 2015/022/r: Crew resource management (crm) training,* https://www.easa.europa.eu/sites/default/files/dfu/ Explanatory Note to Decision 2015-022-R.pdf (2015).

[18] European Aviation Safety Agency, *Decision 2019/005/r: Loss of control prevention and recovery training,* https://www.easa.europa.eu/sites/default/files/dfu/Explanatory Note to ED Decicion 2019-005-R.pdf (2019).

[19] Federal Aviation Administration, *Advisory circular 120-111, upset prevention and recovery training,* (2015).

[20] European Aviation Safety Agency, *Annual activity report 2015,* https://www.easa.europa.eu/sites/default/files/dfu/EASA Annual Activity Report 2015.pdf (2016).

[21] J. Rivera, A. B. Talone, C. T. Boesser, F. Jentsch, and M. Yeh, *Startle and surprise on the flight deck: Similarities, differences, and prevalence,* in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 58(1) (SAGE Publications, 2014) pp. 1047–1051.

[22] P. Ekman, W. V. Friesen, and R. C. Simons, *Is the startle reaction an emotion?* Journal of personality and social psychology **49**, 1416 (1985).

[23] R. I. Thackray, *Performance recovery following startle: A laboratory approach to the study of behavioral response to sudden aircraft emergencies*, Tech. Rep. (Federal Aviation Administration, Washington DC Office Of Aviation Medicine, 1988).

[24] M. W. Eysenck, N. Derakhshan, R. Santos, and M. G. Calvo, *Anxiety and cognitive performance: attentional control theory.* Emotion **7**, 336 (2007).

[25] M. Horlington, *Startle response circadian rhythm in rats: Lack of correlation with motor activity,* Physiology & Behavior **5**, 49 (1970).

[26] C. Landis and W. Hunt, *The startle pattern.* (Farrar & Rinehart, 1939).

[27] A. Damasio, *The feeling of what happens: Body, Emotion And The Making Of Consciousness* (Vintage Publishing, 1999).

[28] T. Hagemann, R. W. Levenson, and J. J. Gross, *Expressive suppression during an acoustic startle,* Psychophysiology **43**, 104 (2006).

[29] N. A. Roberts, J. S. Beer, K. H. Werner, D. Scabini, S. M. Levens, R. T. Knight, and R. W. Levenson, *The impact of orbital prefrontal cortex damage on emotional activation to unanticipated and anticipated acoustic startle stimuli,* Cognitive, Affective, & Behavioral Neuroscience **4**, 307 (2004).

[30] M. I. Foster and M. T. Keane, *Why some surprises are more surprising than others: Surprise as a metacognitive sense of explanatory difficulty,* Cognitive psychology **81**, 74 (2015).

[31] W.-U. Meyer, R. Reisenzein, and A. Schützwohl, *Toward a process analysis of emotions: The case of surprise,* Motivation and Emotion **21**, 251 (1997).

[32] A. Schützwohl, *Surprise and schema strength,* Journal of Experimental Psychology: Learning, Memory, and Cognition **24**, 1182 (1998).

[33] K. H. Teigen and G. Keren, *Surprises: low probabilities or high contrasts?* Cognition **87**, 55 (2003).

[34] M. M. Bradley, *Natural selective attention: Orienting and emotion,* Psychophysiology **46**, 1 (2009).

[35] W. L. Martin, P. S. Murray, P. R. Bates, and P. S. Y. Lee, *Fear-potentiated startle: A review from an aviation perspective,* The International Journal of Aviation Psychology **25**, 97 (2015).

[36] C. D. Wickens, *Attention to safety and the psychology of surprise,* in *Proceedings of the 2001 symposium on aviation psychology* (The Ohio State University Columbus, OH, 2001).

1

[37] A. Rankin, R. Woltjer, J. Field, and D. Woods, *"staying ahead of the aircraft" and Managing Surprise in Modern Airliners,* in *5th Resilience Engineering Symposium: Managing trade-offs, 25-27 June 2013* (Soesterberg, The Netherlands, 2013).

[38] S. Arora, N. Sevdalis, D. Nestel, T. Tierney, M. Woloshynowych, and R. Kneebone, *Managing intraoperative stress: what do surgeons want from a crisis training program?* The American Journal of Surgery **197**, 537 (2009).

[39] A. Bhangu, S. Bhangu, J. Stevenson, and D. M. Bowley, *Effective implementation of "surprise and startle" scenarios in surgical training,* World journal of surgery **37**, 2718 (2013).

[40] W. L. Martin, M. P. S., and P. R. Bates, *The effects of startle on pilots during critical events in aviation,* in *Proceedings of the Human Factors in Transport Conference, Sydney, Australia, November 7, 2011* (2011).

[41] J. N. Field, E. J. Boland, J. M. Van Rooij, J. F. W. Mohrmann, and J. W. Smeltink, *Startle effect management (report nr. nlr-cr-2018-242),* (European Aviation Safety Agency, 2018).

[42] W. L. Martin, *Developing startle and surprise training interventions for airline training programs,* http://pacdeff.com/wp-content/uploads/2017/08/PACDEFF-FC-Forum-Presentation-on-Startle.pdf (2016), accessed: 2019-02-11.

2

A conceptual model of startle and surprise

Today's debate around loss of control following in-flight events has highlighted the importance of pilots' ability to deal with unexpected events. Such events may induce a "startle factor", that may significantly impair performance. The current chapter introduces the problem. Literature on surprise, startle, resilience, and decision-making is reviewed, and findings are combined into a conceptual model. Pilot perception and actions are conceptualized as being guided by "frames," or mental knowledge structures based on previous experiences. Performance issues in unexpected situations can often be traced back to insufficient adaptation of one's frame to the situation. We propose that such reframing processes are especially vulnerable to issues caused by startle or acute stress. Interventions should therefore focus on improving pilot frames, reframing skills, and/or stress management skills.

The introduction section of the paper has been skipped.

2.1. Theoretical framework: a model of startle and surprise

The differences between surprise and startle raise questions regarding ground-based training to prepare flight crew for unexpected events in flight. Would a sudden and loud noise in the simulator be sufficient to simulate difficulties associated with in-flight emergencies [2]? Or should training scenarios primarily involve unexpectedness [3]? To answer these questions, some authors have focused on the causes and effects of surprise (e.g., [4, 5]), and others have described those of startle [6]. In the current paper, we present a conceptual model (Figure 2.1) that brings the existing knowledge about startle and surprise together. The model is a synthesis of elements of the cognitive-psychoevolutionary model of surprise (Meyer et al., 1997), the perceptual cycle model [7], the data/frame theory of sensemaking [8], and literature on startle and acute stress.
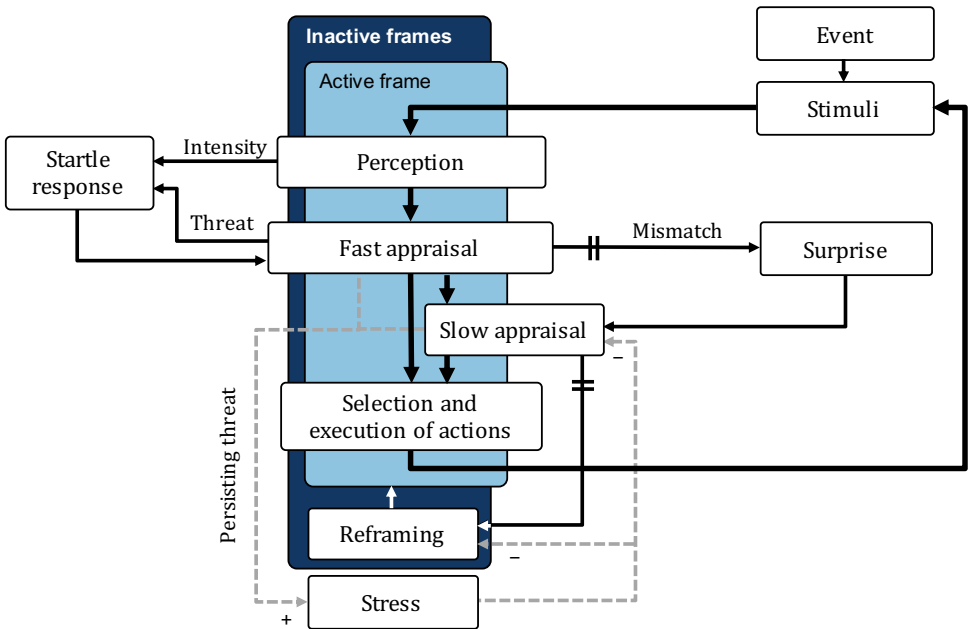
Figure 2.1: Conceptual model of startle and surprise. Solid lines indicate sequenced events. Dashed lines indicate potential influences, with plus signs indicating an increasing effect and minus signs indicating an impairing effect. Double lines indicate thresholds. The model is a slightly more streamlined version of the model presented in [9].

2.2. Elements of the model

2.2.1. The perceptual cycle

The bold lines in the model represent the perceptual cycle: A person perceives stimuli, interprets these stimuli, assesses the situation (appraisal), and selects and executes ac-

tions, which may generate new data. Appraisal is modeled in such a way that it can be fast and highly automatic in some cases, or it may also involve a more slow, effortful, and knowledge-based processing [10, 11]. Action selection (decision making) is modeled so that it is an integral part of the perceptual cycle, which thus represents a continuous process of hypotheses generation and testing [12]. For simplicity, the model does not discern different levels of control at which perceptual cycles may occur in parallel, such as in Hollnagel's extended control model [13].

2.2.2. Startle
On the left side of this perceptual cycle, the startle response is pictured. This response results from a fast, sometimes reflexive, appraisal of a stimulus as threat-related [14]. Startle is modeled to cause a closer examination of the triggering stimulus, which may lead to further increase of stress (dashed line; [6]). If startle occurs in the absence of surprise, only the left loop (startle response) is activated, and the appraisal process will remain relatively fast. However, if the appraisal of a startling stimulus brings momentarily unexplainable information to light, the right loop (surprise) will subsequently be activated. The perceptual cycle then continues, either with actions in response to the threat or by resuming as before in case of a false alarm.

2.2.3. Frames
In order to explain the causes and effects of surprise, the concept of frames is useful. A frame is defined as an explanatory structure, such as a story, map, or plan, which links perceived individual data points together and gives them meaning [8]. Frames synthesize concepts, such as schemata, mental models, scripts, and other types of knowledge structures in long-term memory, that describe generic or specific situations, how things work, how events are sequenced, and which actions are appropriate [7, 15–19]. Frames are created based on previous experiences (i.e., bottom up) so that understanding of a new situation or concept can be achieved and stored in memory (the supply of inactive frames in Figure 2.1). If a situation occurs in which the frame-related knowledge can be applied, a corresponding frame may be activated and applied (see, [20]). Frames are thus instrumental for the achievement of higher levels of situation awareness (i.e., comprehension and projection) based on a lower level of situation awareness (i.e., perception) in the terms of Endsley's [21] model.

Besides being shaped based on incoming data (bottom–up), frames are thought to actively select, filter, and provide meaning to incoming data (i.e., top–down; [7]). They are thought to play a significant role in skilled performance, as frames structure complex stimuli and action sequences into manageable "chunks" based on the existing constraints (see, [22, 23]). This is why, for instance, expert chess players are able to perceive and reproduce chess positions very quickly and accurately, as long as the positions make sense in terms of the game. In our model, we have illustrated the influence of the frame on perception, appraisal, and action by placing it behind these elements of the perceptual cycle, rather than making it an integral part of the perceptual cycle [7]. This way, we indicate that perception and action are still possible—although difficult—when there is no fitting frame activated. The model is simplified in that it represents merely one active frame, distinct from other frames. In reality, people are

thought to use a number of frames at once, which are highly interconnected or nested and have no clear boundaries.

The use of frames to explain performance during surprise events in aviation has recently gained interest (e.g., [5, 24]). In the latter study, pilot performance is modeled as the interaction of a crew with the aircraft and the environment using frames, anticipatory thinking, and expectations. The authors discuss an extensive list of sensemaking activities following surprise event cases in aviation. In our current model, we aim to add to their model by illustrating how the frame interacts with the perceptual cycle and how or why certain performance issues may occur.

2.2.4. Surprise
In the perceptual cycle, hypotheses based on the active frame are continually applied and tested with regard to their practical consequences (abduction; see, [12]). As long as the results are consistent with the hypotheses, the active frame becomes strengthened in memory. However, a mismatch between feedback and the active frame will induce a surprise [25], given that the mismatch exceeds a certain assumed threshold (double intersecting lines before surprise in Figure 2.1; e.g., [26]). This threshold indicates a form of confirmation bias, as events of low salience are more easily missed when they are deemed unlikely within the active frame (see, e.g., [27]).

2.2.5. Sensemaking
Appraisal of a surprise event involves sensemaking activities, or efforts to understand the cause of the mismatch between the encountered data and the active frame [8]. Sensemaking is an explorative process that is active, analytical, conscious, and potentially effortful, characterized by top-down or goal-directed processing [10]. Due to its active nature, it may be particularly problematic when pilots are not mentally prepared, for example, after a long period of automated flight [28]. Sensemaking activities can be categorized into three groups [8, 19]. First, if the surprising data are determined to be the result of a misperception, the active frame can be preserved. Second, if the surprising data are being judged as correct, the active frame may not be detailed enough to account for them, in which case it can be elaborated (i.e., assimilation; [17]). Third, if the data are being judged as correct, and they are fundamentally inconsistent with the active frame (i.e., a fundamental surprise; [29]), a paradigm shift is required and a new frame should replace the active frame (i.e., accommodation; [17]). This sensemaking activity is modeled as the element reframing being connected to the (transformation of the) active frame in Figure 2.1. People were shown to avoid considering a fundamental surprise as being the causal factor for mismatches, perhaps as a mechanism to reduce unnecessary efforts (i.e., frame fixation; [30, 31]), indicated by a threshold toward reframing in Figure 2.1.

2.2.6. Reframing
A frame switch, or reframing, occurs when one restructures the way in which a situation is represented. Previously perplexing information may suddenly "fall into place," and the appropriate responses become obvious. In contrast, the adoption of an inappropriate frame or the loss of a fitting frame may lead to a complete "loss of grip" on

the situation, as there is no frame in place to guide perception, appraisal, and action. This may negatively affect the pilot's ability to track what is going on (loss of Level I situation awareness; [21]) or lead to information overload. Data can no longer be appraised in relation to other data and therefore lose meaning. The selection and execution of actions become reactive and sequential (bottom-up controlled) instead of anticipatory and proactive (top-down controlled), which may lead to tunnel vision or cognitive lockup [32]). The involvement of acute stress may be even more deteriorative, as we will discuss next.

2.2.7. Stress

Both startle and surprise may cause acute stress, which constitutes the appraisal of present demands as taxing or exceeding one's resources and endangering one's well-being ([33] dashed lines with plus signs in Figure 2.1). Startle may increase stress very briefly and rapidly at first, and subsequent appraisal of the startling stimulus as threatening may cause a further increase in stress [6]. Surprise may also cause stress, if it poses a sudden increase in task demands while one becomes at the same time aware of the inadequateness of the active frame.

The function of stress is to facilitate the recruitment of additional resources in order to respond effectively to the –potential– threat. However, aspects of stress, such as emotions of anxiety and frustration, excessive physiological arousal, or narrowed attention, may also impair a pilot's cognitive and motor performance [34–37]. Stress could therefore be modeled to influence perception, slow and fast appraisal, and action in Figure 2.1. However, when a pilot needs to respond to a surprising event and reframe, the influence of stress on slow appraisal and reframing is most relevant (dashed lines in Figure 2.1). Slow appraisal is particularly vulnerable to stress, since stress impairs the processing efficiency of the attentional system [38]. It decreases the capacity of aspects of working memory, as well as the ability to switch attention from one task to another, especially when these tasks are complex [39]. When confronted with an emergency, this means that stress can make it very difficult to perform the necessary complex tasks, maintain a good overview of the situation, or device a strategy while considering all available options.

A second manner in which stress may specifically affect performance in surprising situations is that it appears to interfere with the influence of frames on the perceptual cycle. Under stress, attentional control is shifted from top-down (by frames and goals) to stimulus-driven (by potentially threatening stimuli) [35]. Stressed individuals have difficulty with recognizing relationships between information elements [40]. Stress has also been shown to impair the ability to inhibit attention to irrelevant information[39], which can be seen as a decrease in the filtering function of the frame. Stress seems to shift the influence of frames towards a processing strategy that is more simplistic and risk-avoidant [41, 42]. People in emergency situations strongly prefer familiar and simple solutions over solutions that require increased analysis or seem more uncertain. Examples of this are the tendency to follow authority figures or protocol, and neglecting information that shows this is inappropriate. These effects of stress are modeled as impairing slow appraisal as well as reframing in Figure 2.1.

2.3. Influencing Factors and Intervention Methods

In this section, several factors, which have previously been identified as affecting pilot performance in surprising or startling situations, are described and related to our model.

2.3.1. Domain expertise

One of the factors that facilitate pilot performance in surprising situations is domain expertise, or accumulated knowledge and skills through practice and experience. By repeatedly applying and testing frames, these become more accurate and more fixed in memory (see, [24]), which allows one to easier relate new situations to those that have previously been encountered and to make decisions in a quick manner [23]. In the literature, some results indeed indicate beneficial effects of pilot expertise on problem assessment and flexibility in unfamiliar scenarios [43, 44], whereas other results suggest no effects or even somewhat detrimental effects [24, 44]. Perhaps being highly experienced in normal situations makes one's frames somewhat rigid, making it more difficult to respond to non-normal situations [24].

2.3.2. Judgment skills

Domain-independent judgment skills, such as decision-making skills, cognitive flexibility, and metacognitive skills, were found to improve pilot performance following surprise in one study [24]. Such skills could be tested in the selection process, and certain judgment skills are thought to be trainable as well (see, [24]). Decision-making skills involve capabilities of problem analysis (sensemaking) and action selection. Cognitive flexibility involves reframing abilities. Our model may in particular be useful to increase metacognitive skills in pilots, which include the recognition of frame mismatches cognitive biases and reframing issues. By recognizing such situations, pilots can apply learned coping strategies, such as taking a moment to "breathe" and reflect or returning to more transparent and understandable configurations or autopilot modes.

2.3.3. Variable training

Researchers and aviation safety organizations emphasize the need for training with a variety of situations or scenarios (e.g., [3–5, 45–48]). Training variability can be applied to reduce predictability so as to stimulate sensemaking activities and to improve reframing skills. Training variability is also thought to increase the number and elaborateness of available frames (e.g., [49]). A more elaborate frame is thought to discriminate better between situations, aiding the generation of accurate hypotheses, the detection of data/frame mismatches, and the selection of an appropriate frame based on the available data (see the plus sign on the line from the inactive frames toward reframing in Figure 2.1 [50, 51]). Experiencing examples of a concept in a variety of situations improves the generalized frame, facilitating the transfer of the knowledge and skills to new situations [23]. In contrast, one-sided training of a small number of situations or (combinations of) failures may increase the risk of inappropriately applying th leraned knowledge in stressful situations (the influence of generalized frame on reframing Figure 2.1 [52]).

2.3.4. Practical training

Literature indicates that theoretical training should be enhanced with practical experience and feedback on performance so that the frame-related knowledge is linked to other knowledge, environmental cues, and actions [51]. Our model indicates that action selection in operational practice is an inherent part of the perceptual cycle, meaning that mere theoretical training is likely insufficient. For instance, scenario-based training [53] is based on the concept that knowledge cannot be fully understood independent from its context. This means that training should not be focused on specific maneuvers that are laid out in advance, but on the pilot's own decisions in response to a situation that is presented. Practical training may also be used in combination with exposure to a manageable amount of stress or startle, to make skills more robust to the effects of stress [54]. This may decrease the detrimental effects of stress on other elements in our model (dashed lines with minus signs in Figure 2.1).

2.3.5. Fatigue

Fatigue is known to degrade logical reasoning and accurateness of performance, as well as to increase inattentiveness and the tendency toward preservation [55]. Fatigue can thus be expected to increase confirmation biases (increase the thresholds; Figure 2.1), as well as to impair mentally taxing activities of slow appraisal and reframing.

2.3.6. Flight deck design

Display designs that enhance situation awareness may aid in quicker recognition of anomalies by making mismatching data more salient. Our model suggests that the interpretation of a display system may be straightforward when the appropriate frame is already activated, but this is not the case when a surprise occurs. Thus, interfaces designed for use in surprising situations (e.g., upset recovery display aids) should be tested in conditions in which surprise is simulated in a realistic manner (see Implications for Experimental Design and Simulation section). Transparent automated systems [56, 57] that aim to keep the pilot in the loop may help to update the active frame when a situation changes. Displays can also be designed to aid the sensemaking process (e.g., [58]. For instance, ecological interface design is intended to structure complex relationships between information in such a way that constraints become self-evident, decreasing the need for the pilot to construct frames for these relationships (e.g., [59]).

2.4. Implications for Experimental Design and Simulation

As outlined in the model, startle and surprise have different causes and different effects, which means that different factors should be manipulated depending on whether the aim is to induce mainly startle or mainly surprise. The key element for inducing surprise is to set up a situation that mismatches with a previously activated frame. A mismatch that is not immediately understood would increase the effort required to reframe the situation, which may be useful for training purposes. Surprise and reframing can thus be elicited, for instance, through explicit misinformation, by presenting a

2

number of similar scenarios followed by one that is subtly different, by presenting a situation that is subtly different from one that is well known to pilots, or through variation or novelty.

Although a surprising stimulus can be subtle, a startling stimulus should be highly salient. A startling stimulus can be a loud and abrupt sound or a sudden, uncommanded motion of the aircraft. Unexpectedness may increase salience and perceived threat, but in contrast to the manipulation for surprise, a startling event does not necessarily require sensemaking or reframing (e.g., in the case of a lightning strike). For an extensive list of surprising or startling flight scenarios, see [6].

2.5. Previous Experimental Studies on Startle and Surprise in Aviation

To date, few experimental studies focusing specifically on surprise and startle in the cockpit have been published. The studies indicate that pilot performance may decrease significantly, even when skills and procedures were practiced shortly beforehand. In the concise review that follows, we link the experimental studies to our model. As the reports do not always explicitly mention whether the participating pilots were surprised, startled, or both, we have tried to infer this reaction from the manipulations used.

In two studies, pilots had to detect, recognize, and respond to unannounced problems, such as aerodynamic stalls, wind shears, or automation failures [45, 60]. The results showed that response times were longer after surprising compared with unsurprising events, with some participants responding exceptionally late. Similar results were found in a simulator study by Martin et al. (2016), in which pilots were tasked with flying the same missed approach, once with and once without an unexpected fire alarm and a loud explosion sound. Although the startling stimulus did not require a change of plans, the stimulus resulted in a delayed initiation of the missed approach in one third of the pilots. In regard to our model, the frame-incongruent information in these experiments likely caused a surprise, and the highly salient stimulus in the experiment by Martin et al. [61] was likely startling as well. Our model explains such later responses as being caused by inattentiveness to frame-incongruent information, or by slow appraisal processes delaying or interfering with actions.

Some studies also showed impairments of performance in terms of the incorrect or incomplete application of procedures. Pilots in the study by Casner et al. [45] displayed difficulty with recognizing and responding correctly to an unexpected wind shear compared to an expected one. Schroeder et al. [62] actively misled pilots into expecting a different upcoming event. During final approach, an unexpected aerodynamic stall, induced by a sudden tailwind, was inserted in the scenario. The results indicated that 78% of the pilots made errors in executing the stall recovery template, even though they had applied it many times beforehand. A check of the subjective impact of the manipulation confirmed that all pilots were highly surprised by the event. Whether they were also startled or stressed is not clear. The study did not include a control condition to confirm whether the performance degradation was attributable to the surprise. For this reason, we recently performed a simulator study in which pilots were exposed

twice to an aerodynamic stall: once in a surprise condition and once in an anticipation (control) condition [1]. The results showed that, compared to the control condition, the proportion of pilots adhering to the recovery template decreased by around 25% in the surprise condition, whereas measures of surprise, startle, and mental workload increased significantly. According to our model, this performance impairment would result from reframing efforts, as a frame switch is needed before one can respond accurately to the unanticipated event.

2.6. Applying the Model to Flight Safety Incidents

In this section, we will evaluate four aviation incidents or accidents in the context of our model (see Figure 2.2). These four cases were selected because they seem to demonstrate several different aspects of our model. We focus in particular on potential causes of reframing issues and on the effects of reframing issues on perception, appraisal, and action (see also, [5]).
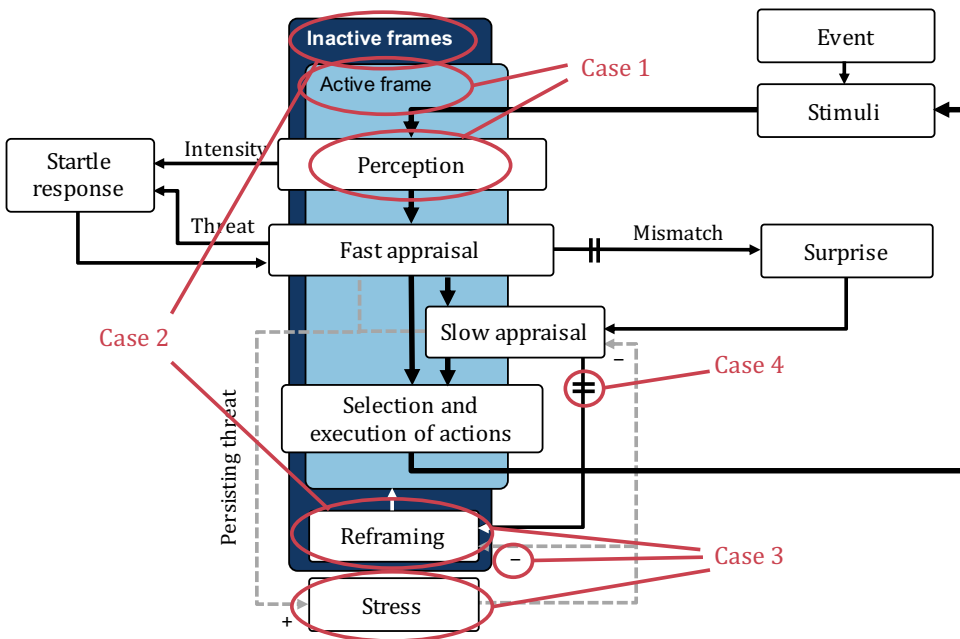
Figure 2.2: Estimated causal factors in the four cases as mapped on to our conceptual model of startle and surprise.

2.6.1. Case 1

The accident of Flash Airlines Flight 604 in 2004 [63] suggests that pilot spatial disorientation [64] of the captain (pilot flying) played a significant role in the development of the event, although other causes of the accident have not been ruled out by all investigating parties. The captain had initiated a long, left climbing turn, during which

the aircraft transitioned from a left bank to a right bank at a rate below the detection threshold of the vestibular system [65]. When the first officer alerted the captain to the right turn ("Aircraft turning right, sir"), the captain expressed surprise ("How turning right?"). Next, he seemed to recognize that the attitude was indeed off ("Ok, come out"). According to our model, there was at that moment likely a mismatch between the captain's still active frame (aircraft turning left) and the first officer's assertion of the aircraft turning right. Next, instead of rolling the wings back to level, the captain gave further roll inputs to the right, leading to an overbank and loss of control. This suggests that reframing did not occur following the surprise, and that the incorrect frame of a left bank remained active to influence perception of the artificial horizon (see, 2.2), causing what is known as a "horizon control reversal" [66]. A similar sequence of events seems to have occurred in the Crossair Flight 498 accident in 2000 and Kenya Airways flight 5Y-KYA accident in 2007, suggesting that frame-induced misinterpretations of the instruments occur more often.

2.6.2. Case 2

The incident with a B-737 near Brisbane, Australia, in 2013 [67] may be an example in which an inactive frame influences the reframing process with negative consequences. While approaching the glide slope beam of Brisbane airport at night, the aircraft unexpectedly began to climb due to an earlier unintended selection of an autopilot mode. The crew quickly noticed this and disconnected the autopilot mode. Later, during the descent, the aircraft began to bank to the left due to a residual rudder deflection that was previously corrected for by the autopilot. This motion was again detected, but the crew incorrectly assumed that it was induced by the autopilot. After 80 seconds, the crew realized that the autopilot was not engaged, and they corrected the deviation manually. In our model (Case 2 in Figure 2.2), these actions are explained as caused by an influence of the previously activated frame on the reframing process. Because of the recent events in the incident, the frame of unintended explaining the events as caused by autopilot activation was perhaps most easily retrievable from memory, such that it was incorrectly applied again to the new situation.

2.6.3. Case 3

The accident of Air France Flight 447 in 2009 [68] seemed to involve a negative spiral of reframing issues and high stress (Case 3 in Figure 2.2). The accident report indicates that there were several signs that the crew were unable to identify an aerodynamic stall situation ([68]; pp. 179–180), which followed unreliable airspeed indication and autopilot disengagement during cruise. Cues indicating stall, such as buffeting and the auditory stall warning, did not lead to a clear problem assessment by the pilots, and potentially led to incorrect reframing to an overspeed situation. The report reads that a lack of training for (high-altitude) aerodynamic stall situations, in contrast to the well-known dangers of overspeed, may have caused the crew to fixate on the overspeed explanation of events. As mentioned previously in section 2.3, stress may create a tendency towards applying the more easily retrievable generalized frame, which is in this case the overspeed frame. The accident report also reads that there were signs of excessive stress, which may have exacerbated the pilots' inability to analyze the avail-

able information. In the model, the situation is explained by the inadequate generalized frame, in combination with excessive stress, impairing the reframing process. Strong initial pitch and roll inputs immediately following the autopilot disengagement suggest that the pilot flying was not only surprised but perhaps also startled by the sudden autopilot disconnect.

2.6.4. Case 4

West Caribbean Airways Flight 708 in 2005 [69] seems to be an example of frame fixation following a switch toward an inappropriate frame (Case 4 in Figure 2.2). Leading up to the accident, the aircraft's anti-icing systems were turned on at a too high altitude, so that sufficient engine performance could not be maintained. Subsequent loss in airspeed, loss in engine power, and autopilot-induced changes in attitude went unnoticed. An aerodynamic stall ensued, causing a further decrease in engine power due to variations of airflow into the engines. According to the voice recorder, the captain (pilot flying) misdiagnosed the problem as an engine flameout (reframed to an incorrect frame) and gave nose-up inputs. It seems that the captain then fixated on this incorrect frame, and disregarded the first officer's two callouts of an aerodynamic stall as well as the stall warnings of the system. It also seems that these reframing issues were not preceded by severe startle. In contrast, the crew seemed to underestimate the gravity of the situation at first, as instead of declaring an emergency they asked air traffic control for lower flight levels.

2.7. Conclusion

We propose an integrated model, which explains the effects of both startle and surprise responses to unexpected events in the cockpit. Examples of flight safety events show that inappropriate crew responses do not always involve startle but can often be traced back to surprise, which indicates a mismatch between what is being perceived and the pilot's active frame. The model explains such inappropriate responses as resulting from reframing issues following the mismatch, issues that can be exacerbated by startle, acute stress, fatigue, or unclear and complex interface designs. Information mismatching with an active frame may also remain unnoticed or be incorrectly interpreted so that a loss of situation awareness may occur.

By explaining inappropriate or absent responses to unexpected situations as reframing issues, we emphasize that intervention methods should be focused on pilots' abilities to reframe under high stress. We suggest that variability and unpredictability is introduced in training scenarios to let pilots practice reframing and enhance their frames of the practiced situations. Additionally, transparent interface designs may aid in framing and reframing. Finally, our model provides an aid to increase pilots' metacognitive skills of recognizing and understanding the hazards involved in frame mismatches and reframing issues.

References

[1] A. Landman, E. L. Groen, M. M. Van Paassen, A. W. Bronkhorst, and M. Mulder, *The influence of surprise on upset recovery performance in airline pilots,* The Interna-

tional Journal of Aerospace Psychology **27**, 2 (2017b).

[2] R. I. Thackray, *Performance recovery following startle: A laboratory approach to the study of behavioral response to sudden aircraft emergencies*, Tech. Rep. (Federal Aviation Administration, Washington DC Office Of Aviation Medicine, 1988).

[3] J. Bürki-Cohen, *Technical challenges of upset recovery training: Simulating the element of surprise,* in *Proceedings of the AIAA Guidance, Navigation, and Control Conference* (2010).

[4] J. A. Kochan, E. G. Breiter, and F. Jentsch, *Surprise and unexpectedness in flying: Database reviews and analyses,* in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 48(3) (SAGE Publications, 2004) pp. 335–339.

[5] A. Rankin, R. Woltjer, and J. Field, *Sensemaking following surprise in the cockpit: a re-framing problem,* Cognition, Technology & Work **18**, 623 (2016).

[6] W. L. Martin, P. S. Murray, P. R. Bates, and P. S. Y. Lee, *Fear-potentiated startle: A review from an aviation perspective,* The International Journal of Aviation Psychology **25**, 97 (2015).

[7] U. Neisser, *Cognition and reality: Principles and implications of cognitive psychology.* (W. H. Freeman and Company, San Francisco, 1976).

[8] G. Klein, J. K. Phillips, E. L. Rall, and D. A. Peluso, *A data-frame theory of sensemaking,* in *Expertise out of context: Proceedings of the sixth international conference on naturalistic decision making*, edited by R. R. Hoffman (Psychology Press, 2007) pp. 113–155.

[9] A. Landman, E. L. Groen, M. M. van Paassen, A. W. Bronkhorst, and M. Mulder, *Dealing with unexpected events on the flight deck: A conceptual model of startle and surprise,* Human Factors **59**, 1161 (2017).

[10] D. Kahneman, *A perspective on judgment and choice: mapping bounded rationality,* American psychologist **58**, 697 (2003).

[11] J. Rasmussen, *Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models,* IEEE transactions on systems, man, and cybernetics , 257 (1983).

[12] J. M. Flach, M. A. Feufel, P. L. Reynolds, S. H. Parker, and K. M. Kellogg, *Decision-making in practice: The dynamics of muddling through,* Applied ergonomics **63**, 133 (2017).

[13] E. Hollnagel and D. D. Woods, *Joint cognitive systems: Foundations of cognitive systems engineering* (CRC Press, 2005).

[14] J. Globisch, A. O. Hamm, F. Esteves, and A. Öhman, *Fear appears fast: Temporal course of startle reflex potentiation in animal fearful subjects,* Psychophysiology **36**, 66 (1999).

[15] F. C. Bartlett, *Remembering: An experimental and social study,* Cambridge: Cambridge University (1932).

[16] W. F. Brewer and G. V. Nakamura, *The nature and functions of schemas,* in *Handbook of social cognition*, edited by R. S. Wyer Jr. and T. K. Srull (Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, 1984).

[17] J. Piaget, *Piaget's theory,* in *Piaget and his school: a reader in developmental psychology*, edited by B. Inhelder, H. Chipman, and C. Zwingmann (Springer, 1976) pp. 11–23.

[18] R. A. Schmidt, *A schema theory of discrete motor skill learning.* Psychological Review **82**, 225 (1975).

[19] K. E. Weick, *Sensemaking in organizations*, Vol. 3 (Sage, 1995).

[20] N. A. Stanton, P. M. Salmon, G. H. Walker, and D. Jenkins, *Genotype and phenotype schemata and their role in distributed situation awareness in collaborative systems,* Theoretical Issues in Ergonomics Science **10**, 43 (2009).

[21] M. R. Endsley, *Toward a theory of situation awareness in dynamic systems,* Human Factors **37**, 32 (1995).

[22] J. Flach, M. Mulder, and M. M. van Paassen, *The concept of the situation in psychology,* in *A cognitive approach to situation awareness: Theory and application*, edited by S. Banbury and S. Tremblay (Ashgate Aldershot, Oxon, UK, 2004) pp. 42–60.

[23] G. A. Klein, *A recognition-primed decision (RPD) model of rapid decision making* (Ablex Publishing Corporation New York, 1993).

[24] J. A. Kochan, *The role of domain expertise and judgment in dealing with unexpected events*, Ph.D. thesis, University of Central Florida Orlando, Florida (2005).

[25] W.-U. Meyer, R. Reisenzein, and A. Schützwohl, *Toward a process analysis of emotions: The case of surprise,* Motivation and Emotion **21**, 251 (1997).

[26] J. W. Senders, *The human operator as a monitor and controller of multidegree of freedom systems,* IEEE Transactions on Human Factors in Electronics , 2 (1964).

[27] C. D. Wickens, B. L. Hooey, B. F. Gore, A. Sebok, and C. S. Koenicke, *Identifying black swans in nextgen: Predicting human performance in off-nominal conditions,* Human Factors **51**, 638 (2009).

[28] M. S. Young and N. A. Stanton, *Malleable attentional resources theory: a new explanation for the effects of mental underload on performance,* Human factors **44**, 365 (2002).

[29] Z. Lanir, *Fundamental surprise,* Eugene, OR: Decision Research (1986).

[30] C. A. Chinn and W. F. Brewer, *The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction,* Review of educational research **63**, 1 (1993).

[31] V. De Keyser and D. Woods, *Fixation errors: Failures to revise situation assessment in dynamic and risky systems,* in *Systems Reliability Assessment. ISPRA Courses (On Reliability and Risk Analysis)*, edited by A. Colombo and A. de Bustamante (Springer, Dordrecht, 1990) pp. 231–251.

[32] T. B. Sheridan, *Man-machine systems,* Information, control, and decision models of human performance (1981).

[33] R. S. Lazarus and S. Folkman, *Stress, appraisal, and coping* (Springer, New York, 1984).

[34] R. K. Dismukes, T. E. Goldsmith, and J. A. Kochan, *Effects of acute stress on aircrew performance: literature review and analysis of operational aspects,* (2015).

[35] M. W. Eysenck, N. Derakhshan, R. Santos, and M. G. Calvo, *Anxiety and cognitive performance: attentional control theory.* Emotion **7**, 336 (2007).

[36] A. Nieuwenhuys and R. R. Oudejans, *Anxiety and perceptual-motor performance: toward an integrated model of concepts, mechanisms, and processes,* Psychological research **76**, 747 (2012).

[37] C. D. Wickens, A. Stokes, B. Barnett, and F. Hyman, *The effects of stress on pilot judgment in a midis simulator,* in *Time Pressure and Stress in Human Judgment and Decision Making*, edited by A. Maule and O. Svenson (Springer, 1993) pp. 271–292.

[38] M. W. Eysenck and M. G. Calvo, *Anxiety and performance: The processing efficiency theory,* Cognition & Emotion **6**, 409 (1992).

[39] N. Derakhshan, S. Smyth, and M. W. Eysenck, *Effects of state anxiety on performance using a task-switching paradigm: An investigation of attentional control theory,* Psychonomic bulletin & review **16**, 1112 (2009).

[40] C. Remmers and T. Zander, *Why you don't see the forest for the trees when you are anxious: Anxiety impairs intuitive decision making,* Clinical Psychological Science **6**, 48 (2018).

[41] F. Ozel, *Time pressure and stress as a factor during emergency egress,* Safety Science **38**, 95 (2001).

[42] N. Schwarz and G. L. Clore, *Feelings and phenomenal experiences,* Social psychology: Handbook of basic principles **2**, 385 (1996).

[43] C. Gillan, *Analysis of multicrew decision making from a cognitive perspective,* in *Proceedings of the12th International Symposium on Aviation Psychology* (2003) pp. 427–432.

[44] E. H. McKinney Jr and K. J. Davis, *Effects of deliberate practice on crisis decision performance,* Human Factors **45**, 436 (2003).

[45] S. M. Casner, R. W. Geven, and K. T. Williams, *The effectiveness of airline pilot training for abnormal events,* Human Factors **55**, 477 (2013).

[46] European Aviation Safety Agency, *Loss of control prevention and recovery training: Notice of proposed amendment 2015-13),* (2015).

[47] Federal Aviation Administration, *Advisory circular 120-111, upset prevention and recovery training,* (2015).

[48] International Civil Aviation Organization, *Manual of evidence-based training,* https://www.icao.int/SAM/Documents/2014-AQP/EBT ICAO Manual Doc 209995.en.pdf (2013).

[49] J. J. Van Merriënboer, *Training complex cognitive skills: A four-component instructional design model for technical training* (Educational Technology, 1997).

[50] D. A. Gioia and P. P. Poole, *Scripts in organizational behavior,* Academy of management review **9**, 449 (1984).

[51] J. K. Phillips, G. Klein, and W. R. Sieck, *Expertise in judgment and decision making: A case for training intuitive decision skills,* (Blackwell Publishing, Malden, 2004) pp. 297–315.

[52] K. M. Kowalski-Trakofler, C. Vaught, and T. Scharf, *Judgment and decision making under stress: an overview for emergency managers,* International Journal of Emergency Management **1**, 278 (2003).

[53] M. M. Summers, *Scenario-based training in technically advanced aircraft as a method to improve risk management,* (2007).

[54] J. E. Driskell, E. Salas, J. H. Johnston, and T. N. Wollert, *Stress exposure training: An event-based approach,* in *Performance under stress*, edited by J. L. Szalma and P. A. Hancock (Ashgate London, 2008) pp. 271–286.

[55] J. J. Caldwell, *Fatigue in the aviation environment: an overview of the causes and effects as well as recommended countermeasures,* Aviation, Space, and Environmental Medicine **68**, 932 (1997).

[56] M. R. Endsley, *Automation and situation awareness,* in *Human factors in transportation. Automation and human performance: Theory and applications*, edited by R. Parasuraman and M. Mouloua (Lawrence Erlbaum Associates, Inc., Hillsdale, NJ, US, 1996) pp. 163–181.

[57] L. Sherry, M. Feary, P. Polson, and E. Palmer, *What's it doing now? taking the covers off autopilot behavior,* in *Proceedings of the 11th International Symposium on Aviation Psychology* (2001) pp. 1–6.

[58] W. J. Muhren and B. Van de Walle, *A call for sensemaking support systems in crisis management,* in *Interactive collaborative information systems*, edited by R. Babuška and F. C. Groen (Springer, 2010) pp. 425–452.

[59] C. Borst, F. Sjer, M. Mulder, M. M. Van Paassen, and J. Mulder, *Ecological approach to support pilot terrain awareness after total engine failure,* Journal of Aircraft **45**, 159 (2008).

[60] D. B. Beringer and H. C. Harris, Jr, *Automation in general aviation: Two studies of pilot responses to autopilot malfunctions,* The International Journal of Aviation Psychology **9**, 155 (1999).

[61] W. L. Martin, P. S. Murray, P. R. Bates, and P. S. Lee, *A flight simulator study of the impairment effects of startle on pilots during unexpected critical events,* Aviation Psychology and Applied Human Factors (2016), 10.1027/2192-0923/a000092.

[62] J. A. Schroeder, J. Bürki-Cohen, D. A. Shikany, D. R. Gingras, and P. Desrochers, *An evaluation of several stall models for commercial transport training,* in *Modeling and Simulation Technology Conference, Washington, DC* (2014).

[63] E. M. of Civil Aviation, *The final report of the accident investigation: Flash airlines 604,* (2004).

[64] F. H. Previc and W. R. Ercoline, *Trends in spatial disorientation research,* Aviation, space, and environmental medicine **72**, 1048 (2001).

[65] R. J. Mumaw, E. Groen, L. Fucke, R. Anderson, J. Bos, and M. Houben, *A new tool for analyzing the potential influence of vestibular illusions,* in *ISASI Forum, January-March, 6-12* (2016).

[66] F. H. Previc and W. R. Ercoline, *The outside-in attitude display concept revisited,* The international Journal of aviation psychology **9**, 377 (1999).

[67] Australian Transport Safety Bureau, *Flight path management occurrence involving boeing 737-838, vh-vye,* https://www.skybrary.aero/bookshelf/books/3153.pdf (2015).

[68] Bureau d'Enquêtes et d'Analyses pour la Sécurité de l'Aviation Civile, *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,* (2012).

[69] J. I. de Accidentes de Aviación Civil, (2010).

3

Performance issues caused by surprise

This chapter describes a study to test the effect of surprise on pilots' performance of the stall recovery procedure. Whereas training settings are usually very predictable, issues in the real world are surprising, which means that performance of well-known procedures may suffer significantly. Using a within-subjects design, stall recovery performance of 20 pilots is tested in an anticipation condition and in a surprise condition in a motion-based simulator with a post-stall aerodynamic model. Pilot performance, as well as subjective and physiological data relating to surprise and startle are measured. The results show that pilots had significantly more difficulties with adhering to the recovery template in the surprise condition compared to the anticipation condition. The subjective and physiological measures confirm that pilots were more surprised and startled in the surprise condition. The results indicate that pilots have more difficulty in managing an upset situation (i.e., an aerodynamic stall) when this situation is presented unexpectedly. This underlines the importance to train specifically for the element of surprise.

3.1. Introduction

Loss of control in-flight currently makes up the largest category of fatal aviation accidents (see, Figure 1.1, [2]). A number of these events have been associated with inappropriate responses of the flight crew [3]. It is commonly suspected that surprise and startle contribute to such inappropriate responses (see e.g., Colgan Air flight 3407; [4]. For this reason, aviation authorities recommend the introduction of surprise and startle in upset prevention and recovery training ([5, 6]).

As outlined in our conceptual model [7], startle and surprise can both impair pilot performance, but in different ways. The negative consequences of startle involve an acute increase in stress [8], which may negatively affect cognitive functioning as well as perceptual-motor control [9]. The negative consequences of surprise include the need for mentally taxing efforts, or "sensemaking activities", to solve the existing cognitive mismatch, [10] before one can take appropriate actions. Current cognitive models propose that interrelated knowledge and procedures are grouped in cognitive structures such as schemata or frames [10]. Information is processed within the context of the currently active frame. If a mismatch arises between perceived information and the active frame, a frame switch may be required (i.e., "reframing" [11]). Reframing is relatively effortful, potentially requiring reasoning and knowledge-based behavior, meaning that it is vulnerable to negative aspects of stress [9]. Difficulties with reframing may express themselves as confusion, loss of "grip" on the situation, or the adoption of an inappropriate frame.

Some recent experimental studies addressed the effects of startle or surprise on pilot performance. The results show that pilots have substantial difficulties with applying learned procedures when they are surprised [12, 13], or when they must recover from an aerodynamic stall without prior refresher of recovery procedures [14]. It was also shown that the time to respond to an event increases when this event comes unexpected [12] or when it is accompanied by a startling stimulus [15].

Although these studies seem to demonstrate the usefulness of familiarizing pilots with unusual flight situations (e.g., upsets, aerodynamic stall), as well as introducing elements of surprise or startle, the design of the studies was not optimized to conclusively show the effects of surprise. For example, the studies of Ledegang and Groen [14] and Martin et al. [15] did not require a reaction to a surprising event, while the study of Martin et al. [15] was focused on startle instead of on surprise. Schroeder et al. [13] did not include a control condition to check if the observed performance decrements were attributable to surprise. Casner, Geven and Willams [12] did not report a manipulation check. The current study was aimed at complementing these previous studies by specifically manipulating surprise, while including a control condition and a manipulation check. As a novel approach, we included not only self-report measures but also physiological measures to check the surprise manipulation. The results of this study should provide an indication of the effects of surprise on pilot performance in a simulated upset event, which should be of interest to those who seek to simulate such events for training or research purposes.

We expected that a mismatch between pilots' expectations and the upset event (surprise), would lead to lower adherence to the recovery procedure. This is because the retrieval of the procedure from memory should be more difficult when pilots are

not in the correct frame. On the other hand, when an upset event is anticipated, sense-making may occur beforehand, eliminating the need to reframe at the moment when the upset occurs.

3.2. Method

3.2.1. Participants

Twenty male airline pilots participated in the study (mean age = 36.3 years, *SD* = 7.88; mean flying experience: 12.4 years, *SD* = 5.05; 6986 flight hours, *SD* = 3804). Experience in operating medium-size twinjet aircraft types was required. Eight pilots had mainly experience with the A330, five with the B737, six with the E190 and one with the A320. All pilots were employed at the time of the experiment, and they had been on duty at least once in the week prior to the experiment. Five were currently employed as captains, eleven as first officers and three as second officers. To limit inter-individual differences in experience, exclusion criteria for participation were: expecting a jetlag at the time of the experiment; having participated in similar experiments; being a simulator training instructor; having experience with flying in the air force, with aerobatics, or with glider flying. The trait anxiety scores (29.0, *SD* = 6.23) of the participating pilots, measured beforehand with the State-Trait Anxiety Inventory [16] were significantly lower than the norm (i.e., 36.7; t = -5.57; $p < .01$), indicating that they were not extraordinary sensitive to threat. The pilots provided written informed consent prior to participation and the ethics committee of the TNO Soesterberg research institute approved the experiment.

3.2.2. Apparatus

The experiment was performed in the Desdemona flight simulator (AMST Systemtechnik; see [14], Figure 3.1a), 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 8 meters laterally on a horizontal track. The centrifuge capability of the simulator was not used to generate g-forces. The cockpit mockup was styled after the Boeing 737NG (see, Figure 3.1b, and included the left-side seat, primary flight display (without pitch limit indicator), navigation display, engine-indicating and crew-alerting system, 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 [17]. The model includes aerodynamic phenomena like buffeting, longitudinal and lateral instabilities, dynamic hysteresis and degradation of control response [18].

(a)   (b)
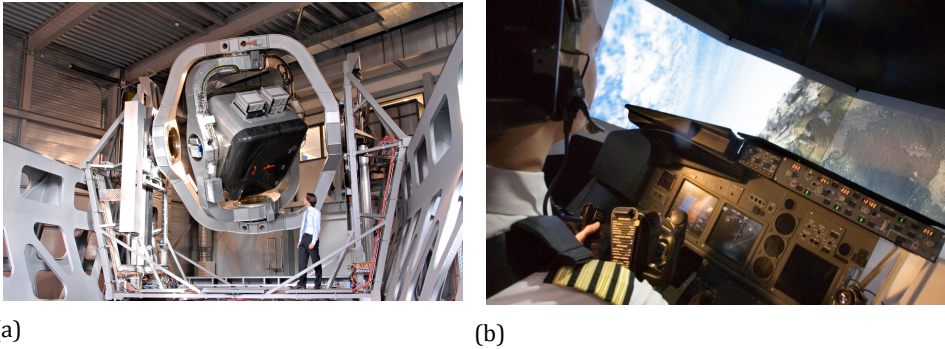
Figure 3.1: (a) The Desdemona flight simulator, located at TNO Soesterberg, the Netherlands. (b) The interior of Desdemona with the mockup as used in the experiment. The picture also shows an overbank situation due to a wingdrop following a stall, which often occurred in the experiment.

3.2.3. Task and conditions

Before the experiment, pilots were informed that the simulator session would comprise two subsequent sections of approximately 20 to 30 minutes. They were told that they would perform recoveries from upsets and stalls to validate the simulator's aerodynamic model in the first section, and that they would judge the fidelity of several simulated spatial disorientation illusions in the second section of the experiment. In reality, the first section was only used for practice, while the second section would not take place as described. It was made up to manipulate the pilots' expectation before test conditions. Figure 3.2 shows an overview of the experimental design.

First, a briefing on aerodynamics and recovery techniques was given in a 20-minutes session to groups of two pilots. They were asked to respond accurately to any simulated situation as if it was real, unless they were explicitly instructed to do otherwise. They were informed that sometimes, they would be asked to fly manually straight and level for a few minutes with the purpose of obtaining a baseline measure of the physiological parameters. During the briefing, pilots received verbal instructions about the simulated aircraft model and the stall recovery template as advised by the FAA ([6], p. 2), which involves the following steps:

1. Disconnect the autopilot and autothrottle / autothrust systems.
2a. Apply nose down pitch control until impending stall indications are eliminated.
2b. Use nose down pitch trim as needed.
3. Roll wings level.
4. Apply thrust as needed.
5. Retract speed brakes or spoilers.
6. Return the aircraft to the desired flight path.

Then, one pilot was outfitted with the physiological measuring equipment and seated in the simulator, while the other pilot waited in a different room. The simulator practice session consisted of a few minutes familiarization with the aircraft model (by performing basic flight maneuvers), followed by practicing recoveries from eight different upsets, in a set order, taken from the Airplane Upset Recovery Training Aid [19].

This practice session was aimed at providing a basic familiarity with the aircraft model outside the normal flight envelope, and to prevent potential excess in stall recovery behavior. The first four upsets involved unusual attitudes, starting with the aircraft in the following states: 1) 35 deg pitch down at 5,000 ft, 2) 22 deg pitch up at 5,000 ft, 3) 35 deg pitch down at 37,000 ft. and 4) 120 deg overbank at 10,000 ft. Next, four recoveries from aerodynamic stalls were exercised: 5) A level flight stall at 20,000 ft, 6) A 15-20 deg pitch up stall at 38,000 ft (the pilots experimented with aileron inputs during the stall until the wing dropped and they recovered), 7) A 15-20 deg pitch up stall at 20,000 ft, and 8) A 20 deg pitch down stall at 7,000 ft. and at low speed. The final exercise 8 was repeated until the pilot was able to push down quickly and forcefully enough to avoid stick shaker events, while avoiding overspeed or excessive g-load. It took pilots on average 2-3 times to succeed, while the maximum number of required attempts was 5 times. Pilots received feedback on their performance from the instructor. The angle of attack (AoA) was displayed during all scenarios except the last. Following exercises 5 and 7, pilots were asked to fly manually straight and level for two minutes in order to habituate them with this task.

Unbeknownst to the pilots, the practice session transitioned into two test conditions in which the same aerodynamic stall scenario (see Figure 3.2) was presented, once in a surprise condition and once in an anticipation condition. The latter served as a control condition. The order of the two test conditions was counterbalanced between subjects. The two resulting groups (with order: anticipation-surprise and order: surprise-anticipation, see Figure 3.2) were added together for analysis.
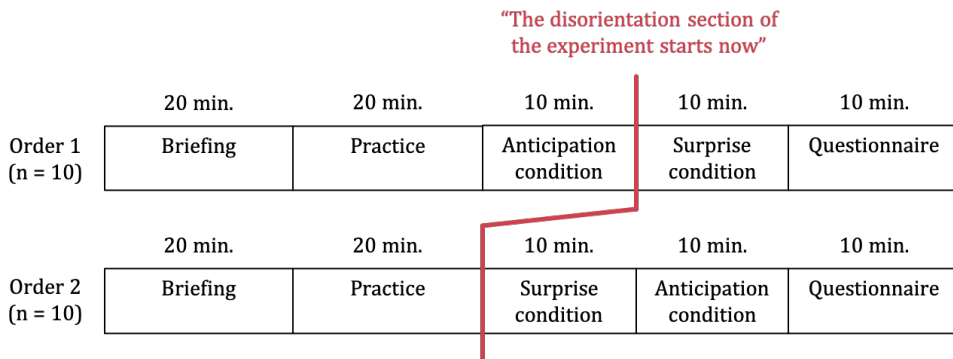
Figure 3.2: Experimental design.

In the anticipation condition, the pilots were told that, when they crossed a landmark after flying for three minutes at 5000 ft., an external factor would bring them into a stall. They were instructed to recover from this stall as safely as possible. The scenario occurred in accordance with the instructions. The stall was induced by creating a strong tailwind (decreasing the calibrated airspeed (CAS) by 75 knots in five seconds), and by simultaneously adjusting the pitch trim up, towards 48% of the maximum, in 3 seconds time. In a post-hoc questionnaire all pilots reported that they had not been aware of any pitch trim adjustment. The simulator aerodynamic model was set to induce a slightly asymmetric stall, so that one wing would stall quicker than the other,

which results in a "wing drop" when the stall is not arrested.

In the surprise condition, exactly the same stall event was induced about five seconds before the landmark was reached. In this case, however, several measures were taken to mislead the pilots and to activate a cognitive frame that would mismatch with the stall situation. First, pilots were made to believe that the experiment would include a section about spatial disorientation. To make this more convincing, the general questionnaire that was taken before the experiment included several questions on the pilots' experience with spatial disorientation. Also, pilots were told that the DESDEMONA simulator is particularly suitable for the reproduction of spatial disorientation illusions (DESDEMONA is an acronym for "disorientation demonstrator" in Dutch). Hence, in the surprise condition, the pilots were asked to do a climb-out above the landmark, and to pay special attention to pitch sensation as part of a potential somatogravic illusion. Finally, to further increase the mismatch between the stall event and the pilot's active frame, their attention was taken away from the displays at the initiation of the stall. This was done by asking them to give a rating on a sickness scale that was displayed in the lower right of the cockpit, next to the throttle levers.
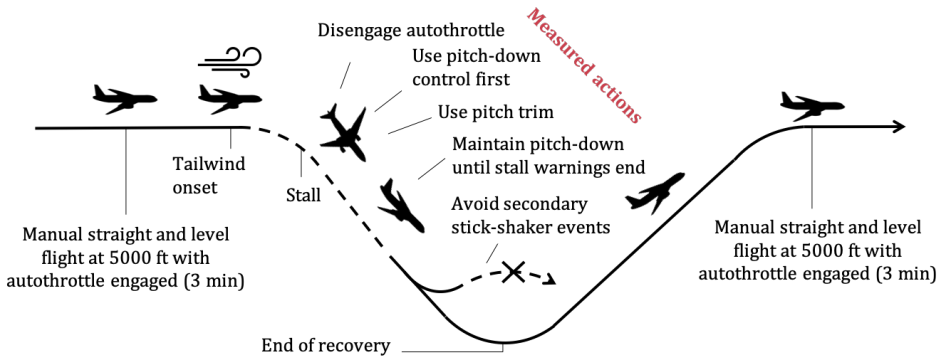
Figure 3.3: The stall recovery test scenario and the measured recovery actions.

3.2.4. Dependent measures
Performance
During the experiment, flight parameters were logged from the simulator at a sample rate of 100 Hz. These flight parameters were twice (forth and back) low-pass filtered using a 2nd order Butterworth filter with a cut-off frequency of 2 Hz. To extract the relevant data, the times of occurrence of several events during the recovery were determined. First, tailwind onset was defined as the start of the externally induced decrease in CAS (and adjusting of the pitch trim). Second, to discern a pitch down control response to the stall from normal fluctuations, the first moment after tailwind onset at which the cumulative sum of the pitch control signal moved beyond 5 standard deviations (SD) of its mean was determined (see e.g. [20]). The mean and *SD* were obtained from two minutes of straight and level flight before tailwind onset. Since the *SD*s of rudder and aileron inputs before tailwind onset was sometimes zero due to the absence

of turbulence, any change in rudder or aileron deflection was counted as a significant control response. Third, moments of interrupting or ending pitch down control inputs were defined as moments at which the signal moved back to within 5 *SD*s from the mean. Finally, the end of the recovery was defined as the moment at which the descent stopped. The data were visually inspected to check whether each of these moments was detected correctly. The start of the significant pitch down control input was manually reset for one pilot who gave a brief 10% pitch down input that lasted for approximately one second before truly starting pitching down (see, Figure 3.5).

As the recovery from upsets can be highly dynamic and complex, it is difficult to determine a single or objective performance criterion. In line with the FAA's ([6], p. 16) recommendations, performance was evaluated by checking four criteria that were derived from the template's principles. These criteria were created in such a way that they could be measured unambiguously in the data. Consequently, our criteria may diverge from those used for proficient recovery training. Table 3.1 shows the four criteria that were checked using the simulator data and the corresponding steps from the FAA template described in the 'tasks and conditions' subsection. The scoring on these performance criteria (met/unmet) was determined from the simulator data using a MATLAB script.

In addition to these binary variables, a number of flight parameters was collected to obtain a general impression of the stall event and the pilots' performance profiles: response time (from tailwind onset to the first significant control input, including autothrottle disengagement, pitch, bank and rudder control), recovery duration (from the first significant control input to the end of the descent), maximum and minimum CAS, maximal rate of descent, maximal vertical g-load (Nz), duration of secondary stick shaker events (see C3) and total altitude loss. It should be noted that these parameters are not necessarily indicative of the quality of the pilots' performance and that the outcomes are likely to be influenced by the distraction manipulation.

Manipulation check and subjective variables

As a measure of acute stress during the test scenarios, ratings of perceived anxiety were collected on a 0-10 point Likert-type version of the Anxiety Scale [21]. The surprise manipulation was first checked by asking the pilots whether they had expected an upset at the landmark (yes/no). To obtain measures of the level of perceived surprise, startle and confusion, similar 0-10 points Likert-type scales as the Anxiety Scale were created. For surprise and startle, the following questions were used: "Were you surprised by the upset?" and, respectively, "Were you startled by the upset?" . These could be rated from: "not at all" (0) to "very much" (10). Confusion was questioned by an inversely-scored item: "Did you immediately know how to respond when the upset occurred?" Overall workload during the recovery was rated on an Overall Workload Scale, ranging from 1 to 20 (see also [22]). To avoid suspicion during the experiment, subjective data were collected after the two conditions had ended. The pilots completed the questionnaires for each condition in the order that the conditions were presented to them and did not visually compare ratings between the conditions.

Table 3.1: Description of the four measured performance criteria, with the corresponding FAA ([6], p. 2) recovery template principles.

Criterion	Corresponding FAA principle(s)	Requirements to meet criterion
C1. Disengage autothrottle early	1	Disengage the autothrottle, no later than 2.0 s after the first significant yoke or pedal input.
C2. Start with pitch down control	2a, 3	Give priority to pitch down control by starting the recovery with pitch down control inputs. Strong aileron or rudder inputs (>50% of max) may not occur at around the same moment (within 2.0 s) of starting pitch down control to meet this criterion.
C3. Sufficient adjustment of loading	2a, 6	Respond (within 2.0 s) to stick shaker events with significant pitch down control and maintain significant pitch down control during stick shaker activation. Or, apply sufficient pitch down control to avoid any stick shaker events. Timing and strength of unloading and reloading should be so that secondary or late stick shaker events are avoided. Stick shaker events were defined as secondary if they occurred subsequent to an earlier stick shaker event, or late if they occurred after the first unloading action, i.e., following the first peak of pitch down control.
C4. Apply pitch down trim	2b	Using the pitch trim to aid in pitch down control during the recovery.

Physiological measurements

The physiological measurements were performed using Shimmer3 sensor units (Shimmer, Dublin, Ireland). ECG was measured at 128 Hz with five electrodes placed on the pilot's chest, which were connected to a portable data collector. RR (heartbeat) interval durations were determined from the ECG signal using a script [23] implementing the algorithm of Pan and Tompkins [24]. Artifacts in the RR intervals were removed with linear interpolation. From the resulting data, mean heart rate (HR) was obtained. The increase in mean HR (Δ mean HR) during the pilot's response was determined by taking the mean HR over the 10 seconds following the pilot's first significant control input, and subtracting a baseline mean HR measured between 130 seconds to 10 seconds prior to tailwind onset.

Skin conductance data were obtained at 8 Hz using two electrodes, placed approximately 4 cm apart on the ventral side of the pilot's left underarm, and using a portable data collector placed on a strap around the pilot's left wrist (Figure 3.4). The data was twice band-pass filtered (forth and back) using a 2nd order Butterworth filter with a

bandwidth of .01 to 2 Hz to eliminate drift and movement artifacts. The peak skin conductance in the 10 seconds after tailwind onset was measured, and standardized by subtracting the mean of the skin conductance in the 10 seconds before tailwind onset [25].

Since HR and skin conductance are indicative of both stress and mental workload, the outcomes were expected to be higher in the surprise condition compared to the anticipation condition.

3
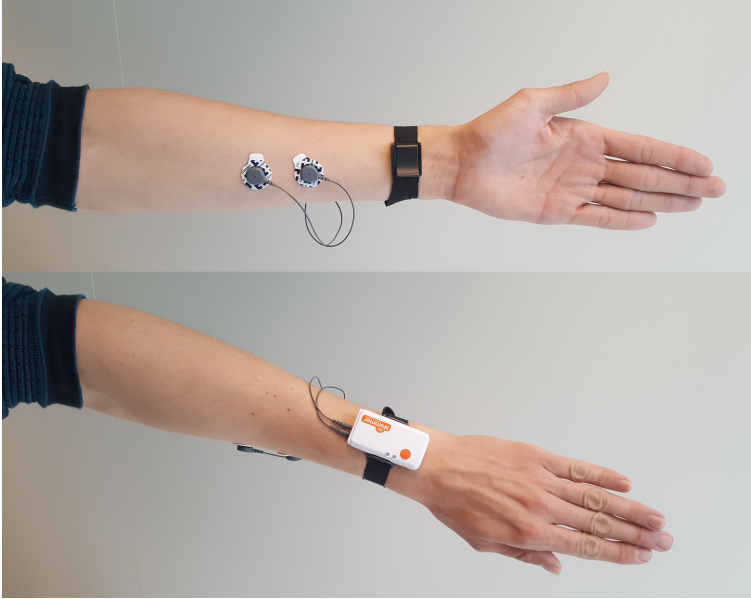
Figure 3.4: The Shimmer3 skin conductance sensor with two electrodes, as mounted on the arm of a participant.

3.3. Statistical analysis

Outliers, defined as values falling outside three times the interquartile range, were excluded from statistical analyses (non-binary measures only). The effect of Condition (anticipation or surprise) on the binary performance variables, i.e., meeting the criteria, was tested using generalized estimating equations (GEE) models of logistic regression. We controlled for the effect of the order of conditions by first entering the order and the order × condition interaction as predictors into the model. Effect sizes of the GEE analyses were calculated by transforming the odds ratio (exp B; cf., [26]).

The effect of Condition on the general flight parameters and on the pilots' subjective and physiological (state) measures was tested with paired-samples t-tests. The significance level of all analyses was set at alpha = .050. To limit potential type-I errors, the outcomes of the template adherence variables and the pilot state measures were checked separately using Holm's sequential Bonferroni [27]. Since the general flight parameters were not measured to test a hypothesis, but instead to describe the

performances, these were not corrected in this manner.

3.4. Results

3.4.1. Performance examples

Figure 3.5 shows the control inputs of participant number 8 and the aircraft's state in the surprise condition. This pilot met all criteria in the anticipation condition, but failed to meet two criteria in the surprise condition. Tailwind onset occurred at t = 0. The pilot's first response was a maximal pitch down control input and autothrottle disengagement at around t = 6 seconds. The bottom plot shows that a bank angle developed during the stall, to which the pilot immediately responded by giving maximum aileron control inputs in the opposite direction (both displayed as positive in accordance with conventions). Since pitch down control and strong aileron inputs occurred at the same time, the pilot did not meet criterion C2 of pitching down first. The pilot also responded late (>2 seconds) to the stick shaker, meaning that criterion C3 of sufficient adjustment of loading was not met. As can be seen in the figure, the pilot met the criteria of using pitch trim and disengaging autopilot early.
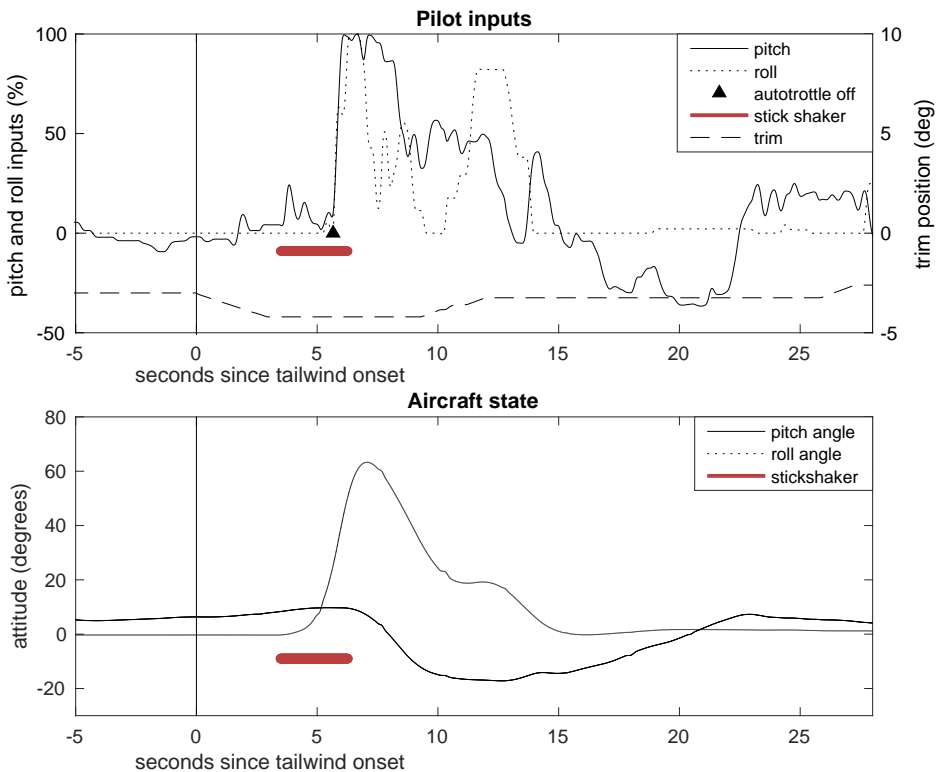
Figure 3.5: Time history of the control inputs of participant 8 (top plot) and the aircraft's state (bottom plot) in the surprise condition.
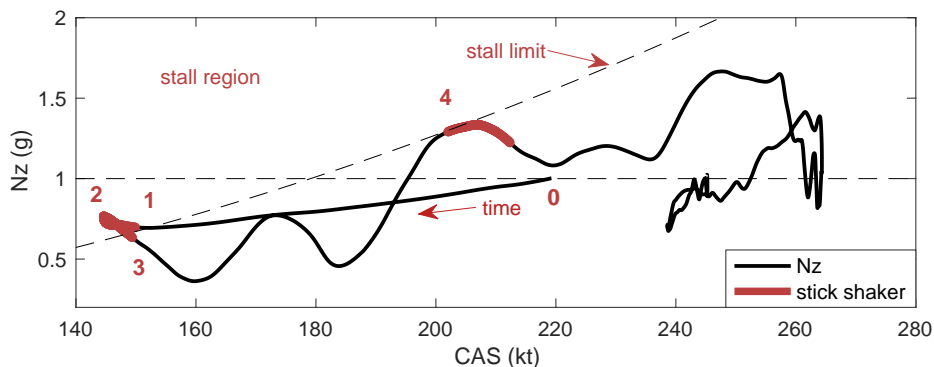
Figure 3.6: V-N diagram of the recovery of participant number 2 in the surprise condition.

Figure 3.6 shows an example of not meeting our criterion of sufficient adjustment of loading, by means of the V-N diagram of participant number 2 in the surprise condition. In a V-N diagram, airspeed (CAS) is plotted against vertical g-load (Nz) in order to display the aerodynamic boundaries. Tailwind onset starts at CAS = 220 kt and Nz = 1 (event 0 in Figure 3.6). The tailwind reduced CAS to approximately 145 kt, which unloaded the aircraft to .7 g and elicited a stick shaker event (1). The brief increase in Nz during the first stick shaker event (2) can be attributed to rolling wings level in response to a wing drop (not shown). This was followed by pitch down control, so that Nz dropped (3). However, when CAS reached 200 kt again, the pilot started loading the aircraft too aggressively and too early (Nz increases), leading to a secondary stick shaker event (4).

3.4.2. Adherence to the recovery template

The bitmap in Figure 3.7 provides an overview of the adherence to the four performance criteria by each pilot in the two conditions. One pilot (5%) met one more criteria in the surprise condition than in the anticipation condition. Three pilots (15%) met an equal number of the criteria in both conditions; nine pilots (45%) met one fewer criterion; three pilots (15%) met two fewer criteria; four pilots (20%) met three fewer criteria, and no pilots met four fewer criteria in the surprise condition than in the anticipation condition.

Table 3.2 provides an overview of the GEE analyses, testing for differences between conditions for each of the performance criteria that were measured. All differences are statistically significant after Holm-Bonferroni correction, with effect sizes (d) varying from medium to large, i.e., in or above the range of .5 to .8. In sum, the surprise manipulation caused a significant decrease in adherence to the criteria that were based on the recovery template.

3.4.3. Flight parameters

Table 3.3 summarizes the statistics of the flight parameters in both conditions. None of the participating pilots approached or exceeded critical safety limits. A number of the parameters suggests that recovering in the surprise condition was more difficult. However, these outcomes were likely influenced by the distraction manipulation and

	Anticipation					Surprise				
	C1	C2	C3	C4	Sum	C1	C2	C3	C4	Sum
PP: 1	□	■	□	□	1	□	■	□	□	1
2	■	■	□	□	2	□	■	□	□	1
3	□	■	■	■	3	□	□	□	□	0
4	□	■	□	■	2	□	■	■	□	2
5	□	□	□	■	1	□	□	□	□	0
6	■	■	■	■	4	□	□	■	□	1
7	■	■	□	□	2	□	■	□	□	1
8	■	■	■	■	4	■	□	□	■	2
9	■	■	■	□	3	□	■	□	□	1
10	■	□	□	□	1	□	□	□	□	0
11	□	□	□	□	0	□	□	□	□	0
12	□	■	■	■	3	□	■	■	□	2
13	■	■	□	□	2	□	■	□	□	1
14	■	■	□	□	2	■	□	□	□	1
15	■	■	□	□	2	■	□	□	□	1
16	■	■	□	□	2	■	■	■	□	3
17	■	■	■	■	4	□	□	□	■	1
18	□	■	■	■	3	□	□	□	■	1
19	□	□	■	■	2	□	□	■	□	1
20	□	■	■	■	3	□	□	□	□	0
Total met (%)	55	80	50	45		30	30	25	15	

□ Criterion unmet   ■ Criterion met
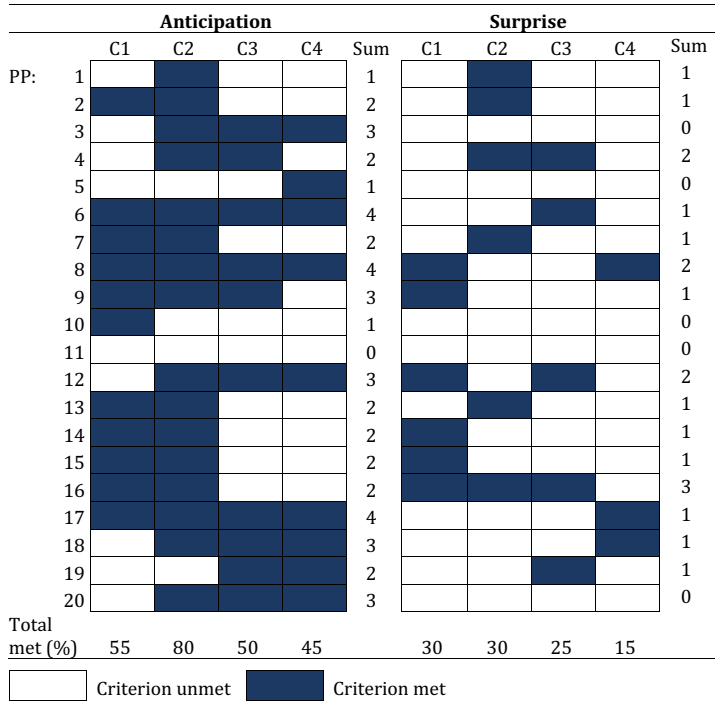
Figure 3.7: Bitmap of the performance criteria of each pilot. PP: participant number, C1: disengage autothrottle early; C2: start with pitch down control; C3: Sufficient adjustment of loading; C4: use pitch trim.

Table 3.2: Criteria met in the two conditions. C1: disengage autothrottle early; C2: start with pitch down control; C3: Sufficient adjustment of loading; C4: use pitch trim.

	Anticipation met/unmet	Surprise met/unmet	N	Δ	χ^2	p	Cohen's d
C1	11/9	6/14	20	-5*	5.10	.024	0.69
C2	16/4	6/14	20	-10*	13.66	<.001	1.26
C3	10/10	5/15	20	-5*	3.96	.046	0.61
C4	9/11	3/17	20	-6*	7.07	.008	0.85

* Difference is significant at $p < .05$ after Holm-Bonferroni correction.

do not necessarily suggest decreased performance.

3.4.4. Subjective ratings

An overview of the results of the subjective and physiological measures is shown in Table 3.4. In the surprise condition, all pilots reported that they did not expect an upset to occur at the landmark. In the anticipation condition, one pilot reported that he did not expect an upset to occur at the landmark, but instead suspected something to occur before reaching the landmark. The ratings of surprise were significantly higher in the

Table 3.3: The means and standard deviations (*SD*'s) of the flight parameters in the two conditions.

	Anticipation Mean (*SD*)	Surprise Mean (*SD*)	Δ	*t*	*p*	Cohen's *d*
Response time (s)	2.02 (.92)	4.12 (1.08)	2.10	8.59	<.001	2.09
Recovery duration (s)	22.83 (2.43)	21.86 (2.90)	-.96	-1.42	.172	.36
Min CAS (kt)	146.9 (2.23)	144.2 (1.16)	2.63	-5.62	<.001	1.53
Max CAS (kt)	254.5 (16.6)	260.7 (17.7)	6.22	1.62	.121	.36
Max descend rate (ft/min)	6502 (1163)	7388 (1309)	886	2.70	.014	.72
Max Nz (g)	1.50 (.09)	1.57 (.16)	.07	2.05	.055	.54
2nd stick shaker (s)	.99 (1.27)	1.69 (1.94)	.70	2.03	.057	.43
Altitude loss (ft)	1508 (361)	1693 (385)	186	2.47	.023	.50

surprise condition compared to the anticipation condition, with a large effect size (*d* > .8). These results indicate that the surprise manipulation was indeed successful.

In addition to surprise, ratings of startle and workload were also significantly higher in the surprise compared to the anticipation condition. The difference in perceived startle constituted a large effect size (*d* > .80), while that of perceived workload was small to medium in strength (.2 < *d* < .5). Although perceived confusion was higher in the surprise condition, this difference did not meet significance after Holm-Bonferroni correction. Similarly, there was no significant difference in perceived anxiety during the recovery, which suggests that surprise did not cause an increase in the participants' level of stress.

Table 3.4: The means and standard deviations (*SD*) of the subjective and physiological measures in the two conditions.

	Anticipation Mean (SD)	Surprise Mean (*SD*)	Δ	N	*t*	*p*	Cohen's *d*
Surprise (0-10)	1.39 (2.00)	8.44 (1.50)	7.05*	20	12.35	<.001	3.99
Startle (0-10)	1.22 (2.00)	4.28 (2.32)	3.06*	20	5.48	<.001	1.41
Confusion (0-10)	2.78 (2.35)	3.50 (1.92)	.72	20	2.16	.044	.34
Workload (1-20)	12.00 (3.18)	13.20 (3.09)	1.20*	20	2.40	.027	.38
Anxiety (0-10)	3.28 (2.35)	4.06 (1.81)	.78	20	1.89	.074	.37
Δ Mean HR (BPM)	14.69 (6.14)	14.18 (6.40)	-.51	15	-.25	.805	.08
Δ Skin conductance (mS)	.05 (.10)	.19 (.31)	.14*	18	2.57	.020	.61

* Difference is significant at *p* <.05 after Holm-Bonferroni correction.

3.5. Discussion

In line with our conceptualization of surprise [7], the results of this simulator experiment show that the mismatch of an aerodynamic stall with expectations effectively surprised the pilots, while it negatively affected their adherence to the FAA stall re-

covery template. In line with previous studies (e.g., [12, 13]), the outcomes show that surprise can be used in simulated environments to cause meaningful challenges to recovery performance. Although our scenarios were somewhat unrealistic in order to achieve highly controlled experimental settings, more realistic scenarios may be created for training purposes. The proportion of pilots meeting each single criterion decreased with around 25 to 50%. A number of pilots also had difficulty meeting the performance criteria in the anticipation condition, suggesting that the instructions and the short practice session were insufficient to create recovery proficiency. Since we were limited in means and qualifications to train pilots to full proficiency, it cannot be ruled out that full proficiency would make performance robust against surprise. Still, all pilots were able to recover without dramatic altitude loss or overspeed, indicating that their overall performance was adequate even when surprised.

The subjective and physiological measures of the pilots' state showed that our manipulation of surprise was effective. They also suggest that the unexpected upset event was more startling and the recovery was mentally more demanding. The absence of a difference in perceived anxiety between conditions suggests that the unexpectedness of the upset event was perhaps not sufficiently threatening, or that the anticipated task caused similar levels of anxiety. In each case, our results do not rule out that excessive levels of stress, which are likely present during an unexpected upset in operational practice, could severely impact recovery performance, especially if pilots are also surprised and need to reframe the situation.

One pilot noted that he "had to think for a moment and regain control" in the surprise condition. Another indicated that he felt "completely unprepared", that he had a different "mental image" of the upcoming task and "had to switch". Pilots also remarked that the distraction method (i.e., being asked a question and turning away from the display) was very realistic and representative of distractions in practice. Finally, it was noted that in particular the conviction that a new phase of the experiment had started took them out of the "performance mode", which made them feel unprepared and surprised by the stall in the surprise condition.

Our outcomes substantiate recommendations of using an element of surprise in the training of upset recovery [5, 6], and indicate the importance of focusing such training on reframing and sensemaking abilities (see also, [11]), so that recovery skills can be made resilient against the effects of surprise. In this respect, the recommendation of using "variations in the types of scenario, times of occurrences and types of occurrence" ([28], section II-1-5), as an alternative to the use of predictable training scenarios, seems to make sense. This approach already has been shown to be beneficial to transfer of training to similar or new situations in the domain of sports [29]. Future research would be necessary to provide evidence of whether variability also produces beneficial effects in upset prevention and recovery training.

References

[1] A. Landman, E. L. Groen, M. M. Van Paassen, A. W. Bronkhorst, and M. Mulder, *The influence of surprise on upset recovery performance in airline pilots,* The International Journal of Aerospace Psychology **27**, 2 (2017b).

[2] B. C. A. Aviation Safety, *Statistical summary of commercial jet airplane accidents, worldwide operations, 1959-2017,* (2018).

[3] C. M. Belcastro and J. V. Foster, *Aircraft loss-of-control accident analysis,* in *Proceedings of AIAA Guidance, Navigation and Control Conference, Toronto, Canada, Paper No. AIAA-2010-8004* (2010).

[4] N. T. S. Board, *Aviation accident report: Loss of control on approach, colgan air, inc., operating as continental connection flight 3407,* (2010).

[5] European Aviation Safety Agency, *Loss of control prevention and recovery training: Notice of proposed amendment 2015-13),* (2015).

[6] Federal Aviation Administration, *Advisory circular 120-111, upset prevention and recovery training,* (2015).

[7] A. Landman, E. L. Groen, M. M. van Paassen, A. W. Bronkhorst, and M. Mulder, *Dealing with unexpected events on the flight deck: A conceptual model of startle and surprise,* Human Factors **59**, 1161 (2017).

[8] W. L. Martin, P. S. Murray, P. R. Bates, and P. S. Y. Lee, *Fear-potentiated startle: A review from an aviation perspective,* The International Journal of Aviation Psychology **25**, 97 (2015).

[9] M. W. Eysenck, N. Derakhshan, R. Santos, and M. G. Calvo, *Anxiety and cognitive performance: attentional control theory.* Emotion **7**, 336 (2007).

[10] G. Klein, J. K. Phillips, E. L. Rall, and D. A. Peluso, *A data-frame theory of sensemaking,* in *Expertise out of context: Proceedings of the sixth international conference on naturalistic decision making*, edited by R. R. Hoffman (Psychology Press, 2007) pp. 113–155.

[11] A. Rankin, R. Woltjer, and J. Field, *Sensemaking following surprise in the cockpit: a re-framing problem,* Cognition, Technology & Work **18**, 623 (2016).

[12] S. M. Casner, R. W. Geven, and K. T. Williams, *The effectiveness of airline pilot training for abnormal events,* Human Factors **55**, 477 (2013).

[13] J. A. Schroeder, J. Bürki-Cohen, D. A. Shikany, D. R. Gingras, and P. Desrochers, *An evaluation of several stall models for commercial transport training,* in *Modeling and Simulation Technology Conference, Washington, DC* (2014).

[14] W. D. Ledegang and E. L. Groen, *Stall recovery in a centrifuge-based flight simulator with an extended aerodynamic model,* The International Journal of Aviation Psychology **25**, 122 (2015).

[15] W. L. Martin, P. S. Murray, P. R. Bates, and P. S. Lee, *A flight simulator study of the impairment effects of startle on pilots during unexpected critical events,* Aviation Psychology and Applied Human Factors (2016), 10.1027/2192-0923/a000092.

3

[16] H. M. Ploeg, P. B. Defares, C. D. Spielberger, P. B. Defares, and C. D. Spielberger, *Handleiding bij de Zelf-Beoordelings Vragenlijst ZBV: een nederlandstalige bewerking van de Spielberger State-trait Anxiety Inventory STAI-DY* (Swets & Zeitlinger, 1980).

[17] E. Groen, W. Ledegang, J. Field, H. Smaili, M. Roza, L. Fucke, S. Nooij, M. Goman, M. Mayrhofer, L. Zaichik, *et al.*, *Supra-enhanced upset recovery simulation,* in *AIAA Modeling and Simulation Technologies Conference* (2012) p. 4630.

[18] M. Goman and A. Khrabrov, *State-space representation of aerodynamic characteristics of an aircraft at high angles of attack,* Journal of Aircraft **31**, 1109 (1994).

[19] Industry Airplane Upset Recovery Training Aid Team, *Airplane upset recovery training aid, revision 2,* (2008).

[20] M. Mulder, J.-M. Pleijsant, H. van der Vaart, and P. van Wieringen, *The effects of pictorial detail on the timing of the landing flare: Results of a visual simulation experiment,* The International Journal of Aviation Psychology **10**, 291 (2000).

[21] I. Houtman and F. Bakker, *The anxiety thermometer: a validation study,* Journal of personality assessment **53**, 575 (1989).

[22] C. R. Anthony and D. W. Biers, *Unidimensional versus multidimensional workload scales and the effect of number of rating scale categories,* in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 41(2) (SAGE Publications Sage CA: Los Angeles, CA, 1997) pp. 1084–1088.

[23] H. Sedghamiz, *Complete pan-tompkins implementation ecg qrs detector,* Matlab Central: Community Profile. Available online at: http://www. mathworks. com/matlabcentral/profile/authors/2510422-hooman-sedghamiz (2014).

[24] J. Pan and W. J. Tompkins, *A real-time qrs detection algorithm,* IEEE Trans. Biomed. Eng **32**, 230 (1985).

[25] W. Boucsein, *Electrodermal activity* (Springer Science & Business Media, 2012).

[26] S. Chinn, *A simple method for converting an odds ratio to effect size for use in meta-analysis,* Statistics in medicine **19**, 3127 (2000).

[27] S. Holm, *A simple sequentially rejective multiple test procedure,* Scandinavian journal of statistics , 65 (1979).

[28] International Civil Aviation Organization, *Manual of evidence-based training,* https://www.icao.int/SAM/Documents/2014-AQP/EBT ICAO Manual Doc 209995.en.pdf (2013).

[29] J. J. Van Merriënboer, *Training complex cognitive skills: A four-component instructional design model for technical training* (Educational Technology, 1997).

4

Performance issues caused by expectation

The current chapter describes two experiments on the effect of expectation on roll reversal errors (RREs) when responding to the attitude indicator (AI). In previous studies, pilots sometimes made RREs. We hypothesized that more RREs would occur when the bank angle on the AI mismatches with the expectation. In a fixed-base simulator, expectation of non-pilots is manipulated with a manual flying task. Participants have to use the AI to roll wings level while expecting a turn. The presented AI often matches this expectation, but it sometimes shows an opposite turn (Opposite condition) or level flight (Level condition). A session is included with no preceding flying task (Baseline). Similar conditions are created in an in-flight experiment, but now by using (misleading) cues while participants are blindfolded. In the simulator, participants make 7.8 times more RREs in the Opposite condition (75 % error rate) compared to Baseline (9.6 %). Additionally, participants make 2.5 times more RREs in the Opposite compared to the Level condition (30 %), indicating that misinterpretations played a role. In-flight, the presence of misleading cues increases the RRE rate by a factor of 2.6, but there is no significant difference between misleading conditions (both ca. 60 %). The results suggest that expectations strongly affect RREs, which should be taken into account when developing spatial disorientation awareness training or upset recovery guidance systems.

4.1. Introduction

In previous studies, pilots were found to sometimes make roll reversal errors (RREs) when referencing the moving-horizon type attitude indicator (AI; see, Figure 4.1), also known as the artificial horizon. An RRE occurs when a roll input is made towards the opposite to the required side. It has been argued that the ambiguity of the presented bank angle on the AI may cause misinterpretations which lead to RREs [2–4]. Due to the relative orientation and motion of the AI aircraft and horizon symbols, they can be confused. This causes the controller to attempt to roll the horizon symbol back to level (i.e., a horizon control reversal; [5]). The AI in Figure 4.1 (left) would in that case be interpreted as indicating a bank angle to the left instead of to the right. Horizon control reversals are thought to be facilitated by several properties of the AI design. People tend to control the moving part of a display, which is in this case the horizon symbol [4, 5]. The horizon also moves in the same direction as the roll control inputs, which may add to the confusion. Furthermore, the horizon symbol is clearly distinguishable, occupies the same space as the aircraft symbol, and it is not interrupted by the aircraft symbol, making it more difficult to correctly interpret the horizon symbol as being the background [5].

Previous research has shown that non-pilots as well as experienced pilots sometimes make RREs when having to respond to an AI that is presented to them. Studies in ground-based simulators, where participants were asked to roll back to wings level from a static bank angle, showed an RRE incidence of 3.9-8.0 % for pilots [6–8] and 15-20 % for non-pilots [7, 9, 10]. In-flight studies showed similar results in non-pilots [11], but better performance in pilots (1.5-3.1 % [6, 12]). Nevertheless, these error rates are still high from a safety perspective.

Figure 4.1: Left is the moving horizon type AI used in the simulator experiment, showing a left turn. The AI is a simplified version of the AI used in large jet airliners. Right is the AI used in the in-flight experiment, showing a right turn.

In these previous studies, the researchers always attempted to prevent the controller from having an expectation with regards to the bank angle. Nevertheless, participants may still have had expectations, for instance due to inadvertent motion cues.

These cues may cause an incorrect assumption of the bank angle, which is the most prevalent form of spatial disorientation known as "the leans" [13–15]. Spatial disorientation involves an erroneous sense of the aircraft's attitude and motion relative to the earth, which is caused by misleading vestibular, visual or proprioceptive cues [15, 16]. The leans is caused by the vestibular system not picking up on low roll accelerations, resulting in the sensation of flying at a bank angle that is incorrect. Spatial disorientation occurs most often in poor visibility conditions, when reading the instruments correctly is most crucial. It continues to be a hazard in commercial aviation, as it was estimated to be a factor in 12 % of loss of control accidents and 24 % of fatalities between 1996 and 2010 [17–19]. Accident reports suggest that leans-induced misinterpretations of the AI may have played a role in the accidents of Flash airlines flight 604 [20], Kenya Airways flight 5Y-KYA [21] and Crossair flight 498 [22]. The accident of Flash Airlines flight 604 occurred shortly following takeoff at night. The first officer alerted the captain of the gradual, unintended turn from left to right, upon which the captain expressed surprise. The captain (pilot flying) followed with a roll input that caused an increase in the bank angle, which led to an overbank and loss of control.

It seems plausible that an incorrect expectation of the bank angle increases the likelihood of an RRE, especially if this expectation coincides with a horizon control reversal (i.e., if the expected bank angle is opposite to the actual bank angle). Expectations and beliefs are known to modulate perceptual experiences (i.e., cognitive penetration [23]). For instance, an object is incorrectly judged as being more deeply red if it has the shape of a heart than if it has the shape of a square [24]. Observers also reported seeing shapes in white noise, or a smile on a neutral face, if experimenters told them that this information was present [25]. And knowledge of the dual-interpretability of figures made it more likely that participants reported both interpretations [26]. According to Bayesian statistical decision theory [27–29], the brain uses presuppositions to create a prediction of what the incoming information is likely to be, which influences perceptual processing on lower levels (i.e., predictive coding). However, the gain or penalty of the resulting decision, as well as the strategy with which one views the information are also recognized as influencing factors.

Although pilots are likely to minimize horizon control reversals by using their extensive experience or specific response strategies, the above-mentioned RRE rates in experienced pilots in-flight (i.e., 1.5-3.1 %) show that they are not impervious to making RREs. It is not yet known to which extent these RREs are caused by expectation, or by expectation-induced horizon control reversals. The current chapter describes two experiments aimed to add to existing literature by testing the effect of expectation and spatial disorientation on RREs. In a fixed-base simulator experiment, a manual flying task was used to simulate the mismatch of an expected bank angle with information on the AI. We predicted that this would make an RRE more likely to occur, especially if a horizon control reversal would match with the direction of the expected bank angle. This simulator experiment was followed by an in-flight experiment, which was set up to test the effect of a true leans, caused by cues of the aircraft motion, on the occurrence of RREs.

4.2. Method

4.2.1. Participants

Simulator experiment

Twenty non-pilots participants were invited at the Aerospace Engineering faculty of the Delft University of Technology (18 men, 2 women, mean age = 25.0, standard deviation, or SD = 3.2). Participants were right-handed, reported no vestibular issues, had (corrected to) normal vision, and reported being well-rested. Five participants had controlled a fixed-wing aircraft (in Visual Flight Rules) before on some occasion (maximum = 6 occasions) and thus had experience with controlling the bank angle. The participants rated experience with fixed-base simulated flying on average 1.9 (around "some") median = 2, SD = 1.6, on a 1 ("very little") to 5 ("very much") point Likert-type scale. This study complied with the tenets of the Declaration of Helsinki, the experiment was approved by the research ethics review board of the university and informed consent was obtained from each participant.

In-flight experiment

Forty non-pilot participants were invited from the Aerospace Engineering faculty of Cranfield University (34 men, 6 women, mean age = 25.1 years, SD = 3.7). Participants reported no vestibular issues, had (corrected to) normal vision, and reported being well rested. Twenty-three participants had previously controlled an aircraft on one or two occasions, while one was in flight training (ca. 20 hours). Participants rated their simulated flying experience on average at 1.93, median = 1, SD = 1.29, on a 1-5 Likert-type scale ranging from "none or very little" (1) to "very much" (5). The experiment was approved by the research ethics review board of the university and participants provided informed consent prior to participating.

4.2.2. Apparatus

Simulator experiment

The experiment was performed in a fixed-base simulator at the faculty of Aerospace Engineering in the Delft University of Technology (see, Figure 4.2). Participants were seated in an adjustable aircraft seat, in front of a liquid crystal display monitor displaying the AI (500 ×500 pixels; 14 ×14 cm; 4.4° visual angle; see Figure 4.1 (left) for a screenshot) and no other instruments. They controlled a control-loaded hydraulic side stick with their right hand, with a length of 9 cm, and 30° roll and 22° pitch excursion space. Three digital light processing projectors presented the outside view rendered with FlightGear on three screens. This resulted in a 180° field of view. The sun and moon were not in view. The aircraft model had a fixed speed of 120 knots, with controllable pitch and roll rate, whereas yaw rate was coupled to the bank angle. No rudder was used. The simulator data were logged at 100 Hz.

In-flight experiment

The experiment took place in a light propeller aircraft (Scottish Aviation Bulldog 122). Participants used a centerstick, and had the AI (Figure 4.1) available in front of them (see, Figure 4.3). Test runs prior to the experiment confirmed that when looking at
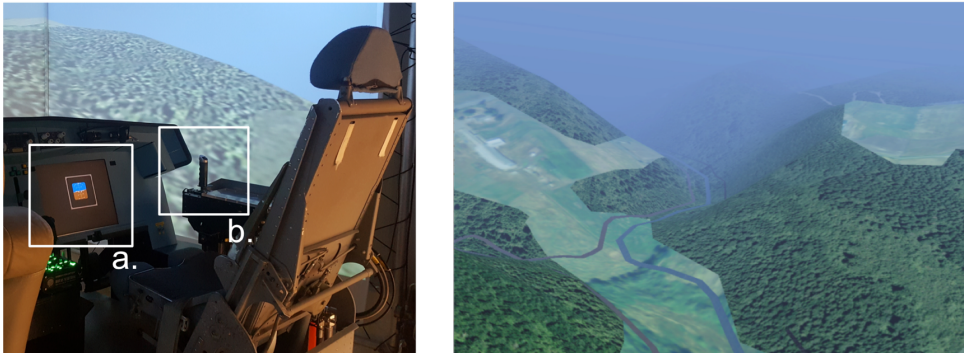
Figure 4.2: The simulator (left) with the AI on the monitor (a), the right sidestick (b) and two of the three screens in the background. The outside view (right), of which only the front screen is shown.
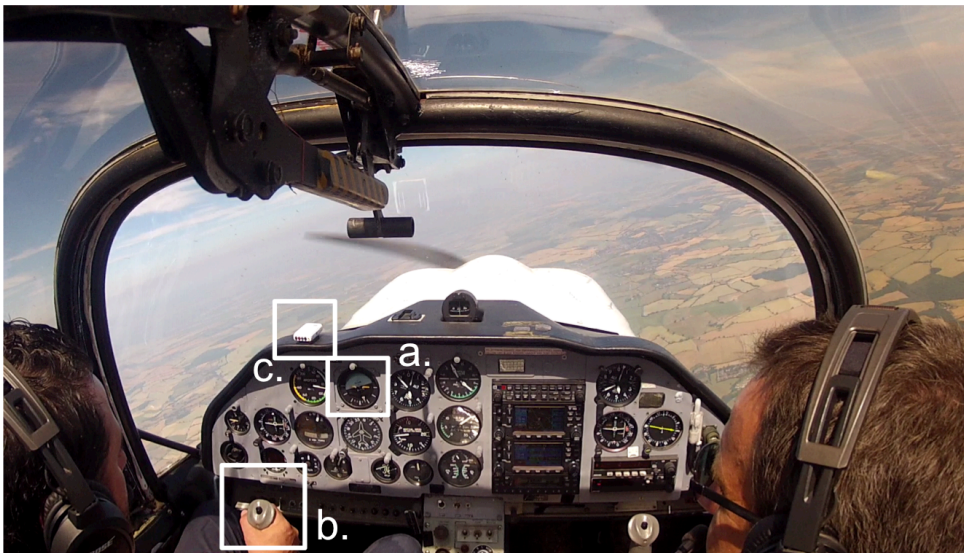
Figure 4.3: A screenshot of the video recording of the in-flight experiment. Left is the participant with the AI (a) the centerstick (b) and the IMU (c) visible, right is the test pilot.

the AI, it would be very difficult to distinguish slight (10°) slopes of the outside horizon.This is because the outside view is relatively bright compared to the instrument panel, which would require adjusting the eyes. Also, the front view is largely obstructed by the instrument panel. Roll rate of the aircraft was logged at 100 Hz using an Inertial Measuring Unit (IMU, Shimmersensing, Dublin, 500°s setting) attached to the top of the instrument panel in front of the participant. Roll rates were corrected by subtracting the mean roll rate of the whole flight. The stick inputs were filmed using a Gopro camera (See Figure 4.3 for a screenshot), placed above and behind the seats, facing the participant's center stick and the instrument panel.

4.2.3. Task and conditions

Simulator experiment

After a briefing, participants familiarized themselves with the simulator and the AI by flying a winding trajectory for three minutes while both the outside view and the AI were visible. Halfway in, participants were reminded to use the AI. They then performed two test sessions in a counterbalanced order: one session without a flying task before each AI presentation (Baseline condition, no manipulated expectation), and one session with a flying task before each AI presentation (other conditions, with manipulated expectation). The session with a flying task consisted of 22 runs, preceded by two practice runs. The order of events in each run is graphically presented in Figure 4.4.

1. The participant performs a flying task without AI.

2. When in a turn, the instructor calls out: "Steady".

3. The outside view fades to black.

4. The AI is shown two seconds later.

5. The participant is tasked to use the AI to roll level.
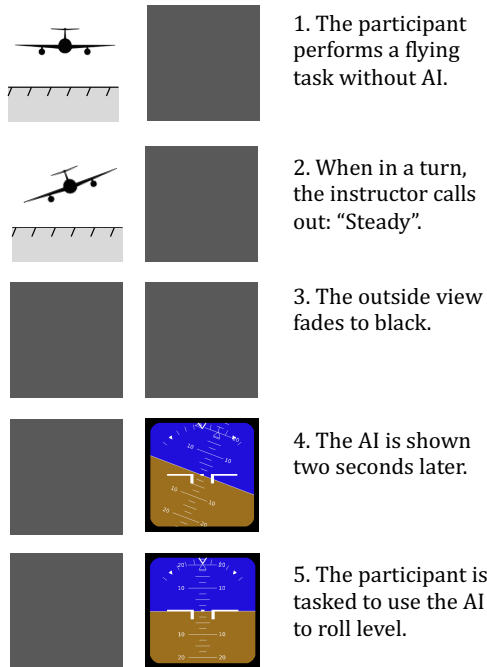
Figure 4.4: A timeline of the events in each run (with preceding flying task) in the simulator experiment. The aircraft is pictured from behind and the run is a matching run.

In each run, the participant was flying along a winding trajectory (indicated by a river) while the AI was not shown. After approximately one minute, when in a turn, the experimenter instructed the participant to hold steady, that is, maintain the flown bank angle by moving the stick to neutral. When flying steady, as checked by the experimenter, the outside view was removed (turned grey). After two s, the AI was shown, and the participant was to roll the wings level based on the AI. The shown AI could either exactly match the previously shown turn (Matching condition), show a bank angle

in the opposite direction (Opposite condition), or show level flight (Level condition, in which case no input was required). The different conditions in the simulator experiment are graphically listed in Figure 4.5.
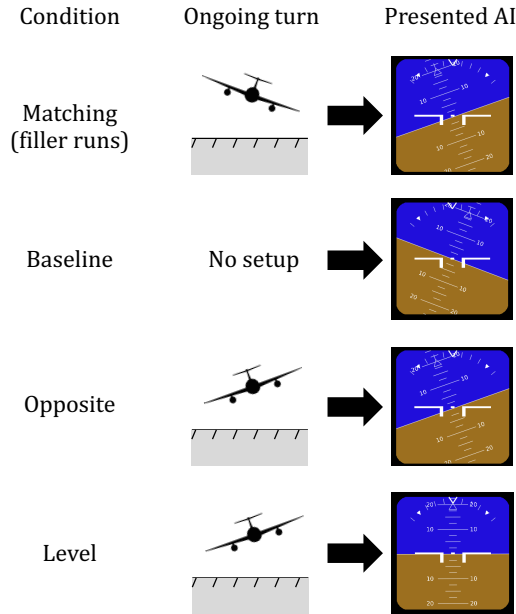
Figure 4.5: Examples of situations and corresponding AI configurations in each condition. The aircraft is pictured from behind.

Several measures were taken to improve the participants' involvement in the flying task, so that a strong spatial model of the assumed bank angle would be present when participants responded to the AI. The participants flew in a challenging mountainous area, whereas the visibility during the start of the each run was low (2000 m, see Figure 4.2, right). The trajectory in each run was different (although each participant performed the same runs), and the moment of the "steady" command was somewhat different between runs (SD = 17 s). The bank angles at which participants flew at the "steady" command are reported in the results. Furthermore, until the outside view disappeared, small disturbances ('gusts') lasting for 0.5 s were added to or subtracted from the roll rate and pitch rate with 1-5 s random time intervals. These disturbances stopped when the outside view disappeared, to make sure that participants responded to a steady AI, instead of to a roll motion. Participants were tasked to react to the shown AI immediately by rolling towards level, in order to offset their increased vigilance due to the experimental setting. It was emphasized that they should trust the AI and wait for it to appear before responding. Pitch angle was to be ignored by participants. To motivate participants to respond quickly and intuitively, their response time (time from AI presentation to steady level flight) was given as feedback after each run.

In 18 of the 22 runs, the AI bank angle matched the bank angle exactly as flown (Matching condition). These runs were used as filler runs to set up an expectation of a normal situation. However, in four of the 18 runs, the AI was manipulated to mismatch with the actually flown turn. In the Opposite condition (two runs), the AI indicated a bank angle of 10° into the opposite direction of the preceding turn (Figure 4.5). In the Level condition (two runs), the AI indicated level flight (Figure 4.5). The number of runs in these test conditions was kept low, so as to increase the element of surprise and reduce potential changes in response strategy. These four runs were distributed semi-randomly over the total 22 runs, with the following rules: 1. One run of each test condition occurred in the first and one in the second half of the 22 runs. 2. The runs of the test conditions were always preceded by at least three matching filler runs (with at least one of these runs ending left and one right). 3. The order of the four test runs was counterbalanced between participants. 4. The final run was a test run. The direction of the final bank angle in the Matching, Level and Opposite conditions was evenly distributed.

The session without a flying task before AI presentation (Baseline condition) consisted of 18 runs, with 5 practice runs at the start. These runs consisted only of the presentation of the AI (4-5 in Figure 4.4). An equal number of runs in this condition were presented with bank angles of 10, 20 and 30° left and right, and pitch angles of 0°, 5° and -5°. The order of these runs was randomized. There was approximately 5-10 s of time in between the runs.

In-flight experiment
After filling in a questionnaire and receiving a briefing, the participant was seated in the left hand seat of the aircraft and the experimenter pilot flew to the test area. The participant was then familiarized with the controls for approximately three minutes by flying left and right turns and leveling the aircraft from bank angles using the AI. Then, the participant performed a number of test conditions, with one run per condition. This run started with the participant putting on a blindfold. The pilot then flew a maneuver to induce a specific vestibular cue (see below). Immediately after, the participant was asked to take the stick with their dominant hand and, after a countdown from three, remove the blindfold and roll the wings level using the AI. The runs took place at an altitude with minimal turbulence and with the sun from behind. Tests were planned on days when the pilot judged the weather calm enough for minimal turbulence.

The maneuvers flown in each condition are listed in Figure 4.6. First, a number of practice runs (at least four, mean = 4.7, *SD* = 1.08) was flown until the test pilot considered the participant's performance to be adequate. In the practice runs, the cues were aimed to set up an expectation that matched the AI (the Matching condition). For the analysis of performance in this condition, the results of the third and fourth practice run were used. More practice runs were performed if the test pilot deemed performance inadequate. The practice session ended with a run in which the instructor waited 30 s before presenting the vestibular cue, to make this matching run similar to the subsequent test runs.

Three test runs followed, one for each test condition (No leans, Leans-opposite and Leans-level, see Figure 4.6). In the No Leans condition, the aircraft was rolled to 10°

bank slowly (at circa 0.3°/s and .01 Hz, which is below the 4.0°/s perception threshold; [30]), while flying in a coordinated turn. The intended expectation here was no bank. In the Leans-opposite condition, the aircraft was rolled similarly slowly to 20° bank, and then quickly back (at circa 5.0 °/s and .25 Hz) to 10° bank on the same side. The intended expectation here was a bank angle opposite to the actual bank angle. In the Leans-level condition, the aircraft was rolled slowly to 20° bank, and then quickly back to level. The intended expectation here was a bank angle, whereas the AI showed level flight. The direction of the fast roll in the test conditions was always the same within each participant, and it was counterbalanced between participants. Two variations of condition sequence were used. The first half of participants followed sequence A (1-2-3-4) and the second half sequence B (1-4-3-2). The numbers here indicate the conditions as numbered in Figure 4.6.

Figure 4.6: The four conditions in the in-flight experiment. The two most right columns show where the expectation is meant to deviate from, or match with, the actual situation.

4.2.4. Dependent measures
Error rate
In the simulator experiment, an input was registered if a stick input in the roll axis exceeded 1.5° In the Level condition, an error was registered if an input in the first 2.0 seconds following AI presentation caused a roll rotation away from level, into the opposite direction of the final turn in the flying task. The error rate was the ratio of the number of errors with the total number of runs in each condition (i.e., two runs in

Opposite and Level, and eighteen in Baseline).

In the in-flight experiment, the experimenter pilot observed the participant's first roll input and registered its direction on a log sheet. An error was registered if the first input caused the aircraft to roll away from level. This error would be an RRE in the Matching, No-leans or Leans-opposite condition, and an undesired input (not an RRE) in the Leans-level condition. The data on the log sheet were checked post-flight by an experimenter using video data, or, if video was not available, with IMU data. The agreement between both observations was high (98.1%). In case of disagreement the video analysis took precedence.

We expected error rates of the Baseline condition to be similar to previous studies in fixed-base simulators [7, 10]. The error rate was expected to be higher in the Opposite condition as well as in the Level condition than in the Baseline condition, due to the presence of an incongruent expectation. The error rate was also expected to be higher in the Opposite condition than in the Level condition, since the oppositely shown AI allows for misinterpretations (i.e., horizon control reversals), whereas the level shown AI does not.

Error durations

In the simulator experiment, After a first input was located in the data (see, 4.2.4), the start of this input was defined as the moment the stick started moving (> .06° /s or .001 rad/s) towards the direction of this input. The duration of an erroneous input was defined as the time from its start to the moment the stick passed neutral again. The mean was obtained for each condition.

In the in-flight experiment, video analysis was used to determine the start of the participant's first input, and the moment the participant started to move the stick back in the opposite direction again. The time between these moments was defined as the duration of an error. This definition was chosen instead of, for instance, the time until reaching level flight, to decrease potential variance due to inter-personal differences in control input strength.

If an erroneous input was based purely on the expectation, the feedback resulting from this input is expected to quickly allow the participant to intervene and roll back to level. If an erroneous input is instead caused by the expectation in combination with a horizon control reversal, which can only occur in the (Leans-)Opposite condition, the misinterpretation may confirm the incorrect expectation. This could increase the estimated likelihood of the incorrect assumption being correct, so that it requires more evidence to adjust. Thus, participants are expected to intervene in errors later in the (Leans-)Opposite condition than in the other conditions.

Response times

In the simulator experiment, the response time was defined as the time from presentation of the AI to the first input (see, 4.2.4).

In the in-flight experiment, the response time was defined as the time from removal of the blindfold until the start of the first input. These were both measured with video analysis.

If a participant is careful to look at the AI before responding, there may be a moment of hesitation before giving a correct response if the AI mismatches with the expecta-

tion. Thus, only response times of correct responses were compared, and we expected longer mean response times in the (Leans-)Opposite than in the Baseline condition. The (Leans-)Level condition cannot be included in this comparison because the correct response in this condition is to make no input.

Learning effect

In the simulator experiment, the predictive effect of the order of each run in the condition on the occurrence of an error in that run as well as its duration, was determined. If participants learned to anticipate the mismatching AI presentations, they would perform better in later runs.

In the in-flight experiment, to check whether there was a training or surprise effect on the occurrence of errors, the predictive effect of the sequence (A or B, Figure 4.6) on the occurrence of an error and the error duration was determined. If participants learned to anticipate the mismatching AI presentations, they may perform better in later runs.

Subjective measures

In the simulator experiment, participants rated their subjective simulator sickness on the 11-point Misery Scale, or MISC, where 0 = no problems, and 10 = vomiting [31] half-way and at the end of the session with a flying task. At the end of the session, participants also reported whether they noticed if the AI was sometimes incorrect. If they had noticed this, they rated whether they changed their response strategy because of this on a 5-point Likert-type scale ranging from 1 ("not at all") to 5 ("very much").

In-flight, participants who performed the No-leans condition last (n = 20) provided verbal feedback of their sensation of the bank angle (left, right or none) before the test pilot started the countdown for the response. The participants who performed the No-leans condition first did not report this, to prevent them from becoming too aware or focused on the goal of the experiment.

Run similarity check

In the simulator experiment, no specific instructions were given to hold a certain bank angle when the "steady" command was given. Thus, there may be inadvertent variations in the run parameters between each run. The roll rate should ideally be zero, and pitch and bank angles, at the end of the flying task and when showing the AI should be similar between the test conditions (Baseline, Opposite and Level). To test whether there were inadvertent differences between conditions that affected the outcomes, these parameters will be reported and compared between the test conditions. To check if any inadvertent variations in these parameters increased the likelihood of making an RRE, the predictive effect of these parameters on making an RRE was checked as well within each condition.

In-flight, we measured two parameters to test if the runs in each condition were similarly set up. The duration participants were blindfolded was measured. The duration of the fast roll cue in the Leans-opposite and Leans-level condition was defined as the time the roll rate exceeded $1.0°/s$ (as measured with the IMU).

4.2.5. Statistical analysis

Results of the Matching or No leans condition (practice and filler runs) are reported, but they are not used for any comparisons with the test conditions.

Simulator experiment

Non-parametric tests for ordinal data were chosen to analyze error rates, as the Opposite and Level conditions featured only three performance categories (0, 50 or 100% error rate). The main effect of Condition (Baseline, Opposite and Level) on error rates was tested using a Friedman test, and post-hoc tests were performed with Wilcoxon Signed Rank, and Holm-Bonferroni correction. The main effect of Condition (Baseline, Opposite and Level) on mean error duration was tested using repeated-measures ANOVAs, and post-hoc t-tests were performed with paired-samples t-tests with Holm-Bonferroni correction. The mean response time of correct responses in the Opposite and Baseline conditions were compared with a paired-samples t-test.

The run parameters were compared between relevant conditions using paired-samples t-tests without correction for multiple comparisons, since the measured parameters should ideally be similar between the conditions.

Learning effect (in all conditions) and the effects of run parameters on errors were tested with binary logistic regression for each condition separately. The run order and the run parameters (absolute and discrete pitch angle, bank angle, if applicable at the end of the flying task and if applicable at AI presentation) were the predictors, whereas occurrence of an error was the dependent measure.

In-flight experiment

The error rates in the three test conditions were analyzed using Chochran's Q for main effects. Post-hoc comparisons between all conditions were performed using McNemar with Holm-Bonferroni correction. The effects of Condition on error duration as well as reaction time was tested using repeated-measures ANOVAs and post-hoc pairwise t-tests with Holm-Bonferroni correction. The reaction times between errors and correct responses were compared for the Leans-opposite and No-leans conditions separately, using independent-samples t-tests, while correcting for two comparisons using Holm-Bonferroni.

Furthermore, training effects were tested by performing a binary logistic regression, with the sequence of conditions (A and B; see Figure 4.6) as predictor, and occurrence of an error (true or false) as dependent measure. The run characteristics were compared between each pair of conditions with paired-samples t-tests without correction to check for differences.

4.3. Results

4.3.1. Performance example

Figure 4.7 (top) shows an example of an RRE in the Leans-opposite condition of the in-flight experiment. The plotted data represents the low-pass filtered (integrated) IMU data. First, there was a sub-threshold roll to 20° bank at around 1°/s (a), followed by a rapid roll back to 10° bank at around 13°/s (b). The pilot then counted down from three (c), after which the participant removed their blindfold at t = 0. After removing

the blindfold, the participant responded by rolling into the opposite direction, i.e., away from level, for about 2 s, before correcting the input towards the correct direction.

Figure 4.7 (bottom) shows a different example of an RRE in the Leans-opposite condition in the in-flight experiment. In this case, the participant made two extra RREs at t = 2.5 and 4 s, before rolling to level flight. The confusion in this example lasted for a total of almost five s. However, the first input briefly stopped at around t = 1.8 s, meaning that the measured error duration was only 0.8 s.
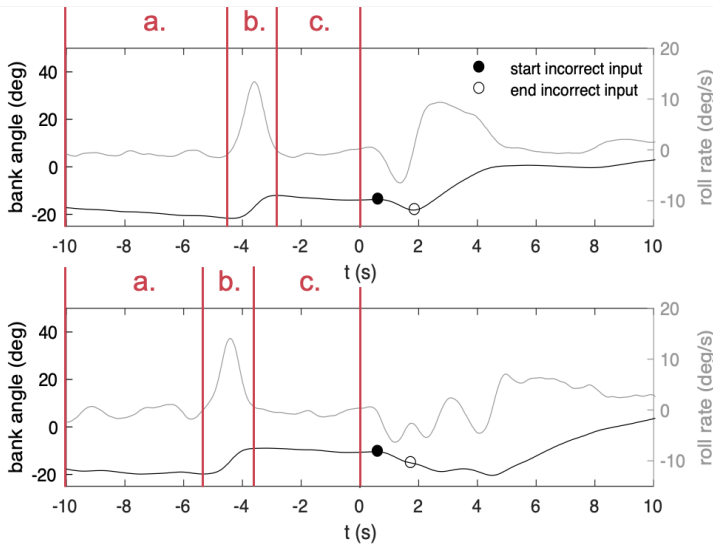
Figure 4.7: Two examples of roll reversal errors in the Leans-opposite condition in-flight. The black line is the bank angle, the grey line is the roll rate. See the text for description.

4.3.2. Error rate

The error rates of both experiments are shown in Figure 4.8. The outcomes of the simulator experiment are listed in Table 4.1, whereas those of the in-flight experiment can be found in Table 4.2. In the in-flight experiment, the video recordings of four participants were lost. For one participant, this was the case for the Leans-opposite and Leans-level condition only. This resulted in missing cases for the error duration, the reaction times and the blindfolding duration. Participants in the video recordings always looked at the AI instead of outside when removing the blindfold. One participant was excluded from the analysis due to prematurely removing the blindfold in a mismatching condition, which gave the participant insight into the maneuvers flown. A new participant was recruited instead.

In the simulator experiment, there was a significant main effect of Condition on error rate, $\chi2(2,20) = 24.54$, $p < .001$. The error rate was significantly higher in the Opposite condition compared to the Baseline condition (7.8 times as high), $Z = 3.92$, $p < .001$, and the Level condition (2.5 times as high), $Z = 3.22$, $p = .001$. All participants

made at least one RRE in the Opposite condition. There were also significantly more errors in the Level condition than in the Baseline condition, $Z = 2.10$, $p = .035$.

In-flight, there was also a significant main effect of Condition on error rate, $Q(2,38) = 14.25$, $p = .001$. Significantly more (2.7 times as many) RREs were made in the Leans-opposite condition than in the No leans condition, $p = .001$. There were also significantly more errors in the Leans-level condition than in the No leans condition, $p = .002$. In contrast to the simulator study, there was no significant difference between the Leans-opposite and Leans-level condition, $p = .832$. All erroneous responses in the Leans-Level condition were towards the opposite site of the fast roll cue.
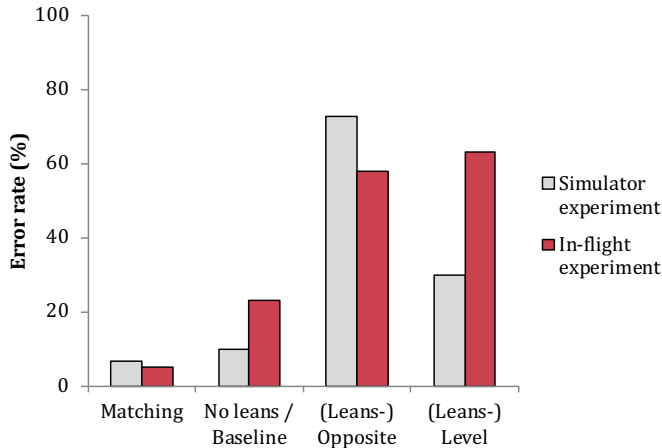
Figure 4.8: The mean error rates per condition, in the simulator and in the in-flight experiment.

Table 4.1: The means and standard deviations (*SD*) of performance variables and the run characteristics in the simulator experiment.

	Matching		Baseline		Opposite		Level	
	Mean (*SD*)	N	Mean (*SD*)	N	Mean (*SD*)	N	Mean (*SD*)	N
Error rate (%)	7.3 (9.7)	20	9.6 (8.9)	20	75.0 (25.6)	20	30.0 (37.7)	20
Error duration (s)	.35 (.25)	9	.28 (.17)	15	.44 (.25)	20	.33 (.18)	9
Response time (s)	.47 (.10)	20	.54 (.12)	20	.64 (.24)	11	-	-
Bank pre AI (°)	17.8 (4.1)	20	-	-	18.8 (4.5)	20	17.2 (5.2)	20
Bank post AI (°)	17.8 (4.1)	20	20.0 (0.1)	20	10.0 (10.0)	20	0.0 (0.0)	20
Roll rate AI (°/s)	.56 (.49)	20	.05 (.11)	20	.44 (.62)	20	.24 (.21)	20
Pitch angle AI (°)	3.6 (1.0)	20	3.7 (3.9)	20	4.8 (3.7)	20	2.9 (2.8)	20

4.3.3. Error duration

Scatterplots of the error duration in both experiments are shown in Figure 4.9. Although Figure 4.9 indicates that, in the imulator experiment, errors lasted generally longest in the Opposite condition than in the other conditions. However, there was no significant effect $F(2,6) = 2.55$, $p = .158$. In the in-flight experiment there was a signifi-

Table 4.2: The means and standard deviations (*SD*) of performance variables and the run characteristics in the in-flight experiment.

	Matching		No leans		Leans-opposite		Leans-level	
	Mean (*SD*)	N	Mean (*SD*)	N	Mean (*SD*)	N	Mean (*SD*)	N
Error rate (%)	5.0 (19.0)	40	23.0 (N/A)	40	58.0 (N/A)	40	63.0 (N/A)	40
Error duration (s)	.70 (.11)	2	.88 (.63)	8	.91 (.76)	19	.76 (.52)	24
Response time (s)	.50 (.39)	34	.77 (.50)	28	.67 (.23)	16	N/A	N/A
Blindfolding duration (s)	27.0 (3.0)	36	31.2 (8.0)	36	34.7 (6.6)	35	33.4 (5.1)	35
Fast roll cue duration (s)	2.1 (.38)	40	N/A	N/A	1.5 (.32)	40	2.0 (.33)	40

4

cant effect, $F(2,2) = 25.27$, $p = .038$. Post-hoc analyses revealed that errors lasted significantly longer in the Leans-opposite condition than in the No-leans condition, $t(1,5) = 3.19$, $\Delta = .53$ s, $p = .024$. Comparing the two experiments, it seems that there were some excessively long error durations in the in-flight experiment, which were not present in the simulator experiment.
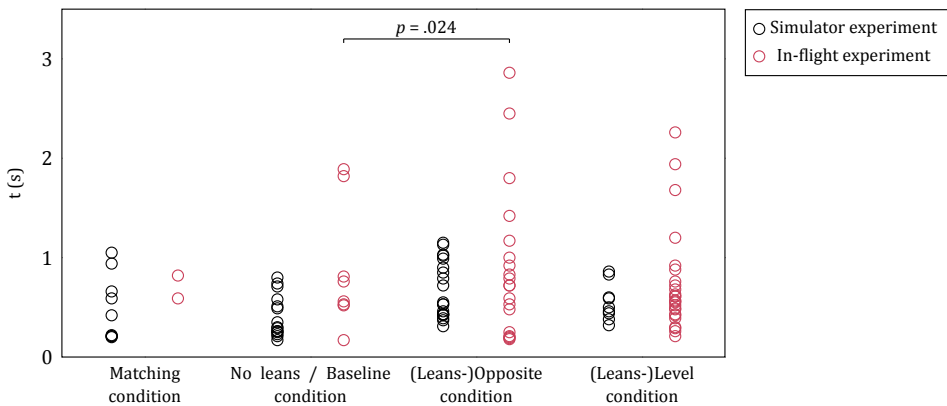
Figure 4.9: Scatterplots of the error durations in the simulator and in the in-flight experiment.

4.3.4. Response time

The response times in both experiments are shown in Figure 4.10. In the simulator experiment, the response time of correct responses was, as expected, longer in the Opposite condition than in the Baseline condition $t(1,10) = 2.27$, $p = .047$, $\Delta = .11$ s. In-flight, there was no significant difference in correct responses between the Leans-opposite and No Leans conditions, $t(1,14) = .16$, $p = .879$. When comparing the response times in the two experiments in Figure 4.10, it seems that there were more early responses in the in-flight experiment, and one outlier (late response) in each experiment.
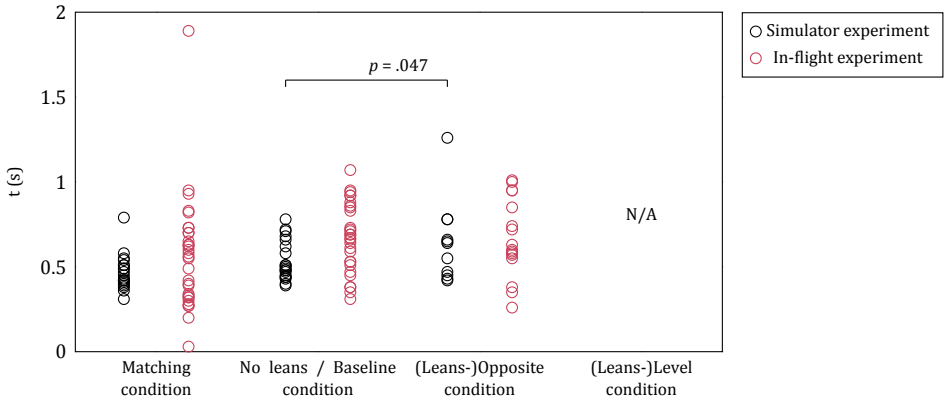
Figure 4.10: Scatterplots of the response times of correct responses in the simulator experiment.

4.3.5. Learning effect
In the simulator experiment, there was no significant predictive effect for the order of the run within its condition, on whether an error was made or on the error duration.

In the in-flight experiment, the order of the conditions (1-2-3 or 3-2-1) significantly predicted whether an error was made in the Leans-level condition only, $B = 2.14$, $p = .006$. Participants were approximately twice as likely to make an error if the Leans-level condition was the first condition, compared to the last. There were no significant effects of the order of the conditions on error duration.

4.3.6. Run similarity check
In the simulator experiment, there were some small roll rates present when the AI was shown after the flying task, meaning that the bank angle was not perfectly stable. The roll rate was significantly lower in the Baseline condition than in the Opposite, $\Delta = .40$ °/s, $p = .005$, and Level condition, $\Delta = .19$ °/s, $p = .003$. There were no differences between the conditions in pitch angle (Opposite, Level, Baseline), or bank angle at the end of the flying task (Opposite and Level). There was also no significant predictive effect of the run parameters (roll rate, absolute bank and pitch angle, discrete pitch angle, both at the end of the flying task and when showing the AI) on whether an RRE occurred.

In the in-flight experiment, participants were blindfolded for a significantly longer time in the Leans-opposite condition than in the No leans condition, $t(1,34) = 2.47$, $p = .019$, $\Delta = 3.4$ s. This difference is small compared to the average blindfolding time (i.e., ca. 30 s). The duration of the fast roll cue was significantly longer in the Leans-level condition than in the Leans-opposite condition, $t(1,39) = 7.37$, $p < .001$, $\Delta = .51$ s.

4.3.7. Subjective variables
In the simulator experiment, the average score of severity of sickness issues on the MISC was .75, $SD = 1.00$, half way into the flying task session, and 1.20, $SD = 1.11$, at the end. The maximum score was 3 in the middle and 4 at the end. Eighteen (95 %)

participants indicated that they noticed that the AI was sometimes different from the preceding turn. These participants rated the extent to which they changed their response strategy on average with 2.1, median = 2, *SD* = 1.2, on the 5-point scale, which is around "very little". Changes in strategy included waiting a moment before responding, and not assuming a roll direction of the AI. The two participants who had not noticed the mismatches reported that they had mentally "decoupled" the flying task from the recovery task.

In the in-flight study, there were two missing cases due to forgetting to question the participant. Four out of the remaining eighteen questioned participants (22 %) indicated that they perceived a bank angle at the end of the blindfold phase during the No leans condition. This perceived bank angle was in the direction of the actual bank angle in two cases, and into the opposite direction in two other cases.

4.4. Discussion

The results of the two experiments show that the expectation, either induced through a flying task (simulator experiment) or though misleading vestibular cues (in-flight experiment) strongly affected the controller's responses, even overruling information on the AI. In the simulator, error rates in the Level condition (30.0 %) and in the Opposite condition (75 %) were 3, respectively, 7.8 times higher compared to the rate in the Baseline condition (9.8 %). The error rate in the Baseline condition was somewhat lower than that in other experiments with a similar task and non-*t*-test-pilot participants, but it is not very far off (15–20 %; [7, 9, 10]).

In the in-flight experiment, a large increase in the error rate was also found between the No Leans condition (23 %) and the conditions with leans manipulation (both ca. 60 %), which is a factor of 2.3 higher. The RRE rate in the No leans condition in-flight was high compared to that in the simulator, possibly due to unintended leans cues. However, this rate coincided with rates found in previous in-flight experiments with non-pilots (i.e., 21.9–23.6 %; [10, 11]).

In contrast to the simulator experiment, the RRE rates were not significantly different between the Leans-level and Leans-opposite condition in-flight. This difference between both experiments suggests that in-flight cues had a stronger influence on the first response than the manipulation of expectation with a flying task in the fixed-base simulator. It is relevant here that participants sometimes indicated surprise by the fact that no input was needed in the Leans-level condition and the in-flight condition featured one run per leans condition. A learning effect was present for the Leans-level condition, meaning that the error rate would have been lower if more runs were included. Participants were also sometimes surprised by the fact that they were flying banked in the in-flight No Leans condition, which was not the case in the simulator Baseline condition. This difference makes it problematic to compare these conditions between the experiments.

The observation that the error rate was significantly higher in the (Leans-)Level conditions than in the No Leans or Baseline conditions suggests that the participants sometimes responded based solely on their expectation and neglected the AI, although their response time was not significantly shorter. The observation in the simulator experiment that the error rate in the Opposite condition was 2.5 times higher than in

the Level condition suggests that, in the Opposite condition, a proportion of the RREs was caused by horizon control reversals instead of neglecting the AI.

There were some other findings indicating that the (Leans-)Opposite condition was the most difficult condition in both experiments. Correct responses took longer than those in the No Leans condition in-flight, and participants required more time to correct erroneous inputs than in the Baseline condition in the simulator.

The results of the simulator experiment are thus in line with Bayesian models of perceptual inference, in that participants were more susceptible to horizon control reversals when this misinterpretation matched with their prior expectation. The results imply that RRE rates that are currently reported in experiments with pilots (1.5-3.1% in-flight; [6, 12]) are likely lower than they would be in conditions of spatial disorientation. Additional factors, such as startle [32] or the presence of a roll rate when responding, may further increase the likelihood of an RRE occurring [33].

However, several factors are also likely to prevent pilots from making RREs, such as additional knowledge (e.g., about spatial disorientation) and the use of different response strategies (e.g., mapping oneself onto the aircraft icon; e.g., [34]). This makes it problematic to extrapolate the current findings to pilots. Nevertheless, the occurrence of RREs in experienced pilots suggests that pilots are also affected by this issue, be it to a lower extent. It would be interesting to repeat this study with pilots in a high-fidelity simulator, or real aircraft, to investigate whether piloting experience prevents misinterpretations. Also, it remains to be investigated which piloting skills and knowledge are particularly useful.

In the simulator experiment, the expectation was set up in an artificial manner so that the responses could be measured in a highly controlled environment. There were no vestibular, tactile or proprioceptive cues of motion, as there would be in-flight. In order to keep the participant's attentional focus and task execution as natural as possible, there were slight variations between the aircraft attitude between each run. This meant that there were small but significant roll rates present in the Opposite and Level condition, but not in the Baseline condition. However, in our regression analysis we could find no significant predictive effect of these variations on making an RRE. The number of test runs in the conditions with a mismatching expectation was kept low, to increase surprise by mismatching runs. Even though 18 of the 20 participants became aware that the AI did not always match with their expectation, we found no learning effect. Thus, it seems that more runs with a mismatch could be included in future experiments, provided that the task is to respond immediately to offset a potential increase in vigilance in later runs.

In-flight, the presented leans cues were more realistic, however there was also some variation in the manually flown maneuvers. The fast roll cue in the Leans-level condition lasted longer than in the Leans-opposite condition due to standardization of the start of this cue (i.e., 20° bank). Based on verbal reports by the participants, it seems that there were sometimes inadvertent leans sensations present in the No-leans condition. Some participants indicated that they were surprised by the bank angle in the No-leans condition in-flight, which may have affected their response and caused more errors in-flight compared to the simulator Baseline condition.

For future in-flight experiments on this topic, the following lessons were learned

with regard to the methodology. First, it may be wise to include at least one run with level flight in the practice runs, to prevent that participants presume that they always need to give an input. Second, although test flights indicated that the outside view wasn't noticeable when focusing on the AI, the information presented to participants could be more tightly controlled by covering the side window or by using training glasses that prevent outside vision (i.e. 'foggles').

In conclusion, the results suggest that pilots experiencing spatial disorientation, or are otherwise surprised by the aircraft attitude, are more prone to make an RRE. Promisingly, a learning effect was present in the later conditions in-flight. This suggests that spatial disorientation awareness training may help pilots to prevent incorrect intuitive responses. It may thus be wise to not only teach pilots to "Believe your instruments." but also to "Control the aircraft to make the instruments read what you want them to." [35].

4

References

[1] A. Landman, S. Davies, E. L. Groen, M. M. Van Paassen, N. J. Lawson, A. W. Bronkhorst, and M. Mulder, *In-flight spatial disorientation induces roll reversal errors when using the attitude indicator,* Applied Ergonomics (advance online publication) **81** (2019).

[2] S. N. Roscoe, *Moving horizons, control reversals, and graveyard spirals,* Ergonomics in Design **12**, 15 (2004).

[3] F. H. Previc and W. R. Ercoline, *The outside-in attitude display concept revisited,* The international Journal of aviation psychology **9**, 377 (1999).

[4] C. D. Wickens, *Aviation displays,* in *Principles and practice of aviation psychology*, edited by P. S. T. . M. A. Vidulich (Lawrence Erlbaum, Mahwah, NJ, 2003) pp. 147–200.

[5] S. L. Johnson and S. N. Roscoe, *What moves, the airplane or the world?* Human Factors **14**, 107 (1972).

[6] D. B. Beringer, R. C. Williges, and S. N. Roscoe, *The transition of experienced pilots to a frequency-separated aircraft attitude display,* Human Factors **17**, 401 (1975).

[7] S. Müller, V. Sadovitch, and D. Manzey, *Attitude indicator design in primary flight display: revisiting an old issue with current technology,* The International Journal of Aerospace Psychology **28**, 46 (2018).

[8] G. Singer and S. Dekker, *The effect of the roll index (sky pointer) on roll reversal errors,* Human Factors and Aerospace Safety **2** (2002).

[9] D. Bauerschmidt and S. Roscoe, *A comparative evaluation of a pursuit moving-airplane steering display,* IRE Transactions on Human Factors in Electronics , 62 (1960).

[10] F. Ince, R. C. Williges, and S. N. Roscoe, *Aircraft simulator motion and the order of merit of flight attitude and steering guidance displays,* Human Factors **17**, 388 (1975).

[11] S. N. Roscoe and R. C. Williges, *Motion relationships in aircraft attitude and guidance displays: A flight experiment,* Human Factors **17**, 374 (1975).

[12] A. H. Hasbrook and P. G. Rasmussen, *In-flight performance of civilian pilots using moving-aircraft and moving-horizon attitude indicators*, Tech. Rep. AD0773450 (Federal Aviation Administration, Oklahoma City, OK, Civil Aeromedical Institute, 1973).

[13] S. R. Holmes, A. Bunting, D. L. Brown, K. L. Hiatt, M. G. Braithwaite, and M. J. Harrigan, *Survey of spatial disorientation in military pilots and navigators,* Aviation, space, and environmental medicine **74**, 957 (2003).

[14] P. D. Navathe and B. Singh, *An operational definition for spatial disorientation,* Aviation Space and Environmental Medicine **65**, 1153 (1994).

[15] F. H. Previc and W. R. Ercoline, *Spatial disorientation in aviation*, Vol. 203 (Aiaa, 2004).

[16] K. K. Gillingham, *The spatial disorientation problem in the united states air force,* Journal of Vestibular Research **2**, 297 (1992).

[17] R. Gibb, B. Ercoline, and L. Scharff, *Spatial disorientation: decades of pilot fatalities,* Aviation, space, and environmental medicine **82**, 717 (2011).

[18] J. C. Neubauer, *Classifying spatial disorientation mishaps using different definitions,* IEEE Engineering in Medicine and Biology Magazine **19**, 28 (2000).

[19] C. M. Belcastro, J. V. Foster, G. H. Shah, I. M. Gregory, D. E. Cox, D. A. Crider, L. Groff, R. L. Newman, and D. H. Klyde, *Aircraft loss of control problem analysis and research toward a holistic solution,* Journal of Guidance, Control, and Dynamics **40**, 733 (2017).

[20] E. M. of Civil Aviation, *The final report of the accident investigation: Flash airlines 604,* (2004).

[21] C. C. A. Authority, *Technical investigation into the accident of the b737-800 registration 5y-kya operated by kenya airways that occurred the 5th of may 2007 in douala,* (2010).

[22] Aircraft Accident Investigation Bureau, *Final report of the aircraft accident investigation bureau on the accident to the saab 340b aircraft, registration hb-akk of crossair flight crx 498 on 10 january 2000 near nassenwil/zh,* (2002).

[23] P. Vetter and A. Newen, *Varieties of cognitive penetration in visual perception,* Consciousness and cognition **27**, 62 (2014).

[24] J. L. Delk and S. Fillenbaum, *Differences in perceived color as a function of characteristic color,* The American Journal of Psychology **78**, 290 (1965).

[25] F. Gosselin and P. G. Schyns, *Superstitious perceptions reveal properties of internal representations,* Psychological Science **14**, 505 (2003).

[26] J. J. Girgus, I. Rock, and R. Egatz, *The effect of knowledge of reversibility on the reversibility of ambiguous figures,* Perception & Psychophysics **22**, 550 (1977).

[27] A. Clark, *Whatever next? predictive brains, situated agents, and the future of cognitive science,* Behavioral and brain sciences **36**, 181 (2013).

[28] D. C. Knill, D. Kersten, and A. Yuille, *Introduction: A bayesian formulation of visual perception,* Perception as Bayesian inference **1**, 1 (1996).

[29] L. T. Maloney and H. Zhang, *Decision-theoretic models of visual perception and action,* Vision research **50**, 2362 (2010).

[30] A. Gundry, *Thresholds of perception for periodic linear motion.* Aviation, space, and environmental medicine (1978).

[31] A. H. Wertheim, J. E. Bos, and W. Bles, *Contributions of roll and pitch to sea sickness,* Brain Research Bulletin **47**, 517 (1998).

[32] A. Landman, E. L. Groen, M. M. van Paassen, A. W. Bronkhorst, and M. Mulder, *Dealing with unexpected events on the flight deck: A conceptual model of startle and surprise,* Human Factors **59**, 1161 (2017).

[33] S. N. Roscoe, *Airborne displays for flight and navigation,* Human Factors **10**, 321 (1968).

[34] D. Cohen, S. Otakeno, F. H. Previc, and W. R. Ercoline, *Effect of" inside-out" and" outside-in" attitude displays on off-axis tracking in pilots and nonpilots.* Aviation, space, and environmental medicine (2001).

[35] W. Bles, *Spatial disorientation training demonstration and avoidance,* (2008).

4

5

The advantage of unpredictable and variable training scenarios

In the experiment described in this chapter, we test whether making existing simulator training scenarios less predictable and more variable could be an effective way to improve performance in situations that are surprising and novel. One group of ten pilots (the U/V group) practice responses to controllability issues in an unpredictable and variable manner. A control group of another ten pilots practice the same failures in a more predictable and invariable manner. After the practice, performance of all pilots is tested in a surprise scenario, in which the pilots have to apply the learned knowledge in a novel situation. The results indicate superior understanding and performance in the U/V group compared to the control group in the surprise scenario. Based on the results, we conclude that the inclusion of unpredictable and variable scenarios in pilot training may improve transfer of training to unexpected situations in-flight.

5.1. Introduction

According to regulations, the initial and recurrent type-rating checks for pilots must involve relatively predictable and standardized scenarios. The training sessions for these checks can be organized in similarly predictable format [2, 3]. However, research suggests that skills taught in this manner are "brittle" instead of adaptive [3, 4]. That is, they transfer well to predictable situations like the tests, but they may not hold up in emergency situations, which are typically novel and unexpected. For this reason, many aviation companies look for ways to extend on the minimally required training. Aviation safety organizations recognize the benefits of such extensions, and offer ways to formalize them (see e.g., [5]).

It is impossible to train pilots for every conceivable abnormal situation to ensure resilience. However, there are common factors in these situations that can be trained for. One of these factors is the crew's ability to deal with startle and surprise, the training of which will become mandatory in the near future [6–8]. Coping with surprise requires effective sensemaking activities, which involves the search for a structured representation, or "frame" of the situation. This frame is needed to direct attention, and to interpret and understand what is going on (see, [9–12]). Sensemaking includes, for instance, seeking information on the instruments, reasoning, or testing out hypotheses. Such activities are particularly difficult when a pilot is startled [1, 13] or fatigued [14]. If an appropriate frame cannot be found under stress, the events may lose meaning and coherence, which may increase stress even further and prevent effective troubleshooting.

Since predictable scenarios require very little sensemaking, they can be expected to be insufficient as a training to deal with surprise. A simple means to increase unpredictability would be to deny (more) information about upcoming events. Secondly, aspects of a problem can be varied between practice iterations, making each scenario somewhat different [15, 16]. Variable practice is thought to enhance a trainee's recognition of relevant features and rules, since the irrelevant features are different in each practice iteration [17]. The resultant skills and knowledge are therefore better applicable in situations that are not explicitly practiced [18]. Finally, different types of problems can be intermixed (and spaced out), so that trainees cannot assume that the same solution is again applicable in each next practice iteration. This is thought to improve their ability to pair the appropriate solution to the problem [19]. Such methods are known as "mixed review" in math education [20], or "contextual interference" in perceptual-motor learning literature (e.g., [21]).

In the current simulator study, we investigated whether arranging simulator training in a more unpredictable and variable manner improves pilot performance in a surprise test. The test required pilots to apply the learned knowledge and skills in a demanding and partly novel situation. Based on the aforementioned literature, it was hypothesized that unpredictable and variable practice would improve the pilots' performance in this test, due to a better understanding of the events and principles involved.

5.2. Method

5.2.1. Participants

Twenty participating airline pilots were assigned either to an unpredictable and variable practice (U/V) group or to a control group in a randomized manner, unless when the groups became unbalanced with regard to the variables listed in Tables 5.1 and 5.2. Thus, two type-rating instructors and one pilot with extended light (i.e., CS-23/FAR part 23) multi-engine piston (MEP) flying experience were re-assigned to the control group. All pilots, except one in the U/V group, indicated that they had at least 25 hours light MEP flying experience. Pilots were mainly recruited from a single airline company. Eight pilots from this company were in the U/V group and seven in the control group. This study complied with the tenets of the Declaration of Helsinki and informed consent was obtained from each participant.

Table 5.1: Characteristics of the participants

	U/V group Mean (*SD*)	Control group Mean (*SD*)	Δ	*p*
Age (yrs)	41.3 (9.0)	41.5 (9.3)	.2	.961
Employed (yrs)	17.2 (8.8)	16.4 (7.3)	.8	.827
Flight hours	9311 (6352)	7571 (4590)	1740	.491

Table 5.2: Characteristics of the participants (cont.)

	U/V group	Control group
Extended MEP exp. (>50 hrs)	2	1
Type rating instructors	2	2
Captains / First officers	4/6	5/5
Men / Women	9/1	9/1

5.2.2. Apparatus

The experiment was performed in the Simona research simulator (SRS), of the Delft University of Technology in the Netherlands (Figure 5.1a). The simulator features a six-degrees-of-freedom hydraulic hexapod motion system, and a collimated 180 degree horizontal by 40 degree vertical field of view display system. For outside visuals, the FlightGear open-source flight simulator was used. Standard washout filters were used for motion cueing [22]. A piston aircraft engine sound was played in mono over the pilot's headphones. Audio pitch was coupled to engine rpm and volume to torque. The cockpit mock-up was styled after a jet airliner, and featured a B747-style primary flight display and a Cessna Citation-style engine display. Controls consisted of a right-hand side-stick with pitch trim control, rudder pedals with force feedback, and thrust, flaps and gear levers. A non-linear aerodynamic model was used of a light twin-propeller aircraft, the Piper PA-34 Seneca III [23, 24] (Figure 5.1b. The aircraft model has certain properties and failure options which allows for the development of challenging flying tasks. The airflow over the wing of each propeller induces extra lift, which causes a

roll moment as well as a yaw moment in case of asymmetric thrust. At low speed, the moments generated by asymmetric thrust will exceed the maximum obtainable opposite moments generated by the control surfaces.

(a)   (b)
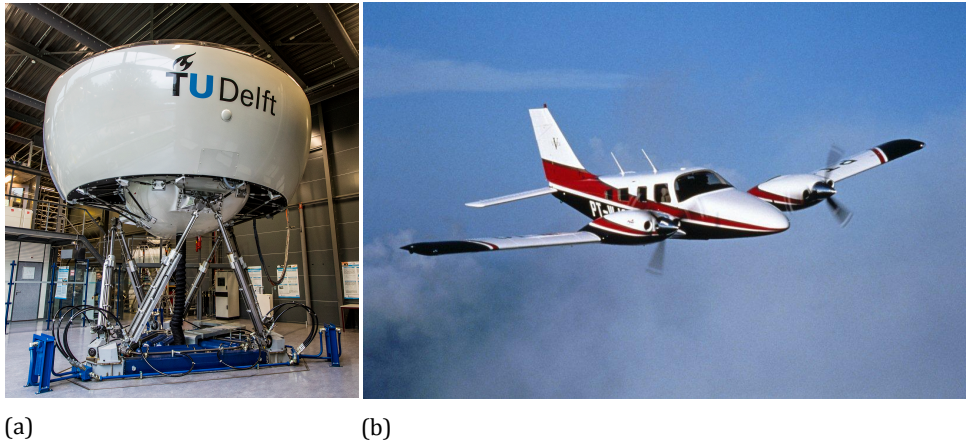
Figure 5.1: (a) Simona research simulator. (b) A Piper PA-34 Seneca.

5.2.3. Tasks

The design of the experiment is illustrated in Figure 5.2. Each of the sessions listed in the figure is described below.
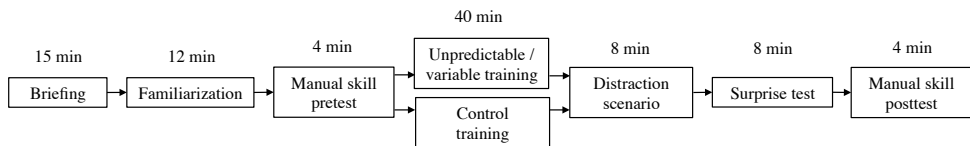
Figure 5.2: Experimental design

Briefing and familiarization

Pilots were informed that they would perform a number of scenarios and respond to malfunctions. They were instructed to complete the task that was given (e.g., perform a landing, fly a circuit), unless a crash was imminent. ATC communication and check-lists were not included. Pilots were instructed to call out any problem as soon as they noticed it. They were informed about the required settings: a circuit would need to be flown at 1,000 ft, with a speed of 130 kt and a power setting of approximately 42 Nm torque. Full throttle was used during takeoff. A flap setting of 25 degrees was required only during landing. The speed at rotate was 80 kt, optimal rate of climb (V_2) was 92 kt, the approach speed was 85 kt, and the minimum control speed with single engine was around 80 kt. Pilots were reminded of the settings if they diverged from them during the familiarization and practice. The wind direction and strength (light or moderate)

was provided verbally before each run, and was indicated by a wind sock located next to the runway. Following the briefing, pilots were seated in the simulator and flew two familiarization circuits, one without wind and one with light crosswind.

Practice session

Pilots then practiced with managing asymmetric thrust by performing six takeoff runs with a single engine failure, four flyby runs with a rudder failure, and four flyby runs with a single engine failure (see description below). Aspects of the runs were varied in the U/V group only (see Table 5.3). The runs were presented in blocks of two (e.g., two takeoff runs, see Table 5.4). These blocks were alternated in the U/V group only. Before the first run of each block, pilots in the U/V group were merely informed that a malfunction would occur. Between the first and the second run of each block, they were informed what the malfunction was, and how they could respond. Pilots in the control group received all this information before the first run of each block, and they were informed that each subsequent run was a repetition.

Table 5.3: Characteristics of the runs

Run ID	Malfunction timing	Side	Wind direction (from)	Wind strength
Engine failure during takeoff				
i	gear lever up	left	left	9.7 kt
ii*	speed 65 kt	right	left	9.7 kt
iii*	rotate	right	left	9.7 kt
iv	altitude 270 ft	right	left	9.7 kt
v	gear halfway up	left	left	9.7 kt
vi	altitude 310 ft	right	left	9.7 kt
Flyby runs with rudder failure				
i	20 s into run	15°	right	13.6 kt
ii	50 s into run	20°	ahead	13.6 kt
iii	50 s into run	25°	left	13.6 kt
iv	30 s into run	10°	right	13.6 kt
Flyby runs with the engine failure				
i	20 s into run	left	left	9.7 kt
ii	40 s into run	right	right	9.7 kt
iii	30 s into run	left	left	9.7 kt
iv	50 s into run	right	right	9.7 kt

* In these runs, the takeoff was to be aborted.
Note: rudder deflection angles are to the right.

The takeoff runs started with the aircraft on the runway (runway 18), with 3,000 ft of runway ahead. The U/V group was instructed to respond as they saw best, whereas the control group was told when the engine failure would occur, and whether the take-off was to be aborted or continued. Following the first run in a block, the U/V group was reminded to pay attention to the minimum control speed (80 kt), below which the takeoff was to be aborted. If pilots continued the takeoff following the engine failure,

they were instructed to continue to climb 100 ft straight ahead, where the run ended. Both groups included a similar number of to-be-aborted takeoffs.

The flyby runs started with the aircraft in approach, approximately 90 seconds from reaching the runway. The task was to fly over the runway, follow the center line as closely as possible and, upon reaching the runway, descend to 100 ft altitude and reduce speed to 85 kt. The gear remained down and flaps remained at 25 degrees. The malfunction occurred before reaching the runway. For the rudder issue, response instructions included that it could be countered by commanding a throttle differential. For the engine failure, pilots were reminded of the minimum control speed of 80 kt. In the U/V group, extra variation was introduced (besides factors listed in Table 5.3) by asking pilots to increase their speed in the second half of each flyby, by adding turbulence, and by reducing the visibility in 50 % of the runs.

Table 5.4: The order of the runs and the variations used in the practice session. FB = flyby.

Block	Run	U/V group		Control group	
		Scenario	Run ID	Scenario	Run ID
1	1	takeoff	i	takeoff	i
	2	takeoff	ii	takeoff	i
2	3	fb rudder	i	takeoff	i
	4	fb rudder	ii	takeoff	ii
3	5	fb engine	i	takeoff	ii
	6	fb engine	ii	takeoff	ii
4	7	takeoff	iii	fb rudder	i
	8	takeoff	iv	fb rudder	i
5	9	fb rudder	iii	fb rudder	i
	10	fb rudder	iv	fb rudder	i
6	11	fb engine	iii	fb engine	i
	12	fb engine	iv	fb engine	i
7	13	takeoff	v	fb engine	i
	14	takeoff	vi	fb engine	i

Related surprise test

After the practice session, two surprise tests were performed: first a control test, which will be described in the next section, and second a "related surprise test", was the main test of the study. This test required the application of practiced skills, i.e., management of asymmetric thrust, in a surprising, demanding and partly novel scenario. The scenario started on the runway of a different airport, featuring a single, 4,000 ft long runway (runway 03), and a line of trees that was to be crossed following takeoff. There was moderate crosswind, coming from 310 (see Figure 5.3). Pilots were instructed to fly a left-handed circuit. As shown in Figure 5.3, the following malfunctions were inserted into this run. First, during takeoff, when the speed reached 55 kt, thrust in the right engine dropped in 20 seconds to 40 %. After a call-out (or else after approximately 30 seconds) pilots were instructed to continue the circuit at 800 instead of 1,000 ft, so as to limit the run's duration. When reaching 490 ft, there was a brief (3 seconds duration) dip in power of the still fully functioning (left) engine. Pilots were immedi-

ately informed that both engines were unreliable, and that they could keep using them both. This event was included to ensure that pilots were able to apply differential throttle later (see below). Finally, the rudder effectiveness decreased to 20 % when pilots rolled out of the turn towards downwind, decreasing their ability to counter the thrust differential due to the engine failure.
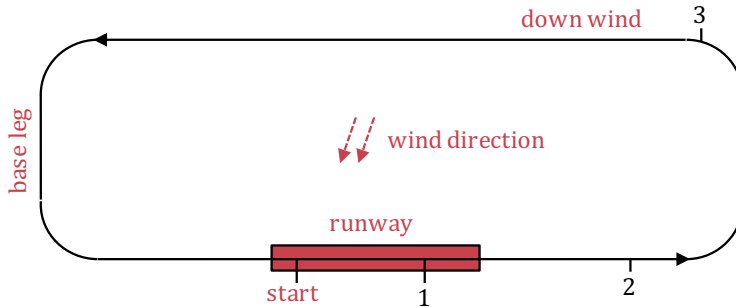
Figure 5.3: The circuit flown in the related surprise test, and the moments at which the malfunctions occur. 1: Right engine starts losing power over 20 seconds. 2: Brief decrease in left engine power, which is immediately restored. 3: Rudder effectiveness decreases to 20 %.

A successful landing was more likely if pilots identified the failures, compensated for the disturbances (potentially by commanding asymmetric throttle), and projected how the decrease in speed during the landing would affect these disturbances. While the first two steps were practiced explicitly in the practice session, finding a solution for the landing was not. At the minimum control speed with single engine (80 kt), the control surfaces can no longer be used to counter the moments resulting from the engines' asymmetric thrust. One solution is to maintain a high speed during the landing, keeping the control surfaces effective. A second solution is to apply little throttle and make a steeper approach. As an additional measure, the thrust asymmetry can be reduced by commanding differential throttle.

Unrelated surprise test
An unrelated surprise test was included as a control test to evaluate whether the groups responded differently to a surprise event that did not feature the practiced principles. Although the groups were balanced as much as possible, inadvertent differences in preexisting skills (e.g., due to company training) or traits could still exist. The test was also included to provide both groups with a surprise, so that potential differences in expectations between the groups are reduced. Finally, the test also served to separate the related surprise test from the practice.

The run started with the same instructions and settings as the related surprise test. From the moment of lift-off, the indicated airspeed decreased with 1 kt/s from the actual airspeed. Pilots could identify the problem by checking whether the apparent decrease in speed corresponded with the pitch angle, vertical speed, control responsiveness or ground speed. If pilots correctly called out the problem, they were instructed to finish the circuit. Otherwise, they were allowed to make an emergency landing.

Manual skill pre- and posttest
After the familiarization section, pilots performed a manual skill pretest, which was repeated as posttest at the end of the experiment. This test was included to check whether manual flying skills in the groups were different at the start, or differently affected by the practice. It contained a precision steering task, requiring comparable manual skills as the related surprise test, i.e., a landing, using ailerons only, in moderate crosswind. The task started in approach, approximately two minutes from touchdown, with moderate crosswind from the left. It was announced that before reaching the runway, the rudder would become fixed in the neutral position. It was also announced that nose wheel steering would become inoperative due to this malfunction, and that the nose should be pitched up as long as possible during the roll-out. Pilots were asked to follow the glide slope (indicated by the or Precision Approach Path Indicator (PAPI) lights) and land on the center line as accurately as possible.

5.2.4. Dependent measures and hypotheses

Practice session
The time to correct call-outs was obtained using the voice recorder. This time was defined as the time from the start of the malfunction to the utterance of the malfunctioning system (i.e., "engine", "rudder" or "speed indicator"). These times were measured in every first run of a block during the practice (see Table 5.4), as these runs were designed to be (maximally) differently predictable between the groups. The hypothesis was that the U/V group would have more difficulties with determining the problems, causing longer correct call-out times.

Further manipulation checks of the practice included the total time participants were controlling the aircraft, and the pilots' interest and enjoyment, measured with the Interest and Enjoyment subscale (seven items) of the Intrinsic Motivation Inventory (IMI; [25]). The outcomes of these should be similar for the two groups.

Tests
In the related surprise test, the main performance measure was whether or not pilots managed to successfully land on the runway. Second, the throttle setting of the fully functioning engine was measured when reaching the minimum control speed of 80 kt. This provided an indication of pilots' awareness of the problem of commanding too much throttle during landing. Finally, the proportion of time in which pilots applied differential throttle (in the effective direction) was computed in the final stage of the run (from the rudder failure to touchdown). A differential was defined as a difference of at least 10 % of the maximum throttle.

Correct call-out times (see, practice session) following the single engine failure and the rudder failure were measured in the related surprise test, and following the airspeed indicator malfunction in the unrelated surprise test. Incorrect or missing call-outs were counted as missing cases.

In the manual skill pre- and posttest, the root mean square of elevator and aileron corrections was obtained in the last phase of the landing (35 seconds to 5 seconds before touch down). To discard low-frequency components (e.g., caused by trim), these inputs were first high-pass filtered (forth and back) using a second order Butterworth

filter with a cutoff frequency of 0.1 Hz.

At the end of the entire simulator session, pilots rated their experiences following the apparent airspeed problem (unrelated surprise test), the single engine failure and the rudder failure (related surprise test). Each pair of ratings following the latter two malfunctions was combined into a mean. Subjective surprise and startle were rated on a 5-point Likert scale in answer to the questions: *"How surprised were you when you discovered the issue?"* and, respectively, *"How startled or shocked were you when you discovered the issue?"* from "not at all" (1) to "extremely" (5). Understanding was rated similarly by answering: *"How difficult was it to understand what had happened?"* from "not difficult" (1) to "very difficult" (5). These scores were then reversed.

Improved performance, including faster correct call-out times and higher reported understanding, was expected in the U/V group in the related surprise test. This group was thus expected to feel less threatened and less confused by the problems, causing them to report less startle [26] and surprise [27] as well. In the unrelated surprise test, the indicators of performance were call-out times and reported understanding. These measures, as well as reported surprise and startle, were not expected to differ between the groups in this test.

The manual skill tests were expected to show an increase in manual skill from pre- to posttest due to familiarization with the controls. No other differences were expected, since the groups should be equally balanced, as well as become equally familiarized with steering and landing due to the practice.

5.2.5. Data analysis

Differences between the groups in the surprise tests were tested separately with independent-samples t-tests, or with Pearson's Chi-squared test in case of binominal data. Differences between the groups in correct call-out times during the practice were tested with Group × Block mixed model ANOVAs. Manual skill in the pretest and posttest was analyzed with Group × Test mixed model ANOVAs. Significant main effects of Group and significant interaction effects were followed-up with group comparisons. The significance level of reported significant results was set at $p < .05$. Holm-Bonferroni correction for multiple comparisons was applied separately to the performance measures, correct call-out times and subjective measures.

5.3. Results

5.3.1. Manipulation checks of the practice

Call-out times

Boxplots of the correct call-out times are shown in Figure 5.4, and the corresponding statistical analyses are listed in Table 5.5. Correct call-out times were overall longer in the U/V group than in the control group, indicating that the U/V group spent more time making sense of the events. In the flyby runs with rudder malfunction, this was the case in both blocks, whereas in those with the engine failure, this was the case in the first block only. The takeoff runs with engine failure were excluded from statistical analysis due to an insufficient number of valid cases in the control group in the first run (n = 1). Missing cases resulted from pilots giving no call-out, giving an incorrect

call-out, or indicating that they did not know the cause of the problem.
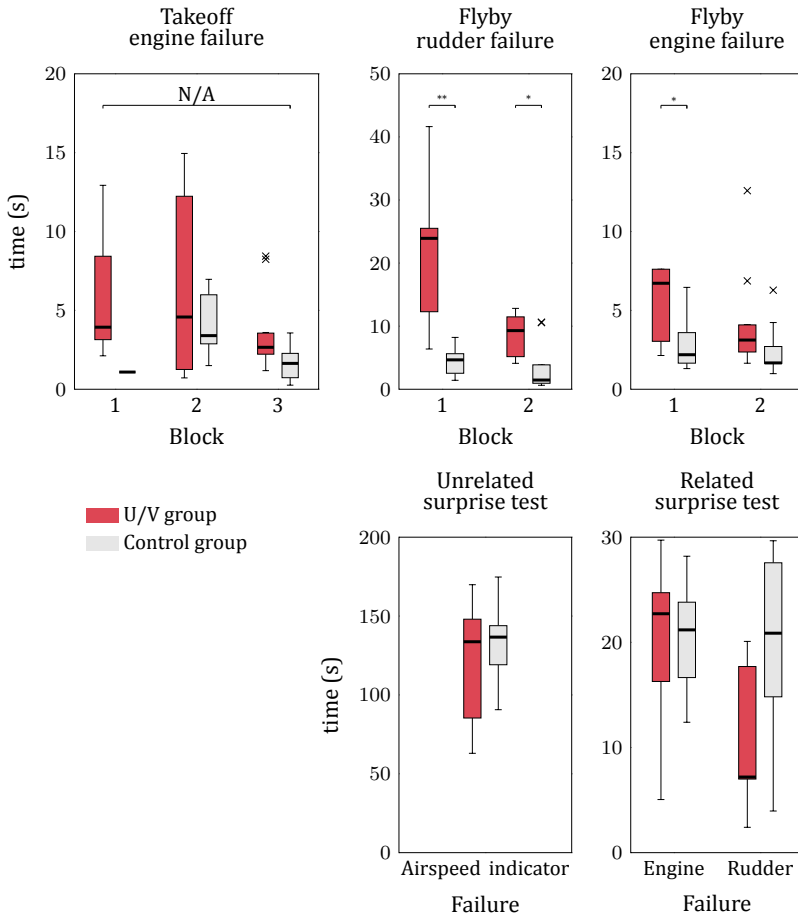
Figure 5.4: The correct call-out times in the practice (top plots), unrelated surprise test and related surprise task (bottom plots). * p < 0.05; ** p < 0.01.

Table 5.5: Statistical analysis of the correct call-out times during the practice

	F	*df*	*p*	Valid cases	
				U/V	Control
Takeoff engine failure	-	-	-	6	1
Flyby rudder failure					
Group	12.86**	1,10	.005	5	7
Flyby engine failure					
Group x Block	4.61*	1,16	.047	9	9

* Significant at *p* <.05 after Holm-Bonferroni correction.
** Significant at *p* <.01 after Holm-Bonferroni correction.

Flight time

The time controlling the aircraft in the practice session did not differ significantly between the groups, p = .670. The mean flight time in the practice session was 29 minutes and 47 s, SD = 1 minute and 14 s.

Interest and enjoyment

There was no significant difference in scores on the Interest/Enjoyment subscale of the IMI, t = .555, p = .586, indicating that the control practice, mean = 43.90, SD = 3.90, was not perceived as less interesting than the U/V practice, mean = 44.70, SD = 2.36. Both groups rated the practice near the maximum score (i.e., 49), suggesting that the pilots generally found the practice interesting and enjoyable.

5.3.2. Related surprise test

Performance

The run was completed with a successful landing on the runway by nine out of ten pilots in the U/V group, and by two out of ten pilots in the control group. This difference was significant, $\chi^2(1,19)$ = 9.90, p = .002. One pilot in the U/V group and four pilots in the control group landed somewhere else. Four other pilots in the control group lost control in-flight, as the altitude was insufficient to recover (<300 ft, see e.g., Figure 5.5). Unsuccessful landings always involved moments of losing aileron authority. Three pilots in the control group responded to loss of aileron authority by increasing throttle, which exacerbated the problems. The average throttle setting when reaching 80 kt at the end of the circuit, was significantly lower in the U/V group than in the control group (Table 5.6), meaning that the U/V group appeared to use a more favourable throttle/speed ratio during the landing. The U/V group also applied differential throttle during a larger proportion of the last part of the run (Table 5.6). Two pilots in both groups did not apply it at all, and two pilots in the control group applied it solely in the opposite direction.

Table 5.6: Group differences in the related surprise test.

	U/V group Mean (*SD*)	N	Control group Mean (*SD*)	N	Δ	*t*	*p*
Throttle at 80 kt (% of max)	19.05 (24.45)	10	56.12 (32.33)	10	37.07*	2.89	.010
Differential throttle used (% of run)	45.62 (33.72)	10	18.08 (20.42)	10	-27.53*	2.21	.040
Correct call-out time engine failure (s)	22.03 (9.81)	9	20.50 (5.14)	8	-1.52	.39	.700
Correct call-out time rudder failure (s)	10.88 (7.61)	5	24.07 (17.48)	7	13.19	1.57	.148
Surprise (1-5)	2.50 (.62)	10	3.10 (.39)	10	.60*	2.57	.019
Startle (1-5)	2.10 (.74)	10	2.65 (.47)	10	.55	1.98	.063
Understanding (1-5)	3.95 (.69)	10	3.10 (.66)	10	-.85*	2.83	.011

* Difference is significant at p <.05 after Holm-Bonferroni correction.

Performance example

Figure 5.5 shows an example in which a pilot lost control. No differential throttle was applied during the run (middle plot). On base leg, flaps were set to 25 degrees and gear down was selected (top plot). This caused the speed to rapidly drop below 85 kt around t = 135 seconds (top plot). The pilot responded to this by increasing throttle (middle plot). When turning towards the runway, aileron authority was lost, as indicated by the increasing bank angle despite maximum inputs in the opposite direction (also positive as per convention) at t = 140 seconds (bottom plot). Altitude was traded for speed at t = 145 seconds (top plot), and gear up was selected again to decrease drag. Despite these efforts, the decreasing speed caused loss of aileron authority again at t = 160 seconds and t = 170 seconds, after which the run was stopped to prevent a crash.
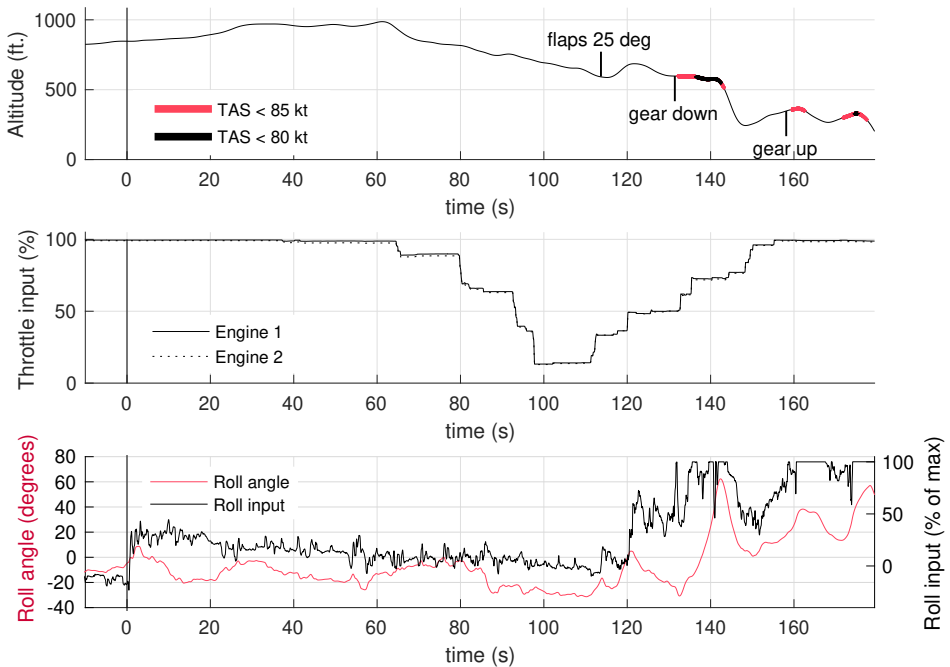
Figure 5.5: The altitude and speed (top), throttle input (middle), roll angle and roll inputs (bottom) during a run with of loss of control in the related surprise test. See the text for description.

Call-out times

There were no significant differences between the groups in correct call-out times (see, Table 5.6 and Figure 5.4). Missing values included four pilots who did not identify the malfunction, and four pilots who did not give any call-out, perhaps due to being too absorbed in the task. None of the missing cases involved incorrect call-outs. When filling in the questionnaire after the test, all pilots indicated that they had noticed that

the plane became more difficult to control in down wind.

Subjective ratings
The events were rated as significantly easier to understand and as less surprising by the U/V group compared to the control group (Table 5.6). Startle scores did not differ significantly between the groups, although there was a trend towards lower scores in the U/V group (p = .063). Pilots were on average moderately (around 3.0) surprised by the events, indicating that the surprise manipulation was successful. The maximum rating of surprise was 4 (very) for all events. Startle ratings were on average slightly (2.0) to moderately (3.0). The maximum startle rating was 5 in the unrelated surprise test, 4 for the rudder failure and 3 for the engine failure. One pilot in the U/V group broke off the takeoff. The scenario was repeated with the instruction to continue the takeoff, and the subjective ratings of the engine failure were obtained with regard to the first run.

5.3.3. Unrelated surprise test
The outcomes of the unrelated surprise test are listed in Table 5.7. There were no significant differences between the groups in any of the measures. One pilot in the U/V group did not identify the problem and performed an emergency landing. One missing case in the control group was caused by a simulator malfunction during the run.

Table 5.7: Group differences in the unrelated surprise test.

	U/V group Mean (*SD*)	N	Control group Mean (*SD*)	N	Δ	*t*	*p*
Correct call-out time (s)	118.46 (38.13)	9	136.36 (26.12)	9	17.90	1.16	.262
Surprise (1-5)	3.40 (.84)	10	3.70 (.48)	10	.30	.976	.345
Startle (1-5)	2.90 (1.10)	10	2.90 (1.10)	10	.00	<.001	>.999
Understanding (1-5)	3.00 (1.63)	10	3.30 (1.06)	10	.30	.487	.632

5.3.4. Manual skill pre- and posttest
In the posttest compared to the pretest, both groups used significantly less inputs on the ailerons, $F(1,18)$ = 7.29, p = .015, and on the elevator, F(1,18) = 23.15, p <.001, indicating a familiarization with the controls. There were no significant differences between the groups in aileron and elevator inputs (p = .522 and .354, respectively), nor were there significant Group $times$ Test interaction effects (p = .421, p = .831, respectively), indicating that the practice had not affected manual skill of the groups differently.

5.4. Discussion
The results of this simulator experiment show that pilots who had received the unpredictable and variable (U/V) practice, used throttle and airspeed more effectively in a

novel and unexpected situation, which resulted in more successful landings. The subjective ratings confirmed that the U/V group found it easier to understand the events in the test, and reported, perhaps as a consequence, also significantly less surprise [27]. The control tests suggested that the results in the related surprise test were not attributable to pre-existing differences between the groups, or to differences in habituation to surprise or familiarization with the controls. Three pilots in the U/V group lost aileron authority during the practice session, which may have affected their performance in the related surprise test. However, the difference in successful landings between the groups is still statistically significant when these pilots are excluded. In addition, three pilots in the control group also experienced loss of aileron control early in the related surprise test (before turning to base leg).

The surprise ratings in the tests were higher than the startle ratings, indicating that the events were foremost unexpected, but that they did not contain very intense or threatening stimuli. Interestingly, the difference between the groups in startle and surprise ratings was similar in magnitude, but that of startle did not reach statistical significance due to a larger variance. Something similar was also observed in a previous study [4], so it may be indicative of a higher interpersonal variation in startle responses, or larger variation in interpretation of the startle rating scale. In this respect, it is worth contemplating the value of averaging surprise and startle responses. Although it is required for the statistical analysis of training effectiveness, individualized pilot training may benefit more from the evaluation of individual surprise and startle responses.

A limitation of the study is that the practice session was very brief and the pilots were not trained to proficiency. The surprise test closely followed the practice, so the study provides no insight into long-term effects of U/V practice. Before implementing U/V into pilot training, factors such as the optimal degree of U/V, which aspects of tasks to be made unpredictable or variable, or the optimal stage of training to introduce U/V, should be carefully considered. Finally, it should be noted that it cannot be ruled out that unintended differences between the groups existed.

The results are interesting in the light of contemporary theories on surprise and sensemaking (e.g., [1, 12, 28]). According to this theoretical framework, the U/V group, compared to the control group, was more stimulated to perform sensemaking activities during the training, which helped them to develop a better "frame" with regard to the effects of asymmetric thrust, control inputs and airspeed on aircraft behavior. This knowledge was applicable to the related surprise test, so it may have helped the U/V group to make sense of the events more quickly and extensively. In line with previous literature on frame or schema construction (e.g., [10, 29]), our results imply that obtaining knowledge about principles that overarch specific training experiences, is essential for building resilient complex skills. The current study also suggests that unpredictable and variable training are means to obtain such resilience (see also, [16]). Future research may be aimed at investigating whether more general problem-solving skills exist (e.g., "flexible procedures"; [30]), which can be effectively applied in situations that are entirely new and unpracticed.

In conclusion, the results show that organizing part of pilot training in a U/V way can be an effective means to improve the generalization of skills to in-flight situations

that are not explicitly trained. Also, they suggest that one-sided and predictable training is insufficient as a means to prepare pilots for unexpected and novel situations.

References

[1] A. Landman, E. L. Groen, M. M. van Paassen, A. W. Bronkhorst, and M. Mulder, *Dealing with unexpected events on the flight deck: A conceptual model of startle and surprise,* Human Factors **59**, 1161 (2017).

[2] Bureau d'Enquêtes et d'Analyses pour la Sécurité de l'Aviation Civile, *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,* (2012).

[3] S. M. Casner, R. W. Geven, and K. T. Williams, *The effectiveness of airline pilot training for abnormal events,* Human Factors **55**, 477 (2013).

[4] A. Landman, E. L. Groen, M. M. Van Paassen, A. W. Bronkhorst, and M. Mulder, *The influence of surprise on upset recovery performance in airline pilots,* The International Journal of Aerospace Psychology **27**, 2 (2017b).

[5] Federal Aviation Administration, *Advanced qualification program; final rule,* (1990).

[6] European Aviation Safety Agency, *Loss of control prevention and recovery training: Notice of proposed amendment 2015-13),* (2015).

[7] Federal Aviation Administration, *Advisory circular 120-111, upset prevention and recovery training,* (2015).

[8] International Civil Aviation Organization, *Manual of evidence-based training,* https://www.icao.int/SAM/Documents/2014-AQP/EBT ICAO Manual Doc 209995.en.pdf (2013).

[9] G. Klein, J. K. Phillips, E. L. Rall, and D. A. Peluso, *A data-frame theory of sensemaking,* in *Expertise out of context: Proceedings of the sixth international conference on naturalistic decision making*, edited by R. R. Hoffman (Psychology Press, 2007) pp. 113–155.

[10] U. Neisser, *Cognition and reality: Principles and implications of cognitive psychology.* (W. H. Freeman and Company, San Francisco, 1976).

[11] A. Rankin, R. Woltjer, and J. Field, *Sensemaking following surprise in the cockpit: a re-framing problem,* Cognition, Technology & Work **18**, 623 (2016).

[12] P. Zhang, D. Soergel, J. L. Klavans, and D. W. Oard, *Extending sense-making models with ideas from cognition and learning theories,* Proceedings of the Association for Information Science and Technology **45**, 23 (2008).

[13] M. W. Eysenck, N. Derakhshan, R. Santos, and M. G. Calvo, *Anxiety and cognitive performance: attentional control theory.* Emotion **7**, 336 (2007).

5

[14] J. J. Caldwell, *Fatigue in the aviation environment: an overview of the causes and effects as well as recommended countermeasures,* Aviation, Space, and Environmental Medicine **68**, 932 (1997).

[15] F. G. Paas and J. J. Van Merriënboer, *Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach.* Journal of educational psychology **86**, 122 (1994).

[16] J. J. Van Merriënboer, R. E. Clark, and M. B. De Croock, *Blueprints for complex learning: The 4C/ID-model,* Educational technology research and development **50**, 39 (2002).

[17] D. Shapiro and R. Schmidt, *The schema theory: Recent evidence and developmental implications,* in *The development of movement control and coordination*, edited by J. KeIso and J. Clark (Wiley, New York, 1982) pp. 113–150.

[18] K. B. Carbonell, R. E. Stalmeijer, K. D. Könings, M. Segers, and J. J. van Merriënboer, *How experts deal with novel situations: a review of adaptive expertise,* Educational Research Review **12**, 14 (2014).

[19] T. D. Lee and R. A. Magill, *Can forgetting facilitate skill acquisition?* in *Differing Perspectives in Motor Learning, Memory, and Control*, edited by D. Goodman, R. B. Wilberg, and I. M. Franks (Elsevier, 1985).

[20] D. Rohrer, *The effects of spacing and mixing practice problems,* Journal for Research in Mathematics Education , 4 (2009).

[21] F. Brady, *Contextual interference: a meta-analytic study,* Perceptual and Motor Skills **99**, 116 (2004).

[22] L. Reid and M. Nahon, *Response of airline pilots to variations in flight simulator motion algorithms,* Journal of Aircraft **25**, 639 (1988).

[23] R. De Muynck and M. V. Hesse, *The a priori simulator software package of the Piper PA34 Seneca III*, Tech. Rep. (TU Delft, 1990).

[24] H. J. Koolstra, C. C. De Visser, and M. J. A., *Effective model size for the prediction of the lateral control envelope of damaged aircraft,* in *AIAA Modeling and Simulation Technologies Conference* (2015) p. 2036.

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

[26] W. L. Martin, P. S. Murray, P. R. Bates, and P. S. Y. Lee, *Fear-potentiated startle: A review from an aviation perspective,* The International Journal of Aviation Psychology **25**, 97 (2015).

[27] M. I. Foster and M. T. Keane, *Why some surprises are more surprising than others: Surprise as a metacognitive sense of explanatory difficulty,* Cognitive psychology **81**, 74 (2015).

5

[28] G. Klein, B. Moon, and R. R. Hoffman, *Making sense of sensemaking 2: A macrocognitive model,* Intelligent Systems, IEEE **21**, 88 (2006).

[29] R. A. Schmidt, *A schema theory of discrete motor skill learning.* Psychological Review **82**, 225 (1975).

[30] J. Field, A. Rankin, F. Mohrmann, E. Boland, and R. Woltjer, *Flexible procedures to deal with complex unexpected events in the cockpit,* in *Proceedings of the 7th REA symposium, Sophia Antipolis Cedex, France* (2017).

5

6

Managing startle and surprise with a checklist

The aim of the current study is to test the effects of a startle and surprise management checklist on pilot performance in a motion-based simulator. Such checklists have been proposed to help pilots cope with startling and surprising events in flight. An experimental group of twelve commercial airline pilots is trained to use a four-item checklist ("COOL"): 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 issue. 4. Lead: formulate a plan of action. A control group of twelve airline pilots receives a control training. Next, all pilots perform four test scenarios with startling and surprising events. Performance measures are obtained, as well as checklist application, pilot appreciation, and other subjective measures. Application of the checklist in the test scenarios is high (90 % full, 100 % partly), and pilots reported high appreciation (median: 4 on a 1-5 point scale). The experimental group makes significantly better decisions, but immediate responses are significantly impaired. Pilots sometimes apply the checklist at inappropriate times. The results thus indicate positive effects on performance, but there is some evidence that it was too distracting. The tested checklist can therefore be improved with further simplification, as well as with practicing the prioritization of other actions that take precedence over the checklist.

6.1. Introduction

Aviation safety organizations have recommended that pilots receive targeted training to manage startle and surprise [1, 2]. 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 [3]. If surprised, one needs to analyze the situation and adjust one's mental model, or "reframe", which is particularly difficult under high stress [4, 5]. This is because stress occupies working memory and impairs top-down or goal-directed attentional processes [6]. In this manner, stress may impair one's ability to frame the situation, so that a perspective on relative relevance of different cues are lost and the transition to a new frame is hindered. It is thought that startle and surprise make in-flight situations significantly more difficult to handle than situations in training [7–9].

It is still unclear how pilots should train for coping with startle and surprise. Recent research suggests that one way to do this is by introducing unpredictability and variability in training is useful [10]. Another way is to train a startle and surprise management procedure structured as a checklist. Decision-making checklists already exist to systemically deal with emergencies, for instance: FOR-DEC [11], DESIDE [12] or DO-DAR [13]. However, these checklists 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 [5]. Therefore, new checklists have been proposed, and are in use by airlines, which aim to "de-startle" pilots before they engage in a problem-solving routine (e.g. Reset-Observe-Confirm, ROC [14]); Breathe-Analyze-Decide, BAD [15]; Unload-Roll-Power, URP; [16]. Until now, there exist no peer-reviewed publications about these checklists. A report about URP [16] indicates good pilot appreciation (on average 8.3 on a 1-10 point scale; page 87) and an increase in the collection of information (page 74) in a simulator scenario at the end of a 3-hour training session. However, participants were not instructed to focus on optimal decision-making and ensuring favorable scenario outcomes. Thus, no data exists yet to indicate whether these checklists lead to better decision-making and performance.

The current experiment aims to change this by testing the effects of a startle and surprise management checklist on pilot decision-making as well as other factors. We developed a new checklist, similar to the existing ones, to have complete control over its design and presentation. Based on the above-mentioned theoretical framework, the primary aims our checklist are to manage (the effects of) stress and to aid pilots in reframing. The checklist is purposefully kept concise for use under high stress. The first step is to manage stress through breathing and muscle relaxation. Similar techniques are applied in the military (e.g., [17]), competitive sports [18], and education (e.g., [19]). The next step is to observe the immediate situation, after which the pilot focuses on the issue and formulates a plan of action. The rationale behind starting with a systematic observation is that it may prevent tunnel vision on the issue and establish a frame of the overall situation. This may be a good starting point for the following steps, in which the pilot attempts to resolve the surprise (reframe) by reaching an understanding of the problem and/or its implications. Once an understanding is reached, the pilot can come up with appropriate countermeasures.

Table 6.1: Characteristics of the participants.

	Experimental group Mean (*SD*)	Control group Mean (*SD*)
Age (yrs)	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)

Table 6.2: Characteristics of the participants (continued).

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

6.2. Method

6.2.1. Participants

Twenty-four Dutch, currently employed, line pilots participated in the experiment. Pilots with military flying experience were excluded because 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 the balance, in terms of the characteristics listed in Table 6.1 and 6.2, 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. 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 [20]. All characteristics were similar between the groups, with all p-values of t-tests or chi-squared tests larger than .300. This research complied with the tenets of the Declaration of Helsinki and informed consent was obtained from each participant.

6.2.2. Apparatus

The experiment was conducted in the SIMONA research simulator at the Delft University of Technology (see, Figure 5.1a in Chapter 5). 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.

The Piper PA-34 Seneca III, a light MEP aircraft, served as the aircraft model in the experiment (see, Figure 5.1b in Chapter 5). None of the participating pilots had the advantage of having more than basic experience on this type. The light MEP aircraft model allowed us to test the pilots' general flying skills, instead of their application of type-specific standard operating procedures. The corresponding software model is a non-linear, six-degrees-of-freedom model developed by De Muynck and Van Hesse [21], which has been adapted to simulate failures by Koolstra (e.g., [22]). The flight deck (see, Figure 6.1) 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 degrees. 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.

Figure 6.1: The flight deck as used in the experiment. Visible are: a. control column and primary flight display; b. pitch trim controls; c. engine display; d. rudder pedals; e. gear lever; f. throttle lever; g. flap lever.

6.2.3. Experimental design and tasks
As can be seen in Figure 6.2, the experimental and control group followed for the most part the same protocol.

Pilots performed the tasks as single-pilot crews. In the familiarization-, most training- and posttest scenarios, pilots were to take off at EHAM (Schiphol, Amsterdam) runway 18C, make two left turns, join a left-handed traffic pattern at 1000 ft, and land

Experimental group
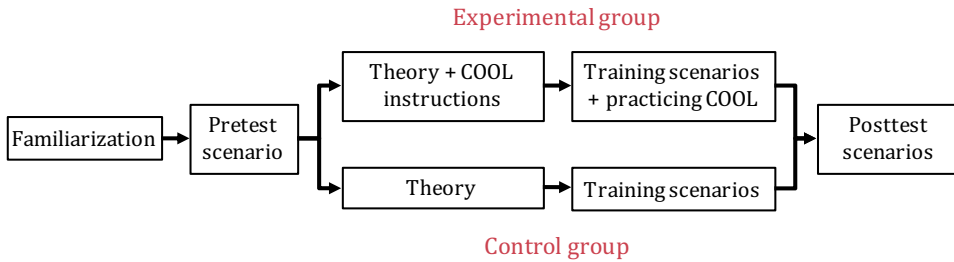
Figure 6.2: Experimental design. See the text for description.

on 18C again (Figure 6.3). This will hereafter be referred to as "standard pattern". Pilots had the required settings, as shown in Figure 6.3, available on a checklist in the cockpit, including the single-engine minimum control speed (Vmca = 80 kt).
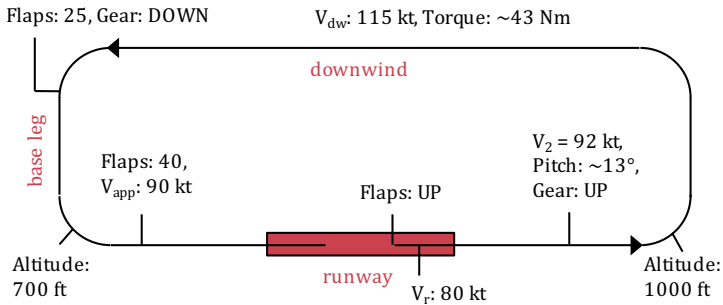
Figure 6.3: The standard traffic pattern with target settings, used in the familiarization, training and, (with some changes) in the posttest scenarios.

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 not make go-arounds, leave the standard pattern, or land on different runways. These limits were set to keep performance comparable and to increase the 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) and the third pattern was used to demonstrate the stall warning and gear-up alert by letting the pilot trigger these on downwind. At the end of the familiarization session, none of the pilots had issues with flying the pattern.

Pretest scenario
Following the familiarization, a pretest scenario was included to compare the groups in terms of performance, surprise or stress responses. Pilots had to perform a precision

landing in strong crosswind conditions. An unannounced left engine failure occurred at ca. 600 ft altitude, 1.5 minute before touchdown.

Theory
Pilots came out of the simulator to receive theoretical training. Both groups received a 10-minute presentation which explained the definitions of startle and surprise with illustrative examples of incidents and accidents. briefing in which the concepts of startle and surprise were defined and introduced (see, Introduction), and the current relevance of the topic for pilot training was explained. This was done to prime both groups equally on startle and surprise, and to convince the control group that the aim of the experiment was to test pilot responses to startling and surprising events. Only the experimental group received a second 10-minute briefing in which they learned about the startle and surprise management checklist and the reasoning behind it (see, Introduction). The checklist was taught using 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 relevant to the 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 the exact cause is not known, the more general issue (e.g., "controllability issue in pitch").

- **L. Lead.** Formulate a plan for action and follow through (e.g., "thus, I'm going to…").

The experimental group was told that the purpose of the experiment was to test the usefulness of the checklist for dealing with startling and surprising events, and they were asked to apply the checklist whenever an unusual event occurred. However, it was emphasized that immediate actions required to fly the aircraft (e.g., recovering an upset) always took precedence over the checklist, which was understood by all pilots in the group. Going back into the simulator, the experimental group now had a note with the *COOL* checklist steps attached to the control column.

Training scenarios
Pilots went back into the simulator to practice the *COOL* checklist with feedback on the execution (experimental group) or to simply respond to the issues (control group) in five training scenarios. 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 (19 kt) crosswind, 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 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 and resolving the situation safely should take precedence over applying *COOL* (experimental group only). Distractions were included to increase workload and stress, such as 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 standardized operating procedures and ATC communication were not included in the task. To make sure that pilots were in principle able to recognize the issues and respond to it, three out of four scenarios (all except the mass shift) featured issues which the participating pilots should be familiar with through their own training. The order in which the scenarios were presented was counterbalanced using the Latin square method.

The "flap asymmetry" scenario (FLAP, see Figure 6.4) consisted of a standard pattern, but with low visibility. The runway was just visible when turning towards the base leg. When selecting flaps 25, the left flap remained up. This caused a roll as well as a yaw moment. Appropriate decisions would be to land with flaps up, or to leave flaps at 25 degrees. When selecting flaps 40, the asymmetry would increase and landing would become very difficult.
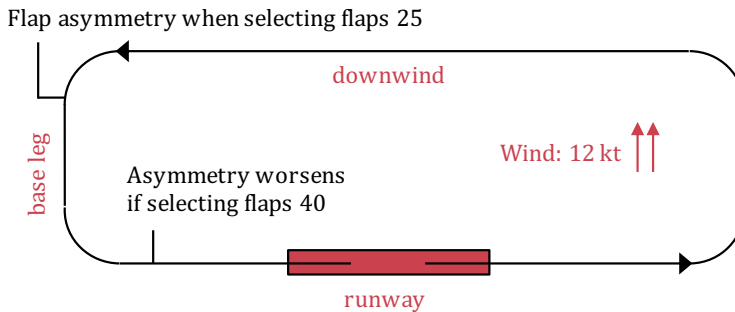
Figure 6.4: An overview of the pattern flown in FLAP.

The "mass shift" scenario (MASS, see Figure 6.5) consisted of a standard pattern. Upon rotation, a piece of cargo broke loose and shifted towards the tail, with a loud scraping noise coming from the back. The center of gravity (CoG) shifted backwards, producing a violent pitch-up moment. Pitch down controllability was significantly impaired when the pitch angle exceeded 20 degrees. In that case, the upset could be recovered by reducing thrust or rolling away from level. Since selecting flaps 25 would cause a pitch up moment as well (balloon effect), appropriate responses included early configuration in downwind, or landing with flaps up.

For the "false stall warning" scenario (STALL, see Figure 6.6) 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 stick shaker and stall audio alarm. Pilots were expected to first respond to the stall alarms by unloading (i.e., decreasing their rate of

Ballooning effect at flaps 25 causes excessive pitching up again

Wind: 13 kt ↑↑

downwind

base leg

Controllability further
impaired if selecting
flaps 40

At rotate, violent pitch-up
moment due to mass shift. Pitch
control henceforth impaired if
exceeding 15 degrees.

runway
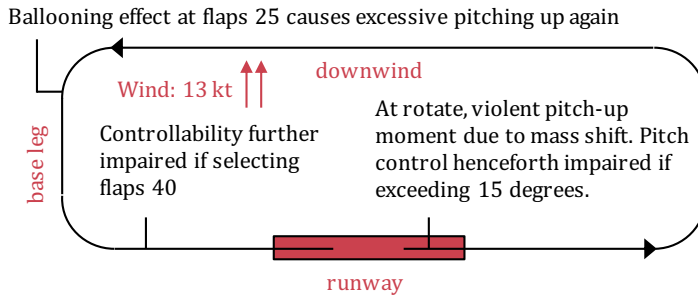
Figure 6.5: An overview of the pattern flown in MASS.

climb or descending), then figure out that the alarm was false and resume the pattern.
The scenario was stopped in downwind.

altitude
2000 ft.

runway

Bird strike and
continuous stall
alarm at 1500
ft. altitude

Wind: 6 kt ↑↑
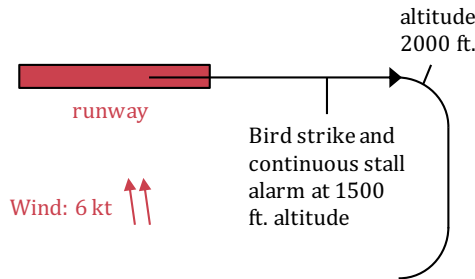
Figure 6.6: An overview of the pattern flown in STALL.

The "airspeed unreliable" scenario (ASU, see Figure 6.7) featured a standard pattern at a different airport (EHLE 05) to increase workload. Upon rotation, the indicated airspeed decreased by 1 kt every second from the actual airspeed. After realizing that the airspeed was unreliable, standard operating procedures (see e.g., [23], Appendix 6) dictate reverting to the known pitch and power settings for the required speed. This was provided on a checklist in the cockpit (see also, Figure 6.3).

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

Application and usefulness of the checklist
Pilots in the experimental group reported which steps of the *COOL* checklist they had applied. This was confirmed by checking the audio recording for the application of Observe, which was most clearly identifiable. If applied, pilots rated perceived usefulness of the checklist, after each test scenario, on a 1-5 Likert-type scale ranging from "very little" to "very much".
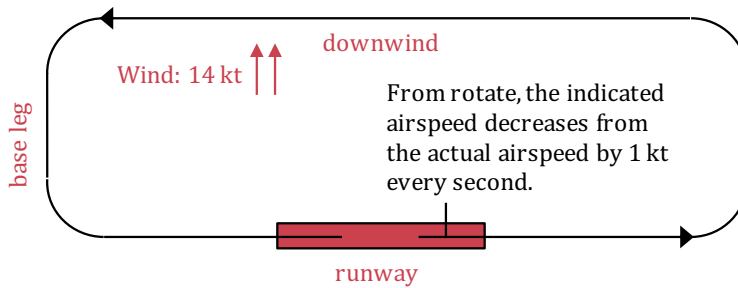
Figure 6.7: An overview of the pattern flown in ASU.

Performance

Pilot performance in the pretest was measured by checking whether the speed fell below Vse (i.e., 80 kt), the duration at which it remained below Vse, whether loss of aileron control occurred, 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 and optimal responses by the pilot. The criteria were selected based on being clearly identifiable in the logged flight parameters and in collaboration with 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 the pilot adhered to. This clustering done to avoid having too many separate outcomes, and to decrease variance in the outcomes due to chance.

The first aspect, *Aviate* (Table 6.3), 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., [24]), outlining the order of importance of piloting actions to ensure safety. Only immediate responses were analyzed, as the *COOL* checklist may cause an impairment here due to an inappropriate timing of the application. Thus, for *Aviate*, we tested whether the *COOL* checklist did not cause a performance impairment. Immediate responses were in particular necessary in FLAP (i.e., stopping or recovering the roll motion; A1 in Table 6.3), in MASS (i.e., stopping or recovering the pitch motion; A3 in Table 6.3), and to a lesser extent, in STALL (i.e., unloading; A5 in Table 6.3). Therefore, additionally to Aviate, we checked with the audio recordings if pilots started to execute the Observe step before completing these actions.

Diagnosis referred to the pilots' ability to identify the cause of the problem (see, Table 6.4), which was questioned following each scenario. Our hypothesis was that the Observe and Outline steps of the *COOL* checklist could improve *Diagnosis*. In ASU, we expected that the gradually changing instrument readings would eventually be identified by all pilots (see also, [10]). Therefore, the time it took pilots to identify the problem was obtained, indicated by a change of the power setting (D4 in Table 6.4). In addition to the criteria listed in Table 6.4, pilots rated how much difficulty they had with comprehending the issues on a 0-10 point Likert-type scale, ranging from "very little" to "extremely".

Table 6.3: Criteria defined for the performance aspect *Aviate* (A).

Criterion	Scenario	Action	Description
A1	FLAP	Prevented excessive bank angle	After selecting flaps 25, the pilot responds quickly enough to prevent the bank angle from exceeding 40 degrees. This limit was set lower than the FAA's [25] definition (i.e. 45 degrees) as pilots are flying in-the-loop.
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 Vmca (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. This limit was set lower than the general definition (i.e., 25 degrees; [25], since pilots are flying in-the-loop.
A4	MASS	Recovered quickly	When an excessive pitch angle occurs (A3), 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 (lowering pitch > 4 × the SD from the mean pitch angle, both taken over the 20 seconds before the bird strike).

Table 6.4: Criteria defined for the performance aspect *Diagnosis* (D).

Criterion	Scenario	Description
D1	FLAP	Identified a flap asymmetry or malfunctioning flaps.
D2	MASS	Identified a cargo or mass shift.
D3	STALL	Identified a false stall alarm.
D4	ASU	Reverting to a lower than full throttle setting within two minutes. This was the average identification time for the same scenario in [10].

Decision-making referred to the decisions pilots made to ensure safety in the rest of the flight (see, Table 6.5). This involved taking precautions, increasing safety margins, or being alert for the reoccurrence of the problem. We expected that the *COOL* checklist would improve situation awareness through *Calm down* and *Observe*, which could lead to better comprehension of the situation, and better projection of how the situation would or could evolve [26]. This was expected to help the experimental group to obtain higher scores in *Decision-making*.

Manipulation checks and stress response
To test if the scenarios were challenging, pilots rated startle and surprise on a 0-10 Likert-type scale ranging from "very little" to "extremely". Pilots also rated their perceived anxiety during the scenario on a visual-analogue scale ranging from 0-10 [27].

Table 6.5: Criteria defined for the performance aspect *Decision-making* (DM).

Criterion	Scenario	Action	Description
DM1	FLAP	Refraining 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 and 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 (> 1200 ft at downwind, heading 330 to 030). To limit inadvertent altitude increases, those who selected flaps in downwind are excluded as this is likely to cause an inadvertent altitude increase.
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/MASS	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 from the runway compared to the last familiarization pattern. This would require planning in downwind, and is therefore only applicable in ASU and MASS.

Although startle and surprise are involuntary initial responses, it can be imagined that better startle management in the experimental group may lead to lower overall anxiety during the scenario. Mental effort was scored on the Rating Scale Mental Effort [28], to check if the checklist did not increase mental workload in the experimental group.

Interest and enjoyment during the theory and practice sessions were rated on the Interest and Enjoyment subscale (seven items) of the Intrinsic Motivation Inventory [29]. Ideally, there would be no difference between the groups, so that the only difference is having learned the *COOL* checklist.

Data analysis

For each performance aspect, 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 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. The performance outcomes for which we predicted a positive effect of the *COOL* check-

list (i.e., *Diagnosis* and *Decision-making*) were corrected for multiple (2) comparisons using Holmes-Bonferroni correction.

6.3. Results

6.3.1. Application and perceived usefulness of the *COOL* checklist

The application of the *COOL* checklist was high according to self-report (Table 6.6). The full checklist was executed by 89.6 % of pilots on average over the scenarios. Observe was executed most consistently, with 100 % of pilots reporting application. This could be confirmed for nine pilots in the audio recordings. For another pilot this could not be heard, and the recordings of the remaining two pilots were lost. However, three out of the nine confirmed pilots 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 value (e.g., "Speed is 100"). Pilots found the checklist generally useful, mostly in FLAP and STALL, and least in MASS (Table 6.6).

Interestingly, 60 % 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, these callouts were generally very basic and specific to the failure. Around 25 % in the control group remained (nearly completely) silent during the scenarios.

Table 6.6: Self-reported application of the *COOL* checklist items by the experimental group, and perceived usefulness of the checklist.

	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 checklist (n)	11	9	11	12
Perceived usefulness median (1-5)	4	3	4	3-4

6.3.2. Examples of application of the *COOL* checklist

Tables 6.7 and 6.8 present two examples of pilots applying the *COOL* checklist. The transcripts are partly translated from Dutch. The first example (Table 6.7) is in ASU. *Observe* was performed by calling out the parameters' meaning instead of the values. The pilot appropriately interrupted the checklist multiple times to aviate or navigate. The second example in MASS (Table 6.8) 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 checklist seemed to increase attention to the possibility to control pitch with thrust. Afterwards, the pilot indicated not knowing what the issue was.

Table 6.7: An audio script showing an example of the *COOL* checklist being applied in ASU. Author comments are in [captions].

Category	Pilot comments
(Aviate/navigate)	*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.*
Calm down	[Pilot breathes]
Observe	*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.*
(Aviate/navigate)	*1000 ft* [altitude]. *Initiating slight turn.*
Outline	*It... It could also be an airspeed indicator failure, but I'll check that later.*
(Aviate/navigate)	*Let's see,* [heading] *230, there's the runway. Let's not climb too much.*
Outline	*Airspeed is... Might be indication failure. This is difficult, it's in the stall* [region] *but it still doesn't stall.*
Lead	*Let's try to stall then. Or approach stall.*
Outline	*Yeah this isn't possible at 50 knots. ASI* [air speed indicator] *problem. So then...*
(Aviate/navigate)	*Ho! We must keep it at 1000 ft.*
Lead	*Then I'm just going to fly pitch-power.*

6

6.3.3. 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 Vse (80 kt). The time flown with speed below Vse did not differ significantly ($U = 64$, p = 0.638). Three (experimental) versus four (control) experienced loss of aileron control, and one (experimental) versus two (control) did not land on the runway (p's > .500). There were also no significant differences in surprise, $U(22) = 69.0$, $p = .860$, startle, $U(22) = 54$, $p = .289$, anxiety, $U(22) = 64.0$, $p = .644$, and perceived difficulty to understand the issues, $U(22) = 41.0$, $p = .069$, although there was a slight trend in the latter towards higher scores in the control group.

6.3.4. Performance in the posttest

Aviate

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 6.9). Of the eight pilots who experienced an excessive bank angle in FLAP, two audio recordings were missing. Of the remaining six pilots, none executed *Observe* before recovering. Of the ten pilots 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 6.8). 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 *Observe* before unloading.

Table 6.8: An audio script showing an example of the *COOL* checklist being applied in MASS. Author comments are in [captions].

Category	Pilot comments
Calm down	[Mass shift occurs. Pilot breathes]
Observe	*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].
Outline/Lead	*I want to recover after the stall warning, but my aileron seems unresponsive. Maybe if I decrease power a bit.* [Pilot recovers.]
Lead	*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.*
(Aviate/navigate)	*I'm going to go back to 1000 ft* [Pilot selects flaps 25 in base leg. Nose pitches up again.]
Observe	*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.*
Outline	*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.*
Lead	*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]

Table 6.9: Adherence to the criteria defined for Aviate. Adherence to A3 or non-adherence to A5 resulted in exclusion from respectively A4 or A6.

	Experimental group		Control group	
	N	%	N	%
A1. FLAP: prevented bank angle > 40	8	66.7	12	100.0
A2. FLAP: kept speed in base leg	7	58.3	11	91.7
A3. MASS: prevented first upset	2	16.7	4	33.3
A4. MASS: recovered in under 10 seconds	4 (/10)	40.0	6 (/8)	75.0
A5. STALL: unloaded	10	83.3	11	91.7
Overall adherence median	N/A	60.0*	N/A	80.0*

* $p < .05$

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 6.10). In FLAP, those who did not identify a flap issue guessed that it was an aileron issue. In MASS, most pilots thought the issue was with the elevator or pitch trim. However, three pilots in the experimental group stated that they did not know what the issue was, due to the reoccurrence of the problem 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.

Table 6.10: Adherence to the criteria defined for *Diagnosis*

	Experimental group		Control group	
	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 2 min.	6	50.0	6	55.0
Overall adherence median	N/A	60.0	N/A	60.0

Decision-making

The experimental group scored significantly higher than the control group on the performance aspect *Decision-making*, $U(22) = 28.0$, $p = .007$ (see, Table 6.11). In MASS, one pilot in the control group selected flaps 40 for landing, which led to loss of control in-flight and a prematurely ending of the scenario. Increasing final was most often applied in ASU, whereas one pilot in each group increased final in MASS.

All comparisons between the groups per performance aspect are shown in Figure 6.8.

Table 6.11: Adherence to the criteria defined for *Decision-making*. Selecting flaps in downwind resulted in exclusion from DM3, whereas not selecting flaps resulted in exclusion from DM4.

	Experimental group		Control group	
	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	75.0	7	58.3
DM5. ASU/MASS: increased distance final	8	66.7	5	41.7
Overall adherence median	N/A	50.0*	N/A	40.0*

* $p < .05$

6.3.5. Manipulation checks and stress response

Table 6.12 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 pressure. Although not significant, the trends in surprise, perceived anxiety and mental effort were in the direction of higher scores in the experimental group. For anxiety, this was in contrast to the expectation. The experimental group rated the theory and practice as significantly less interesting and enjoyable compared to the control group.
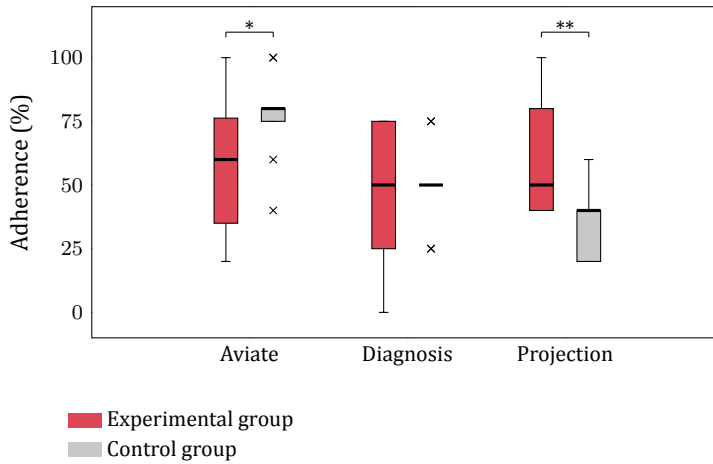
Figure 6.8: Tukey boxplots of adherence to the criteria defined for the three performance aspects. * *p* < .05.

Table 6.12: The means and standard deviations of the subjective measures in the posttests and in the practice session.

	Experimental group Mean (*SD*)	Control group Mean (*SD*)	*t*	*p*
Startle (0-10)	5.85 (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
Practice interest & enjoyment	39.4 (5.60)	44.3 (3.55)	2.57	.018

6.4. Discussion

The startle and surprise management checklist tested in the current experiment had significant positive effects on pilot decision-making under startle and surprise. Taking a moment to step-by-step relax oneself, observe the whole situation, and then analyze the problem, therefore appears to be a good approach to enable effective troubleshooting. However, the pilots' "aviation" actions (e.g., upset prevention and recovery), were somewhat delayed. This indicates that the checklist 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 workload, which was counter to the hypotheses. A cause for the distraction effect may be that a significant number of pilots started the procedure too soon in some scenarios (62.5% in MASS), instead of ensuring a safe flight path first. This happened despite explicit instructions given beforehand. It could be that inappropriate prioritizing of the *COOL* checklist was an artefact caused by the experiment, as pilots were perhaps unnaturally focused on having to execute the checklist. Still, the finding shows that it is advisable that pilot's recognition of when to execute a startle and surprise management checklist is sufficiently practiced and tested.

The application of the checklist in the experiment was high and pilots generally rated the checklist as "very useful". Some pilots remarked that the checklist 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 in a large jet aircraft. The *Calm down* and *Observe* steps were regarded as being the "core" of the checklist. A criticism was that the checklist was somewhat elaborate and could be too mentally tasking when startled in a real situation. Some improvements to reduce workload suggested by the pilots were to: "Call out the instruments' meaning instead of the absolute values in the *Observe* step, for a more natural feel", "Reduce the number of parameters to call out", "Drop the *Outline* and *Lead* steps", and: "Let only the pilot monitoring perform *Observe* in a two-pilot crew". The *Observe* step can also be combined with reverting to less dynamic and known settings, as suggested by [30], which may further decrease workload and stress.

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 appropriately to the presented scenarios, which were selected to feature familiar controllability and instrument issues that were not type-specific. Although not all issues were easy to identify (e.g., MASS), the 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 checklist was trained in this experiment was designed to allow for a comparable practice session for the control group, and therefore does not reflect optimal training for operational practice. This may have caused the lower interest and enjoyment ratings of the training by the experimental group. In operational practice, training can be optimized by practicing the checklist at a higher frequency within the training session.

In conclusion, it appears that a startle and surprise management checklist can positively influence pilot decision-making when startled and surprised. The most useful elements of the tested checklist seemed to be that it stimulated stress management and observing the overall situation, so that appropriate decisions can be made. Improvements to the tested checklist can be made to reduce workload and to ensure proper prioritization of immediate issues over checklist execution.

6

References

[1] European Aviation Safety Agency, *Loss of control prevention and recovery training: Notice of proposed amendment 2015-13),* (2015).

[2] Federal Aviation Administration, *Advisory circular 120-111, upset prevention and recovery training,* (2015).

[3] J. Rivera, A. B. Talone, C. T. Boesser, F. Jentsch, and M. Yeh, *Startle and surprise on the flight deck: Similarities, differences, and prevalence,* in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 58(1) (SAGE Publications, 2014) pp. 1047–1051.

[4] G. Klein, J. K. Phillips, E. L. Rall, and D. A. Peluso, *A data-frame theory of sensemak-*

ing, in *Expertise out of context: Proceedings of the sixth international conference on naturalistic decision making*, edited by R. R. Hoffman (Psychology Press, 2007) pp. 113–155.

[5] A. Landman, E. L. Groen, M. M. van Paassen, A. W. Bronkhorst, and M. Mulder, *Dealing with unexpected events on the flight deck: A conceptual model of startle and surprise,* Human Factors **59**, 1161 (2017).

[6] M. W. Eysenck, N. Derakhshan, R. Santos, and M. G. Calvo, *Anxiety and cognitive performance: attentional control theory.* Emotion **7**, 336 (2007).

[7] S. M. Casner, R. W. Geven, and K. T. Williams, *The effectiveness of airline pilot training for abnormal events,* Human Factors **55**, 477 (2013).

[8] A. Landman, E. L. Groen, M. M. Van Paassen, A. W. Bronkhorst, and M. Mulder, *The influence of surprise on upset recovery performance in airline pilots,* The International Journal of Aerospace Psychology **27**, 2 (2017b).

[9] W. L. Martin, P. S. Murray, P. R. Bates, and P. S. Lee, *A flight simulator study of the impairment effects of startle on pilots during unexpected critical events,* Aviation Psychology and Applied Human Factors (2016), 10.1027/2192-0923/a000092.

[10] A. Landman, P. van Oorschot, M. van Paassen, E. L. Groen, A. W. Bronkhorst, and M. Mulder, *Training pilots for unexpected events: A simulator study on the advantage of unpredictable and variable scenarios,* Human factors **60**, 793 (2018).

[11] H.-J. Hörmann, *Training of aircrew decision making,* in *AGARD*, Vol. 19 (1996).

[12] S. R. Murray, *Deliberate decision making by aircraft pilots: A simple reminder to avoid decision making under panic,* The International Journal of Aviation Psychology **7**, 83 (1997).

[13] A. J. Walters, *Crew resource management is no accident* (Aries, 2002).

[14] E. Boland, *Managing startle & surprise,* http://pacdeff.com/wp-content/uploads/2016/11/PACDEFF-Startle-Surprise-Management.pdf (2016), accessed: 2019-02-11.

[15] W. L. Martin, *Developing startle and surprise training interventions for airline training programs,* http://pacdeff.com/wp-content/uploads/2017/08/PACDEFF-FC-Forum-Presentation-on-Startle.pdf (2016), accessed: 2019-02-11.

[16] J. N. Field, E. J. Boland, J. M. Van Rooij, J. F. W. Mohrmann, and J. W. Smeltink, *Startle effect management (report nr. nlr-cr-2018-242),* (European Aviation Safety Agency, 2018).

[17] U. S. M. Corps, (2010).

[18] M. Pelka, J. Heidari, A. Ferrauti, T. Meyer, M. Pfeiffer, and M. Kellmann, *Relaxation techniques in sports: A systematic review on acute effects on performance,* Performance Enhancement & Health **5**, 47 (2016).

[19] G. Paul, B. Elam, and S. J. Verhulst, *A longitudinal study of students' perceptions of using deep breathing meditation to reduce testing stresses,* Teaching and learning in medicine **19**, 287 (2007).

[20] C. Spielberger, R. Gorsuch, R. Lushene, P. Vagg, and G. Jacobs, *State-trait anxiety inventory. palo alto,* (1970).

[21] R. De Muynck and M. V. Hesse, *The a priori simulator software package of the Piper PA34 Seneca III*, Tech. Rep. (TU Delft, 1990).

[22] H. J. Koolstra, C. C. De Visser, and M. J. A., *Effective model size for the prediction of the lateral control envelope of damaged aircraft,* in *AIAA Modeling and Simulation Technologies Conference* (2015) p. 2036.

[23] Bureau d'Enquêtes et d'Analyses pour la Sécurité de l'Aviation Civile, *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,* (2012).

[24] FAA Safety Team, *Fly the aircraft first. (ga safety enhancement topic fact sheets),* https://www.faa.gov/news/safety_briefing/2018/media/SE_Topic_18-07.pdf (2018).

[25] Federal Aviation Administration, *Upset prevention and recovery training (advisory circular no. 120-111),* (2017).

[26] M. R. Endsley, *Toward a theory of situation awareness in dynamic systems,* Human Factors **37**, 32 (1995).

[27] I. Houtman and F. Bakker, *The anxiety thermometer: a validation study,* Journal of personality assessment **53**, 575 (1989).

[28] F. R. H. Zijlstra, *Efficiency in work behaviour: A design approach for modern tools,* http://repository.tudelft.nl/view/ir/uuid3Ad97a028b-c3dc-4930-b2ab-a7877993a17f/ (1993).

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

[30] M. W. Gillen, *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, 2016).

6

7

Pitfalls when implementing a startle and surprise training intervention

Little information is still available on useful training interventions for startle and surprise. Therefore, a training intervention method, proposed by an airline company, is tested during airline recurrent simulator training. The method consists of a slow visual scan from the side-window, over the instruments, ending with facing the other pilot. Following a recorded video instruction, 38 airline pilots in two-pilot crews perform a training scenario in which they had the option to apply the method. Data on application and evaluation of the method are obtained from each pilot. Few pilots actually apply the method in the training scenario (18.4 %), and many give low ratings to the applicability of the method in the scenario, as well as in operational practice. Results show that a startle management method, as well as manner in which it is trained, should be carefully evaluated, and if necessary adjusted, before being implemented in training practice.

7.1. Introduction

Aviation safety organizations have recently recommended that pilot training should include specific means to deal with startle and surprise. Although startle is commonly used to designate both startle and surprise, strictly seen, startle refers to a reflexive stress response, whereas a surprise occurs when information is encountered that does not fit within one's mental model of the situation [2]. If both are experienced simultaneously, there needs to be an adjustment of the mental model under high stress, which can be very difficult [3]. This may result in panic, cognitive lockup and total confusion. Training interventions that have been proposed include teaching pilots certain actions to "break out" of this state. An example of this would be a checklist specifically focused on relaxation, problem analysis and decision-making. The application rate of such a method was high in an experimental setting and pilots generally appreciated it, however, some also indicated that it was too distracting or complex [4].

The current study tested a simpler startle management method, consisting of a slow scanning motion of the head from the side window, over the instrument panel, ending with facing one the other pilot. The reasoning behind this method were as follows. First, it may help one consider the overall situation, including the other pilot's state, instead of immediately zooming in on the problem. Looking out the side window, which is also used in initial training and aerobatics, can be used to obtain natural sense of the aircraft's attitude. Second, the method buys time and prevents intuitive reactions to a problem that is not fully understood yet. In a similar manner, standard procedures to recover from spatial disorientation include a first step of recognizing and confirming the spatial disorientation, before attempting to recover (see e.g., [5]). Third, performing a slow, conscious motion may instill a sense of control and stimulate goal-directed processing, as high stress is known to shift attentional control towards being more stimulus-driven [6]. Potential advantages that this method may have compared to a checklist, are that it is faster, simpler, more active and more specific (compared to e.g. the command to "Observe"). The current paper describes an early-stage trial of this method, to obtain data on its application and perceived usefulness in a representative sample of airline pilots in a standard training setting.

7.2. Method

7.2.1. Participants

Data were collected of 38 B737 pilots (19 captains and 19 first officers) and 18 Bombardier Q400 pilots (9 captains and 9 first officers). For privacy reasons, no personal data besides rank was collected. The experience level of the B737 pilots was generally higher than that of the Q400 pilots, with circa 2,500-25,000 hours compared to 600-12,000 hours. Pilots were informed that their data would be processed anonymously. They were also free to refrain from filling in the questionnaire, but there were no refusals.

7.2.2. Training intervention

The experiment took place during a recurrent simulator training session at Luxair, Luxembourg Airlines. The training intervention consisted of an 8-minute instructional

video, in which a type rating instructor gave information about startle and surprise, and outlined the intervention method:

1. Turn your head to the outside shoulder, look out of the side window.

2. Turn your head back in a continuous movement, check your flight instruments.

3. Continue turning and see your colleague's flight instruments.

4. Continue turning and have a look at your colleague.

5. Now turn back and evaluate the situation.

The total duration of the method can be under 10 seconds. The video demonstrated execution of the method from a first-person view in the cockpit.

7.2.3. Tasks

The B737 training session was a Line Oriented Flight Training (LOFT), which consisted of a complete flight from Tenerife (TFS) to Luxembourg (LUX). In cruise, the crew received warnings from ATC about an explosive device being on board. Sharing workload with the first officer, the commander would need to order a search and prepare the cabin for descent. During descend, the device would trigger, causing an elevator runaway. Since the explosion in the B737 scenario was most startling, this scenario was expected to be the most suitable for applying the startle management method.

The Q400 training session consisted of practicing several flight situations. The scenario that was used for the experiment involved a double engine malfunction, one after the other. The standard procedure in this case would not be adequate, as it would cause both engines to be shut down simultaneously. The inadequacy of standard procedures was expected to be surprising and stressful.

7.2.4. Dependent measures

During the debriefing of the training session, the pilots filled in questionnaires, which were collected in sealed envelopes. As a manipulation check of the scenarios, the following variables were rated on a 1-5 scale, ranging from "very little" (1) to "very much" (5): Surprise by the ATC warning (B737) or engine malfunctions (Q400) and Startle by the device explosion (B737) or engine malfunctions (Q400). Anxiety following the events was rated on a 10 cm horizontal visual-analogue scale ranging from none at all to maximum [7]. Mental demand and perceived time pressure following the ATC message (B737) or engine malfunction (Q400) were rated on the NASA-TLX mental demand and temporal demand subscales [8]. Finally, pilots also indicated whether they were informed by colleagues about the events in the scenario.

Next, pilots were asked if they applied the training intervention during the scenario. If confirmed, they were asked at which moments they applied it, and to what extent they felt that it helped them, as rated from "very little" (1) to "very much" (5). On a similar scale, all pilots rated how useful the method would be in operational practice. If pilots did not apply it, they indicated if this was mainly because they forgot, because they didn't find it applicable to the situation, or because they used a different method to manage their startle.

7.3. Results

7.3.1. Manipulation check

The manipulation check shows that pilots found the scenarios moderately surprising and stressful, scoring on average around the midpoint on the scales (Table 1 and 2). It is interesting that startle and surprise scores spread from the lowest to the highest endpoints, indicating that pilots may experience the same scenario very differently. Anxiety levels are similar between the groups, while the Q400 group reported somewhat higher surprise and the B737 group more startle. In the B737 group, 45 % (17) of the pilots were informed about the scenario, whereas 54 % (20) were not, and one skipped the question. The Q400 pilots all reported not being informed.

Table 7.1: Pilots' subjective experience of the B737 bomb threat scenario.

	Mean (*SD*)	Minimum	Maximum
Startle (explosion) (1-5)	3.05 (1.21)	1	5
Surprise (message) (1-5)	3.11 (1.13)	1	5
Mental demand (message) (5-100)	51.7 (16.1)	15	75
Time pressure (message) (5-100)	57.2 (20.0)	20	95
Anxiety (message) (0-100)	4.5 (2.3)	0.0	7.5
Anxiety (explosion) (0-100)	5.1 (2.3)	0.0	10.0

Table 7.2: Pilots' subjective experience of the Q400 double engine malfunction scenario.

	Mean (*SD*)	Minimum	Maximum
Startle (1-5)	2.61 (.92)	1	4
Surprise (1-5)	3.33 (.91)	1	5
Mental demand (5-100)	58.6 (17.2)	35	100
Time pressure (5-100)	54.4 (17.2)	25	90
Anxiety (0-100)	5.1 (1.9)	1.7	8.0

7.3.2. Application of the startle management method

In the B737 group, 9 out of 38 pilots (24 %) applied the method in the scenario. Eight when the explosion occurred, and one as an extra scan to check for issues. Of those not applying the method, most indicated that they forgot (37 %), or found it not applicable (37 %). Others reported they used a different method to manage startle (26 %).

In the Q400 group, 4 out of 18 pilots (22 %) applied the method in the scenario. Most pilots did not find in applicable in the scenario (56 %), some forgot (16.7 %) and one used a different method (5.6 %). All in all, the application rate of the method was low and it was similar in the different scenarios.

7.3.3. Perceived usefulness of the startle management method

The perceived usefulness of the method in the scenarios is shown in Figure 7.1. As can be seen in the figure, there were many in the B737 group who rated the method of very little use, whereas those in the Q400 group rated it little to moderately useful.
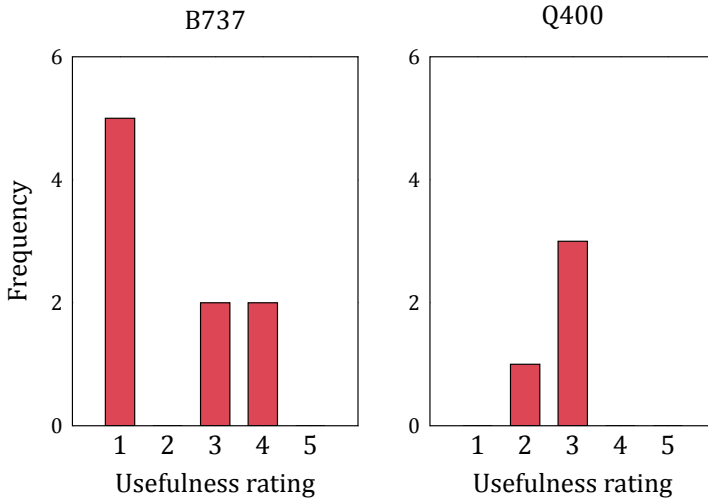
Figure 7.1: Pilots' perceived usefulness of the startle management method in the scenario. Only those who applied the method are included.

The perceived usefulness of the method in operational practice is shown in Figure 7.2. It was similar to the ratings of usefulness in the scenario. Although the Q400 group seemed a little more positive towards the method, both groups included a relatively large proportion of pilots who rated the method of "very little" or "little" use in operational practice.
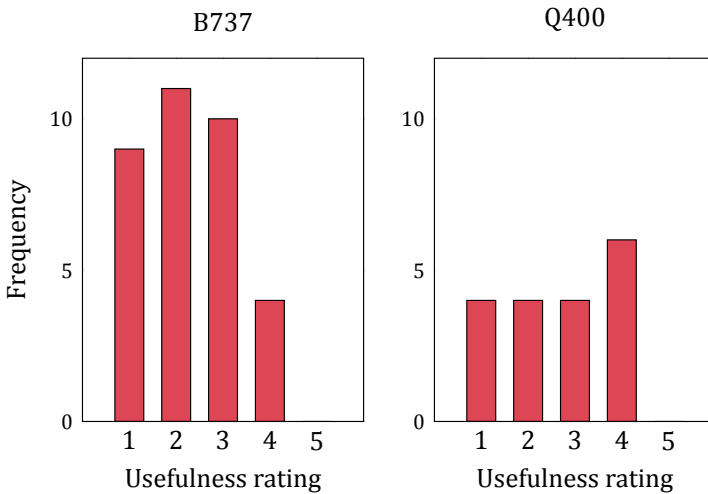

Figure 7.2: Pilots' perceived usefulness of the startle management method in operational practice.

7.4. Discussion

This experiment shows that pilots found the tested startle management method generally of little to moderate use in the scenarios and in operational practice. It is most notable that the intervention had a very low application rate (22-24 %), and a relatively large proportion of pilots (22-24 %) rated the method to be of very little use in practice. This is in contrast to the checklist-based startle management method, tested in a different study in chapter 6, which received a low rating by 8 % of the participating pilots [4].

There are some methodological aspects of the current study that may have caused lower ratings and application compared to the other study. First, there was very little time available in the experimental training session to explain the reasoning behind the method. With more time, the reasoning behind the method can be explained and there would be more room for discussing the method beforehand, which may improve the pilots' openness towards it. Second, the tested startle management method was introduced during a mandatory training session. On the one hand, this mandatory setting could make the pilots more resistant to accepting the method. On the other hand, the current sample group is more representative of the general pilot population, compared to a group who participated in an experimental study based on invitation [4]. Third, many pilots, especially those in the Q400 engine malfunction group indicated that the scenario was not startling enough for the method to be applicable. One remarked that it might be more useful "in cruise, when not mentally prepared for a malfunction as we are in the simulator."

Besides the manner in which the instructions of the method were given, there are some aspects of the method itself that can be adapted to improve it. First, most pilots who applied the method in the experiment, applied it together with their fellow pilot, indicating that if one pilot takes the initiative to execute the method, the other pilot is likely to join. The application rate might thus be improved by adding a callout at the start (e.g., "Let's do a scan"). Second, pilots indicated that they particularly experienced looking out of the side window as unhelpful. Some remarked that they thought it would be disorienting in-flight; that it seemed senseless; and that it caused them to lose time. Some of these objections can potentially be tacked with an explanation of the purpose behind the "senseless" and counter-intuitive actions. However, these objections may also indicate that the tested method may benefit from including actions that are more task-focused. Task-focus is known as an effective coping mechanism against performance stress (see e.g., [9, 10]). Consciously working on (part of) a solution to the stressful problem, even if that means systematically gathering information or simplifying the situation, may give a sense of control and instill confidence. Perhaps placing more emphasis on a structured scan of the instruments and checking verbally with one's fellow pilot would improve the acceptance and effectiveness of the method.

In conclusion, whereas the current experiment had a strong practical approach, this made it difficult to accurately measure pilots' evaluations of the method. In order to obtain a more accurate picture, pilots could be tasked with executing and evaluating a method in a more experimental setting. Also, the experiment shows the importance to reserve time and resources for the development, training and testing of a startle

management method, so that the end product is an effective method that pilots will apply in practice.

References

[1] A. Landman, E. L. Groen, M. Frank, G. Steinhardt, M. M. Van Paassen, A. W. Bronkhorst, and M. Mulder, *Pilot evaluations of a non-verbal startle and surprise management method, tested during airline recurrent simulator training.* in *Proceedings of the 20th International Symposium on Aviation Psychology* (2019).

[2] J. Rivera, A. B. Talone, C. T. Boesser, F. Jentsch, and M. Yeh, *Startle and surprise on the flight deck: Similarities, differences, and prevalence,* in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 58(1) (SAGE Publications, 2014) pp. 1047–1051.

[3] A. Landman, E. L. Groen, M. M. van Paassen, A. W. Bronkhorst, and M. Mulder, *Dealing with unexpected events on the flight deck: A conceptual model of startle and surprise,* Human Factors **59**, 1161 (2017).

[4] A. Landman, H. Van Middelaar, Sophie, E. L. Groen, M. M. Van Paassen, A. W. Bronkhorst, and M. Mulder, *Testing the applicability of a checklist-based startle management method in the simulator,* in *Proceedings of the 20th International Symposium on Aviation Psychology* (2019).

[5] F. H. Previc and W. R. Ercoline, *The outside-in attitude display concept revisited,* The international Journal of aviation psychology **9**, 377 (1999).

[6] M. W. Eysenck, N. Derakhshan, R. Santos, and M. G. Calvo, *Anxiety and cognitive performance: attentional control theory.* Emotion **7**, 336 (2007).

[7] I. Houtman and F. Bakker, *The anxiety thermometer: a validation study,* Journal of personality assessment **53**, 575 (1989).

[8] S. G. Hart and L. E. Staveland, *Development of nasa-tlx (task load index): Results of empirical and theoretical research,* Advances in psychology **52**, 139 (1988).

[9] R. F. Baumeister, *Choking under pressure: self-consciousness and paradoxical effects of incentives on skillful performance.* Journal of personality and social psychology **46**, 610 (1984).

[10] G. Matthews, E. J. Hillyard, and S. E. Campbell, *Metacognition and maladaptive coping as components of test anxiety,* Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice **6**, 111 (1999).

7

8

Conclusions

The first research objective of this thesis was to obtain insight into the mechanisms that cause pilot performance issues in startling and surprising situations. The second research objective was to identify effective simulator training interventions for startle and surprise. For both research objectives, we attempted to answer a number of key questions, which are listed below.

8.1. Key question 1

The first key question was: How do startle and surprise cause pilot performance issues in unexpected situations according to the literature? Based on existing models of human perception, performance, surprise and effects of stress, a conceptual model was developed on the effects of startle and surprise (see Figure 2.1 in Chapter 2).

The key points of this model can be summarized as:

- Dealing with surprise requires a frame adaptation (reframing). Performance issues when surprised often stem from the inability to activate an appropriate frame for the situation (e.g., not recognizing a situation at all), or from the activation of an inappropriate frame (e.g., misrecognizing a situation).
- Since reframing is typically a top-down or goal-directed process that requires significant working memory, it is highly vulnerable to stress (i.e. state anxiety).
- Stress does not necessarily originate from (acute) startle, but it may increase more gradually when the situation is slowly being perceived as a more serious threat.
- Besides stress, other factors can complicate reframing as well, such as fatigue, automation complacency, insufficient knowledge or suboptimal interface design.

With these key points, we address two potential misconceptions. The first potential misconception is that stress is the primary factor of confusion when an unexpected event occurs in flight. Such confusion is sometimes referred to as a "hijack", as the rational brain is said to be taken over by the emotional limbic system (e.g., [1–3]). This

term can be very useful when introducing the topic to pilots, but it is, in our view, important to make sure that reframing issues are not underestimated. If a pilot is merely startled but not surprised, a moment of active relaxation might be beneficial, but no reframing help is needed.

The flight safety incidents described in Chapter 2 illustrated this point. Often, there are signs of high stress or even panic in these incidents, but this often appears to *follow* a reframing problem instead of *precede* it. The sequence of events in accidents often appears to start with an underestimation and disbelief of the severity of the issue, instead of with a startle. The findings in our experiments confirm that frame mismatches, not stress, are at the root of certain perceptual and response issues. In Chapter 3, a frame mismatch led to a significant decrease in pilot performance but not to a significant increase in stress. In Chapter 4, a frame mismatch affected perception and responses, whereas surprise, and perhaps startle, were secondary.

A second potential misconception in our view, is that stress in unexpected situations necessarily entails a startle. As we outlined in our model as well as with the cases in Chapter 2, stress (i.e., anxiety) can also increase slowly, and it does not require pilots to immediately perceive an event as threatening. Therefore, in Chapters 5, and 6, we also included subtle failures that were not immediately noticed.

8.1.1. Applying the model

The conceptual model was used to design the subsequent experiments, and to identify effective training interventions. Figure 8.1 shows how the experiments in Chapters 3-7 map onto the model. The effect on pilot actions of activating a frame that mismatched with a subsequent event, was investigated in Chapter 3. How the active frame influences display perception and fast appraisal processes, and wether it induces a confirmation bias, was investigated in Chapter 4. In Chapter 5, the effect of unpredictable and variable training was tested, under the assumption that this would improve pilot knowledge frames and reframing abilities. In Chapters 6 and 7, we tested two interventions which pilots could apply when startled and/or surprised. These methods started with stress management to decrease the negative effects of stress. More importantly, they stimulated active observation of the overall situation, under the assumption that a surprised pilot does not have the appropriate frame activated to guide attention to relevant information. Finally, the checklist tested in Chapter 6 was also aimed at bringing structure to the slow appraisal process by analyzing the problem step-by-step and out loud.

8.1.2. New insights considering the model

Over the course of this research, new insights were obtained with regard to limitations and potential improvements to the conceptual model. One inconsistency is that the "inactive frame" that was manipulated in the experiments refers in some cases to a situational frame ("situational model" [4], e.g., of the aircraft bank angle in Chapter 4). In other cases, it refers to a more permanent knowledge frame (e.g., knowledge about managing asymmetric thrust in Chapter 5). The model does not differentiate between the two types of frames, and it is therefore not entirely accurate in representing the processes that are investigated in the different experiments. A more accurate
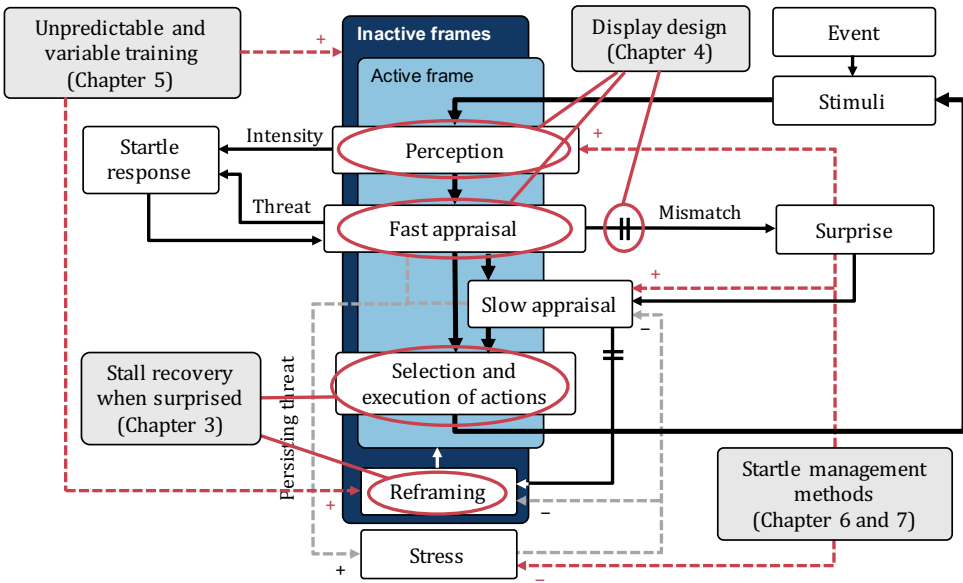
Figure 8.1: The investigated aspects of the problem of startle and surprise, as well as the aimed effects of training interventions, mapped onto the conceptual model of startle and surprise.

representation would be to place a situational frame in the model as an extra layer on the activated knowledge frame. The general frame would coincide with the concept of "genotype schemata", and the situational frame with that of "phenotype schemata" [5]. The situational frame mismatching with incoming information would constitute a situational surprise, whereas a mismatch with the general frame would constitute a fundamental surprise [6]. This change in model design, however, might make the model less clear. The active frame as shown in the model could also be interpreted as consisting of both situational (short-term) knowledge and general (long-term) knowledge.

Another insight is that, when describing the theory behind the experiments in the thesis, the distinction between frame elaborations (of the same frame) and true reframing (i.e., switching frames), was not deemed very relevant in the context of this thesis. As mentioned in Chapter 2, frames can be highly interconnected and nested within each other, which makes it difficult to distinguish where one frame ends and the other frame begins. Thus, although reframing would likely require significantly more mental effort than smaller frame adaptations, and although the realization that an entire switch of frames is necessary would likely generate more surprise, all frame adaptations were referred to as "reframing".

Small adaptations have been made to the model shown in Chapter 2 and in Figure 8.1, compared to the model published in the corresponding paper [7]. Lines indicating negative effects of stress on perception and fast appraisal were removed, to emphasize the effects of stress *when surprised*, instead of showing the effects of stress in general.

Reframing was placed within the inactive frames, which was a cosmetic improvement. This representation replaced a somewhat unclear line indicating potential positive as well as negative effects of the inactive frames on reframing.

Finally, the effects of stress on performance when surprised could perhaps have been better represented by the element of stress affecting the *influence of the frames* on the elements of perception, appraisal and action, instead of on these elements directly. There exists some evidence that shows that holistic and associative processes are typically impaired by stress [8], and stress impairing frame influence would coincide with increased bottom-up control of attention as theorized in the attentional control theory [9]. This would further indicate that stress is particularly hazardous when a frame mismatch is present.

8.1.3. Recommendations

- We recommend that experiments on managing startle and surprise should firstly manipulate the subject's expectations with regard to the situation (surprise). High pressure (e.g., time pressure) can be used to make responses more difficult and stressful.
- It is important to not only focus on startling (intense) events when training for startle and surprise. If this happens, an important part of the issue (i.e., recognizing slowly emerging problems) may be overlooked.
- One should keep in mind that reframing issues can also occur in the absence of high stress, and that the inappropriate activation of a frame may even occur unnoticed. Practically, an underestimation of reframing issues may lead to intervention methods that focus solely on stress management, resulting in a blind spot for aiding pilots with reframing. If an intervention method successfully calms a pilot down in an emergency situation, but it does not aid in sensemaking, the confusion may remain and stress may quickly return in force. In addition, when reframing issues are overlooked when explaining accident cases, pilots who do not feel vulnerable to stress may not acknowledge that something similar could happen to them.

8.2. Key question 2

In Chapter 3, we performed a simulator experiment to answer key question 2: Can we induce performance issues in responding to a critical situation by surprising pilots in the simulator? An aerodynamic stall was selected as the critical situation, as this allowed us to both surprise the pilots and measure performance in a highly controlled manner. Compared to more dynamic scenarios of longer duration, such as the scenarios in Chapter 6, the required procedure in the stall recovery task is straight-forward. The evaluation of the performance in the recovery task was based on adherence to the recovery procedure, which is more objective than the performance evaluations in Chapter 6. The pilots were familiar with the required procedure from their training and received a refresher before the test.

We found a significant decrease in adherence to the procedure in a surprising condition, compared to an anticipated condition. This indicates that surprise impairs the

correct application of the stall recovery procedure. The significant decrease in performance in the simulator suggests that, if a similar situation were to occur in reality, the performance impairments would likely be more serious. When the pilots were surprised, aspects of the procedure were skipped, and pilots seemed preoccupied with correcting the bank angle at the expense of unloading.

Based on our conceptual model, we expected that the discovery of the frame mismatch in combination with task demands would induce (secondary) startle. The mental unpreparedness of pilots to the situation was also expected to cause higher workload compared to the control condition. This was confirmed by the pilots' subjective ratings. The surprise ratings in the surprise condition were the highest of all experiments in this thesis, indicating that instilling an expectation of a completely different event was highly effective. In contrast to our expectation, the pilots' ratings of stress were not significantly higher in the surprise condition. This was likely caused by the experimental setting, as there was a social stressor (namely, that of being monitored) in both conditions.

8.2.1. Recommendations

- The results indicate that merely training a procedure in anticipated situations may not transfer to a surprising situation in operational practice. Targeted training for surprise is recommended.
- Pilots may have a tendency to fall back on intuitive responses even if these are inappropriate. Extensive practice might be necessary to make the procedure a natural response in surprising situations.
- For research purposes, surprise was manipulated by truly misleading the pilots about the upcoming events. Although this manipulation was effective and would be comparable to a surprising situation in operational practice, similar misleading measures are not advisable for training purposes. Pilots should not feel that they are being set up to fail, or made to mistrust the instructors.

8

8.3. Key question 3

In Chapter 4 we dealt with key question 3: Can we induce interpretation and response errors by inducing an inappropriate expectation through spatial disorientation? We tested whether the activation of a frame that mismatches with reality can cause misinterpretations of the artificial horizon and response errors (see Figure 8.1). Thus, instead of confronting participants with an event that clearly mismatched with their frame (i.e., a surprise), we attempted to instigate a misinterpretation of what was happening based on their frame (i.e., a confirmation bias). In a fixed-based experiment, the spatial disorientation was simulated by creating a mismatch of the expectation and the situation through a flying task. In an in-flight experiment, true spatial disorientation was induced with vestibular cues.

The results show that the incorrect responses with mismatching expectations increased by a factor of 7.8 (simulator) and 2.6 (in-flight) compared to a control condition without expectation. Participants were highly likely to respond based on an activated inappropriate frame, even if they were presented with information that indicated that a

frame adjustment was required. There was evidence that the activated frame caused a misinterpretation of the display, but this was in the simulator experiment only. Participants had much difficulty with preventing incorrect responses due to the mismatching expectation in the in-flight experiment, which suggests that vestibular cues are a powerful determinant of the situational frame of the aircraft attitude.

These results are in line with the conceptual model (Figure 8.1), as they show that inappropriate responses and misinterpretations can be explained through the divergence of the actual situation from the pilot's frame. In most cases, the incorrect response was quickly corrected by the participant. However, in one in-flight case, the confusion lasted for approximately 4 seconds, which reflects the extent of confusion that appeared to be present in real accident cases involving roll reversal errors (see also, 2.6.1; [10–12]).

8.3.1. Recommendations

- The experiments in Chapter 4 were a first investigation of the effect of expectation and disorientation when responding to the artificial horizon. The experiments were therefore designed to require little time and resources, and employed non-pilots to magnify potential effects of expectations on erroneous responses. Pilots were previously found to have lower error rates (1.5-3.1 % in-flight; [13, 14]). The next step would therefore be to investigate if, and to what extent, pilots are similarly affected by expectations. If such an effect is found, it would be recommended that pilots are made aware of it in spatial disorientation awareness training or upset prevention and recovery training.

8.4. Key question 4

Using existing knowledge from training literature in other fields, the first intervention that was tested was the use of unpredictable and variable training scenarios (Chapter 5). Key question 4 was: Does variable and unpredictable training help pilots to solve startling and surprising events in a simulator scenario? Instead of using repetitious exercises in which the pilots always foresee what will happen, it may be wise to make scenarios more variable and unpredictable and let pilots in some cases find out by themselves what is going on. In this way, surprise is introduced into the training scenarios, which is in accordance with the recommendations of aviation safety organizations. The intervention focuses on the manner in which piloting skills are trained, instead of teaching pilots a new method aimed at startle and surprise management. An advantage of this approach is that this does not require more pilot training hours.

The results show that introducing variability and unpredictability in a practice session improved the application of the practiced skills in a novel and surprising situation. This indicates that through variable practice, the pilots developed more accurate and versatile knowledge frames of the trained issues (see, Figure 8.1). Also, the unpredictability could have forced pilots to constantly reframe and analyze the trained situations, which promotes active learning (see, Figure 8.1). In contrast, pilots who trained the scenarios in a one-sided and predictable manner seemed to have developed more rigid knowledge frames that were highly specific to the trained situations. Such rigid

frames are more difficult to apply in situations that are different from the situations that are trained. A recent publication [15] explains pilot ability to recognize situations in a similar manner, that is, through the degree of overlap of the event with the trained situations (i.e., categorization theory). Thus, the key points of our conceptual model are reflected in this intervention, in that it is not about *exposing* pilots to severe surprise or startle, but that low level surprise (unpredictability) can be used to improve reframing capabilities.

8.4.1. Recommendations

- To increase pilots' abilities to deal with startle and surprise, it is important to organize scenario-based training so that the timing of events, their context and the types or combinations of events are varied and unpredictable.
- Unpredictable and variable training limits the possibilities to standardize the training and accurately compare performance data between pilots. However, dedicating a part of the training specifically to practicing, instead of using training time for testing only, has important benefits for the learning process.
- The introduction of variability and unpredictability does not exclude the possibility of using repetitious and predictable exercises as well. Such exercises would be especially useful in the earlier training phase, when the execution of procedures still needs to be learned.

8.5. Key question 5

The second training intervention that was tested was a four-item checklist to manage startle and surprise (Chapter 6). Key question 5 was: Does a startle and surprise management checklist help pilots to solve startling and surprising simulator scenarios? Compared to unpredictable and variable training scenarios, this can be seen as a more general "catch-all" tool for pilots to manage startling and/or surprising events that offer time to execute a checklist. We chose to test this type of intervention method because it was proposed by several authors [3, 16], pilots evaluated it positively in previous tests [3], and because elements of the method align with our conceptual model. The tender of EASA, which we mentioned in Chapter 1, resulted in a three-item checklist [3]. A test indicated that pilots applied the method in simulator scenarios after having received instructions and practice, but there were no indications that the method also helped pilots to make better decisions. In the current experiment, we specifically focused on such performance markers.

The checklist that we tested consists of four steps: Calm down, Observe, Outline and Lead (COOL). The first step was active relaxation, aimed at reducing the disruptive effects of stress. This intended effect is represented by a line towards "stress" in Figure 8.1. The second step forced the pilot to look around and observe the overall situation (the line to "perception" in Figure 8.1). This was done to first establish a (situational) frame of the situation and to make sure that no other threats were overlooked. According to our model, looking around would be especially important when surprised, as there may not yet be a fitting frame to guide attention toward the relevant information. Moreover, establishing an overview could have a calming effect, since the pilot

may sense that the overall situation is understood and under control. From observing the overall situation, the next step was to zoom in on the problem and diagnose what was wrong (line to "slow appraisal" in Figure 8.1). Based on this diagnosis, the last step was to decide on and execute a response.

The effects of this training intervention were less pronounced than those of the unpredictable and variable training intervention. A group that trained with the checklist showed more positive indications of performance in terms of more long-term decision-making. However, the immediate response to the situation was significantly impaired, and some pilots complained that they found the method distracting or overtaxing their attentional capacities. According to some pilots, the last two steps of the checklist were unnecessary, as analyzing the problem and devising a solution would happen naturally.

The results also show that it is not a straightforward matter to get pilots to execute a startle management method in the correct manner when they are startled. While it was made clear beforehand that immediate threats should be dealt with before executing the checklist ("Aviate first"), many pilots (60 %) executed steps of the checklist while they were still not fully recovered from an upset situation. Moreover, steps of the method were also sometimes skipped, even though several measures had been taken in the experiment to make execution of the checklist easier than it would be in operational practice.

8.5.1. Recommendations

- A startle and surprise management method should primarily support the reframing process. A first step of active relaxation may help in obtaining access to the cognitive capacity that is required for reframing.
- Our startle and surprise management checklist could be improved by increasing focus on dealing with immediate issues first, and by further simplification. Based on the results, an effective alternative could be, for instance, Aviate - Breathe - Check (ABC). Here, the latter two steps are kept the same as the first two steps of the COOL checklist, while an "Aviate" step is added to focus on stabilizing the flight path before the active relaxation phase (Breathe) and observation phase (Check).
- Since the experimental tasks were executed by single-pilot crews, an important aspect of startle and surprise management remains to be investigated, which is teamwork and communication. It remains relevant to investigate the advantages and pitfalls of startle and surprise management methods in two-pilot crews.

8.6. Key question 6

The end of the previous section already touched on key question 6: What are potential pitfalls when implementing a startle and surprise training intervention in practice? Chapter 6 already indicated that if a startle and surprise management method is taught to pilots, but is insufficiently practiced, pilots are unlikely to correctly apply the method when an emergency event happens. Stress makes us fall back on responses and solutions that are most natural, familiar and easy to us. This may cause a startle and surprise management method to be applied too soon or not at all.

Another important factor determining the application of an intervention method is whether pilots evaluate it as being useful. The checklist in Chapter 6 was generally positively evaluated. This was not the case for the non verbal startle and surprise management method which we tested in Chapter 7. This method consisted of a slow, deliberate scan from the outside window, over the instruments, ending with facing one's fellow pilot. The advantages of this method are that it is very simple and non verbal, and that it facilitates checking in with one's colleague. Similar to the checklist-based method, it aims to prevent immediate responses, and to facilitate establishing an overview of the situation first before troubleshooting (see, Figure 8.1).

In the experiment, very few pilots executed this intervention method, which made it impossible to check the effects of the method on performance. The outcomes are nevertheless interesting from a practical perspective. The difference in appreciation of the methods could indicate that pilots prefer a problem-solving oriented method (the checklist) over the non verbal method, which some indicated had no clear purpose to them. However, there were some important differences between the experiments. Pilots did not receive an extensive explanation of the reasoning behind the non verbal method, as they did for the checklist-based method. They were under more peer-pressure when applying the non verbal method, and participation to the experiment was not based on invitation. Another important difference between the two experiments is that pilots were not encouraged to try out the non verbal method in the scenarios, while they were encouraged to do so with the checklist. Finally, the pilots had not practiced the non verbal method beforehand to make it part of their repertoire.

8.6.1. Recommendations

- The different outcomes in Chapter 6 and 7 suggest that developers should closely monitor whether pilots accept a startle and surprise management method and apply it correctly. We recommend that any intervention method is first validated by testing its effect on pilot performance in startling and surprising simulator scenarios. Even if a method seems reasonable in theory, it may have disadvantageous effects on performance in practice and lead to negative transfer of training. A next relevant step would be to retrospectively obtain information from pilots who applied the method in situations in operational practice. Since the level of stress in a real situation cannot be achieved in the simulator, this would be the ultimate test to confirm if a startle and surprise management method truly works as intended.
- A method that is to be applied when startled and surprised will need to be practiced thoroughly so that application is natural when it is needed. Any startle and surprise management method is likely to feel unnatural when first applying it, as its aim is to prevent or adapt certain intuitive response tendencies. With practice, a method may become less counterintuitive, and pilots may become more likely to remember to apply the method when surprised.
- If not sufficiently encouraged to try a method in the simulator, the application of a startle and surprise management method can be perceived as a sign of weakness, especially if scenarios are not experienced as very stressful.
- Resistance to a new method can be prevented when this discomfort is acknowl-

8

edged and the purpose of it is clarified. The results of this thesis can be used to show pilots that an intervention method is based on sound theory and evidence.

8.7. Final conclusions

- When dealing with startle and surprise in-flight, pilots must solve an unexpected situation which they perceive as threatening. This requires an adaptation or a switch of the active frame under high stress.
- Performance issues in these situations can often be traced back to either the inability to activate an appropriate frame, or to the activation of an inappropriate frame. Stress may exacerbate these issues by impairing top-down attentional control.
- Training interventions should focus forst of all on strengthening pilot reframing abilities, so that these are resistant to high stress. We recommend that training exercises are made more unpredictable so that reframing is practiced. A higher variety in training scenarios may additionally improve pilots' recognition and response repertoire (i.e., the available frames). Further, a startle and surprise management procedure can be used to structure the reframing process under high stress.
- A startle and surprise management procedure will likely feel unnatural at first, as it interferes with the pilots' own natural response tendencies. To prevent resistance to a procedure, it is important that its purpose is explained and that it is evidence-based. To ensure correct application in situations in operational practice, knowing of the method is not enough. It should be practiced sufficiently to become a natural response.
- It is important to evaluate any training intervention that is implemented, to make sure that it is indeed effective instead of counter-effective in helping pilots respond to startling and surprising situations.

8

References

[1] W. L. Martin, P. S. Murray, P. R. Bates, and P. S. Y. Lee, *Fear-potentiated startle: A review from an aviation perspective,* The International Journal of Aviation Psychology **25**, 97 (2015).

[2] S. Woods, *Surprise! combating the startle effect,* FAA Safety Briefing (march/april 2016) , 18 (2016).

[3] J. N. Field, E. J. Boland, J. M. Van Rooij, J. F. W. Mohrmann, and J. W. Smeltink, *Startle effect management (report nr. nlr-cr-2018-242),* (European Aviation Safety Agency, 2018).

[4] M. R. Endsley, *Theoretical underpinnings of situation awareness: A critical review,* in *Situation awareness analysis and measurement*, edited by M. R. Endsley and D. Garland (Lawrence Erlbaum Associates, Mahwah, NJ, 2000) pp. 3–32.

[5] U. Neisser, *Cognition and reality: Principles and implications of cognitive psychology.* (W. H. Freeman and Company, San Francisco, 1976).

[6] Z. Lanir, *The reasonable choice of disaster,* in *Distributed decision making: Cognitive models for cooperative work*, edited by J. Rasmussen (Wiley, Oxford, England, 1991) pp. 215–230.

[7] A. Landman, E. L. Groen, M. M. van Paassen, A. W. Bronkhorst, and M. Mulder, *Dealing with unexpected events on the flight deck: A conceptual model of startle and surprise,* Human Factors **59**, 1161 (2017).

[8] C. Remmers and T. Zander, *Why you don't see the forest for the trees when you are anxious: Anxiety impairs intuitive decision making,* Clinical Psychological Science **6**, 48 (2018).

[9] M. W. Eysenck, N. Derakhshan, R. Santos, and M. G. Calvo, *Anxiety and cognitive performance: attentional control theory.* Emotion **7**, 336 (2007).

[10] E. M. of Civil Aviation, *The final report of the accident investigation: Flash airlines 604,* (2004).

[11] C. C. A. Authority, *Technical investigation into the accident of the b737-800 registration 5y-kya operated by kenya airways that occurred the 5th of may 2007 in douala,* (2010).

[12] Aircraft Accident Investigation Bureau, *Final report of the aircraft accident investigation bureau on the accident to the saab 340b aircraft, registration hb-akk of crossair flight crx 498 on 10 january 2000 near nassenwil/zh,* (2002).

[13] D. B. Beringer, R. C. Williges, and S. N. Roscoe, *The transition of experienced pilots to a frequency-separated aircraft attitude display,* Human Factors **17**, 401 (1975).

[14] A. H. Hasbrook and P. G. Rasmussen, *In-flight performance of civilian pilots using moving-aircraft and moving-horizon attitude indicators*, Tech. Rep. AD0773450 (Federal Aviation Administration, Oklahoma City, OK, Civil Aeromedical Institute, 1973).

[15] R. Clewley and J. Nixon, *Understanding pilot response to flight safety events using categorisation theory,* Theoretical Issues in Ergonomics Science , 1 (2019).

[16] W. L. Martin, *Developing startle and surprise training interventions for airline training programs,* http://pacdeff.com/wp-content/uploads/2017/08/PACDEFF-FC-Forum-Presentation-on-Startle.pdf (2016), accessed: 2019-02-11.

8

Acknowledgements

The completion of this PhD has been a wonderful and valuable experience for me. I've learned a lot, not only about the subject of startle and surprise in aviation, but also about the cooperative process of doing research. The dissertation that lies before you is the result of many people's input and work. Therefore, I'd like to express my appreciation and gratitude to those who supported me.

First of all, I'd like to thank Eric Groen and René van Paassen, my promotors. I have many fond memories of the conferences we visited. Eric, you have always read my manuscripts the most thoroughly and gave me valuable directions and feedback with so much care and enthusiasm. All PhD students would be jealous of a promotor who is so involved and interested in their work. Thank you very much for making sure that the methodology was sound, my message was presented well, and for advertising my work to other researchers and the industry. Doing a PhD can be quite solitary work, but you've shown me the importance as well as the joy of making and maintaining connections with others. I look forward to keep working with you at TNO.

René, thank you for putting in so much effort in the creation and design of the simulator scenarios. It was not an easy matter to translate the theory of this topic into experiments with which we could test our hypotheses. It requires great insight into human psychology, as well as into the technical aspects of simulation, and you definitely possess both. Thank you for answering all my questions. I very much enjoyed our discussions and our supervising the master students together. I wish you all the best in your sabbatical in the US.

Second, I'd like to thank the other members of the consortium: Max Mulder and Adelbert Bronkhorst. Max, thank you for giving me the opportunity to explore this topic at the TU Delft. Despite your busy schedule, you always made time to review my manuscripts, answer my questions, give valuable input, and fight for my work with reviewers and editors... I've learned a great deal from you academically as well as professionally. Adelbert, thanks for advising me with your great enthusiasm on the process of converting an idea into a model, a model into an experiment, and an experiment into a paper. Your help with developing my ideas and writing them down are very much appreciated.

I am very grateful to have worked with Peter van Oorschot and Sophie van Middelaar, who graduated by developing and executing two of my experiments. Your great ideas, creativity and commitment has made these experiments possible. Thank you Sophie for showing me what a glider can do. I very much enjoyed working with you both and wish you all the best with your further career.

Other people I'd like to thank: Herman Koolstra, Matthijs Ledegang, Olaf Stroosma and Dirk van Baelen, for helping us design and tweak our surprise scenarios in the simulator. Hans Mulder for your advice on appropriate pilot responses. My friends of the startle and surprise project team: Geoffroy Bemelmans, Gunnar Steinhardt and Marc

Curriculum Vitæ

Annemarie Landman

09-03-1986   Born in Bunschoten-Spakenburg, Netherlands.

Education

2015–2019   PhD Candidate
Delft University of Technology, Delft, Netherlands
Thesis: Managing startle and surprise in the cockpit
Promotors: prof. dr. E. L. Groen, dr. ir. M. M. van Paassen

2010–2011   Master Human Movement Sciences: Sports, Exercise & Health
VU University Amsterdam, Netherlands
Thesis: Effects of anxiety on the execution of police arrest
                        and self-defense skills

2004–2010   Bachelor Human Movement Sciences
VU University Amsterdam, Netherlands

1998–2004   Atheneum
Guido de Brès, Amersfoort, Netherlands
Subject cluster: Science & Technology

Recent work experience

2012–2014   Junior Researcher
VU University Amsterdam, Netherlands
Activities: Designed and performed experiments to investigate
                        factors influencing performance under pressure by
                        the Dutch police and specialized arrest unit.

Awards

2010   Best presentation award for presenting the Bachelor dissertation.
2019   ISAP Stanley Roscoe best student's paper award

List of Publications

Journal papers

12. **A. Landman**, S. H. van Middelaar, E. L. Groen, M. M. van Paassen, A. W. Bronkhorst and M. Mulder, *Keep it cool! – Testing the effectiveness of a checklist-based startle management method for pilots*, Manuscript submitted for publication. (2019).

11. **A. Landman**, E. L. Groen, M. M. van Paassen, A. W. Bronkhorst and M. Mulder, *Expectation causes misinterpretation of the attitude indicator and roll reversal errors: a fixed-base simulator experiment*, Manuscript submitted for publication. (2019).

10. **A. Landman**, S. Davies, E. L. Groen, M. M. van Paassen, N. Lawson, A. W. Bronkhorst and M. Mulder, *In-flight spatial disorientation induces roll reversal errors when using the attitude indicator*, Applied Ergonomics, **81** (2019) Advance online publication.

9. **A. Landman**, P. van Oorschot, E. L. Groen, M. M. van Paassen, A. W. Bronkhorst and M. Mulder, *Training pilots for unexpected events: A simulator study on the advantage of unpredictable and variable scenarios*, Human Factors, **60**, 6 (2018).

8. Fu, W., **A. Landman**, M. M. van Paassen, and M. Mulder, *Modeling human difference threshold in perceiving mechanical properties from force*, IEEE Transactions on Human-Machine Systems, **48**, 4 (2018).

7. **A. Landman**, E. L. Groen, M. M. van Paassen, A. W. Bronkhorst and M. Mulder, *The influence of surprise on upset recovery performance in airline pilots*, The International Journal of Aerospace Psychology, **27**, 1-2 (2017).

6. **A. Landman**, E. L. Groen, M. M. van Paassen, A. W. Bronkhorst and M. Mulder, *Dealing with unexpected events on the flight deck: a conceptual model of startle and surprise*, Human Factors, **59**, 8 (2017).

5. P. G. Renden, **A. Landman**, N. R. Daalder, H. P. de Cock, G. J. P. Savelsbergh and R. R. D. Oudejans, *Effects of threat, trait anxiety and state anxiety on police officers' actions during an arrest*, Legal and Criminological Psychology, **22**, 1 (2017).

4. **A. Landman**, A. Nieuwenhuys and R. R. D. Oudejans, *Decision-related action orientation predicts police officers' shooting performance under pressure*, Anxiety, Stress, & Coping, textbf29, 5 (2016).

3. **A. Landman**, A. Nieuwenhuys and R. R. D. Oudejans, *The impact of personality traits and professional experience on police officers' shooting performance under pressure*, Ergonomics, textbf59, 7 (2016).

2. P. G. Renden, **A. Landman**, G. J. P. Savelsbergh and R. R. D. Oudejans, *Police arrest and self-defence skills: Performance under anxiety of officers with and without additional experience in martial arts*, Ergonomics, **58**, 9 (2015).

1. P. G. Renden, **A. Landman**, S. F. Geerts, S. E. Jansen, G. S. Faber, G. J. P. Savelsbergh and R. R. D. Oudejans, *Effects of anxiety on the execution of police arrest and self-defense skills*, Anxiety, Stress, & Coping, **27**, 1 (2014).

Conference papers

3. **A. Landman**, S. H. van Middelaar, E. L. Groen, M. M. van Paassen, A. W. Bronkhorst and M. Mulder, *Testing the applicability of a checklist-based startle management method in the simulator*, Proceedings of the 20th International Symposium on Aviation Psychology (2019).

2. **A. Landman**, E. L. Groen, M. Frank, G. Steinhardt, M. M. van Paassen, A. W. Bronkhorst and M. Mulder, *Pilot evaluations of a non-verbal startle and surprise management method, tested during airline recurrent simulator training*, Proceedings of the 20th International Symposium on Aviation Psychology (2019).

1. **A. Landman**, E. L. Groen, M. M. van Paassen, A. W. Bronkhorst and M. Mulder, *The effect of surprise on upset recovery performance.*, Proceedings of the 19th International Symposium on Aviation Psychology (2017).

Books

1. **A. Landman**, A. Nieuwenhuys and R. R. D. Oudejans, *Effectief omgaan met acute stress: Effecten van aanleg en trainingservaring op de schietprestatie onder druk*, Apeldoorn / Amsterdam: Politie & Wetenschap / MOVE, Vrije Universiteit (2015).