Quantifying Startle and Surprise: Development of Measuring Instruments and Validation in an Aviation Context

Chen, J.

DOI
10.4233/uuid:9d02fccd-8c3a-45d9-b10a-f84806b370f9

Publication date
2025

Document Version
Final published version

Citation (APA)
Chen, J. (2025). *Quantifying Startle and Surprise: Development of Measuring Instruments and Validation in an Aviation Context*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:9d02fccd-8c3a-45d9-b10a-f84806b370f9

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

Quantifying Startle and Surprise

Development of Measuring Instruments and Validation in an Aviation Context

Jiayu Chen

陈佳钰

Quantifying Startle and Surprise

Development of Measuring Instruments
and Validation in an Aviation Context

Quantifying Startle and Surprise

Development of Measuring Instruments
and Validation in an Aviation Context

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, prof. dr. ir. T. H. J. J. van der Hagen,
chair of the Board for Doctorates,
to be defended publicly on
Wednesday 10 December 2025 at 12:30 o'clock

by

Jiayu Chen

Master of Engineering in Aeronautical Engineering,
Northwestern Polytechnical University, China,
born in Xi'an, China.

This dissertation has been approved by the promotors and copromotor.

Composition of the doctoral committee:

Rector Magnificus,	chairperson
Prof. dr. ir. M. Mulder,	Delft University of Technology, promotor
Dr. ir. M. M. van Paassen	Delft University of Technology, promotor
Dr. H. M. Landman	Delft University of Technology, copromotor
Independent members:	
Dr. ir. J. Ellerbroek	Delft University of Technology
Dr.-Ing. B. R. Korn	German Aerospace Center (DLR), Germany
Prof. dr. S. C. Pont	Delft University of Technology
Prof. dr. ir. J. C. F. de Winter	Delft University of Technology
Prof. dr. G. C. H. E. de Croon	Delft University of Technology, reserve member

Ir. O. Stroosma has contributed greatly to the realization of this dissertation.

Keywords: Aviation; Pilots; Performance; Psychometric measures; Stress; Cognition; Emotions; Evidence-based training

Printed by: Ipskamp Printing

Front & Back: Jiayu Chen

ISBN 978-94-6518-166-0

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

Summary

Unexpected in-flight safety events often require pilots to rapidly assess the situation and respond appropriately. Such events may also elicit startle and surprise, which can severely impair pilot cognitive performance. A significant scientific gap remains, however, in the ability to measure and quantify startle and surprise in a systematic manner, specially within ecologically-valid task environments. This dissertation addresses this gap by developing and validating self-report measuring instruments designed to quantify startle and surprise. To achieve this, the research focused on investigating three key components: 1) the conceptualization of cognitive processes underlying pilot decision-making in unexpected situations; 2) the investigation of how existing self-report measures of startle and surprise, relate to personality traits and flight experience; and 3) the development of new self-report measures for startle and surprise and their validation in an ecologically-valid setting.

To conceptualize and describe cognitive processes underlying pilot decision-making and actions in unexpected situations, Chapter 2 describes pilots' decision-making and actions in two real-world incidents, Loganair Flight 6780 and US Airways Flight 1549. Both incidents were examined through the lens of three cognitive models: the Landman model of startle and surprise, the perceptual cycle model and the three-level model of situation awareness. The strengths and limitations of each model in explaining pilot decision-making in emergency situations were critically assessed. In addition, the role of schema/frame selection and modification, stress, and pilot flight experience in shaping decision-making during the incidents were explored. The application of these cognitive models reveals key distinctions that carry implications for pilot training in high-stakes, critical situations.

In Chapter 3, we examined whether existing self-report measures of cognitive and affective responses relate to personality traits, across seven startling and surprising scenarios performed in motion-based simulators. A dataset of 89 airline pilots from four studies was analysed. The personality traits included trait anxiety, decision-related action orientation, and failure-related action orientation. Pilot cognitive and affective responses in scenarios were standardized by obtaining Z-scores of perceived startle, surprise, stress, and mental workload. Results show that pilots with higher trait anxiety reported higher level of stress in response to startling and surprising events, highlighting a potential vulnerability in cognitive performance and emotional regulation under pressure. No significant effects of action orientation, or flight hours were found on pilots' responses. This could indicate that existing, non-validated measures of startle and surprise are lacking in accuracy. And the possible benefit from targeted training interventions for mitigating effects of high stress, startle or surprise could be obtained for both novice and experienced pilots.

To establish psychometrically sound self-report measures for startle and surprise, a structured and iterative process of development and validations was conducted. Chapter 4 outlines the three-phase construction of the Startle and Surprise Inventories (Startle-I; Surprise-I) and Visual Analogue Scales for Startle and Surprise (Startle-VAS; Surprise-VAS). In Phase 1, 14 items for surprise and 7 items for startle were derived from foundational

and applied literature. These items were evaluated for content validity by seven experts in the field, with retention determined by a minimum of 50% agreement on the relevance. In Phase 2, 81 participants rated the retained 19 (of 21) items nine times, each time immediately after watching a video clip. Construct validity was assessed using multilevel exploratory factor analysis (ML-EFA) with an oblique, direct oblimin rotation. In Phase 3, the concurrent validity of the Startle-VAS and Surprise-VAS was tested by comparing with the Startle-I and Surprise-I scores, respectively. The first two phases yielded an 11-item two-factor solution, corresponding to the constructs of Startle-I and Surprise-I, with Cronbach's α ranging from $\alpha = 0.714$ to $\alpha = 0.929$ for Startle-I, and $\alpha = 0.843$ to $\alpha = 0.955$ for Surprise-I, indicating acceptable to excellent internal consistency. These results support the Startle-I and Surprise-I as validated and reliable self-report measures of startle and surprise, with the single-item Startle-VAS and Surprise-VAS as efficient alternatives.

Building on the factor structure, the construct validity of the Startle and Surprise Inventories was further investigated in Chapter 5, in a more ecologically-valid setting, using a multilevel confirmatory factor analysis (MCFA). A sample of 26 professional pilots was exposed to eight different in-flight startling and surprising scenarios in a motion-based simulator. Results show that the two-factor model, comprising the factors Startle and Surprise, demonstrates a superior and acceptable fit compared to a one-factor model. Additionally, all items demonstrated significant factor loading at both the within-scenario and between-scenario level in the two-factor model. Reliability estimates, as measured by McDonald's ω, ranged from $\omega = 0.88$ to $\omega = 0.96$ for the Startle-I and $\omega = 0.77$ to $\omega = 0.96$ for the Surprise-I across scenarios, indicating acceptable to excellent internal consistency. These findings provide the first evidence for the validity and reliability of the Startle and Surprise Inventories in a relatively ecologically-valid setting, corroborating the findings in Chapter 4.

To establish criterion-related validity, the effects of startle and surprise on pilots' information-processing performance during simulated in-flight events was investigated in Chapter 6. Pilot information-processing performance was assessed using a secondary auditory cognitive task administered within the same experiment used to examine the factor structure of the Startle-I and Surprise-I in Chapter 5. Linear mixed-effects models were used to analyse the relationships between self-report startle and surprise, and information-processing performance, while accounting for individual differences and scenario variability. Results show that heightened startle significantly impaired the secondary task performance, whereas surprise did not show a significant effect. This finding suggests that, within the tested scenarios, startle imposed a more immediate and disruptive influence on pilot information processing than surprise. Since the expected impact of surprise was not confirmed, criterion-related validity of the Surprise-I remains to be established.

In conclusion, the Startle and Surprise Inventories represent the first systematically-validated self-report measuring instruments for startle and surprise in response to specific stimuli or events. This offers a robust foundation for quantifying startle and surprise in an operationally-relevant aviation context. The application of the Startle and Surprise Inventories within structured simulation environments enables systematic tracking of operators' responses to unexpected events in aviation and other safety-critical domains. This, in turn, can inform the refinement of training protocols or interventions to better support cognitive resilience and emotional regulation in high-stakes situations involving unexpected events.

Contents

1

Introduction

Since the advent of commercial jet air transport in 1958, the rate of fatal accidents has steadily declined, despite a significant growth in air traffic over the past 65 years [1] (Figure 1.1). Safety has improved tremendously and many causes for accidents have been eliminated. Currently, Loss of Control In-flight (LOC-I) is the leading cause of fatal accidents in the worldwide commercial jet fleet between 2004 and 2024 [1] (Figure 1.2). This highlights the need for understanding the causes of LOC-I and developing targeted interventions to reduce the number of fatal accidents by addressing the contributing factors of LOC-I [2].

LOC-I refers to flight conditions that are outside the normal operating envelopes, not

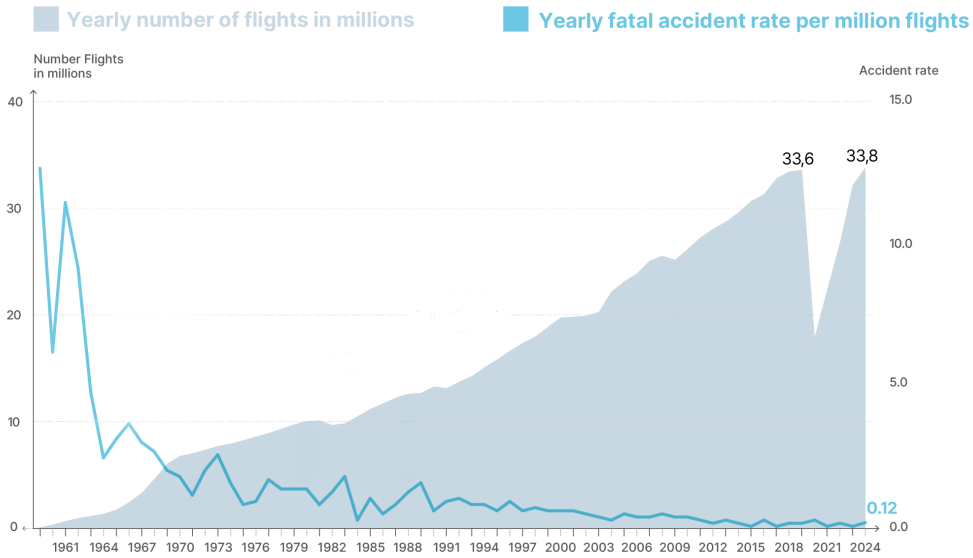


Figure 1.1: Yearly rates of fatal accident per million flights (adapted from [1]).

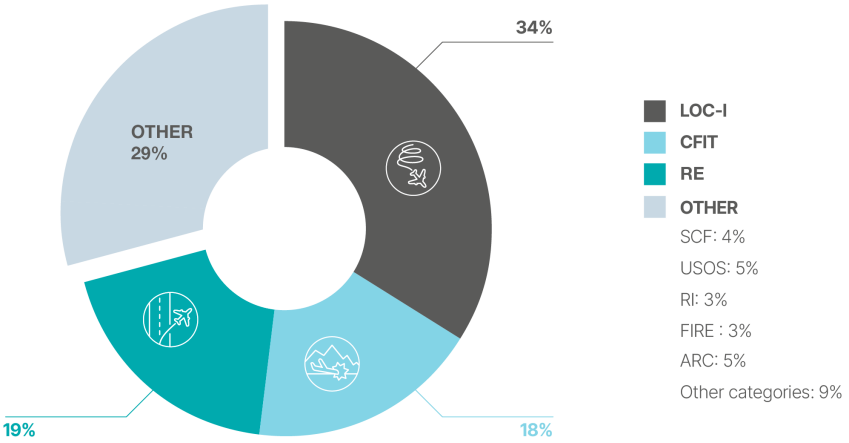


Figure 1.2: Fatal accident distribution per accident category 2004-2024. Source: [1] (CFIT = Controlled Flight Into Terrain; RE = Runway Excursion; SCF = System/Component Failure or Malfunction; USOS = Undershoot/Overshoot; RI = Runway Incursion; ARC = Abnormal Runway Contact).

predictably altered by pilot inputs, and characterized by nonlinear dynamic effects, high angular rates, and the inability to maintain stable flight [3]. It is quantitatively identified when an aircraft exceeds three or more of five defined flight envelopes relating to aerodynamics, attitude, structural integrity, and control response (pitch and roll control).

Despite substantial advancements in flight systems and automation, unexpected events such as system failures, software anomalies, bird strike and adverse weather still require human intervention [4–6]. Such situations, when they occur, are cognitively demanding and emotionally stressful. Accident investigations have frequently cited inappropriate or delayed pilot control inputs or actions as contributing factors in LOC-I [7].

Among the various operational challenges, a particularly pervasive issue in modern cockpits is automation surprise [8]. Unlike external disturbances, automation surprise reflects mismatches between the pilot's mental model and automation behaviour, rather than a system malfunction per se. A field survey involving 200 airline pilots reported an average of three automation surprise events per pilot per year [9]. While 90% of the 180 recorded automation surprise cases did not result in undesired aircraft states and only 1% led to consequential damage, the frequency of such events underscores persistent cognitive mismatches between pilots mental models and automation behaviour. Notably, 89% of these events were self-discovered, with pilots often becoming aware of the issue *after* it had affected flight operations or even safety. The cognitive mismatches could increase cognitive workload and diminish trust in automation systems, especially under high-stress or high-demand conditions [8, 10, 11].

While automation surprise represents a cognitive disruption unique to highly-automated environments, it shares essential characteristics with other in-flight unexpected events. In particular, such anomalies require pilots to rapidly assess the situation, interpret abnormal sensor data and system information, and take actions to maintain control of the aircraft. These

events may induce startle and surprise responses, which could impair situation recognition and delay recovery [12], thereby contributing significantly to LOC-I. The cognitive demands associated with startle and surprise could also elevate stress levels, further degrading pilot cognitive performance. This degradation may manifest itself as attentional narrowing, breakdowns in crew coordination, rushed or unstructured decision-making, and neglect of essential tasks [10, 13]. The definitions and operational implications of startle and surprise are examined in the following section.

1.1. Startle and surprise

Recognizing the impact of startle and surprise on pilot performance, the European Union Aviation Safety Agency (EASA) and the Federal Aviation Administration (FAA) have integrated startle and surprise into the regulatory framework [12, 14] for Upset Prevention and Recovery Training (UPRT). Their goal is to help pilots develop both the necessary flight handling skills and the resilience to cope with sudden, high-stress events:

> *"The training focuses on applying correct and timely recovery strategies to return the aeroplane to safe flight, whilst building the pilot's resilience against the associated psychological and physiological human factors (to better cope with the startle and surprise effect)"* (Page 12 [14])
>
> *"Because upsets that occur in normal flight operations are unplanned and inadvertent, pilots may be startled or surprised, adversely impacting recognition or recovery. Instructors need to plan scenarios to balance potential for startle or surprise while applying sound judgment with respect to realism and fidelity..."* (Page 9 [12])

Although startle and surprise have been widely recognized as critical human factors affecting pilot performance during unexpected situations, there remains a significant gap in how these responses can be systematically measured and integrated into operational contexts. While validated self-report measures exist for constructs such as stress [15, 16] and (mental) workload [17, 18], there are currently no equivalent standardized tools for assessing startle or surprise. In the context of enhancing resilience in response to startle and surprise, it is challenging to holistically evaluate the effectiveness of targeted training or interventions for UPRT, without clear methods to quantify the onset, duration and intensity of startle and surprise.

Reliable and validated measuring methods would enable deeper insights into physiological correlates, recovery dynamics, and task-specific vulnerabilities to startling or surprising stimuli. In aviation, such methods are essential for identifying key factors that affect pilot performance during unforeseen emergencies. By capturing real-time responses to unexpected events, these methods could support the development of evidence-based strategies to prevent and mitigate adverse effects. Moreover, insights derived from validated measures could inform the design of more effective pilot training protocols by identifying critical performance limitations and adaptive capacities under stress. Together, these methodological applications contribute to the advancement of training systems tailored to high-stakes environments, enhancing safety and operational performance.

A critical step toward developing reliable and validated measuring methods is to clarify the conceptual distinctions and overlaps between startle and surprise, which are often

used interchangeably despite representing distinct cognitive, affective, and physiological processes [19]. Following this, an overview of existing approaches to measuring startle and surprise will highlight current limitations and inform the rationale for developing more targeted measuring methods.

1.2. Definitions

Startle refers to a coinciding emotional and physiological response elicited by a sudden, threatening, intense stimulus [20]. The startle reflex, evolving typically within 20-50 milliseconds after a stimulus [21], involves the involuntary physiological reflexes and muscular activities (e.g., eye-lid-closure, contraction of facial, neck and skeletal muscles [19, 20]), which can prepare the body for protection against adverse circumstances [22]. If the threat persists, it is followed by the generalized stress response activated within the autonomic nervous system [23], including the release of cortisol, activation of the autonomic nervous system, rapid breathing, accelerated heart rate, sensory arousal, increased systolic blood pressure and dilation of the pupils [21, 24–26]. The response is more severe when an individual's arousal or stress level is already high (i.e., fear-potentiated startle [23]). Startle can be triggered by acoustic (e.g., sudden noises), electrical (e.g., cutaneous shock), tactile (e.g., air puff), or visual (e.g., lightning flash) stimuli.

Evidence suggests that, in aviation, the immediate psychomotor impact of startle may induce brief disorientation and short-term psychomotor impairments [27]. The startle response can inhibit cognitive processing and muscular activities, causing deterioration of task performance with increased response time and lower response accuracy [28]. A skilled motor task will be momentarily disrupted by the startle reflex but return to normal within five to ten seconds. Although there are some short-term physiological changes, there is little evidence on the impact of startle response on cognitive functions [29], especially in operational aviation contexts.

Surprise is a cognitive and emotional response triggered by unexpected, schema-discrepant events that are (momentarily) difficult to explain [30, 31]. Its primary evolutionary function is thought to monitor the accuracy of one's understanding of the world [32, 33]. The ability to predict relevant events is essential for survival, and surprise reflects a failure in anticipation and a state of unpreparedness. The unexpectedness interrupts ongoing mental processes and redirects attention to the surprising events, which can be disruptive and distressing due to the innate need for predictability [30, 31]. The experience of surprise serves as an alert to cognitive discrepancy, and motivate deeper information processing or schema revision, leading to improved cognitive flexibility and adaptability [34].

One's understanding of the situation needs to shift in order for the unexpected new information to make sense again. These so-called reframing efforts require effortful, goal-directed attentional processing [35], which is more vulnerable under acute high stress [36]. However, if the discrepancy remains unresolved, the situation may be perceived as poorly understood, hindering the ability to focus on the information related to the task, make accurate projections, and execute appropriate actions.

Thus, although often used interchangeably, startle and surprise are distinct phenomena that differ in their temporal dynamics, eliciting conditions, and effects on human performance. Recognizing these differences, Landman et al. [35] proposed an integrated conceptual model to explain how pilots respond to unexpected events on the flight deck, with

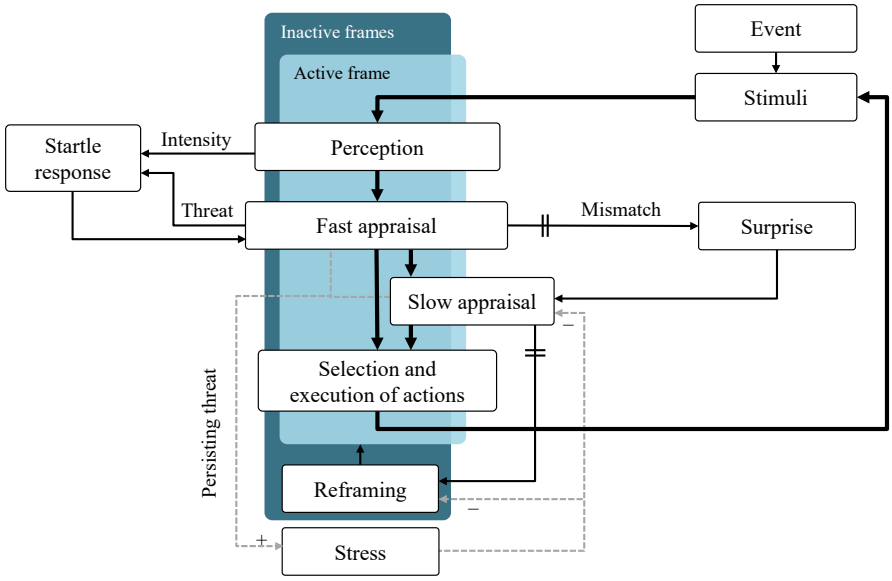


Figure 1.3: Landman model of startle and surprise [35].

particular attention to the roles of startle and surprise (Figure 1.3). Drawing on insights from perceptual psychology, sensemaking theory, and acute stress response, the model outlines how startling or surprising stimuli can disrupt perception, appraisal, and action selection. Building on this theoretical framework, the research presented in this dissertation extends prior work by investigating reliable and validated measures of startle and surprise, with the aim of more accurately characterizing their respective impacts on human performance.

1.3. Measures of startle and surprise

1.3.1. Measures of startle

The intensity of the startle response is generally measured through the eye-blink reflex. This can be done with surface electromyographic (EMG) [37], by which action potentials generated within the orbicularis oculi muscle can be detected. Other methods for capturing eyelid movement include potentiometric sensors [38], photoelectric systems [39], vertical electro-oculographic (vEOG) [40], and magnetic search coil techniques [41]. In addition to the blink reflex, pupil dilation has been identified as a physiological correlate of startle arousal [42]. Autonomic responses, such as increased heart rate and blood pressure, have also been documented within 10 seconds of an acoustic startle, offering complementary cardiovascular markers [26]. Despite the availability of objective physiological measures, no self-report instrument to assess startle has yet been developed or validated [43].

1.3.2. Measures of surprise

Surprise can be physiologically measured through the EEG P300 event-related potential (ERP) [44], pupil dilation, and activation in subcortical regions associated with dopamine [45]. The P300 ERP, originates from the anterior cingulate cortex, and peaks in ACC activity occur in aversive defence responses in general [46]. Pupil dilation has been considered an index of the brain's global state during cognitive processing under stress, rather than being exclusively elicited by surprise [47]. Nevertheless, pupil dilation related to surprise (at maximum around 500 ms after the stimulus) has also been reported to be slower than pupil dilation due to startle [48].

For subjective measures, the Differential Emotions Scale (DES-IV) [49] and the Positive and Negative Affect Schedule-Expanded Form (PANAS-X) [50] have been developed with subscales for measuring surprise. These two subscales consist of the same three items (i.e., How do you feel "surprised", "amazed", and "astonished"?), which are rated on 5-point Likert scales pertaining to feelings in this moment or to a certain past time frame. In DES-IV, surprise is considered a positive emotion among twelve fundamental affective states. Factor analysis support the distinctiveness of the surprise construct [49, 51], with test-retest reliability over six months reported at $r = 0.61$ [52]. However, internal consistency is moderate, Cronbach's $\alpha = 0.65$, likely due to the limited number of items [53].

The PANAS-X conceptualizes surprise as a discrete, valence-neutral emotion. Its internal consistency is slightly higher than that of the Surprise subscale in the DES-IV, with Cronbach's ranging from $\alpha = 0.72$ to $\alpha = 0.80$. The test-retest coefficient over a two-month interval was lower when referring to the time frame "past week", $r = 0.23$, and higher to "in general", $r = 0.52$ to $r = 0.56$ [50]. The mean scores on Surprise were the lowest compared to other affects in the PANAS-X over different samples [50]. Surprise was also the only subscale for which self-ratings did not significantly correlate with peer-ratings, $r = 0.14$. Besides these multi-item subscales, researchers have employed single-item measures to assess self-report surprise [54, 55], however, at the time of writing this dissertation, such measures have not been validated in a systematic way.

1.3.3. Limitations to current measures

While physiological and current self-report measures have been widely used to assess startle and surprise, several critical limitations remain. These limitations can be broadly categorized into two areas, those related to physiological measures and those associated with existing self-report instruments.

Physiological measures provide an objective, real-time assessments and overcome certain inherent biases, such as socially desirable answering pattern [56]. However, the above-mentioned physiological measures are not specific to startle or surprise, but reflect broader autonomic, neural activation, and affective responses [57]. While these techniques offer high temporal resolution, they cannot distinguish between a startle response and similar defensive reactions triggered by fear or stress [58].

In addition, physiological measures are also uneconomical for application to large numbers of participants [59], and some are impractical and invasive to apply in operational settings. The measures obtained are often only meaningful in relation to an individual's own baseline, and were found to be inconsistent with the subjective experience of the responses [4]. For example, in a study testing the effect of surprise on pilots' performance [4],

participants were found to show nearly similar levels of heart rate yet reported significantly different levels of startle and surprise in two conditions. Thus, similar to the literature on experienced challenge and threat [60], validated self-report measures on startle and surprise are necessary to complement physiological measures and contribute to the study of relationships between physiological data, subjective experience and performance.

Self-report measures of surprise offer valuable insight into subjective experiences and can complement physiological assessments [61]. However, there is no information available on the loadings of the individual items within the Surprise subscale, nor is there peer-reviewed reporting on a systematic methodology of items selection in DES-IV or PANAS-X. Furthermore, these scales were not originally developed to assess responses to a specific stimulus, even though the experience of surprise, unlike other affective states such as fatigue or shyness, is inherently stimulus-dependent. The absence of specific eliciting events likely contributes to the low reliability and low mean scores [50], and makes questions like whether participants felt surprised "in general" difficult to interpret in a meaningful way.

Concluding, for measuring startle there has been no systematic attempt to develop and validate a self-report measure. For measuring surprise, items in existing self-report scales were not selected in a systematic manner and scales were not developed nor validated to specific stimuli. Moreover, current approaches remain largely limited to physiological measures, which may be insufficient or impractical in more applied contexts, outside the laboratory. These limitations underscore the need for developing systematic, validated methods that are both sensitive to the characteristics of startle and surprise and also feasible for application in operational settings.

1.4. Research objectives, approach and outline

The overarching aim of this dissertation is to investigate the following main question: **How can startle and surprise be quantified in an aviation operational context?** Three research objectives and five corresponding key questions have been formulated.

Main question **How can startle and surprise be quantified in an aviation operational context?**	
Research objective 1 Conceptualize the cognitive processes underlying pilot decision-making in unexpected situations.	
Key questions	**Chapters**
1. How do different cognitive models represent pilot decision-making and actions in unexpected situations?	Chapter 2
Research objective 2 Investigate the relationships between existing self-report measures of startle and surprise, and personality traits and flight experience.	
2. How do personality traits and flight experience influence pilot cognitive and affective responses to simulated in-flight hazards?	Chapter 3
Research objective 3 Develop and validate psychometrically-sound self-report measures for startle and surprise.	
3. How can startle and surprise be quantified with self-report measures? *This question is addressed through the development of Startle and Surprise Inventories, resulting in the following research questions:*	Chapter 4
4. How valid are the Startle and Surprise Inventories as measures of startle and surprise in an ecologically-valid aviation context?	Chapter 5
5. How well do the Startle and Surprise Inventories predict pilot information-processing performance?	Chapter 6

1.4.1. Research objective 1

The first research objective is to conceptualize and describe cognitive processes underlying pilot decision-making in unexpected situations. This objective addresses Key question 1: How do different cognitive models represent pilot decision-making and actions in unexpected situations? To address this, in **Chapter 2**, pilot decision-making and actions are analysed using two flight safety events, Loganair flight 6780 and US Airways flight 1549, with three cognitive models: the Landman model of startle and surprise [35], the perceptual

cycle model [62] and the three-level situation awareness model [63]. The analysis demonstrates how each model captures different facets of pilot decision-making and actions during unexpected events, offering theoretical insights into the potential cognitive mechanisms underlying startle and surprise. Furthermore, the comparison highlights each models' differential explanatory power and applicability, providing a conceptual basis for developing reliable, validated, and non-obtrusive measures of startle and surprise.

1.4.2. Research objective 2

The second research objective is to investigate how existing self-report measures of startle and surprise, relate to personality traits and flight experience. This leads to the Key question 2: How do personality traits and flight experience influence pilots' cognitive and affective responses to simulated in-flight hazards? To address this, in **Chapter 3**, the relationships between stable characteristics: trait anxiety, decision-related action orientation (AOD), failure-related action orientation (AOF), flight experience (flight hours), and pilot situational responses are investigated. The situational responses include perceived startle, surprise, stress and mental workload during startling and surprising flight scenarios.

The analysis draws on a dataset of 89 airline pilots from four previous studies in startle and surprise, providing a solid empirical foundation. Findings reveal how individual differences among pilots affect their susceptibility to startling and surprising events, underscoring the need for measuring instruments that can capture these responses in operational contexts.

1.4.3. Research objective 3

Building on insights from the first two objectives, the third research objective centres on the development and validation of psychometrically-sound self-report measures for startle and surprise. This objective is guided by three key questions: "How can startle and surprise be quantified with self-report measures?", "How valid are the Startle and Surprise Inventories as measures of startle and surprise in an ecologically-valid aviation context?", and "How well do the Startle and Surprise Inventories predict pilot information-processing performance?".

To answer Key question 3, **Chapter 4** presents the development and psychometric validation of self-report measures for startle and surprise. The question is answered through a systematic process in which the Startle Inventory, Surprise Inventory, as well as the more time-efficient Visual Analogue Scales for Startle and Surprise are developed and preliminarily validated.

Three sequential phases are identified, each designed to progressively refine and validate the measuring instruments. First, an initial item pool is formulated based on a comprehensive review of both fundamental and applied literature, with content validity established through expert evaluation. Second, construct validity is examined using multilevel exploratory factor analysis to identify underlying dimensions of the retained items, resulting in the Startle Inventory and Surprise Inventory. Third, concurrent validity of the Visual Analogue Scales (VAS) for Startle and Surprise is assessed by comparing VAS ratings with inventory scores. Data from 81 participants, each of whom rated nine varied stimuli designed to elicit varying levels of startle and surprise, provide a robust empirical basis for these analyses. Through this three-phase process, reliable and valid self-report measures are developed that are capable of capturing variations in the responses of startle and surprise.

Key question 4 is addressed in **Chapter 5**, in which the construct validity of the Startle

and Surprise Inventories are evaluated in a more ecologically-valid operational context. To this end, the factor structure identified in Chapter 4 is tested using multilevel confirmatory factor analysis, based on data collected from 26 professional pilots exposed to eight flight simulator scenarios varying in levels of startle and surprise.

In **Chapter 6**, Key question 5 is addressed by investigating the criterion-related validity of the Startle and Surprise Inventories. The study examines whether self-report startle and surprise are associated with the pilots' information-processing performance, as indicated by their performance on a secondary auditory task conducted concurrently with in-flight events in the flight simulator. To assess these relationships, linear mixed-effects models were employed, providing empirical support for the ability of the Startle and Surprise Inventories to predict pilot information-processing performance in an operational aviation context.

The development and validation of the Startle and Surprise Inventories follow a structured, iterative process involving three types of validity assessment [64]. The validation starts with the establishment of content validity through expert evaluation. Building upon this foundation, construct validity is examined using multilevel exploratory analysis and multilevel confirmatory factor analysis across different samples and experimental contexts. After establishing a stable factor structure, criterion-related validity is assessed by analysing the relationships between pilots' self-report startle and surprise and their information-processing performance. This systematic, iterative approach ensures the psychometric robustness and operational relevance of the Startle and Surprise Inventories.

1.4.4. Synthesis and outline

Finally, in **Chapter 7** the key findings from Chapters 2 to 6 are synthesized, systematically addressing each research question and reflecting on theoretical and practical implications of quantifying startle and surprise. Limitations of the current research project are identified and recommendations for the future application of the Startle and Surprise Inventories in both research and operational contexts are given.

This dissertation consists of three phrases and is structured into seven chapters (Figure 1.4). Each chapter addresses one of the key questions as stated in Section 1.4. Phase I includes the exploratory phrase of the research, including an initial conceptual analysis to understand the role of startle and surprise (Chapter 2), and a review of existing measures of startle and surprise (Chapter 3). Phase II contains the development phrase of the self-report measuring instruments for startle and surprise (Chapter 4). Phase III includes the construct validation and criterion-related validation of the measuring instruments in an ecological-valid aviation context (Chapters 5 and 6).

1.5. Research scope

This dissertation focuses on the quantification of startle and surprise specifically in the context of civil aviation. While startle and surprise can affect individuals across many safety-critical domains [65–67], this research is confined to the aviation domain, where pilots are responsible for managing unexpected emergencies. The dissertation is further scoped to assess only individual pilot performance, with all experimental tasks performed in single-pilot settings. By focusing on the single-pilot setting, the research engages with the cognitive, emotional, and operational challenges pilots face when they are required to

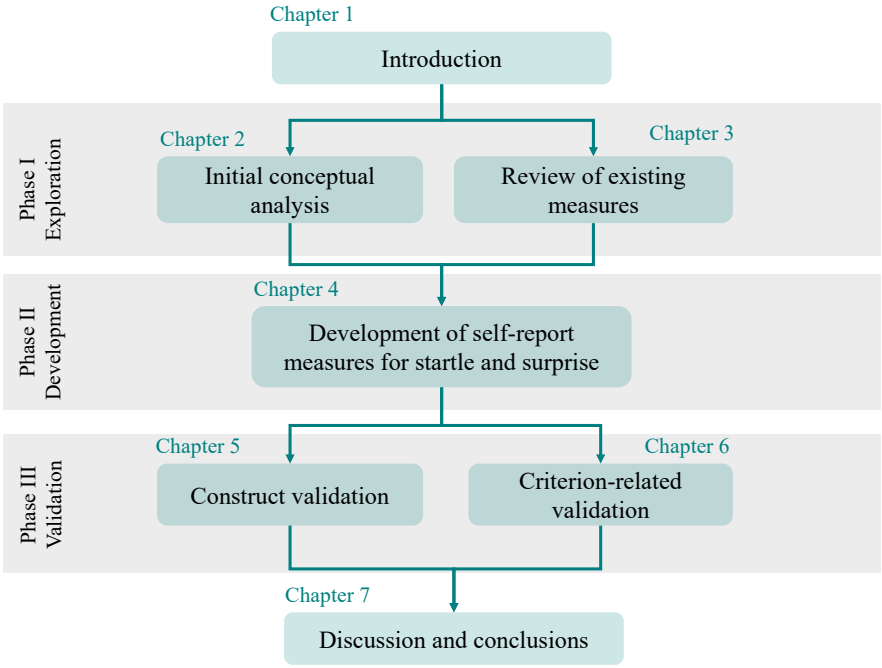


Figure 1.4: Dissertation outline.

handle surprising and startling events in the high-demanding environment all by themselves.

Given the aviation focus, the validation of the Startle and Surprise Inventories is conducted in an ecologically-valid aviation context, rather than in highly-controlled but less realistic laboratory environments. Conducting experiments in a high-fidelity, motion-based flight simulator enhances ecological validity and allows researchers to capture pilot performance under conditions that closely approximate real-world flight demands. This approach provides a more accurate reflection of how startle and surprise manifest in the aviation context and improves the practical applicability of the measures.

This research focuses on the development and validation of self-report measures to capture pilots' subjective experiences of startle and surprise. While physiological [37, 68, 69] and behavioural measures [70] are valuable to capture automatic and observable responses, they provide limited insight into the cognitive and emotional dimensions [71, 72]. Self-report measuring instruments [73] are therefore essential for capturing components such as the appraisal of unexpectedness, emotional intensity, and the lingering cognitive effects of startle that may persist even after physiological responses have subsided.

Furthermore, the validation of the self-report measures in this research does not include direct comparisons with physiological indicators. Although comparisons could contribute to the relationships between subjective experience and physiological responses [71, 74], this aspect lies beyond the scope of the present research. The validation efforts are therefore limited to the psychometric evaluations of self-report measures.

2

Conceptualization of pilot cognitive processes in flight incidents

To conceptualize the cognitive processes underlying pilot decision-making in unexpected situations, pilots' decision-making and actions in two real-world incidents, Loganair Flight 6780 and US Airways Flight 1549 are examined. By analysing pilots' decision-making and actions through the lens of the model of startle and surprise, the perceptual cycle model, and the three-level situation awareness model, this chapter aims to evaluate the practical strengths and limitations of each cognitive model. In doing so, it sheds light on how frame/schema activation/modification, stress, and flight experience influence pilot behaviour in high-stakes environments.

The chapter is structured as follows: Section 2.1 provides an overview of the three cognitive models. Section 2.2 and Section 2.3 present case analyses of the Loganair Flight 6780 and US Airways Flight 1549, respectively. A summary of the models and implications for pilot training are discussed in Section 2.4.

2.1. Introduction

To better understand the cognitive processing underlying pilot decision-making and actions that may contribute to or prevent Loss Of Control In-flight (LOC-I), several cognitive models have been effectively applied to this domain due to their relevance to human cognitive performance, each with their own emphasis and target applications.

This chapter focuses on three cognitive models, the Landman model of startle and surprise [35], the perceptual cycle model [62] and the three-level situation awareness model [63]. We explore their applicability to two real-world aviation incidents, to illustrate how these frameworks capture different facets of pilot cognition and decision-making. These models have been widely employed in the literature to analyse pilot decision-making and actions, including the identification and interpretation of cognitive errors [35, 75, 76].

It is worth noting that all three models contain a similar cognitive structure called *frame* or *schema* (or the plural form as *schemata*). The *frame* in the model of startle and surprise is presented as:

> *"an explanatory structure that defines entities (data) by describing their relationship to each other"* (Pages 118-120 [77])

While, the definition of *schema* in the perceptual cycle model is considered as:

> *"an organized mental pattern of thoughts or behaviors to help organize world knowledge"* (Page 230 [62])

From the perspective of obtaining and maintaining situation awareness, *schemata* in the three-level situation awareness model are defined as:

> *"coherent frameworks for understanding information, encompassing highly complex system components, states, and functioning"* (Page 317 [78])

In general, *frame* synthesizes concepts from *schemata*, mental models, scripts, and other types of knowledge structures in long-term memory. These structures represent both generic and context-specific situations, including how things work, how events are sequenced, and indicate which actions and behaviour are appropriate. In the model of startle and surprise, *frames* modification (i.e., re-framing) is triggered by the mismatch between expectations and environment cues [11] and are further shaped by the sensemaking process [79–81]. The three-level situation awareness model emphasizes *schema* activation based on pattern recognition. The perceptual cycle model does not explicitly account for any mechanism underlying *schema* modification. Nevertheless, *frame* and *schema* share the following characteristics in the analysis of pilot decision-making and actions:

1. Cognitive structures guide how individuals perceive information.
2. Cognitive structures influence how individuals interpret and integrate information.
3. Cognitive structures can be activated through pattern-matching between environmental cues and elements within an internal model (e.g., connecting data with a frame).
4. Cognitive structures can be activated through situation comprehension.
5. Cognitive structures are formed through past experiences and prior knowledge.
6. Cognitive structure are dynamic and can be expanded or modified over time.

This chapter aims to evaluate the potential merits and limitations of applying cognitive models to pilot decision-making and actions in aviation operational contexts, and to outline potential implications for pilot training.

2.1.1. Model of startle and surprise

The Landman model of startle and surprise [35] conceptualizes pilot performance as a dynamic process influenced by physiological startle and cognitive surprise, and the guidance of cognitive frames. As shown in Figure 2.1, the model emphasizes the distinction between startle and surprise, concepts which are often used interchangeably in aviation literature [19]. The model draws upon the data-frame theory of sensemaking [77] and the cognitive-psychoevolutionary model of surprise [30].

Frames can be represented in various meaningful forms, such as stories, maps or organizational diagrams, guiding both sequential and parallel cognitive processes. The model of startle and surprise poses that the information processing is directed by active frames. When an hypothesis derived from active frames aligns with perceived information, highly automatic and low-effort information processing will be triggered (i.e., fast appraisal). However, if a mismatch arises that exceeds a relevance threshold [30, 82], the individual experiences surprise, prompting the need to update or modify the active frame (i.e., reframing).

The reframing process is cognitively demanding and is characterized by switching from a fast, automatic appraisal loop to a slower, more deliberate processing loop. The slow appraisal involves sensemaking activities [77] and knowledge-based reasoning [83] to interpret the cause of the mismatch between the observed data and expectations. In addition to resolving inconsistencies, slow appraisal enables anticipation of potential challenges,

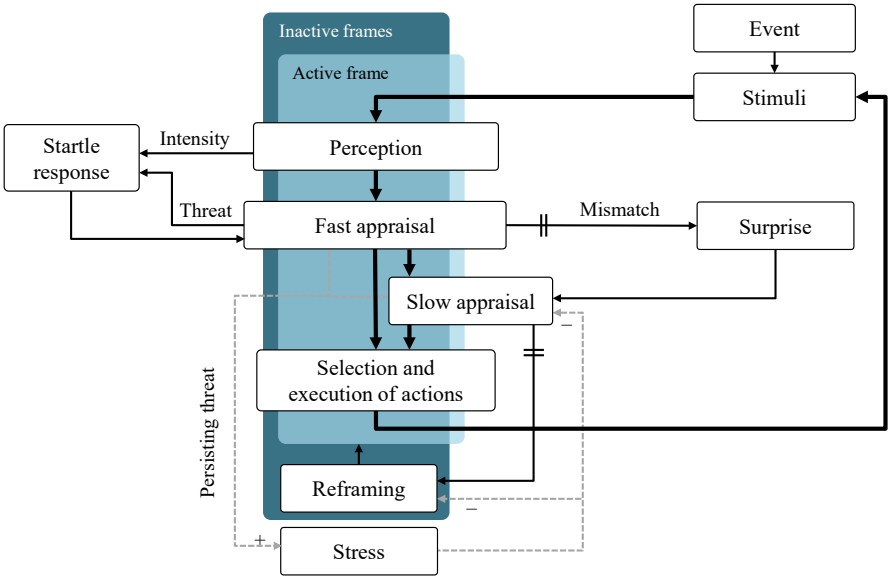


Figure 2.1: Landman model of startle and surprise [35].

supporting decision-making and effective actions. In the model, frames are positioned "behind" the perception-appraisal-action cycle rather than being embedded within it, signifying that frames influence the cycle to varying extents. When reframing is initiated, previously inactive frames are activated as hypothetical or tentative structures, which need to be tested in the slow appraisal loop.

The startle response typically includes the startle reflex, which is a physical reflex to intense stimuli [19], and the appraisal of the potential threatening stimulus [84]. These processes are mediated by two distinct neural pathways, the subcortical and cortical routes [85]. The subcortical pathway is rapid and automatic, transmitting sensory information unconsciously from the thalamus to the amygdala. This route is also responsible for initiating the "fight or flight" response, a rapid hormonal and autonomic response to potential danger [23]. In contrast, the cortical pathway, often referred to as the slow pathway, involves more complex processing within the hippocampal and cortical circuits. It requires additional synaptic transmission and conscious awareness, enabling a more refined appraisal of the stimulus, albeit with a longer latency [86]. Moreover, the intensity of startle response can be heightened under condition of fear or stress, known as fear-potentiated startle [23].

2.1.2. Perceptual cycle model

The perceptual cycle model (PCM) is a cognitive and decision-making model [62], grounded in the schema theory [78] and interpretations of information-processing mechanisms. It comprises three core elements, schema, exploration and world, which interact within both an inner perceptual cycle circle and an outer exploratory circle, as illustrated in Figure 2.2. The model proposes that information processing operates in a cyclical manner, where existing schemata guide exploration, exploration samples information from the environment, and the information in turn modifies the schemata. The PCM highlights the dynamic, reciprocal relationships between the external environment and internal cognitive structures.

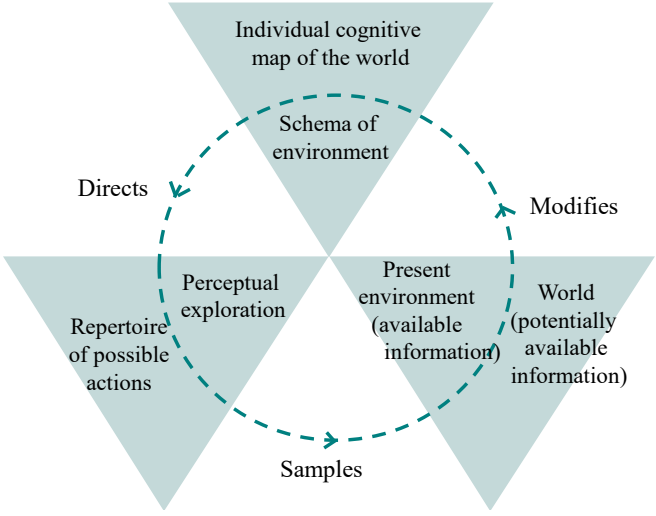


Figure 2.2: Perceptual cycle model (adapted from [62]).

In the inner perceptual cycle, environmental information can trigger the activation of the schemata, a process typically understood as bottom-up. In contrast to the model of startle and surprise, the PCM was not designed to account for responses to unexpected events, and thus does not explicitly distinguish between surprising/startling and routine situations. The PCM mentions automatic, continuous adjustments of schemata in response to incoming information, without the explicit involvement of surprise. This also contrasts with the model of startle and surprise, which centres on cognitive disruption caused by stimuli. In the PCM, schemata exert a top-down influence on perceptual exploration [87], guiding how individuals direct attention, interpret environmental cues, and integrate new information. These perceptual actions, in turn, shape the interaction between the individual and the environment. The information gathered through exploration is treated as perceptual samples that contribute to the construction of situational understanding.

The model's outer cycle represents a more general exploratory process, involving a broader repertoire of actions aimed at acquiring information not immediately available in the environment. This process draws on the individual's broader cognitive knowledge and understanding of the external world and its possibilities [88].

2.1.3. Three-level situation awareness model

The three-level situation awareness (SA) model, originally proposed for dynamic decision-making contexts [63], conceptualizes SA as an information-processing framework (Figure 2.3). In this model, situation awareness is defined as a cognitive state of knowledge achieved through a hierarchical process at three levels. First, the perception of relevant elements' status, attributes, and dynamics (Level 1 SA). Second, the comprehension of the situation based on this perceived information (Level 2 SA). Third, the projection of future status and behaviour of the elements in the environment (Level 3 SA).

Salient environmental characteristics are initially processed in parallel through pre-attentive processing, which provides cues for attention allocation [89] and serves as the foundation for Level 1 SA. Attention allocation is essential not only for accurate perception and comprehension of information, but also supporting later decision-making and action execution. However, the capacity for attentive processing is inherently limited, constraining an individual's ability to manage complex decisions or multitask effectively in dynamic environments. To mitigate potential overload from attention demands, strategies such as information sampling [90], physiological arousal mechanisms [91], and automaticity [92] have been suggested.

In the three-level SA model, schemata function as essential long-term memory structures that can be activated by salient environmental cues, guiding the selection of information to be perceived. Scripts, derived from schemata, represent sequenced, schema-based patterns of appropriate decision-making and actions [93]. These scripts allow individuals to respond automatically to familiar situations, eliminating the need for deliberate action planning at every iteration of the cognitive cycle. Schemata and scripts are dynamic and can be updated based on feedback from actions, provided that the feedback is accurately perceived and interpreted. The core processes of situation awareness all occur within working memory, encompassing the perception of environmental information, comprehension of its integrated meaning, projection of future states of relevant elements.

Building on the three-level situation awareness model, SA could be influenced by a

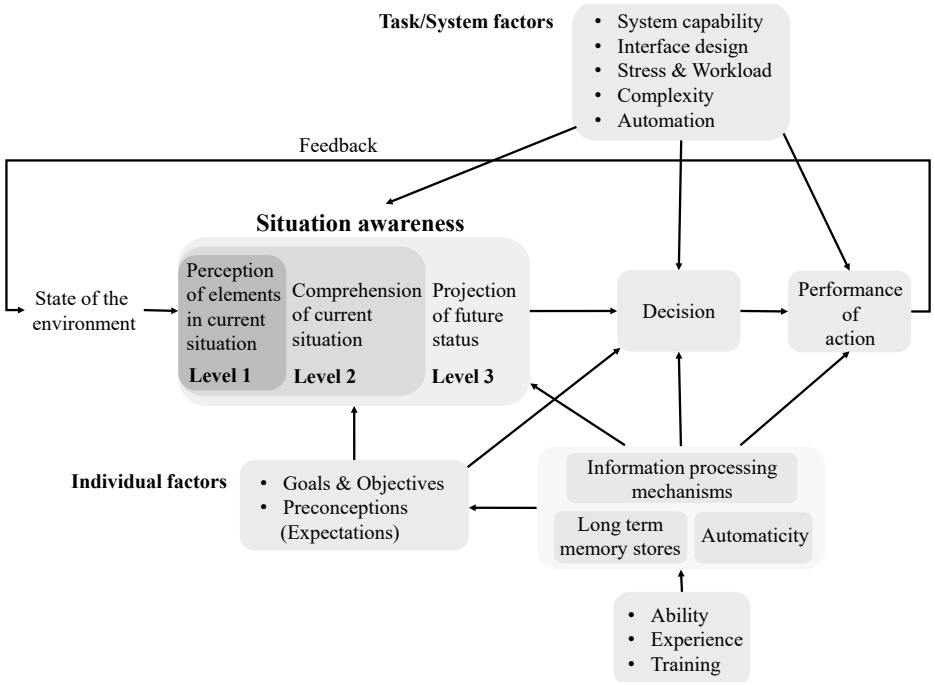


Figure 2.3: Three-level situation awareness model (adapted from [63]).

complex combination of individual, system, and task-related factors. At the individual level, attributes such as prior experience, training, cognitive capacity, and goals significantly shape the effectiveness of one's ability to perceive, comprehend, and project information. System design elements, such as interface design, automation level, information accessibility, and workload, play a critical role to support or hinder SA. Additionally, task-related factors, such as stress, time pressure, and task complexity impose cognitive burdens that compromise SA, particularly in high-stakes environments.

To obtain more insights into the differences between these models and their potential application in analysing pilots' decision-making and actions in startling and surprising situations, the three models were applied to examine two real-world aviation incidents, Loganair flight 6780 [94] and US Airways flight 1549 [95]. These two incidents were selected because they involve distinct cognitive processes, and in both cases, the flights were successfully recovered by the pilots.

The analyses in the following sections are based on official investigation reports. Some inferences concerning the pilots' internal cognitive or affective states are necessarily interpretive. These interpretations are guided by the introduced theoretical frameworks, and are intended as plausible reconstructions rather than psychological conclusions.

2.2. Loganair Flight 6780

2.2.1. Synopsis

On 15th December 2014, a Saab AB 2000 was inbound to land on Runway 27 at Sumburgh Airport following a routine flight from Aberdeen. During the approach phase at an altitude of 2,000 ft, the aircraft was struck by lightning (indicated by the vertical dashed line on the left in Figure 2.4). In responses, the captain aborted the approach, and initiated a climb using nose-up pitch trim and manual pitch inputs. However, during the climb, the captain noticed that his pitch control inputs did not produce the expected aircraft response. Upon reaching 4,000 ft, the pitch attitude suddenly shifted toward a nose-down orientation, initiating a rapid descent with a peak rate of 9,500 ft/min. Although the captain applied maximum manual control input in an attempt to recover due to the control forces he felt were higher than normal, the airplane continued to pitch nose-down and descend. The descent was finally stopped, and the aircraft began to climb again when it reached just 1,100 ft above sea level.

It turns out that the unique design of the Saab 2000's autopilot system prevented both the lightning strike and the captain's manual pitch inputs from disengaging the autopilot. It was later determined that the abnormal control forces experienced by the captain were caused by the autopilot continuously counteracting his inputs. At the time of the lightning strike, the autopilot was engaged in both heading select and altitude hold modes, which commanded the pitch trim to maintain the preselected altitude of 2,000 ft. The autopilot remained engaged until the aircraft descended to 1,100 ft, at which point an unintended Air Data Computer (ADC) fault triggered autopilot disengagement. Only then did the captain's nose-up pitch trim inputs take effect, allowing the aircraft to recover.

At the time of the incident, the captain had accumulated 4,640 flight hours on the Saab 340 and only 143 flight hours on the Saab 2000, a stretched derivative of the Saab 340. Despite their visual similarity, this generation upgrade from the Saab 340 to the Saab 2000 introduced substantial changes in aircraft systems and performance. Notably, one of the most

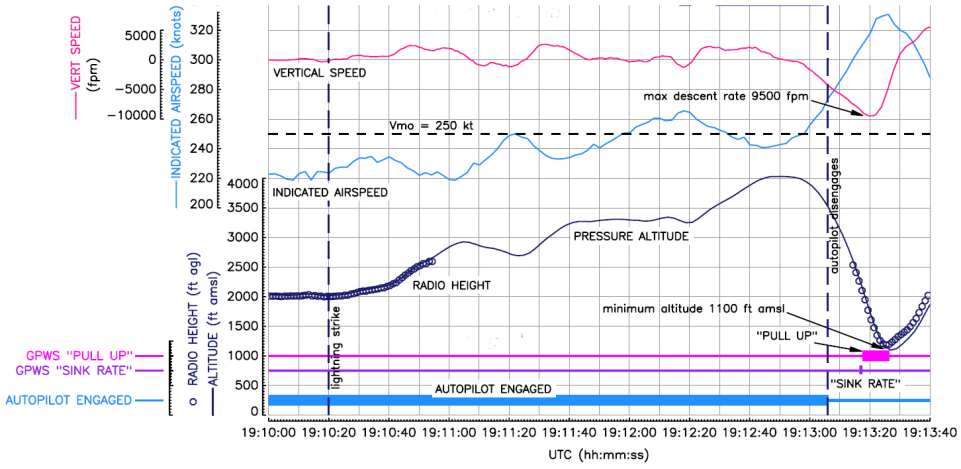


Figure 2.4: Relevant flight data from Loganair 6780 (adapted from [94]).

critical differences is that, in the Saab 2000, neither force applied to the control column nor pitch trim inputs can override the autopilot. During previous training on the Saab 340, the captain had completed a Line-Oriented Flying Training (LOFT) exercise involving a simulated lightning strike that triggered a temporary loss of engine-driven generators, resulting in automatic autopilot disengagement. However, after transitioning to the Saab 2000, the captain had neither encountered an actual lightning strike nor received any training scenarios involving such an event in the context of the new model.

2.2.2. Thematic analysis

Model of startle and surprise

The Loganair Flight 6780 incident illustrates how high levels stress could impair cognitive performance, which leads to cognitive tunnelling [96], inattentional deafness [97] and disruptions in the reframing process [98]. These effects are consistent with research indicating that evaluated stress narrows the range of perceptual cues, resulting in a constricted attentional field [10, 90, 99–101].

From the perspective of the model of startle and surprise, the simulated lighting strike training for the Saab 340 may have contributed to a frame on lightning strike effects that is not applicable to the Saab 2000. The captain might assume that a lightning strike would:

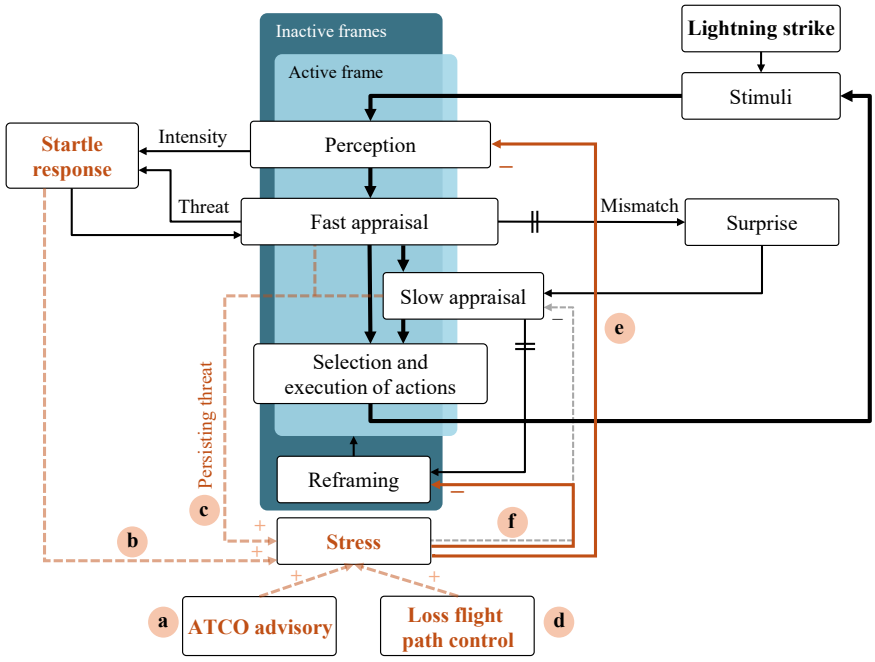


Figure 2.5: Model of startle and surprise (adapted from [35]) annotated to illustrate how stress may impair perception and reframing processes in Loganair 6780 (letters refer to processes associated with the incident, as described in the text).

1) lead to autopilot disconnection, as he experienced during the Saab 340 training, and 2) potentially cause system damage, an expectation which was reinforced by the abnormal control forces. This incorrect frame could have been corrected (i.e., reframe) by cross-checking cockpit instruments to assess the actual status of the aircraft. In this case, a visual "mistrim" caution message on the Primary Flight Display (PFD), accompanied by single aural chimes, indicating that the autopilot remained engaged while the pilot was exerting manual control in opposition to the autopilot's pitch servo. Ideally, such cues should have prompted the pilot to reconsider the situation.

However, if the captain's active frame assumed that the autopilot was indeed disconnected, there may have been no impetus to allocate additional cognitive resources to reading information on the PFD [90], leading to the inadvertent filtering of conflicting cues. Furthermore, if the captain believed that the flight path was unmanageable due to presumed system damage, his attention was likely to be directed toward resolving the perceived issue rather than reassessing the autopilot engagement.

The high level of stress may have further impaired the captain's ability to reframe the situation. During the initial approach, the air traffic controller (ATCO) informed the flight crew that the antenna for ATIS[1] at Sumburgh had been struck by lightning. This message may have reinforced preconceptions that lightning could be encountered, thereby increasing the stress levels. According to the flight crew's recollection, the captain reported that his briefing following the communication with ATCO included considerations for dealing with a potential lightning strike. While the briefing probably increased the crew's preparedness, it may also have elevated their stress levels (Number ⓐ in Figure 2.5).

The lightning strike might have triggered a startle response due to both its sudden loudness and brightness. The fear-potentiated startle may have been exacerbated by pre-existing stress related to the weather conditions. The intensified startle reaction (Number ⓑ), combined with the perceived threat of the lightning strike (Number ⓒ), might in turn contribute to a heightened stress level [23]. Additionally, the aircraft's lack of responsiveness to control inputs, resulting in the inability to manage the flight path, may further elevate the captain's stress (Number ⓓ).

Taken together, stress arising from the ATCO advisory, the lightning-induced startle, the perceived threat, and the loss of flight path control might have significantly impaired the perception of display cues and aural alerts (Number ⓔ) and the reframing process (Number ⓕ). In summary, the application of the model of startle and surprise on the Loganair 6780 case could suggest that the cognitive processes involved in responding to the startling stimulus are highly susceptible to the effects of stress, increasing the likelihood of cognitive failures and impaired decision-making.

Perceptual cycle model

From the perspective of the perceptual cycle model, Figure 2.6 illustrates the possible logical causal relationships between the captain's actions and environment inputs in the Loganair 6780. The primary cause of the incident was the captain's misjudgment regarding the engagement status of the autopilot system. This misjudgment likely resulted from a discrepancy between the encountered situation and the captain's prior training and experiences. The captain appeared to have initially activated an incorrect schema related to the autopilot

[1] ATIS stands for Automatic Terminal Information Service.

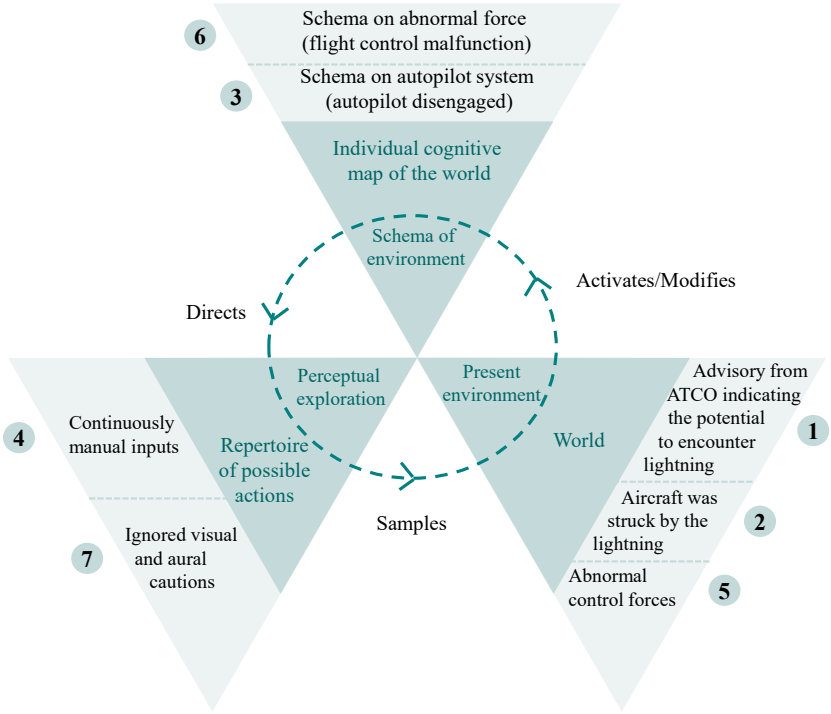


Figure 2.6: Perceptual cycle model (adapted from [62]) annotated to show inferred causal relationships between the captain's actions and environment inputs in Loganair 6780 (Arabic numerals refer to processes associated with the incident, as described in the text).

system through a data-driven pattern-matching process, likely triggered by the ATCO's preliminary warning about the weather conditions, and reinforced by the subsequent lightning strike (Steps ① to ③ in Figure 2.6). During his Saab 340 training, the captain had developed an understanding of how lightning strikes typically affected aircraft systems. When the lightning strike occurred during the actual flight in a Saab 2000, he may have assumed that the autopilot system had disconnected, reflecting expectations shaped by his prior training experience. However, the real-world situation diverged from his prior experiences. Although the aircraft was struck by the lightning, it had no impact on aircraft performance, and the autopilot remained engaged. This resulting situation, partly novel due to the unexpected autopilot behaviour and partly familiar based on prior training and experiences, may lead the captain to activate an inaccurate schema regarding the autopilot's status. The captain's inaccurate schema subsequently resulted in inappropriate decision-making and actions, as illustrated from Steps ③ to ④ in Figure 2.6. Believing incorrectly that the autopilot had disengaged, the captain manually intervened with flight controls (Step ④ in Figure 2.6) and then perceived abnormal forces from the control column (Step ⑤ in Figure 2.6).

The captain's limited experience with the Saab 2000 could prompt him to activate an inaccurate schema regarding the abnormal forces felt through the control column (Steps ⑤ to ⑥ in Figure 2.6). He misinterpreted these abnormal forces as indicative of a flight control

malfunction triggered by the lightning strike. In reality, the unnatural forces were due to the autopilot actively counteracting his manual inputs. The captain demonstrated a tendency to selectively seek information that confirmed his expectations, known as confirmation bias [102], and to be unresponsive to evidence contradicting his established beliefs (i.e., anchoring heuristic [103, 104]). In this case, the captain maintained a preconceived notion that the lightning strike had compromised the aircraft, causing the autopilot to disengage and result in flight control malfunctions. Thus, the schema concerning the flight control malfunction could misdirect his perceptual exploration, causing him to overlook critical visual and auditory warning cues (Steps ⑥ to ⑦ in Figure 2.6).

Applying the perceptual cycle model to this incident highlights its explanatory strengths and limitations in the context of human-environment interactions under complex operational conditions. The model effectively captures the causal relationships between environment cues and schema activation, which in turn shapes perception and guides exploratory behaviour. However, a key limitation lies in its lack of an explicit account for cognitive and affective responses, particularly acute stress triggered by startle. Such responses could transiently impair cognitive functioning, disrupt schema-guided perception, and hinder effective interaction with the environment in high-stakes, critical situations.

Three-level situation awareness model

Figure 2.7 illustrates the way to interpret the Loganair 6780 with the three-level situation awareness model. The cues from the lightning strike together with the advisory from ATCO were possibly matched to the schema, in which a lightning strike could cause system failures and lead to autopilot disengagement (Number Ⓘ in Figure 2.7). The schema, likely established during the captain's Saab 340 training, may have been retrieved from long-term memory. The captain's incorrect interpretation of the environmental information appeared to have been influenced by the activated schema and its associated scripts. The captain responded with a manual nose-up pitch input (Numbers Ⓘ Ⓘ in Figure 2.7). He also allocated a disproportionate share of attentional resources to executing control actions. This narrowed attentional focus limited his capacity to perceive other critical information, such as cues from the mode control panel which could have corrected his misjudgment regarding the autopilot's status. However, the captain's pitch control input did not produce the expected aircraft response, and the aircraft continued to descend (Numbers Ⓘ Ⓘ Ⓘ in Figure 2.7).

Unlike the other two models, the three-level situation awareness model offers a valuable lens for analysing pilots' decision-making from the perspective of SA and attention distribution. Incomplete perception and inaccurate comprehension of real-time information could undermine SA, hindering the captain's ability to form an accurate mental representation of the system state. When the captain encountered abnormal control forces through the control column, he misinterpreted it as flight control malfunctions due to the lightning strike. The lack of expected feedback upon this input from the aircraft further reinforced his belief that the lightning strike had severely affected the aircraft and manual intervention was required, deepening the mismatch between his mental model and the actual system state.

Given the limited capacity of attention resources, the captain allocated more attention to manual control issues that were driven by stimuli [98]. As a result, less attention was directed toward task-relevant cues, such as visual and auditory warnings on autopilot status presented in the cockpit. This over-allocation of attention to aircraft attitude further degraded

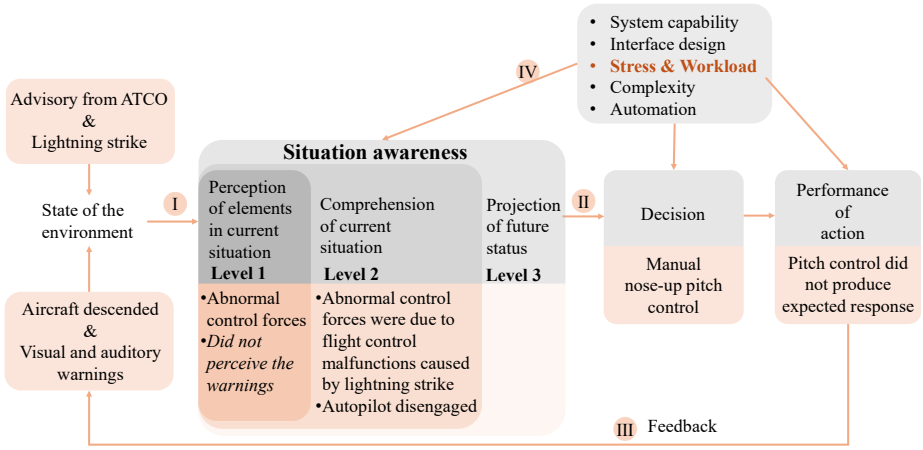


Figure 2.7: Three-level situation awareness model (adapted from [63]) annotated to illustrate inferred situation awareness in Loganair Flight 6780 (Roman numerals refer to processes associated with the incident, as described in the text).

the captain's overall situation awareness.

Stress, potentially triggered by the captain's uncertainty or a startle response (Numbers (IV) in Figure 2.7), may have disrupted the process of building and maintaining situation awareness, impairing attentional control [98, 101], reducing working memory capacity [36, 105], and interfering with effective pattern-matching for schema selection.

Notably, the mechanisms of schema modification are not explicitly represented within the structure of the three-level situation awareness model. While the model of startle and surprise suggests that stress can hinder interaction within the environment by obstructing information perception and schema modification (i.e., reframing).

2.3. US Airways Flight 1549

2.3.1. Synopsis

On January 15 2009, US Airways Flight 1549, operated by an A320, departed from LaGuardia Airport (LGA) and experienced an almost complete loss of thrust in both engines after encountering a flock of birds. The aircraft was subsequently ditched into the Hudson River. The bird strike was accompanied by audible thumps and airframe vibrations.

Approximately 12 seconds after the strike, upon realizing the loss of both engines, the captain assumed control of the aircraft, stating "*my aircraft*" to which the first officer responded, "*your aircraft*". The first officer initiated the Engine Dual Failure Checklist and attempted to restart the engines, following the captain's instruction to "*get the QRH*[2] *loss of thrust on both engines*". Simultaneously, the captain communicated with LaGuardia Departure Control to identify a feasible landing site. The LaGuardia Departure controller stated, "*if we can get it for you do you want to try to land runway one three [in Teterboro*

[2]QRH stands for Quick Reference Handbook.

Figure 2.8: Passengers standing on the wings of US Airways Flight 1549 after its emergency landing in the Hudson River. Source: [106].

Airport (TEB)]" to which the captain responded, "*we're unable. We may end up in the Hudson*". After evaluating the situations, the captain ultimately chose the water landing, which was successfully executed (Figure 2.8).

At the time of the incident, the captain had accumulated 19,663 flight hours, including 4,765 hours on the A320, and also had prior military flying experience. His flight simulator training included dual-engine failure scenarios, however, these did not encompass failures resulting from bird strikes. Standard dual-engine failure training for the A320 typically occurs at cruising altitudes around 25,000 feet, in accordance with Airbus recommendations and industry norms. At such altitudes, pilots have sufficient time to assess the situation, determine a landing location, and complete required checklists. Under these training conditions, pilots are also able to maintain the aircraft at the optimum relight airspeed of approximately 300 knots before deciding on an emergency landing strategy. Additionally, in all training scenarios, it was assumed that at least one engine could be successfully restarted, meaning that pilots never reached the point of having to execute a forced landing or ditching. While the flight crew received theoretical instruction on ditching procedures during ground school, practical ditching scenarios were not included in the simulator training curriculum.

2.3.2. Thematic analysis

Model of startle and surprise

US Airways Flight 1549 may serve as a compelling example of how extensive experience and training influence the selection and execution of decision-making and actions guided by an adaptive frame under startling and surprising situations (see Figure 2.9). After observing a flock of birds through the windshield, the captain immediately called out "*Birds*" and noticed a change in engine noise and frequency. The crew was likely startled by the bird strike, as suggested by the first officer's spontaneous remark, "*uh oh*". However, the captain's subsequent actions suggest a rapid recovery from the initial startle and surprise responses, likely supported by his previous experience. He appeared to shift attention to the Engine Indicating System and to initiate perceptual and cognitive appraisal processes to assess the situation. Through more deliberate, reflective appraisal, he subsequently recognized that both engines had lost thrust resulting from the bird strike.

The unexpected dual-engine failure caused by a bird strike likely elicited a surprise response, prompting the captain to reframe the situation, from managing a routine climb in a normally functioning aircraft to confronting a dual-engine failure situation (Number ⓐ in Figure 2.9). This cognitive shift of frame may have facilitated the captain's rapid assessment of the emergent scenario, his communication of feasible landing options with the LaGuardia departure controller, and his coordination with the first officer in executing the Engine Dual Failure Checklist. The potential alternatives included returning to LaGuardia ("...*Cactus*

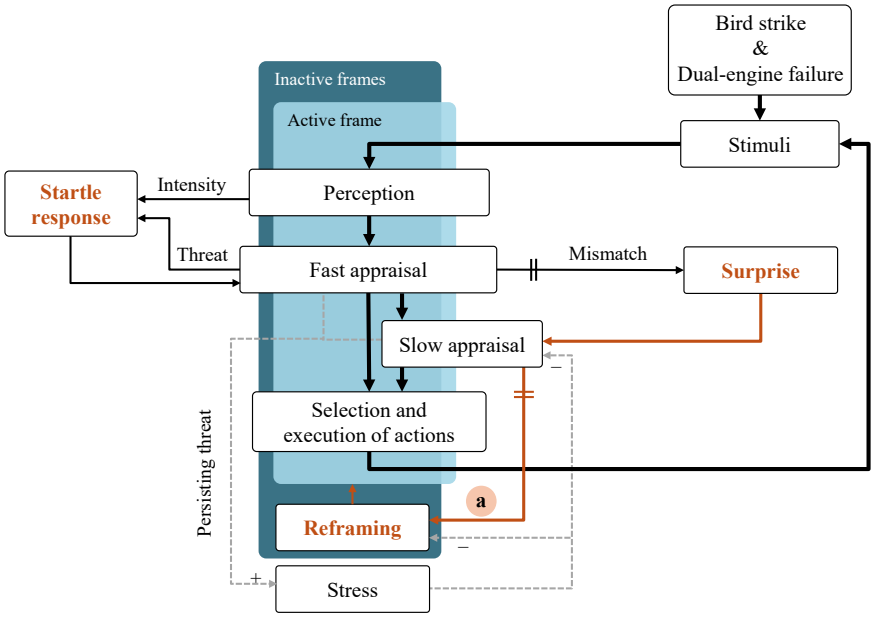


Figure 2.9: Model of startle and surprise (adapted from [35]) annotated to illustrate potential reframing process in US Airways Flight 1549 (letter refers to process associated with the incident, as described in the text).

fifteen thirty-nine, hit birds. We've lost thrust in both engines. We're turning back towards LaGuardia"), diverting to Teterboro Airport, (a response to "*you wanna try and go to Teterboro?*"), and ditching in the Hudson River ("*we're gonna be in the Hudson*"). After evaluating the aircraft's position, heading, altitude, and sink rate, the captain determined that ditching in the Hudson to be the safest course of action.

The bird strike and sudden loss of thrust in both engines at low altitude likely elicited startle and surprise responses in the captain, given the abruptness and severity of the event. Combined with intense time pressure, these factors created conditions typically associated with elevated stress levels, which may have disrupted with cognitive processes such as slow appraisal and reframing. However, the captain's extensive experience and apparent sense of control over the situation may have mitigated the cognitive impairments typically associated with high-stress conditions [107]. Notably, the captain's substantial flight experience, comprising 4,765 hours as pilot-in-command in A320 and military experience, combined with training on dual-engine failures and ditching scenarios, likely contributed to his ability to assess the situation and to take appropriate actions within an adaptive frame. The captain also explicitly acknowledged the role of experience in his performance, stating: "*...I've been making small regular deposits in this bank of experience: education and training, and on January 15, the balance was sufficient so that I could make a very large withdrawal.*" [108].

The case of US Airways Flight 1549 illustrates that the reframing process, as outlined in the Landman model of startle and surprise, could be triggered by sudden and unexpected events. More importantly, it demonstrates that flight experience and training could significantly influence decision-making and actions, enabling the individual to make sense of the situation and respond effectively under extreme time pressure.

Perceptual cycle model

Figure 2.10 indicates the interaction between captain's exploration and the environment, as represented within the framework of the perceptual cycle model. This interaction was likely guided by a schema related to the dual-engine failure. After the pilots noticed the birds strike, a noticeable reduction in engine noise and frequency, and engine-related numerical changes on the dashboard, the captain quickly identified the situation as a dual-engine failure (Steps ① to ③ in Figure 2.10).

Although the captain had received dual-engine failure training in A320 simulators, the context of the real-world event differed from the training scenarios. In this case, the dual-engine failure caused by the bird strike occurred at an altitude of only 2,818 ft, considerably lower than the 25,000 ft typically applied in a simulator training. After the bird ingestion, the aircraft's maximum airspeed was 214 knots, making it impossible to maintain the prescribed 300 knots or to strictly follow the procedures in the QRH.

Rather than adhering mechanically to the checklist, the captain appears to have integrated his existing schema with real-time situation assessment. He initially followed prescribed procedures by completing engine ignition (Step ④) and activating the APU[3] (Step ⑤). However, he then bypassed the procedural step of achieving optimal airspeed and instead immediately coordinated with ATCO to identify a viable landing location (Steps ⑥ to ⑪). When it became evident that an emergency landing on the Hudson River was the only viable option, the captain appears to have drawn upon a ditching schema for situations without

[3] APU stands for Auxiliary Power Unit.

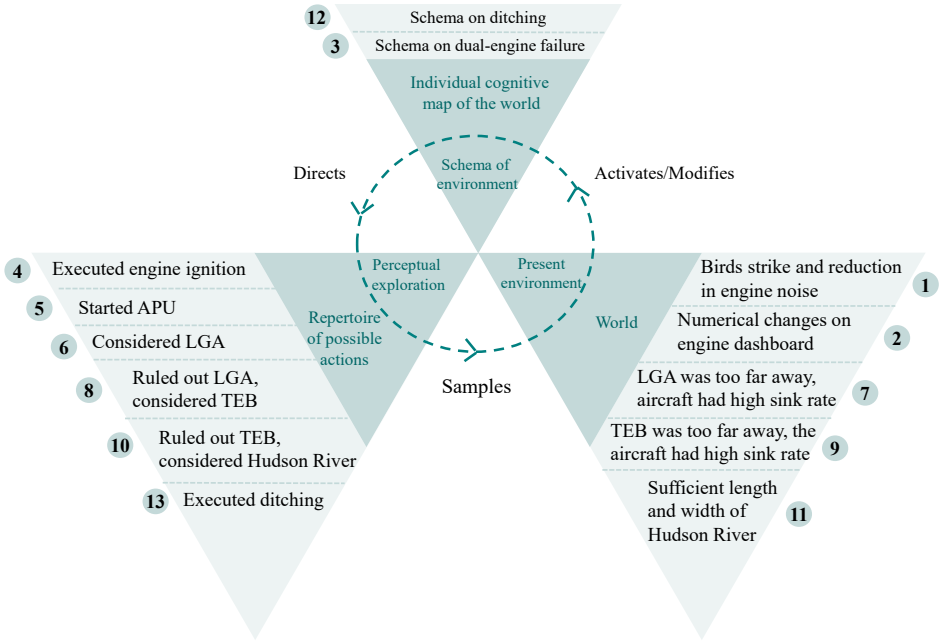


Figure 2.10: Perceptual cycle model (adapted from [62]) annotated to illustrate inferred interactions between the captain's exploration and environment inputs in US Airways Flight 1549 (Arabic numerals refer to processes associated with the incident, as described in the text).

engine power (Steps (11) to (12)). This schematic knowledge may have supported his ability to carry out the water landing (Steps (12) to (13)).

This case underscores a limitation of the perceptual cycle model, which does not explicitly account for factors that may influence schema activation. For example, in the Loganair incident, the pilot's limited experience and training on the specific aircraft type may have led to the activation of an inaccurate schema and the execution of inappropriate actions. In contrast, the captain of US Airways Flight 1549, drawing on his extensive experience, was likely able to rapidly activate accurate schemata and make practical, adaptive decisions beyond the constraints of standard procedures.

Three-level situation awareness model

From the perspective of situation awareness, the US Airways 1549 could underscore the importance of maintaining a high level of situation awareness for making effective decisions in startling and surprising situations (Figure 2.11). An accurate schema may have been rapidly activated after the captain perceived and categorized critical environmental cues into a coherent mental representation, such as a flock of birds visible through the windshield and a noticeable decrease in engine noise and frequency (Number Ⓘ in Figure 2.11).

Guided by the activated schema, the captain appeared to focus selectively on relevant information from the engine instruments and the PFD. This perception of key system information likely supported Level 1 SA. The captain appeared to integrate these perceived data

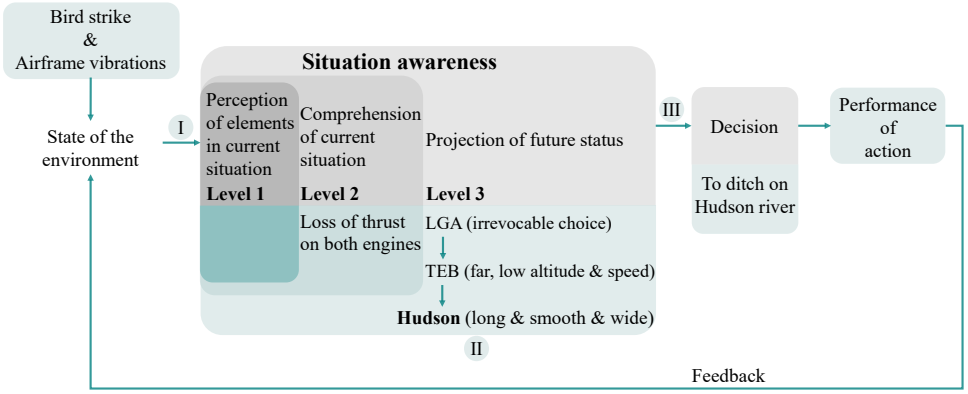


Figure 2.11: Three-level situation awareness model (adapted from [63]) annotated to show inferred projection of landing strategies in US Airways Flight 1549 (Roman numerals refer to processes associated with the incident, as described in the text).

into a comprehensive picture of the aircraft condition, reflecting Level 2 SA, including the knowledge of the aircraft's position, altitude, airspeed and engine status. This understanding, possibly shaped by the activated dual-engine failure schema, supported the recognition that the standard checklist could not be applied under the given conditions.

To determine the most appropriate landing site, the captain appears to have projected the aircraft's future state under different landing strategies (Level 3 SA). This projection was likely informed by his perception and comprehension of environmental information (Number (II) in Figure 2.11). He considered three potential options, each associated with distinct risks. Returning to LGA was deemed unfeasible due to the aircraft's current sink rate. The captain assumed that he would not be able to line up the aircraft with the runway or might risk collapsing the landing gear, potentially resulting in a post-crash fire. Similarly, a forced landing at TEB was dismissed, as the aircraft's altitude and airspeed were too low to land safely. In contrast, the Hudson River appeared to be a viable option. Its length, width, and smooth surface offered a realistic chance for a successful water landing. After rapidly projecting the potential outcomes of each options within the limited time available, the captain chose and executed the ditching on the Hudson River (Number (III) in Figure 2.9).

In addition to guiding the formation of situation awareness in the dynamic environment, the activated schema may have activated relevant scripts. These scripts provided a framework for action, including taking over the aircraft, contacting the ATCO, coordinating with the first officer, and performing the ditching. The captain's extensive experience and training likely supported the integration and comprehension of information, particularly under conditions of extreme time pressure and stress. Through flight training and accumulating experience, he may have developed a repertoire of schemata and associated scripts for a wide range of potential scenarios. This knowledge base, combined with a strong pattern-matching ability, may have enabled him to recognize and respond to unfolding events with efficiency.

Overall, a high level situation awareness may have enabled the captain to comprehend information in an integrated form and accurately project the aircraft's future state under different strategy, probably guided by an activated schema and associated scripts. His

extensive experience and training may have further supported this process by facilitating rapid pattern-matching and information integration under intense time pressure [109, 110].

2.4. Discussion

The captain's decision-making and actions in the incidents involving Loganair Flight 6780 and US Airways Flight 1549 have been analysed through the lens of three cognitive models. Each model is rooted in broader theories of cognitive functioning, yet emphasizes distinct aspects. The model of startle and surprise incorporates the effects of startle and surprise into the perception–appraisal–action loop. The model particularly describes the effect of startle and surprise on perception, appraisal and actions guided by active frames. It highlights the process of frame modification (i.e., reframing) triggered by mismatches between perceived information and the active frame. However, the model does not account for the initial frame selection process, which is essential in situations where no surprise occurs. Additionally, the model does not depict the possibility of multiple perception–appraisal–action loops that may emerge when hypotheses derived from several potential frames are compared [111].

The perceptual cycle model offers a process-oriented framework that causally explains how internal schema interact with environment information in a cyclical manner, and how exploration emerge from the activated schema. Its elegance lies in its conceptual parsimony, as it omits the influence of external factors such as stress or fatigue. However, unlike the other two models propose bi-directional interactions between schema/frame and cognitive processes like perception and comprehension or appraisal, PCM presents the influence of schema as a one-directional process. As a result, it does not clearly differentiate between schema selection through pattern-matching and schema modification based on information processing. Furthermore, schema selection or modification driven by internal factors, such as individual experience also appear to be absent from the model.

The three-level situation awareness model integrates the role of situation awareness, working memory, long-term memory and attention into the decision-making process. Besides, the influence of internal as well as external factors affecting these cognitive components can be inferred directly. Unlike the perceptual cycle model, the situation awareness model does depicts schema selection. However, it does not explicitly account for potential schema modification at Level 2 or Level 3 SA. The schema serves only to guide comprehension and projection at these levels, without accounting for the possibility of schema modification based on the evolved level of SA. This stands in contrast to the model of startle and surprise, whose premise is that higher levels of SA (i.e., slow appraisal) are vulnerable to the effects of stress. Consequently, schema selection or reframing processes that depend on these levels may be significantly hindered under stress circumstances.

The application of different cognitive models has revealed distinctions with implications for training practices. In particular, sensemaking in response to startle and surprise can be categorized into two distinct processes. The first involves rapid pattern-matching, where perceived information is quickly aligned with existing schema. Through this mechanism, individuals recognize the situation and associated action possibilities based on prior experience and knowledge. Training interventions that support this capability include repeated simulator exposure to varied scenarios [112], the review of past occurred incidents in a simulated environment, and didactic approaches that encourage exploratory or problem-based learning [113]. However, this form of sensemaking may lead to errors when training sce-

narios are overly stereotyped, with unrelated events consistently co-presented, potentially distorting pattern recognition [114].

The second encompasses more deliberate, conscious cognitive processes, such as evaluating potential options, projecting outcomes, or systematically applying troubleshooting procedures. Target training should focus on enhancing pilots' conscious knowledge and meta-cognitive skills. This may include the use of memorized mnemonic strategies for managing stress [115], as well as theoretical instruction or stress exposure training designed to improve stress recognition and coping abilities [116, 117].

3

Effects of personality traits and flight experience on perceived startle and surprise

Building on the theoretical insights from Chapter 2, in this chapter, it is investigated if and how personality traits and flight experience influence pilots' cognitive and affective responses. The responses of startle and surprise are collected across seven in-flight startling and surprising scenarios conducted in motion-based simulators. The analysis draws on a dataset comprising 89 airline pilots across four experimental studies. The measured personality traits include trait anxiety, decision-related action orientation and failure-related action orientation. The pilot's self-report responses, perceived startle, surprise, stress, and mental workload, are standardized by obtaining Z-scores for each scenario.

Section 3.1 outlines the relevant personality traits, and summarizes the findings from prior research regarding their influence on performance under stress circumstances, which lead to the hypotheses investigated in this chapter. Section 3.2 details the research design and data analysis methods, followed by a presentation of results in Section 3.3. Section 3.4 discusses the theoretical and practical implications of the findings, addresses the limitations, and offers recommendations for future research.

The contents of this chapter have been published as: J. Chen, A. Landman, O. Stroosma, M. M. van Paassen, and M. Mulder. "The Effect of Personality Traits and Flight Experience on Pilots' Cognitive and Affective Responses to Simulated In-Flight Hazards". In: *Aviation Psychology and Applied Human Factors* 14.2 (2024), pp. 104–113.

3.1. Introduction

Individuals exhibit considerable variability in their perception and response to stressful stimuli [118], which could affect their ability to perform under stress. Neuroticism, one of the Big Five personality traits, is characterized by emotional instability and sensitivity to negative emotions [119]. This was found to be strongly related with severity of the startle reflex [120], symptoms of prolonged state anxiety [121], psychological stress [122], and impaired decision-making performance under pressure [123]. Conscientiousness, the Big Five personality trait of being responsible, diligent and careful [119] was found to be positively correlated with electrodermal stability when pilots encountered social stress [124]. Not surprisingly, a combination of low Neuroticism and high Conscientiousness has shown to be a favorable personality profile when it comes to coping with stress [125].

Compared to Neuroticism, trait anxiety represents a more specific facet of sensitivity to negative emotions, as it reflects a consistent tendency to appraise situations as threatening and to respond with heightened stress [126]. A meta-analysis found that pilots exhibiting low levels of both Neuroticism and trait anxiety were more likely to succeed in military aviation training, although additional factors are likely involved [127].

Another more specific trait than Neuroticism and Conscientiousness, which could be relevant to pilot performance in startling and surprising situations, is trait self-control. This is defined as one's ability to alter or override dominant response tendencies and to regulate behaviour, thoughts and emotions [128]. Trait self-control is associated with stress-reducing coping styles [129]. Individuals with high trait self-control strength are able to achieve desirable responses and inhibit undesirable responses, and are more successful in achieving their goals [130]. Kuhl developed measures for action and state orientation. Action orientation refers to the tendency to detach from irrelevant concerns, initiate goal-related actions more effectively, and more persistently focus on tasks until these are completed. In contrast, state orientation refers to the tendency to be distracted by alternative goals and affective states, and have difficulty in initiating actions to achieve goals. Highly action-oriented individuals were found to show increased down-regulating of stress and were able to maintain more control over behaviour and attention in demanding situations [132–135].

The current chapter focuses on the effects of trait anxiety and self-control on pilots' cognitive and affective responses to startling and surprising events. These traits could be highly relevant, as these may affect pilots' stress, and coping mechanisms in high demanding situations. Flight experience may also be an important mitigator, therefore it is included in the analysis. More insights into these relationships is useful for the development of personalized training interventions for pilots. Three hypotheses were formulated:

H1: Higher trait anxiety is positively correlated with pilots' perceived startle and stress in simulated in-flight emergency events, due to increased sensitivity to threat.

H2: Higher action orientation is negatively correlated with pilots' perceived startle, surprise, stress and mental workload, due to an increased focus on tasks and goals and less on emotional states, and distractions.

H3: More extensive flight experience is negatively correlated with perceived startle, surprise, stress and mental workload, as unexpected situations would be less novel for pilots, less demanding, and therefore less threatening.

The second, more exploratory goal of the current chapter is to analyse the relationships between pilots' perceived surprise, startle, stress and mental workload during simulated startling and/or surprising events. Insights into the strength of these relationships are useful to better understand the processes associated with the responses of startle and surprise, and for the development of measuring instruments for startle and surprise. Three additional hypotheses were formulated and tested for this second objective:

H4: Startle was hypothesized to be positively correlated with stress, as startle is expected to initiate a generalized stress response.

H5: Perceived mental workload was hypothesized to be positively correlated with stress, as highly perceived demand is likely to induce stress, and because stress may cause the mobilization of more mental effort.

H6: Perceived surprise was hypothesized to be positively correlated with perceived mental workload, as solving a mismatch following surprise is expected to be effortful.

3.2. Method

3.2.1. Participants

The dataset was established from four previous experiments, which involved a total of 89 commercial airline pilots. These experiments will be referred to as Study 1 [4], Study 2 [112], Study 3 [115], and Study 4 [136]. The relevant characteristics of all participating pilots are summarized in columns 2 to 4 in Table 3.1. All participants were required to hold a valid commercial pilot license. All studies complied with the tenets of the Declaration of Helsinki and informed consents were obtained from all participants.

3.2.2. Tasks and apparatus

An overview of the test scenarios is shown in Table 3.2. More detailed descriptions of the apparatus, tasks and conditions can be found in the respective publications. All tasks were performed in motion-base simulators, namely the Desdemona flight simulator in Study 1 and the SIMONA Research Simulator in Studies 2, 3 and 4. Experimental procedures included briefing, familiarization, pre-test, a ground theory session on startle and surprise (only for Studies 3, 4), a training session (only for Studies 2, 3, 4), a test session and a debriefing.

Table 3.1: Characteristics of the participants.

Study	**Number of participants**	**Age (years) Mean (*SD*)**	**Flight hours (hrs) Mean (*SD*)**
Study 1	20	36.3 (7.9)	6,987 (3,804)
Study 2	20	41.2 (8.7)	8,441 (5,467)
Study 3	24	38.5 (12.0)	7,358 (5,580)
Study 4	25	43.2 (9.2)	9,930 (6,281)

Some studies divided participants in an experimental group and a control group (Studies 2, 3 and 4), but since no significant effects of these treatments were found on the dependent measures for the current chapter, participants in these studies are regarded as one group.

During the briefing, participants were instructed about the flight tasks, the aerodynamic model, the simulator and its features. After this, they completed questionnaires on demographic information, flight experience and personality traits (see Section 3.2.3). In the training session for Study 2, participants in the experimental group were required to complete tasks under more variations in the mixed order, which includes various wind directions, wind strengths, and malfunction timings. The control group repeatedly conducted the same task under one of the variations and performed task in a more repetitive sequence. During the ground theory session of Studies 3 and 4, participants in the experimental group were introduced about the theory of startle and surprise, and the rationale behind the experimental training intervention. The control group received only the introduction on startle and surprise. After each test scenario in each study, participants were asked to indicate their ratings on startle, surprise, stress and mental workload (see Section 3.2.4).

Table 3.2: Test scenarios and personality traits in four studies.

Study	Between-subject factor	Scenario	Event description	Scale on personality trait
Study 1	None	1-1 Stall	In the presence of a strong tailwind, the pitch trim was adjusted toward 48% of its maximum capacity in 3 s.	STAI Form Y-2; AOF subscale; AOD subscale
Study 2	High variability and unpredictability training session versus low variability training session	2-1 Airspeed indicator failure	Upon rotation, the indicated airspeed decreased with 1 kt/s from the actual airspeed.	None
		2-2 Single engine failure	When the speed reached 55 kt, thrust in the right engine dropped in 20 s to 40%.	
		2-3 Rudder failure	The rudder effectiveness decreased to 20% as the pilot rolled out of the turn towards the downwind leg.	
Study 3	Training session with the experimental training intervention or without	3-1 Flap asymmetry	When selecting Flaps 25, the left flap remained UP.	STAI Form Y-2
		3-2 False stall warning	When reaching 1,500 ft, a bird struck the angle of attack vane and the stall warning appeared.	
		3-3 Airspeed indicator failure	Same as Scenario 2-1.	
		3-4 Mass shift	Upon rotation, a piece of cargo broke loose and shifted towards the tail.	
Study 4	Training session with the experimental training intervention or without	4-1 Flap asymmetry	Same as Scenario 3-1.	STAI Form Y-2; AOF subscale; AOD subscale
		4-2 Mass shift	Same as Scenario 3-4.	

Table 3.3: Overview of personality traits and flight hours in four studies.

Study	STAI-trait Mean (*SD*)	AOD Mean (*SD*)	AOF Mean (*SD*)	Flight hours Mean (*SD*)
Study 1	29.0 (6.2)	9.2 (2.4)	8.2 (3.0)	6,987 (3,804)
Study 2	N/A	N/A	N/A	8,441 (5,467)
Study 3	27.0 (9.2)	N/A	N/A	7,358 (5,580)
Study 4	29.1 (5.0)	9.0 (3.0)	8.2 (3.4)	9,930 (6,281)

Note. AOD = decision-related action orientation; AOF = failure-related action orientation; STAI = State–Trait Inventory; N/A = not available.

3.2.3. Independent measures

An overview of the measured personality traits in each study, and mean values and SDs obtained in the different samples, is shown in Table 3.3.

Action vs. state orientation

As measures of trait self-control, two subscales of the action control scale (ASC-90) were used [132], namely: decision-related action orientation (AOD), and failure-related action orientation (AOF). The third subscale of the ACS-90, on performance-related action orientation (AOP), was not included. This scale relates more to intrinsic motivation to persevere in tasks, and not to dealing with high demand and threat.

The AOD and AOF each consists of twelve items describing self-regulatory situations. For each situation, participants indicated which of two alternatives best describes how they would usually respond: an action-oriented or state-oriented option. To analyse the results, scores are assigned (1 for action-oriented, 0 for state-oriented) and summed. Higher scores indicate a stronger disposition toward action orientation, with a score of seven or higher typically reflecting trait action orientation and a score of six or lower indicating trait state orientation. Action-oriented and state-oriented individuals have been shown to be equally well represented in general population [131]. The ACS-90 was reported to have sufficient construct validity [137] and good internal consistency, Cronbach's $\alpha > 0.70$ [138].

For the current sample (N = 45 in Studies 1 and 4), the AOD and AOF subscales had high internal consistency, $\alpha = 0.844$ and $\alpha = 0.821$, respectively. Overall, our sample significantly scored above the norm on AOD, M = 9.1, $SD = 2.7$, $t = 4.59$, $p < 0.001$, and on AOF, M = 8.2, $SD = 3.2$, $t = 7.62$, $p < 0.001$, which indicates that pilots in this sample were considerably more action-oriented than the general population.

Trait anxiety

Trait anxiety was measured by the Y-2 From (Trait scale) of the State-Trait Inventory (STAI; [139]) in Studies 1, 3, and 4. Trait anxiety is defined as a relatively stable behavioural disposition to respond anxiously to a wide range of threatening stimuli. Participants were required to indicate how they generally feel on twenty statements on four-point Likert scales. In previous studies, Cronbach's α for the scale ranged from 0.86 to 0.95 [140], the test-retest reliability coefficients were found to range from 0.86 to 0.73 over a retest interval of 20 days and 104 days, respectively [141].

For the sample (N = 69 in Studies 1, 3 and 4), the trait anxiety scores, M = 28.3, SD = 7.0, were significantly lower than the general population (i.e., 36.7), t = -9.93, $p < 0.001$.

Flight experience
Pilots listed their flight hours on large jet aircraft after briefing.

3.2.4. Dependent measures

Perceived startle and surprise
Startle and surprise were measured in Studies 1, 3, 4 using non-validated 0-10 Likert scales ranging from 0 ("*not at all*") to 10 ("*extremely*") by answering the questions: "*How **startled** were you by [the stimulus]*?" and "*How **surprised** were you by [the stimulus]*?" (*[the stimulus]* was substituted by the potentially startling/surprising event in the scenario). In Study 2, a nonvalidated 5-point Likert rating scale was used instead to collect responses to the same questions, with 1 representing "*not at all*" and 5 representing "*extremely*".

Perceived stress
Ratings of acute stress were measured using the anxiety scale [142]. The anxiety scale applied in Studies 1 and 4 was the 11-point Likert-type version ranging from 0 to 10, while a continuous visual analogue scale version was applied in Studies 2 and 3. The visual analogue version scale was 10 cm long horizontal line, with tick marks at 1 cm intervals labelled 0 and "*not at all*" at the left endpoint, and 10 and "*extremely*" at the right endpoint.

Perceived mental workload
In Study 1, mental workload was measured using a unidimensional scale ranging from 0 ("*very low workload*") to 100 ("*very high workload*") [143]. Considering that the task (e.g., stall recovery) required little physical effort, the score was used as an indication of mental workload. In Study 2, mental workload was rated using the mental demand subscale of the NASA-TLX [18], a 21-point scale ranging from 0 ("*low*") to 100 ("*high*"). In Studies 3 and 4, the English version of Rating Scale Mental Effort (RSME; [17]) was used as an indication of perceived mental workload. The RSME consists of a 150 mm line marked with nine anchor points, each accompanied with a descriptive label indicating a degree of effort. Participants were instructed to indicate their invested effort by placing a cross on the continuous line, resulting in a score between 0 to 150.

If participants invest less mental effort than the workload required for completing the task successfully, mental effort can differ from mental workload. In all four experiments, however, all pilots declared beforehand that they would do their best to perform well in the test, which leads us to assume that their invested mental effort coincides with one's perceived workload imposed by the task. Similar conclusions were found in NASA-TLX validation study [18], where the factor "mental effort" was consistently related to overall workload from single cognitive laboratory tasks to simulations in motion-based simulators.

3.2.5. Statistical analysis

For each of the dependent measures, Z-scores were calculated per participant per scenario. This means that each score reflects how a pilot responded relatively to other pilots in the same scenario, and that we corrected for different ranges of different scales. To investigate the

effects of personality traits and flight hours on dependent measures, the averaged Z-scores of startle, surprise, stress and mental workload were obtained for each pilot by averaging scores in different scenarios. Then, we calculated Spearman's correlations between the independent measures: STAI-trait, AOD, AOF, flight hours, and the dependent measures: Z-scores of perceived startle, surprise, stress and mental workload.

With regard to relationships between the dependent measures obtained in the repeated-measures conditions, the "between and within formulation" [144] was applied. The working principle of this formulation is that the total sample variances can be decomposed into within-individual variance and between-individual variance. Both between-individual and within-individual correlation matrices were obtained from the Z-scores of perceived startle, surprise, stress and mental workload per participant per scenario.

3.3. Results

3.3.1. Effects of personality traits and flight hours

Table 3.3 shows the averaged AOD, AOF, STAI and flight hours in the four studies. Table 3.4 summarizes the means and SDs of pilots' responses in each scenario performed in four studies. The measures show that most of the scenarios were experienced as startling, surprising, or both, as most of the mean scores are above the midpoint of the scales.

Table 3.5 lists the Spearman's correlations between STAI-trait, AOD, AOF scores, flight hours, and the Z-scores of the pilot responses, and Table 3.6 shows the Spearman's correlations between STAI-trait, AOD, AOF scores and flight hours. The STAI-trait score was significantly positively correlated with perceived stress. Pilots with higher trait anxiety levels reported higher stress with regard to the simulated events. Neither AOD nor AOF was significantly correlated with the dependent measures. No significant correlations were observed between flight hours and any of the pilot responses.

Table 3.4: Means and standard deviations of the cognitive and affective responses.

Study	**Scenario**	**Surprise Mean (*SD*)**	**Startle Mean (*SD*)**	**Anxiety Mean (*SD*)**	**Mental workload Mean (*SD*)**
Study 1	Scale range	[0-10]	[0-10]	[0-10]	[0-100]
	1-1	8.0 (1.8)	3.9 (2.1)	3.7 (1.6)	66.0 (15.4)
Study 2	Scale range	[1-5]	[1-5]	[0-10]	[0-100]
	2-1	3.6 (0.7)	2.9 (1.1)	5.1 (2.1)	67.3 (19.4)
	2-2	2.4 (0.6)	2.0 (0.7)	3.6 (2.0)	57.0 (16.7)
	2-3	3.2 (1.0)	2.8 (0.9)	5.8 (2.0)	72.3 (14.9)
Study 3	Scale range	[0-10]	[0-10]	[0-10]	[0-150]
	3-1	6.3 (2.5)	5.3 (2.3)	4.4 (2.1)	56.8 (20.7)
	3-2	6.4 (2.3)	7.0 (2.0)	4.2 (2.3)	51.5 (18.8)
	3-3	6.3 (2.4)	4.8 (2.5)	3.7 (2.2)	55.1 (20.8)
	3-4	7.2 (2.3)	6.3 (2.3)	5.5 (2.1)	69.8 (23.2)
Study 4	Scale range	[0-10]	[0-10]	[0-10]	[0-150]
	4-1	5.5 (2.2)	4.6 (2.0)	3.7 (1.8)	64.8 (17.8)
	4-2	7.2 (1.7)	6.7 (2.0)	5.7 (2.1)	80.3 (16.7)

Table 3.5: Correlations between STAI-trait, AOD, AOF scores, flight hours, and Z-scores of the cognitive and affective responses.

		Z(Startle)	Z(Surprise)	Z(Stress)	Z(Mental workload)
STAI-trait	ρ	0.183	-0.048	0.332*	0.188
	p	0.132	0.694	0.005	0.122
AOD	ρ	0.045	0.067	-0.116	-0.019
	p	0.771	0.660	0.448	0.899
AOF	ρ	0.005	0.175	-0.070	-0.115
	p	0.975	0.249	0.650	0.452
Flight hours	ρ	-0.141	-0.095	-0.051	-0.161
	p	0.188	0.373	0.638	0.132

* $p < 0.05$ (2-tailed).

Table 3.6: Correlations between STAI-trait, AOD, AOF scores, and flight hours.

		STAI-trait	AOD	AOF	Flight hours
STAI-trait	ρ	1.000			
	p	-			
AOD	ρ	-0.444**	1.000		
	p	0.002	-		
AOF	ρ	-0.447**	0.347*	1.000	
	p	0.002	0.019	-	
Flight hours	ρ	-0.069	-0.035	0.189	1.000
	p	0.576	0.821	0.215	-

** $p < 0.01$ (2-tailed). * $p < 0.05$ (2-tailed).

The STAI-trait scores were significantly negatively correlated with the AOD and AOF scores. These results suggest that pilots with higher trait anxiety also tended to have lower decision-related and failure-related action orientation Moreover, the AOD and AOF scores were significantly positively correlated. Pilots who were more action-oriented in AOD are more likely to be action-oriented in AOF, and vice versa.

3.3.2. Correlations between pilot responses

The pooled within-individual correlation matrix is shown in the top panel of Table 3.7. The within-individual correlation matrix shows the average correlations between four cognitive and affective responses for each pilot. The estimated between-individual correlations presents the correlations between responses based on the average responses across all pilots, as shown in the bottom panel of the Table 3.7.

For within-individual correlations, the strongest significant correlations were observed between Z-scores of mental workload and Z-scores of stress, and between Z-scores of mental workload and Z-scores of surprise. This means that pilots who rated a certain scenario as more mentally demanding, also were likely to rate it as more stressful and surprising. For between-individual correlations, the highest significant correlations were observed between Z-scores of surprise and Z-scores of startle, and between Z-scores of mental workload

Table 3.7: Correlation matrices of the cognitive and affective responses.

	Z(Startle)	Z(Surprise)	Z(Stress)	Z(Mental workload)
Within-individual correlation matrix				
Z(Surprise)	0.316**	-		
Z(Stress)	0.265**	0.236**	-	
Z(Mental workload)	0.258**	0.454**	0.469**	-
Between-individual correlation matrix				
Z(Surprise)	0.673**	-		
Z(Stress)	0.569**	0.536**	-	
Z(Mental workload)	0.402**	0.492**	0.695**	-

** $p < 0.01$ (2-tailed).

and Z-scores of stress. This means that pilots who generally scored higher than others on surprise, also generally scored higher than others on startle, and pilots who scored generally higher than others on mental workload, also scored generally higher than others on stress.

3.3.3. Missing values

In Study 2, three cases of stress ratings were missing. In Study 4, three scenarios were presented incorrectly for three respective participants, leading to loss of all their responses' data in these scenarios. All missing values were replaced by the mean value of the available responses of the rest of participants in the corresponding scenario. The substituted values were 2.65% of stress ratings and 1.33% of surprise, startle, and mental workload ratings, with regard to the total number of data.

3.4. Discussion

For the main objective, in line with hypothesis H1, trait anxiety was found to correlate significantly and positively with perceived stress during simulated in-flight events. Pilots with higher trait anxiety experienced more stress during these events. This relationship supports the basic hypothesis from the Interaction model of stress [145], in that trait anxiety could interact with the stressful situation (i.e., the unexpected failure) leading to an increase in acute stress. Given that increased stress could disrupt the balance between a pilot's goal-directed and stimuli-driven system [98], it could be more difficult for pilots with higher trait anxiety to manage their attention effectively.

This finding highlights the importance of personalizing pilot training to individual differences. For instance, pilots with higher trait anxiety may benefit from specialized stress interventions, such as repeated exposure to simulated high-stress scenarios, to reduce sensitivity to unexpected events [146, 147]. No evidence was found, however, that trait anxiety affected mental workload in the presented situations.

An analysis of pilot performance was considered beyond the scope of the current chapter, as differences between scenarios made it difficult to standardize and pool pilots' performance metrics. In addition, most of the included four studies did not involve a control condition where baseline measures of pilot performance without startle or surprise were obtained.

Contrary to hypothesis H2, our findings did not indicate that higher action orientation was associated with lower ratings of startle, surprise, stress, or mental workload in the scenarios. One possible explanation is that action orientation perhaps did not impact the pilots' responses themselves, but instead contributed to better coping mechanisms to these responses. However, the absence of significant correlations may also be caused by the homogeneity of our sample (see Table 3.1). The participated pilots exhibited lower trait anxiety and higher action orientation (both AOD and AOF) compared to the general population, and standard deviations were relatively small. All participants were active commercial airline pilots, and they volunteered for a study they knew would assess their ability to cope with in-flight failures. This self-selection likely resulted in a sample with generally high action orientation, even relative to the average pilot. Similarly, contrary to hypothesis H3, we found no significant correlation between flight experience and surprise, startle, stress, or mental workload. This could indicate the existing non-validated measures of startle and surprise are lacking in accuracy. And the possible benefit from targeted training interventions for mitigating effects of high stress, startle or surprise could be obtained for both novice and experienced pilots.

For the secondary objective, in line with hypotheses H4, H5 and H6, significant correlations were found between all dependent measures, both within individuals and between individuals. The strongest within-individual correlations emerged between stress and mental workload, and between surprise and mental workload. This indicates that pilots who rated a certain scenario as more mentally demanding, were also highly likely to rate it as more stressful and surprising. One explanation is that stress may increase mental effort invested in task performance, enhancing focus on task-relevant rather than threat-related stimuli [98]. Also, if certain pilots experience difficulty with responding to the failures, this may increase both high mental workload and stress. In the case of surprise, the cognitive effort required to make sense of the situation and reframe it following an unexpected event likely contributed to increased mental workload [35].

Interestingly, the within-individual correlation between startle and stress was not one of the strongest correlations. It seems that stress in the scenarios was affected by other factors besides startle, such as task difficulty. For the between-individual correlations between dependent measures, the strongest correlations were observed between startle and surprise and between stress and mental workload. The strong correlation between startle and surprise implicates that the propensity to be startled is possibly related to the propensity to be surprised. It could also be caused by the fact that we did not manipulate startle and surprise separately in the presented scenarios. The strong correlation between stress and mental workload implies that pilots who tend to experience more stress are also those who experience the highest workload, possibly due to being less skilled. However, the latter was not substantiated by significant correlations between flight hours and mental workload.

When considering the findings of the current chapter, a number of limitations need to be mentioned. First, a variety of different scenarios were used to obtain the dataset. Events in the scenarios were considered, on average, to be moderately startling or surprising by the pilots (i.e., scored around the midpoint of the scales). However, no complex flight system failures or checklists were included, and all scenarios were flown manually at a relatively low altitude. This limits the generalizability of results to those types of events. Also, many of the presented scenarios differed from the pilots' daily operational tasks. All scenarios were

performed in a single-pilot setting and using a simplified twin-prop aircraft model that most pilots had limited experience with. Apart from the unfamiliarity, high workload was induced by requiring pilots to fly manually, instead of simulating complex tasks involving system management, higher levels of automation, crew teamwork, and resource management or emphasizing planning and navigation. Second, regarding the measures of the cognitive and affective responses, the rating scales used for surprise and startle were not psychometrically validated to provide insights for future research. Outcome responses were measured on unidimensional scales, which might be generally less accurate than multidimensional scales.

A third limitation of this chapter is that we did not apply formal corrections to the correlation analyses. While this decision was intentional to avoid overly conservative adjustments that might mask meaningful relationships, it increases the risk of Type I errors. This means that some of the significant correlations reported in this chapter could have occurred by chance rather than representing true underlying relationships.

To address these possible biases, future research could possibly be performed in an actual training environment, employing a simulated aircraft type that pilots are familiar with. This allows for more complex, high-demand tasks. It is also recommended to include a larger and more diverse sample of pilots to improve the generalizability of the findings. Additionally, future studies should incorporate objective or real-time physiological measures to investigate the potential causal relationships between startle, surprise, stress, and mental workload. Moreover, correction methods are recommended to be applied, such as Bonferroni correction, to strengthen the reliability of current findings.

In conclusion, the current chapter provides data on pilot responses for different simulated emergency events which are useful for applications in future research. Within the aviation context, data on effects of pilot personality traits on reactions in surprising situations are scarce. The current chapter contributes to the literature by providing insights in effects of trait anxiety and trait self-control.

4

Development and preliminary validation of the Startle and Surprise Inventories

Theoretical frameworks and empirical evidence presented in Chapters 2 and 3 underscore the need for reliable, validated, and non-obtrusive measuring methods to assess startle and surprise. In response to this gap, this chapter details the development and preliminary validation process of the Startle and Surprise Inventories (Startle-I; Surprise-I) and the Visual Analogue Scales for Startle and Surprise (Startle-VAS; Surprise-VAS), following a structured three-phase process. In Phase 1, 14 items for surprise and 7 items for startle are derived from foundational and applied literature. These items are evaluated for content validity by seven experts in the field, with retention determined by a minimum of 50% agreement on the relevance. In Phase 2, 81 participants rated the retained items nine times, each time immediately after watching a video clip. Construct validity is assessed using multilevel exploratory factor analysis with an oblique, direct oblimin rotation. In Phase 3, the concurrent validity of the Startle-VAS and Surprise-VAS are tested by comparing with the scores from Startle-I and Surprise-I, respectively.

The structure of the chapter is as follows: Section 4.1 introduces the scientific gap concerning self-report measures of startle and surprise. Section 4.2 outlines the methodology, including participants' characteristics, experimental procedures, video stimuli, apparatus, and statistical analysis. Section 4.3 presents the results of three validation phases. Section 4.4 interprets the findings, discusses operational implications of the newly developed measures, acknowledges limitations, and provides recommendations for future research. Section 4.5 concludes the chapter with a summary of key insights.

The contents of this chapter have been published as: J. Chen, A. Landman, A. Derumigny, O. Stroosma, M. M. van Paassen, and M. Mulder. "Development and Validation of the Startle and Surprise Inventories and Visual Analogue Scales". In: *Ergonomics* (2025), pp. 1-14.

4.1. Introduction

Both startle and surprise produce measurable physiological responses, and various methods have been proposed to detect these responses. The intensity of a startle response is commonly assessed via the eye-blink reflex, which can be measured using surface electromyography (EMG) by detecting action potentials in the orbicularis oculi muscle [37]. Alternative techniques for measuring eyelid movement include potentiometric [38], photoelectric [39], vertical electro-oculographic (vEOG) [40], and magnetic search coil methods [41]. In addition to the blink reflex, pupil dilation has been identified as a physiological correlate of startle [42]. Cardiovascular indicators, such as increased heart rate and blood pressure, have also been observed within ten seconds of an acoustic startle, providing complementary autonomic indicators [26].

Surprise can be physiologically measured through the EEG P300 event-related potential (ERP) [44], pupil dilation, and activation in subcortical regions associated with dopamine [45]. The P300 ERP originates in the anterior cingulate cortex (ACC), and peaks in ACC activity occur in aversive defence responses in general [46]. Pupil dilation was considered to represent the global state of cognitive processing as an effect of stress rather than solely elicited by surprise [47]. Pupil dilation related to surprise (maximum around 500 ms after the stimulus) was also reported to be slower than pupil dilation due to startle [48].

Physiological measures provide objective, real-time assessments and can overcome certain inherent biases, such as socially desirable answering patterns [56]. However, the above-mentioned physiological measures are not specific to startle or surprise, but reflect broader autonomic, neural activation, and affective responses [57]. While these techniques offer high temporal resolution, they cannot distinguish between a startle response and similar defensive reactions triggered by fear or stress [58]. They are also uneconomic for application to large numbers of participants [59], and are impractical and invasive to apply in operational settings. These measures are also often only meaningful in relation to individual's own baseline, and were found to be inconsistent with the subjective experience of the responses. For example, in a study testing the effect of surprise on pilots' performance [4], participants were found to show nearly similar levels of heart rate, yet reported significant different levels of startle and surprise between conditions. Thus, similar to the literature on experienced challenge and threat [60], validated self-report measures on startle and surprise are necessary to complement physiological measures and contribute to the study of relationships between physiological data, subjective experience and performance.

For surprise, the Differential Emotions Scale (DES-IV)[49] and the Positive and Negative Affect Schedule-Expanded Form (PANAS-X) [50] have been developed with subscales for measuring surprise. These two subscales consist of the same three items (i.e., How do you feel "surprised", "amazed", and "astonished"?), which are rated on 5-point Likert scales pertaining to feelings at the moment or to a certain past time frame. DES-IV measures twelve fundamental emotions, in which surprise belongs to positive emotional factors. A principal component analysis with orthogonal varimax rotation [49] and a confirmatory factor analysis [51] supported that these three items loaded on a separate construct referring to other affects. The set of three items was stable across time, with a test-retest coefficient of $r = 0.61$ over a six-month interval [52]. However, the subscale only showed moderate internal consistency ($\alpha = 0.65$), likely due to the small number of items [53].

The same three items are also included in the PANAS-X to measure surprise, except that

surprise is treated as a specific affect, neither positive nor negative. Internal consistency was found to be slightly higher than the Surprise subscale in DES-IV, with $\alpha = 0.72$ to $\alpha = 0.80$. Test-retest coefficient over a two-month interval was lower, $r = 0.23$, referring to the time frame "past week", and higher, $r = 0.52$ - 0.56, to "in general" [50]. The mean scores on Surprise were the lowest compared to other affects in the PANAS-X over different samples [50]. The Surprise subscale was also the only one for which self-ratings did not correlate significantly with peer-ratings, $r = 0.14$. The relatively low stability and mean scores for surprise may be attributed to its nature as a transient emotional state typically triggered by stimuli, which were not provided during these studies. For this reason, the question asked whether participants felt surprised "in general" appears difficult to answer in comparison to items in other subscales such as "Fatigue" or "Shyness", which reflect more stable, trait-like experiences. In addition, there is no peer-reviewed report on a systematic methodology for the items selection in DES-IV or PANAS-X. Besides the use of these multi-item subscales, researchers have used single-item scales to measure self-report surprise [54, 55], however, these measures have not been validated in a systematic way yet.

For measuring startle, there has been to date no systematic attempt to develop and validate a self-report measure [43]. For surprise, items in existing scales were not selected in a systematic manner and scales were not developed nor validated to stimuli, even though the concept of surprise, in contrast to other affects, only makes sense referring to a stimulus or event. The goal of this chapter is therefore to systematically develop measures for self-report startle and surprise. Accordingly, this chapter consists of three sequential phases which describe the development and preliminary validation of the multi-item Startle and Surprise Inventories (Startle-I; Surprise-I) as well as the single-item Visual Analogue Scales for Startle and Surprise (Startle-VAS; Surprise-VAS).

4.2. Method

4.2.1. Participants

Participants were recruited from the student and employee population of Delft University of Technology (N = 82). They were invited via flyers, and received a compensation gift worth 5 euros. Data of one of the participants was excluded because this participant did not read the items accurately enough, and missed the reverse-coding of some items (remaining N = 81). The participants ranged in age from 19 to 63 years, in which 74.1% were male (M = 27.2, $SD = 7.9$) and 25.9% were female (M = 25.0, $SD = 3.1$). All participants declared that they possessed basic proficiency in English reading. The Research Ethics Committee of Delft University of Technology approved the research design (No. 2718). Informed consent was obtained from all participants.

4.2.2. Procedure

Phase 1: items set generation and content validity

An initial set of 14 items for surprise, and 7 items for startle (see Table 4.1) were formulated based on a review of fundamental and applied literature on startle [20, 23, 148–150] and surprise [19, 30, 35, 49, 77]. Among these, Items 1 and 8 were derived from the DES-IV [49] and PANAS-X [50]. The item "amazed" was excluded due to its frequent association with positive valence, as the term "amazing" is commonly interpreted as a positive expression.

Table 4.1: Initial set of items and their relevance scores.

Item	%
Initial set of items for surprise	
1. It surprised me.	85.7
2. It was consistent with my expectation.[a]	85.7
3. I was taken aback by it.	85.7
4. I did not understand why it happened.	100.0
5. I predicted it beforehand.[a]	85.7
6. Initially, it made no sense to me.	71.4
7. I did not see it coming.	100.0
8. It astonished me.	85.7
9. Initially, I was confused about it.[b]	85.7
10. It bewildered me.	57.1
11. It made my jaw drop.[c]	42.9
12. I was not mentally prepared for it.	85.7
13. It was unexpected.	100.0
14. It made me feel wide-eyed.[c]	28.6
Initial set of items for startle	
15. It startled me.	85.7
16. It immediately made me feel scared or angry.[b]	71.4
17. It shocked me.	85.7
18. It stunned me.	85.7
19. It made me physically flinch.	85.7
20. It caused my heart to suddenly beat harder or faster.	100.0
21. It immediately caused stress or frustration to me.[b]	85.7

[a] Item is reverse-coded. [b] Item was rephrased.
[c] Item was removed in Phase 1.

The goal was to construct a measure that could be applied consistently across surprises of positive, negative, and neutral valence.

The content validity of each item was examined by letting seven independent experts in the fields of Cognitive Science and Psychology review the items. Experts indicated whether each item is relevant for measuring the experience of startle or the experience of surprise. The relevance score for each item was then calculated as the percentage of experts who rated the item as relevant. Experts were also invited to provide open comments on the formulation of the set of items. An item was retained if at least 50% of the experts considered that item to be relevant for its construct [64].

Phase 2: multilevel exploratory factor analysis

In Phase 2, the factor structure of the 19 items remaining from Phase 1 was explored by letting participants complete these items nine times, each time immediately after watching one of the nine video clips (see Tables 4.2 and 4.3). The retained 19 items were arranged and presented in a randomized order, so that participants had no information on whether items were intended to measure startle or surprise.

Participants sat in a secluded room, received verbal instructions outlining the experiment

procedure, the concepts of startle and surprise, and the list of items. Startle was explained as "a rapid, involuntary reaction to an abrupt and intense stimulus, that is typically perceived as a threat", and surprise was described as "a cognitive-affective response evoked by an unexpected stimulus or event". These definitions were supplemented with practical, real-world examples to clarify the underlying mechanisms and implications.

Participants were instructed to indicate their level of agreement with each statement by circling a number on a 5-point Likert scale. The response options were: (1) "*Strongly disagreed*", (2) "*Disagreed*", (3) felt "*Neutral*", (4) "*Agreed*" or (5) "*Strongly agreed*". Responses were recorded using pen on the paper-based version of the measures. The specific stimulus or event referred to by "it" in each item corresponded to one of the nine video clips, as detailed in the third columns of Tables 4.2 and 4.3.

Table 4.2: Description of the video clips.

ID	Video description	The stimulus[d]	Intended responses	Duration	URL
Tumbler	From the perspective of looking out of a washing machine tumbler, a man fills the machine and turns it on. Instead of the tumbler, the room starts to spin with objects falling, and the man holds on to the tumbler.	The room beginning to spin	Surprise	16 s	https://youtube.com/shorts/kdjlmnsJQvc?feature=share
Clouds	A man appears to jump off a ridge above the clouds into the depths, but then hits the surface of a pond. The clouds were a reflection in the pond.	The man jumping into the water	Surprise	10 s	https://youtu.be/7p_iFLK9Idg
Monster1	A car drives down a mountain road and disappears behind trees. A zombie-like monster suddenly appears on the screen with a loud scream.	The appearance of the monster	Startle and surprise	17 s	https://youtu.be/fMPnWl0o4Yc
Monster2	A repetition of Monster1, whereas participants are informed that the same video is shown again.	The appearance of the monster	Startle	17 s	https://youtu.be/fMPnWl0o4Yc
Pill	White pills are shown laying on a table. A screwdriver appears and unscrews one of the pills out of the table. The pill was apparently a screw.	The "pill" being a screw	Surprise	9 s	https://youtu.be/U3VxQSUioMU
Spider	A man is trying to catch a huge spider on the wall with a pan. The spider is shown to suddenly jump at the camera using computer-generated imagery, which coincides with a loud scream.	The spider jumping at you	Startle and surprise	14 s	https://youtu.be/6em7aIoF5fI

[d] The stimulus as specified in the measures of startle and surprise.

Table 4.3: Description of the video clips (continued).

ID	Video Description	The stimulus[d]	Intended responses	Duration	URL
Puppy	A puppy gives a high-five to a person with its right paw and then with its left paw.	The dog giving a second high five	Neither startle nor surprise	12 s	https://youtu.be/ZeJEVbpn8d8
Baseball	A man seemingly swings a baseball from a stand in slow motion. When connecting, the ball falls to the ground in normal speed. The man was apparently just moving very slowly, and the video was filmed in normal speed.	The baseball not flying away	Surprise	8 s	https://youtu.be/Df_sk92u4lM
Window	A boy is repetitively kicking a football against a wall of a house, just missing the windows. After a few kicks, he hits and breaks a window.	The football hitting the window	Neither startle nor surprise	12 s	https://youtu.be/IkXMQznN5ck

[d] The stimulus as specified in the measures of startle and surprise.

Phase 3: Visual Analogue Scales for Startle and Surprise
After completing the items following each video, participants also provided ratings on the Startle-VAS and Surprise-VAS by answering the question "*How **startled** were you by [the stimulus]?*" and "*How **surprised** were you by [the stimulus]?*", referring to the specific stimulus or event in the preceding video clip (as presented in the third columns of Tables 4.2 and 4.3). The Startle-VAS and Surprise-VAS, each consisted of a 10 cm horizontal line with tick marks at 1 cm intervals, ranging from 0 to 10. The left endpoints were labelled with, "*not startled at all*" and "*not surprised at all*", respectively. The right endpoints were labelled with, "*extremely startled*" and "*extremely surprised*", respectively. Participants were required to place a cross on the line as answer to the question and the resulting score was the distance of the centre of the cross to the left endpoint in centimetres.

4.2.3. Video stimuli
To induce a variety of startle and surprise responses, nine video clips were selected from the internet. Tables 4.2 and 4.3 summarize the description of the video clips and corresponding links. Predictably surprising videos were selected based on instilling an incorrect expectation with regards to upcoming events. Startling videos were aimed to increase attentional focus on a location or object in the video, and then induced a jump-scare by the sudden appearance of something possibly fear-inducing, which coincided with a loud noise. Videos aimed at neither startling or surprising did not contain jump scares and showed a sequence of events that were predictable. More videos were included to induce surprise (n = 6) than startle (n = 3), as we expected that surprise would be less reliably induced than startle.

The order in which the video clips were presented was counterbalanced between participants using the Latin square method [151] to reduce systematic error, except for Monster1 and Monster2 which were always presented in sequence to ensure that Monster2 was not surprising. A 120-second recovery period was imposed following the completion of the scales after each intended startling video clip (i.e., Monster1, Monster2 and Spider).

4.2.4. Apparatus
In Phases 2 and 3, participants were presented the video clips on a desktop computer screen (Dell P2414HB) with noise-cancelling headphones (Sony WH-XB910N). The sound volume was set to a fixed level for all participants at the start of each video clip.

4.2.5. Statistical analysis
A full set of data was obtained in Phase 2 and was preprocessed by reversing the scores on items that were reverse-coded (Items 2 and 5). To examine the suitability of the dataset for factor analysis, Kaiser–Meyer–Olkin measure of Sampling Adequacy, KMO = 0.93, and Bartlett's test of Sphericity were checked, $p < 0.001$.

A two-way ANOVA was conducted for each item to examine the proportion of variance that was attributable to differences between participants (Factor Participant) and differences between videos (Factor Video). The outcomes are summarized in Table 4.4.

Table 4.4: Two-way ANOVA results and estimated intraclass correlation coefficient (ICC) for each item.

Item	Factor	Sum of Square	Variation (%)	F	ICC
Item set for surprise					
1. It surprised me.	Participant	58.02	10.97	1.61	0.33
	Video	495.33	34.38	50.32	
2. It was consistent with my expectation.	Participant	156.99	9.73	1.45	0.39
	Video	592.49	36.72	54.85	
3. I was taken aback by it.	Participant	268.40	26.07	3.54	0.16
	Video	154.49	15.00	20.37	
4. I did not understand why it happened.	Participant	367.59	29.83	4.18	0.21
	Video	161.54	13.11	18.37	
5. I predicted it beforehand.	Participant	140.06	7.79	1.07	0.37
	Video	613.99	34.13	47.02	
6. Initially, it made no sense to me.	Participant	345.51	28.93	4.13	0.16
	Video	180.17	15.09	21.56	
7. I did not see it coming.	Participant	162.11	10.12	1.38	0.35
	Video	497.84	31.08	42.29	
8. It astonished me.	Participant	289.33	26.88	3.63	0.18
	Video	150.09	13.94	18.85	
9. Initially, I was confused about it.	Participant	215.39	18.18	2.36	0.22
	Video	239.81	20.24	26.29	
10. It bewildered me.	Participant	309.90	35.30	4.90	0.09
	Video	61.85	7.05	9.78	
12. I was not mentally prepared for it.	Participant	296.10	27.09	3.57	0.14
	Video	133.19	12.19	16.05	
13. It was unexpected.	Participant	160.00	10.01	1.53	0.38
	Video	601.70	37.63	57.50	
Item set for startle					
15. It startled me.	Participant	183.28	12.77	2.89	0.50
	Video	744.79	51.91	117.58	
16. It immediately made me feel scared or angry.	Participant	178.25	19.09	3.02	0.33
	Video	283.90	30.40	48.15	
17. It shocked me.	Participant	277.65	27.34	4.56	0.25
	Video	251.41	24.75	41.33	
18. It stunned me.	Participant	334.10	35.11	5.74	0.18
	Video	152.30	16.00	26.19	
19. It made me physically flinch.	Participant	174.44	12.31	2.53	0.47
	Video	692.17	48.83	100.51	
20. It caused my heart to suddenly beat harder or faster.	Participant	225.14	17.50	3.52	0.41
	Video	550.00	42.75	86.01	
21. It immediately caused stress or frustration to me.	Participant	221.57	24.46	3.85	0.28
	Video	223.59	24.68	38.83	
Average	Participant		**20.50**		
	Video		**26.84**		

The factor structure of the items set was then analysed by performing a multilevel exploratory factor analysis (ML-EFA) [152] with a repeated-measures design which was clustered per video clip. The factor analysis was performed both on the within (video)-level (i.e., variation from differences between participants) and the between (video)-level (i.e., variation from differences between video clips). An oblique, direct oblimin rotation was used to allow the factors to be correlated [153].

Factor extraction at the within- and between-level was conducted based on: (a) eigenvalues greater than 1.0 (Kaiser's criterion) [154], (b) unique loadings of 0.400 and above, and (c) exclusion of items with cross-loadings > 75%. In addition, the scree plot was examined to help inform a decision about the number of factors to retain. Items were removed one at a time until the loadings of all remaining items were > 0.400, and cross-loadings were < 75%. Also, the proportion of the total variance explained by the retained factors must bes greater than 50% [155]. Items were excluded from the final inventory if they showed insufficient loading on factors in either the within or between-level.

The goodness of model fit was assessed using the χ^2 test [156], Comparative Fit Index (CFI) [157], Tucker–Lewis Index (TLI) [158], Root Mean Square Error of Approximation (RMSEA) [159] and Standardized Root Mean Square Residual (SRMR) [160]. Acceptable model fit was indicated by a non-significant χ^2, CFI and TLI values above 0.95, as well as RMSEA and SRMR (both within- and between-level) below 0.10 [161].

All analyses were performed using the Mplus software version 8.10 [162]. Observations on Likert scales were set as ordered categorical (ordinal) variables instead of continuous for factor analysis on both levels [163].

To test concurrent validity of the Startle-VAS and Surprise-VAS, Spearman correlations were computed by comparing with the averaged scores of the two factors retained in Phase 2, respectively. Correlations were computed over predicted startling or surprising stimuli, considering wider startle or surprise range of observations. Predicted non-surprising or non-startling stimuli were not included in this analysis because the expected low variation in observations would bias the correlation results. Spearman's correlations $\rho > 0.30$, $p < 0.01$, were considered significant for establishing validity [164].

4.3. Results

4.3.1. Phase 1: items set generation and content validity

The initial set of formulated items, along with the percentage of experts indicating their relevance, is presented in Table 4.1. Two items, Item 11 ("It made my jaw drop.") and Item 14 ("It made me feel wide-eyed."), were removed based on a relevance threshold of 50%, as both were originally intended to capture the construct of surprise. In response to open-ended expert comments, several items were revised, Item 9 (originally "I was confused about why it happened."), Item 16 (originally "It made me suddenly feel scared or angry."), and Item 21 (originally "It caused a quick burst of stress or frustration in me."). For detailed expert comments, please refer to Appendix A.

Table 4.5: Final factor loadings from the multilevel exploratory factor analysis (ML-EFA).

Item	Within (video)-level		Between (video)-level	
	Factor 1	Factor 2	Factor 1	Factor 2
15. It startled me.	**0.867**	-0.018	-0.053	**0.990**
19. It made me physically flinch.	**0.944**	-0.054	-0.070	**0.983**
20. It caused my heart to suddenly beat harder or faster.	**0.841**	0.006	-0.101	**0.971**
16. It immediately made me feel scared or angry.	**0.894**	0.002	-0.027	**0.994**
17. It shocked me.	**0.689**	0.284	0.356	**1.013**
21. It immediately caused stress or frustration to me.	**0.857**	-0.019	-0.212	**0.945**
1. It surprised me.	0.155	**0.774**	**0.999**	0.143
2. It was consistent with my expectation.[a]	-0.049	**0.925**	**0.969**	-0.122
5. I predicted it beforehand.[a]	-0.153	**0.970**	**0.935**	-0.201
7. I did not see it coming.	0.087	**0.828**	**1.000**	-0.015
13. It was unexpected.	0.116	**0.850**	**1.011**	0.058

Note. Factor loadings above 0.400 are in bold.
[a] Item is reverse-coded.

4.3.2. Phase 2: multilevel exploratory factor analysis

Two-way ANOVA and ICCs

The two-way ANOVA results, shown in Table 4.4, reveal that the variation explained by Factor Video was generally larger than that by Factor Participant (26.84% > 20.50%). As a consequence, data were then clustered over video clips for the ML-EFA.

Intraclass correlation coefficients (ICCs) [165] were estimated for each item, which indicate the proportion of variation in responses to each item that is due to differences between videos. The ICCs are shown in the rightmost column in Table 4.4. Although most of the variation in these items was due to differences within videos, rather than between video clips (i.e., all ICCs < 0.5), there was considerable variation caused by different video clips (i.e., ICC > 0.05) [152]. The differences between ICCs also hint at possible differences in the outcomes of the following within- and between-level exploratory factor analysis.

Multilevel exploratory factor analysis

The final ML-EFA solution with two within-level factors and two between-level factors is shown in Table 4.5. In the within-level analysis, Items 8 and 10 were removed as they loaded on more than one factor with loadings exceeding 0.400. Item 12 was excluded due to high cross-loading. Items 4, 6 and 9 were removed due to loading on a third factor. In the between-level analysis, Items 3, 8, 10, 12 and 18 were removed because these items loaded on more than one factor with loadings exceeding 0.400. The remaining items loaded on the two expected factors, which were the same as found in the within-level. Factors 1 and 2 from the within-level factor analysis mapped on to the constructs of Startle and Surprise.

In this solution, the largest factor loading for each item at the within-level ranged from 0.689 to 0.970, and at the between-level from 0.935 to 1.013, suggesting meaningful and

Table 4.6: Cronbach's α for Startle-I and Surprise-I across video clips.

ID	Startle-I	Surprise-I
Tumbler	0.815	0.911
Clouds	0.859	0.909
Monster1	0.914	0.955
Monster2	0.929	0.843
Pill	0.714	0.935
Spider	0.884	0.910
Puppy	0.903	0.856
Baseball	0.754	0.936
Window	0.845	0.862

significant factor loadings. Further evidences for the goodness of model fit are the non-significant χ^2 test with $\chi^2 = 51.508$, $df = 68$, $p = 0.932$, CFI = 1.000, TLI = 1.211, RMSEA = 0.000 and SRMR of 0.046 and 0.008 at within-level and between-level, respectively.

At the within-level, an 11-item two-factor solution explained a total of 78.57% of the variance, with Factor 1 (Startle) contributing to 53.26% and Factor 2 (Surprise) contributing to 25.31% of the variance. The correlation between these two factors was positive and significant, $\rho = 0.316$, $p < 0.001$. At the between-level, the 11-item two-factor solution explained a total of 96.72% of the variance, with Factor 1 (Surprise) contributing to 62.26% and Factor 2 (Startle) contributing to 34.46% of the variance. The correlation between these two factors was not significant, ρ = -0.199, $p = 0.637$. Items loading on Factors 1 and 2 at the within-level (i.e., Factors 2 and 1 at between-level) will henceforth be referred as the Startle Inventory (Startle-I) and Surprise Inventory (Surprise-I).

Data from the 81 participants were also used to assess the reliability of the multi-item inventories. Cronbach's α coefficients suggested acceptable to excellent internal consistency across video clips, ranging from $\alpha = 0.714$ to $\alpha = 0.929$ for the Startle-I, and from $\alpha = 0.843$ to $\alpha = 0.955$ for the Surprise-I (see Table 4.6).

Another ML-EFA was conducted treating the data as continuous rather than ordinal, using the same factor extraction criterion. The results were consistent with the ordinal analysis, with the exception that Item 18 loaded onto the Factor Startle at the between level. Notably, this item had been just below the inclusion threshold in the ordinal analysis with a factor loading of 0.409, which is only slightly larger than the exclusion threshold of 0.400.

4.3.3. Phase 3: Visual Analogue Scales for Startle and Surprise

One participant's VAS scores for the Tumbler and Clouds videos were missing, resulting in one fewer data point for these video clips. Table 4.7 presents the Spearman's correlations between the Startle-VAS and Startle-I, and Surprise-VAS and Surprise-I for stimuli predicted to elicit startle and/or surprise responses.

The Startle-VAS ratings highly significantly correlated with the Startle-I ratings, ranging from $\rho = 0.778$ to $\rho = 0.877$, $p < 0.001$. Similarly, for surprise responses, high correlations were observed between the Surprise-VAS and Surprise-I ratings, ranging from $\rho = 0.681$ to $\rho = 0.903$. All correlations were highly significant, $p < 0.001$.

Table 4.7: Correlations between Startle-VAS and Startle-I, and between Surprise-VAS and Surprise-I.

ID	Startle-VAS vs. Startle-I	Surprise-VAS vs. Surprise-I
Tumbler	-	0.729***
Clouds	-	0.743***
Monster1	0.797***	0.903***
Monster2	0.877***	-
Pill	-	0.681***
Spider	0.778***	0.723***
Puppy	-	-
Baseball	-	0.747***
Window	-	-

*** $p < 0.001$ (2-tailed).

4.3.4. Manipulation checks

For the inventories, the scores of all items in each inventory were averaged to obtain a total score ranging from 1 to 5. The responses of startle and surprise measured by inventories and VASs over nine stimuli are shown in Figure 4.1 as pirate plots. These plots present the mean values (square markers with labels), interquartile range (IQR) in black lines, and distribution of ratings across different scenarios. The plots illustrate that the ratings for startle were consistent across video stimuli on the inventory and the VAS. However, surprise ratings on the VAS were systematically lower than those on the inventory.

The mean startle and surprise ratings (with standard deviations) for each video across participants are shown in Figure 4.2a based on the Startle and Surprise Inventories, and in Figure 4.2b derived from the Visual Analogue Scales for Startle and Surprise. The selected stimuli vary in the level of startling and surprising as intended, indicating the selection of stimuli to evoke the desired responses was generally successful. The high variation in responses to each video facilitates the application of ML-EFA.

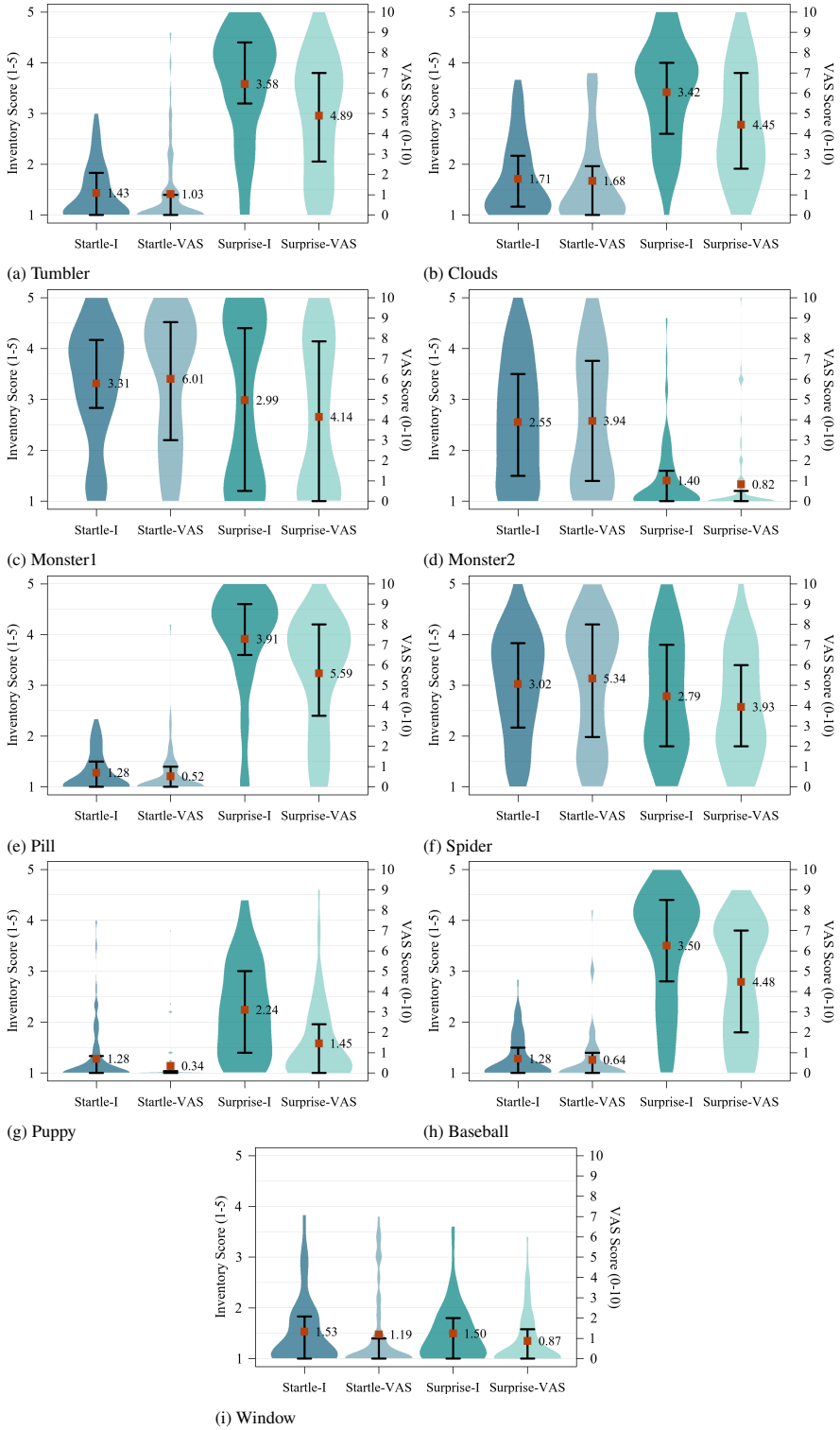
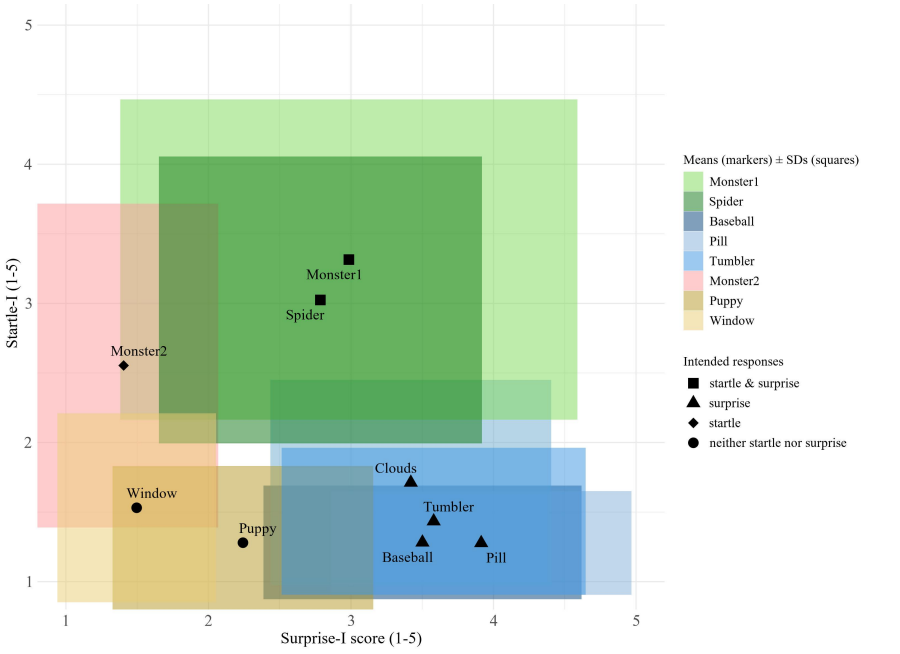
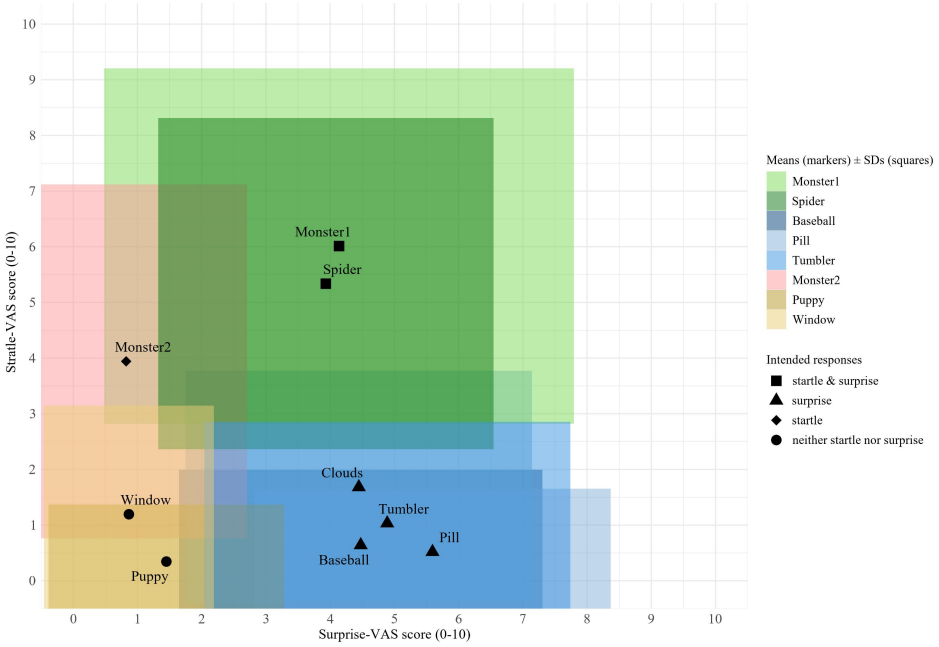


Figure 4.1: Startle and surprise responses measured by inventories and VASs across nine video clips (square markers indicate means, whiskers indicate interquartile range).



(a) Startle and Surprise Inventories



(b) Visual Analogue Scales for Startle and Surprise

Figure 4.2: Mean values (markers) and SDs (squares) of startle and surprise ratings across nine video clips.

4.4. Discussion

The purpose of this chapter was to develop and validate self-report measures of startle and surprise in human factors research. From three sequential phases, a Startle Inventory, a Surprise Inventory, and the more efficient Visual Analogue Scales for Startle and Surprise were developed and preliminarily validated. The initial items set was formulated based on fundamental and applied literature, content validity was tested by asking seven experts to rate the items' relevance (Phase 1). The construct validity of the retained items (Phase 2) and the concurrent validity of the visual analogue scales (Phase 3) were tested by obtaining ratings (N = 81) for nine stimuli which varied in extent of being startling or surprising.

The construct of Startle was supported by the within-level and between-level exploratory factor analysis, and contains six items: "It startled me.", "It made me physically flinch.", "It caused my heart to suddenly beat harder or faster.", "It immediately made me feel scared or angry.", "It immediately caused stress or frustration to me." and "It shocked me.". In line with literature [20, 22, 23], these items refer to physiological as well as psychological aspects of startle. In Phase 1, two items were worded slightly differently based on the experts' feedback. In Phase 2, only the Item 18 "It stunned me." was removed. Items in the construct of Startle possessed high internal consistency, $\alpha_{\text{Startle}} = 0.714$ to $\alpha_{\text{Startle}} = 0.929$.

The construct Surprise stemmed from items loading in both the within-level and between-level analysis, containing five items: "It surprised me.", "It was consistent with my expectation." (reverse-coded), "I predicted it beforehand." (reverse-coded), "I did not see it coming.", and "It was unexpected.". In Phase 1, Items 11 and 14 referring to (sensed) facial expressions of surprise were removed. In Phase 2, Items 3, 4, 6, 8, 9, 10, 12 were removed. Differences in interpreting the wording may have caused relatively high contribution of individual differences in Item 3 ("taken aback"), Item 8 ("astonished"), Item 10 ("bewildered"), Item 17 ("shocked") and Item 18 ("stunned"), as seen from the relatively low ICCs in Table 4.4. These items, except for Item 17, were removed based on the factor extraction criterion. "Feeling not mentally prepared" (Item 12) was removed due to high cross-loading. Items in the construct of Surprise possessed high internal consistency, $\alpha_{\text{Surprise}} = 0.843$ to $\alpha_{\text{Surprise}} = 0.955$. Interestingly, the item "astonished", which was derived from the existing DES-IV and PANAS-X, was removed in our analysis.

Item 4 "I did not understand why it happened.", Item 6 "Initially, it made no sense to me.", and Item 9 "Initially, I was confused about it." were removed as they loaded on a third factor at the within-level, even though they clustered with the construct Surprise at the between-level. Apparently, these three items were related to the experience of surprise when different videos were compared, but not when different participants within one video were compared. This finding conflicted with our hypothesis that surprise would be characterized by a (brief) moment of confusion and requirement to reframe [35, 44, 77]. Possibly, some participants were more likely than others to indicate incomprehension by surreal videos (e.g., Monster1) rather than surprise. Others may have rated such videos as easy to understand as they took the scripted nature of these videos into account instead of the realism of the events. Therefore real (non-scripted) events are recommended in future research to better control for this possibility. Another potential cause of this third factor is that some participants may have experienced cognitive impairment due to startle. Some participants were observed to reflexively jump up and raise their hands when watching the Monster1 or Spider, and were possibly likely to indicate high initial incomprehension as well as high startle.

In Phase 3, the Startle-VAS and Surprise-VAS were developed and tested as efficient alternatives to the Startle-I and Surprise-I. High consistencies were observed between the VAS and inventory measures for both constructs. However, without further validation, it remains uncertain whether the VASs are more or less accurate than the inventories in assessing startle and surprise. While the scores for inventory and VAS over stimuli were similar for startle, surprise was systematically rated lower on the VAS total range compared to the inventory total range (Figure 4.1). This could be caused by the two reverse-coded inventory items (Items 2 and 5) measuring the experienced predictability. An event may be experienced as unsurprising but also as unpredictable, leading to a higher score on the VAS than on the inventory. A visual analogue scale ranging from "highly expected" to "highly surprising" could possibly be more aligned with the inventory scores.

In this chapter, we examined responses to an item set administered across multiple video stimuli designed to elicit varying levels of startle and surprise, while considering dependence between responses from the same participant. From ML-EFA, within-level and between-level variations were properly taken into account. The outcomes of the two-way ANOVA and ICCs illustrate the necessity of ML-EFA, in which the within- and between-level factor analysis may capture different constructs. The structure of the items was explored and the items set was reduced until satisfactory loading on factors was achieved using data collected in a repeated-measures context. The video stimuli were generally successful in eliciting the desired startle or surprise responses (Figure 4.2) and leading to sufficiently high variation between participants and videos, such that the correlation structure between inventory items could be analysed in an extensive manner.

From a compositional standpoint, the Startle and Surprise Inventories are the first that ground the experience of startle and surprise to specific stimuli or event. Since startle and surprise responses have a potential impact on performance and negatively affect safety, the developed measures may help to better distinguish the definitions of these responses as used in operational practice, for instance in the domain of aviation [19]. In addition, the measures are useful to further test and explain some of the (ambiguous) findings in the literature.

For the Startle-I and Surprise-I, content validity and construct validity have been examined, and high internal consistency was found in both inventories. For the Startle-VAS and Surprise-VAS, concurrent validity was tested by comparing with the Startle-I and Surprise-I, respectively. Future research could explore the criterion-related validity of both measures by comparing outcomes with those of objective measures, such as physiological responses (e.g., electromyography, gaze behaviour [166]) or behavioural markers (e.g., reaction time, micro-expressions). Additionally, operational relevance should be further tested by stimuli presented in more ecologically-valid contexts. Test-retest reliability could be performed for the startle measures, with sufficient time between stimuli to account for habituation. For surprise, such checks do not seem feasible, at least not with the same stimuli, as surprise depends by definition on novelty and unexpectedness.

4.5. Conclusion

Previous self-report measures of startle and surprise lacked systematic development and psychometric validation, resulting in suboptimal assessments. To address this gap, we introduced the Startle-I and Surprise-I, which were designed using a systematic construction process aimed at improving the validity and reliability of self-report startle and surprise.

These new measures provide a more robust foundation for the quantitative assessments. The Startle-VAS and Surprise-VAS were developed as rapid and efficient alternatives. The developed measures can be applied to test the effects of startle and surprise on performance, to check the effectiveness of startle and surprise exposure training or testing scenarios.

5

Multilevel confirmatory factor analysis of the Startle and Surprise Inventories

Following the initial development and preliminarily validation of the Startle and Surprise Inventories in Chapter 4, further psychometric evaluations in aviation ecologically-valid settings are necessary. This chapter discusses an attempt to validate the factor structure of the Startle and Surprise Inventories using a multilevel confirmatory factor analysis. The relatively ecologically-valid setting is used, rather than the highly-controlled but less realistic laboratory conditions applied in Chapter 4. In addition, the two-way ANOVA is performed and intraclass correlation coefficients are calculated, to assess within- and between-level variability. Manipulation checks are conducted to ensure the effectiveness of the experimental conditions. A sample of 26 professional pilots is exposed to eight scenarios with varied levels of startling and surprising in the SIMONA simulator.

The chapter is organized as follows: Section 5.1 introduces the objectives and rationale for conducting multilevel confirmatory analysis. Section 5.2 describes the methodology, and Section 5.3 presents the results. Section 5.4 discusses the findings in relation to theoretical and psychometric considerations, and Section 5.5 concludes the chapter with a summary of key results and implications.

The contents of this chapter have been accepted for publication in *Journal of Cognitive Engineering and Decision Making* [167].

5.1. Introduction

The previous chapter discussed the development of the multi-item Startle and Surprise Inventories (Startle-I; Surprise-I) and the single-item Visual Analogue Scales for Startle and Surprise (Startle-VAS; Surprise-VAS) [168]. The Startle-I consists of six statements and the Surprise-I consists of five statements, in which each describe a response to the targeted stimulus. Individuals indicate to what extent they agree with the statement on a 5-point Likert-type scale (1 = "*Strongly disagree*", 2 = "*Disagree*", 3 = "*Neutral*", 4 = "*Agree*", 5 = "*Strongly agree*"). The more time-efficient Startle-VAS and Surprise-VAS require individuals to answer the question, "*How **startled/surprised** were you by [the stimulus]*?". Each VAS consists of a 10 cm horizontal line with tick marks at each 1 cm interval, ranging from 0 to 10. The left endpoint is labelled "*not startled/surprised at all*" and the right endpoint is labelled "*extremely startled/surprised*", respectively.

The development of self-report measures started with a content validation phase, in which seven experts in the fields of Cognitive Science and Psychology assessed the relevance of 21 items, which were derived fundamental and applied literature on startle [20, 23, 148–150, 169] and surprise [19, 30, 35, 49, 77]. Based on this assessment, 19 items retained for multilevel exploratory factor analysis (ML-EFA). A group of 81 participants were exposed to nine video stimuli, designed to elicit a wide range of startle and surprise responses. The ML-EFA resulted in an 11-item two-factor structure. The structure delineated the constructs of Startle and Surprise and demonstrated high internal consistency, with Cronbach's α ranging from 0.714 to 0.928 for Startle-I, and from 0.843 to 0.955 for Surprise-I, across the nine video stimuli. Further concurrent validity testing revealed significant correlations between the scores from the Startle-VAS and Surprise-VAS, and the Startle-I and Surprise-I, respectively. These correlations ranged from 0.778 to 0.877 for Startle-VAS, and from 0.681 to 0.903 for Surprise-VAS across the video stimuli, providing empirical support for the visual analogue scales.

Both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) aim to determine latent factors which optimally account for the variance-covariance among observed variables or indicators. The EFA is instrumental in uncovering the latent factor structure within a set of observed variables without a predefined structure. As a data-driven method EFA requires no a priori assumptions, but it cannot be applied to test hypothesized factor structures against the observed data [170]. In contrast, CFA adopts a hypothesis-driven approach to test hypothesized factor structures while assessing the model fit to the data. This analysis incorporates various goodness-of-fit indices, including the χ^2 test [156], Comparative Fit Index (CFI) [157], Tucker–Lewis Index (TLI) [158], Root Mean Square Error of Approximation (RMSEA) [159] and Standardized Root Mean Square Residual (SRMR)[171]. Construct validity is established through CFA by evaluating whether the observed data align with the hypothesized factor structure, ensuring that the measures adequately represent the theoretical constructs they are designed to measure.

Multilevel confirmatory factor analysis (MCFA) extends confirmatory factor analysis to accommodate hierarchical or clustered data structures [160, 161, 172]. Such structures often arise from individuals nested within groups, or repeated measures nested within individuals. In these contexts, the assumption of independent observations is violated, leading to inflated Type I error rates if clustering is ignored. The "clustering effect" is typically quantified using the intraclass correlation coefficient (ICC), which estimates the proportion of total variance

Table 5.1: Characteristics of the participants (N = 26).

	Mean	***SD***	**Min**	**Max**
Age (yrs)	43.8	13.0	23.0	67.0
Employed time (yrs)	17.7	13.3	0.5	44.0
Flight hours (hrs)	8,633.9	7,082.1	280.0	25,500.0
Large aircraft	6,566.1	6,607.8	0.0	22,000.0
Business jet	1,257.7	2,803.7	0.0	10,000.0
Small aircraft	810.1	1,258.4	0.0	5,000.0

attributable to between-cluster variance [152]. In intervention studies with nested designs, an ICC as low as 0.05 can significantly impact statistical power [173]. By accounting for the clustered nature, MCFA provides more precise and reliable estimates of factor loadings and variances, minimizing biases introduced by data non-independence [174]. This method is essential for validating constructs across clustered levels, ensuring that the measures accurately reflect the constructs at both the within- and between-cluster levels [160].

In this chapter, we aim to validate the factor structure of the Startle and Surprise Inventories in an ecologically-valid setting through MCFA. Based on the two-factor model identified through ML-EFA in Chapter 4, we hypothesize that the two-factor structure will exhibit good model fit across diverse scenarios, both within and between clusters. A representative sample group of professional pilots was recruited, and simulated in-flight scenarios designed to elicit a wide range of startle and surprise responses, were employed. Events in most of the scenarios were found to be effective in eliciting startle and surprise responses in pilots, as reported in Chapter 3, although those findings were based on subjective, non-validated self-report measures. The present construct validation aims to address this limitation by providing a more rigorous empirical foundation for the Startle and Surprise Inventories, with the broader aim of informing the development of evidence-based safety protocols and enhancing the effectiveness of intervention training in aviation contexts.

5.2. Method

5.2.1. Participants

26 currently-employed professional pilots participated in the experiment, comprising 25 males and 1 female. The characteristics of participants are summarized in Table 5.1. Figure 5.1 illustrates the distribution of pilots' flight hours across different aircraft types. All participants possessed either an Airline Transport Pilot License (ATPL) or a frozen ATPL[1] at the time of the experiment. Among them, fourteen worked as captains, eight as first officers, three as second officers, and one employed in a non-airline aviation position. This experiment complied with the American Psychological Association Code of Ethics and the Research Ethics Committee of the Delft University of Technology approved the research design (No. 4056). Informed consents were obtained from all participants.

[1] A frozen ATPL refers to a state in which a pilot has passed all theoretical examinations required under the EASA framework for the ATPL, but has not yet accumulated the 1,500 flight hours required for the licence to be issued.

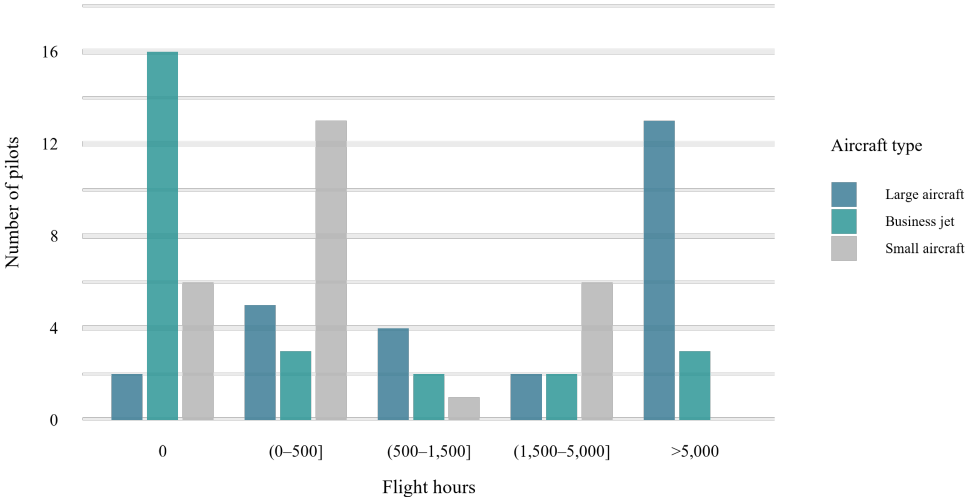


Figure 5.1: Distribution of pilots' flight hours across aircraft types (N = 26). Each pilot may have experience in multiple aircraft types. No pilot reported > 5,000 flight hours on small aircraft.

5.2.2. Apparatus

The experiment was conducted in the SIMONA Research Simulator (see Figure 5.2) at the Delft University of Technology [175]. 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 potential alarms, flaps, retractable gear, aerodynamic noise, ground rumble and engines, which is beneficial for establishing a highly credible operational environment. During the

Figure 5.2: SIMONA Research Simulator.

(a) Daytime setting (b) Night setting

Figure 5.3: Flight deck in the experiment.

experiment, participants wore single-ear intercom headsets (ClearCom CC-110-X4).

The experiment employed an generic aerodynamic model of the Piper PA-34 Seneca III, a light twin-propeller aircraft. The flight deck (see Figure 5.3 for daytime and night settings) featured flight controls including a control column with pitch trim, rudder pedals with force feedback, throttle, gear, and flaps with three settings: 0° (UP), 25° and 40° (LAND). The avionics consisted of a primary flight display (PFD) similar to a G1000 PFD, a backup primary flight display, and a multi-function display for engine, configuration, and navigation data. Information on airspeed, altitude, attitude, engine parameters, flap position, and gear status was available via the avionics displays.

5.2.3. General procedure

An overview of the within-subjects experimental procedure is illustrated in Figure 5.4. Pilots performed the experiment on a single day and as single-pilot crew. The total duration of the experiment per participant was approximately two hours, including the briefing, familiarization, test session (comprising eight test scenarios), and debriefing.

All pilots were briefed about the aircraft model, simulator features, experimental tasks,

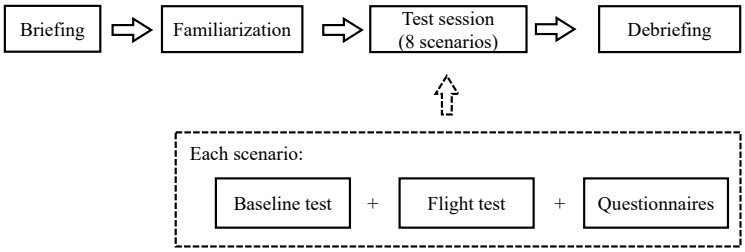


Figure 5.4: Experimental design.

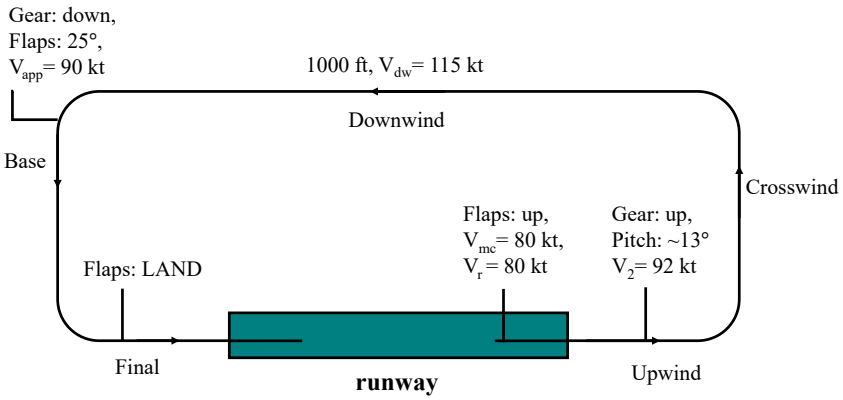


Figure 5.5: Standard traffic pattern with target settings.

and definitions of startle and surprise. Each familiarization and test scenario required the pilot to fly (part of) a left-handed traffic circuit (Figure 5.5) for runway 18C, Schiphol Airport (EHAM). The circuit would need to be flown at 1,000 ft with a speed of 115 kt. A flaps setting of 0° (UP) was required for take-off, 25° during the base leg, and 40° (LAND) in the final leg. The circuit required rotate speed of 80 kt, minimum control speed of 80 kt, best rate of climb speed of 92 kt, and landing approach speed of 90 kt. These configurations were also available on the pilot's kneepad.

During familiarization, pilots also practised one auditory task (details in Section 6.2.3 in Chapter 6) once on the runway, and once while performing the circuit. At the end of the familiarization, all pilots confirmed that they could handle the aircraft model, none required help in determining the turn points of the circuit, and none had difficulty with performing the auditory task.

Pilots then proceeded with the test session, which consisted of eight test scenarios (see Section 5.2.4), presented in a semi-counterbalanced order defined by a Latin square [151]. Test scenarios began either from the take-off position on runway 18C, or in-flight position at 800 ft ahead of runway 18C, with an airspeed of 99 kt. In all cases, participants were required to complete the circuit and land safely on the same runway.

Before each scenario, participants received a briefing on wind strength, direction, and weather code through a Meteorological Aerodrome Report (METAR). Before and during each test scenario, pilots also performed blocks of the auditory task (see Chapter 6). Immediately following each scenario, pilots completed a questionnaire that included the Startle and Surprise Inventories (see Section 5.2.5). After completing all scenarios, pilots were informed about all simulated events in a debriefing.

5.2.4. Startle and surprise events

We aimed to include events that would elicit a comprehensive range of startle and surprise responses. Table 5.2 provides the description of the test scenarios, with the fourth column indicating whether the events in each scenario were designed to elicit startle, surprise, both,

Table 5.2: Description of preset events and intended startle/surprise responses across scenarios.

ID	Event description	Event inserted[1]	Intended responses
LTS	While flying at night, lightning struck the aircraft with a bright flash and loud thunder sound.	The lightning strike	Startle
FLAP	When selecting Flaps 25 in the base leg, the left flap remained UP.	The aircraft response when you selected Flap 25	Surprise
ENF	The right engine failed shortly after takeoff.	The right engine failure during take-off	Surprise
PFDF	The primary flight display screen turned black.	The malfunction of the Primary Flight Display	Surprise
CARGO	Cargo moved backward during takeoff with a loud sound and pitch up motion.	The pitch-up motion (and noise) after rotation	Startle and surprise
STALL	A bird strike triggered a false stall alarm with stick shaker.	The stall alarm	Startle and surprise
NTO	Normal take-off without preset malfunction.	The level-off manoeuvre	Neither startle nor surprise
NLO	Normal landing without preset malfunction.	The effect of crosswind	Neither startle nor surprise

[1] The specific content that replaced the placeholder "*[the stimulus]*" in the Startle and Surprise Inventories and Visual Analogue Scales for Startle and Surprise.

or neither. Events that were rare, unfamiliar, or difficult to immediately explain, were used to induce relatively high surprise [30, 35, 49, 77]. Events that were sudden, loud, or immediately threatening were used to induce relatively high startle [20, 23, 149, 150]. More "high surprise" than "high startle" scenarios were included because we expected difficulty with surprising a sufficient proportion of pilots to induce sufficient variation for analysis.

The lightning strike scenario (LTS) was designed to be highly startling due to the sudden bright flash and loud thunder sound, but not (limited) surprising due to the stated weather conditions. This scenario took place at night and began in the in-flight position. During landing (5 s after descending 500 ft), a lightning strike was simulated, accompanied by a loud thunder sound and a flash of light. The thunder was simulated using a surround sound system in the simulator, playing the sound at 99 dB (including ambient noise). The lightning flash was simulated using a strobe light mounted on the projection system. It overpowered the regular night-time outside visual with a strong flash over the entire field of view of the out-the-window view. To limit surprise, the METAR for this session included "TSRA" (thunderstorm with rain), signalling the potential for thunderstorms.

In the flap asymmetry scenario (FLAP), the left flap remained in the UP position when pilots selected 25° during the turn to the base leg. This caused an unexpected roll and yaw moment, which need to be counteracted using the control column. The pilots' unfamiliarity with the specific aircraft response to the flap failure was expected to induce surprise. Since there was no immediate threat nor intense stimuli, a limited startle response was expected.

In the engine failure scenario (ENF), the right engine failed during takeoff (5 s after reaching 900 ft), causing a roll and yaw moment that could be counteracted using the column and pedals. The pilots' unfamiliarity with the specific aircraft response to the engine failure was expected to induce surprise. Since there was no immediate threat nor intense stimuli, a limited startle response was expected.

The primary flight display failure scenario (PFDF) started at the in-flight position. The PFD malfunctioned and went black during landing at 600 ft. Pilots could use the outside view or the backup flight display to continue landing. Due to its unfamiliar nature, this scenario was intended to elicit a high level of surprise. However, in the absence of intense stimuli or immediate threat, a limited startle response was expected.

In the cargo shift scenario (CARGO), a simulated piece of heavy cargo broke loose and shifted towards the tail after take-off (10 s after reaching 200 ft), with a loud scraping and collision noise coming from the back of the aircraft. This moved the centre of gravity backward temporarily, resulting in a violent pitch-up moment that was difficult to counteract. The novelty of the event and difficulty in explaining it were expected to induce high levels of surprise, while the sudden (upset) motion of the aircraft and the accompanying loud scraping noise were expected to induce startle response.

In the false stall warning scenario (STALL), a bird strike occurred during the climbing (20 s after reaching 800 ft), impacting the angle of attack vane with a sharp noise. This triggered a continuous false stall alarm consisting of a stick shaker and stall aural warning. Due to the lack of context for a stall event, it was expected to be surprising, and due to the sudden loud auditory stall alarm and stick shaker it was expected to be startling.

Finally, two more scenarios, normal take-off (NTO) and normal landing (NLO), were included to present events that were expected to induce low levels of startle and surprise. NTO started at the take-off position and NLO started at the in-flight position, with pilots performing a landing under a 5 kt crosswind from the east.

5.2.5. Measures of startle and surprise

Participants were instructed to indicate their experienced startle and surprise after each scenario on the Startle-I, Surprise-I, Startle-VAS and Surprise-VAS (see Appendix B). The six items from the Startle-I and the five items from the Surprise-I were randomized in order (Table 5.3) so that participants had no information on whether items were intended to measure startle or surprise. The total scores for the Startle-I and Surprise-I were calculated by averaging the scores of all items within each inventory (ranging from 1 to 5). Participants were required to place a cross on the Startle-VAS and Surprise-VAS as ratings and the resulting scores were the distance of the cross to the left endpoint, measured in centimetres (ranging from 0 to 10). The preset event in each scenario, as inserted in the inventories and VASs, is stated in the third column of the Table 5.2.

Table 5.3: Restructured Startle and Surprise Inventories.

Items
1. It startled me.
2. It surprised me.
3. It immediately made me feel scared or angry.
4. I predicted it beforehand.[a]
5. It made me physically flinch.
6. It was consistent with my expectation.[a]
7. It caused my heart to suddenly beat harder or faster.
8. I did not see it coming.
9. It shocked me.
10. It was unexpected.
11. It immediately caused stress or frustration to me.

Note. Items 1, 3, 5, 7, 9, 11 are from the Startle Inventory.
Items 2, 4, 6, 8, 10 are from the Surprise Inventory.
[a] Item is reverse-coded.

5.2.6. Statistical analysis

Two-way ANOVA and ICCs

First, the scores on items 4 and 6 were reversed. To determine whether the data from Startle-I and Surprise-I for the MCFA should be clustered over scenarios or participants, a two-way ANOVA was performed to examine the relative amount of variance on each item attributed to the factor Scenario and the factor Participant.

To assess the proportion of between-level variance relative to the total variance, ICCs were calculated for each item. The ICCs served to evaluate whether a MCFA was necessary instead of a single-level CFA, as sufficiently high between-level variance [173] (ICC > 0.05) justifies the use of MCFA to account for clustered data structure.

Multilevel confirmatory factor analysis

The factor structure of the 11 items in the Startle-I and Surprise-I was analysed using MCFA for the clustered dataset [176]. MCFA was performed using the `lavaan` package in R to test and compare two models: an 11-item, two-factor model comprising the factors Startle and Surprise as identified in a ML-EFA [168], and an 11-item, one-factor model in which all 11 items are considered as variables of a single factor.

Model goodness-of-fit was considered acceptable if the χ^2 value was non-significant [156], the CFI and TLI exceeded 0.90 [157], and RMSEA [159] and SRMR (at both within and between level) are below 0.10 [171]. To further investigate the relationship between the factors Startle and Surprise in the two-factor model, standardized covariances between the two factors were computed at both levels of analysis. Internal consistency was assessed by calculating McDonald's ω [177] for the Startle-I and Surprise-I for each scenario, with values greater than 0.70 indicating acceptable internal consistency [178, 179].

Manipulation checks
To check whether designed scenarios induced the intended responses of startle and surprise, a manipulation check was conducted. Two linear mixed-effects models were applied to account for the repeated-measures design, with heteroscedasticity accounted for in the residuals. The scores obtained from the Startle-I (*startle_inventory*) and the Surprise-I (*surprise_inventory*) were modelled as functions of the stimulus (*stimulus*) in the test scenarios, a categorical fixed effect with eight levels. Participants with assigned identifier numbers (*ID*) were included as a random effect to account for the individual differences. The function `lme` from the `nlme` package in `R` was applied to fit the following models:

$$Response = 1 + stimulus + (1|ID),$$

where the variable *Response* was *startle_inventory* or *surprise_inventory* during the corresponding analysis. The significance level was set as $p < 0.05$ and the p values were adjusted for multiple comparisons with the Tukey method, for each model separately. Heteroscedasticity across scenarios was modelled using the `varIdent` function, allowing variance components to differ by scenarios.

5.3. Results

5.3.1. Two-way ANOVA and ICCs

Results from the two-way ANOVA revealed that the average amount of variance for each item explained by the factor Scenario was larger than that explained by the factor Participant for both the Startle-I (21.10% > 0.37%) and Surprise-I (18.47% > 0.09%) as shown in Table 5.4. Thus, data were clustered over scenarios for the MCFA.

The ICCs ranged from 0.34 to 0.74 (rightmost column in Table 5.4). Notably, *all* of the items have ICCs greater than 0.05, indicating considerable variance is due to the between-scenario differences. To be specific, given that nearly half of the variance on some items is between-scenario variance, MCFA was needed to properly investigate the factor structure.

5.3.2. Multilevel confirmatory factor analysis

Model fit tests demonstrated that the two-factor 11-item model, comprising the factors Startle and Surprise, provided an adequate goodness-of-fit to the data across all indices except for the χ^2 test, with $\chi^2 = 153.760$, $p < 0.001$, CFI = 0.939, TLI = 0.922, RMSEA = 0.062, $SRMR_{within} = 0.089$, $SRMR_{between} = 0.082$. In contrast, the one-factor, 11-item model showed significantly lower fit indices and failed to meet the criteria for goodness-of-fit on all indices, with $\chi^2 = 530.969$, $p < 0.001$, CFI = 0.602, TLI = 0.503, RMSEA = 0.156, $SRMR_{within} = 0.210$, $SRMR_{between} = 0.133$.

In the MCFA of the two-factor model (as shown in Table 5.5), all items loaded significantly on their respective factors (i.e., absolute Z values were greater than 1.96 at a 95% confidence level). For the factor Startle, standardized loadings ranged from 0.493 to 0.694 at the within level, and 0.490 to 0.955 at the between level. For the factor Surprise, standardized loadings ranged from 0.347 to 0.855 at the within-scenario level, and ranged from 0.913 to 1.019 at the between-scenario level.

The standardized covariance indicated a non-significant low to moderate positive relationship $Cov. = 0.171$, $p = 0.067$, between the factors Startle and Surprise at the within-

Table 5.4: Two-way ANOVA results and estimated intraclass correlation coefficient (ICC) for each item.

Item	Factor	Sum of square	Variation (%)	F	ICC
Startle-I					
Item 1	Participant	1.01	0.29	0.87	0.62
	Scenario	114.43	32.42	98.77	
Item 3	Participant	0.27	0.14	0.36	0.34
	Scenario	31.34	16.83	41.55	
Item 5	Participant	0.29	0.10	0.25	0.39
	Scenario	62.38	20.95	54.40	
Item 7	Participant	1.96	0.69	1.87	0.44
	Scenario	66.02	23.36	63.06	
Item 9	Participant	0.25	0.09	0.23	0.44
	Scenario	57.69	20.77	53.79	
Item 11	Participant	2.29	0.88	2.07	0.42
	Scenario	31.86	12.25	28.90	
Average	Participant		**0.37**		
	Scenario		**21.10**		
Surprise-I					
Item 2	Participant	0.04	0.01	0.03	0.74
	Scenario	107.11	31.00	92.11	
Item 4	Participant	0.60	0.14	0.33	0.48
	Scenario	57.23	13.36	31.68	
Item 6	Participant	0.21	0.06	0.14	0.52
	Scenario	40.96	11.77	27.37	
Item 8	Participant	0.32	0.08	0.19	0.57
	Scenario	71.28	17.19	42.58	
Item 10	Participant	0.53	0.14	0.36	0.62
	Scenario	71.03	19.03	48.27	
Average	Participant		**0.09**		
	Scenario		**18.47**		

scenario level, representing that pilots who tended to report high startle did not necessarily report high surprise within the same scenario, and vice versa. A high but marginally-significant covariance value between the factors Startle and Surprise at the between-scenario level $Cov.$ = 0.902, p = 0.063, indicated that scenarios that were rated higher in startle also tended to be rated higher in surprise, and vice versa.

McDonald's ω testing indicated acceptable to excellent internal consistency for both inventories across scenarios, with values of ω = 0.88 to ω = 0.96 for the Startle-I, and ω = 0.77 to ω = 0.96 for the Surprise-I (Table 5.6).

Table 5.5: Factor loadings, standard errors and Z values from MCFA of the two-factor model.

Factor	Item	Within (scenario)-level			Between (scenario)-level		
		Loading[c]	SE	Z	Loading[c]	SE	Z
Startle	Item1	0.493			0.955		
	Item3	0.623	0.190	6.658	0.490	0.034	15.153
	Item5	0.694	0.205	6.871	0.665	0.058	11.994
	Item7	0.647	0.186	7.050	0.699	0.054	13.451
	Item9	0.646	0.189	6.942	0.702	0.032	22.794
	Item11	0.694	0.200	7.049	0.625	0.098	6.662
Surprise	Item2	0.347			1.019		
	Item4	0.723	0.375	5.555	0.913	0.084	10.690
	Item6	0.657	0.335	5.651	0.856	0.078	10.811
	Item8	0.855	0.402	6.133	0.996	0.079	12.426
	Item10	0.801	0.378	6.106	0.980	0.081	11.849

[c] Standardized loading.

Table 5.6: McDonald's ω for Startle-I and Surprise-I across scenarios.

ID	Startle-I	Surprise-I
LTS	0.93	0.96
FLAP	0.92	0.90
ENF	0.88	0.86
PFDF	0.91	0.77
CARGO	0.90	0.89
STALL	0.90	0.80
NTO	0.96	0.94
NLO	0.94	0.93

5.3.3. Manipulation checks

The ratings from the Startle-I, Surprise-I, Startle-VAS, and Surprise-VAS are shown in pirate plots per scenario (Figure 5.6). Each plots represents the mean values (square markers with labels), interquartile range (IQR) in black lines, and distribution of the ratings. The two left beans in each plot represent ratings from the Startle-I and Surprise-I, referring to the left-hand axis (ranging from 1 to 5). The two right beans represent ratings from the Startle-VAS and Surprise-VAS, corresponding to the right-hand axis (ranging from 0 to 10).

The plots illustrate that, across all test scenarios, the Startle-I scores were consistently lower than the Startle-VAS scores, whereas the Surprise-I scores were consistently higher than the Surprise-VAS scores. This suggests that the inventories differentiated better between experienced startle and surprise than the VASs. In addition, the selected scenarios elicited a wide range of startle and surprise levels within and between scenarios, demonstrating the overall effectiveness of the scenarios in provoking the intended responses. The high variation in responses also facilitates the application of MCFA.

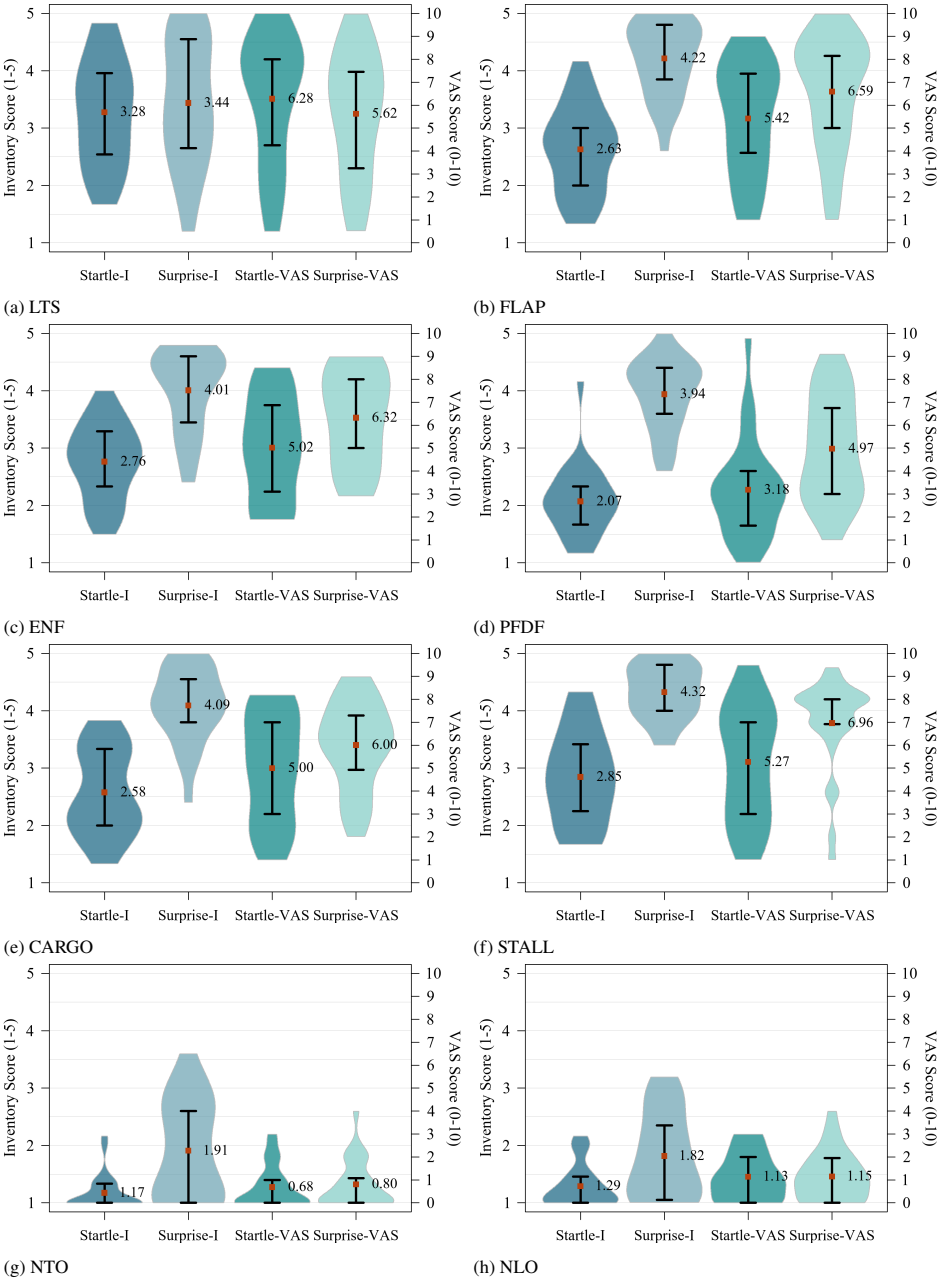


Figure 5.6: Ratings from Startle-I, Startle-I, Startle-VAS and Surprise-VAS across scenarios (square markers indicate means, whiskers indicate interquartile range).

Figure 5.7 illustrates the levels of startle and surprise across scenarios. Dashed circles highlight groups of scenarios that did not differ significantly in elicited startle and surprise responses, as determined by the linear mixed-effects model analysis (see Table 5.7). Individual data points represent the mean values of the Startle-I and Surprise-I ratings for the scenario with markers (squares, triangles, diamonds and circles) indicating different intended responses. Additionally, shaded rectangles around the data points indicate the response variability across individuals, where the width and height of each rectangle correspond to twice the standard deviation of the Startle-I and the Surprise-I in that scenario.

The findings from the linear mixed effect models confirm that most scenarios elicited the intended effects on experienced startle and surprise. However, scenarios ENF and LTS were found to have no significant difference between both their startle, $\text{ENF}_{startle}$ =2.76, $\text{LTS}_{startle}$ =3.28, and surprise average ratings, $\text{ENF}_{surprise}$ =4.01, $\text{LTS}_{surprise}$ =3.44. No significant differences were found on startle ratings between scenarios ENF and CARGO, $\text{CARGO}_{startle}$ = 2.58, scenarios ENF and STALL, $\text{STALL}_{startle}$ = 2.85, scenarios FLAP and CARGO, $\text{FLAP}_{startle}$ = 2.63, scenarios FLAP and STALL. No significant differences were found on surprise ratings between scenarios LTS and CARGO, $\text{CARGO}_{surprise}$ =4.09, scenarios LTS and PFDF, $\text{PFDF}_{surprise}$ =3.94.

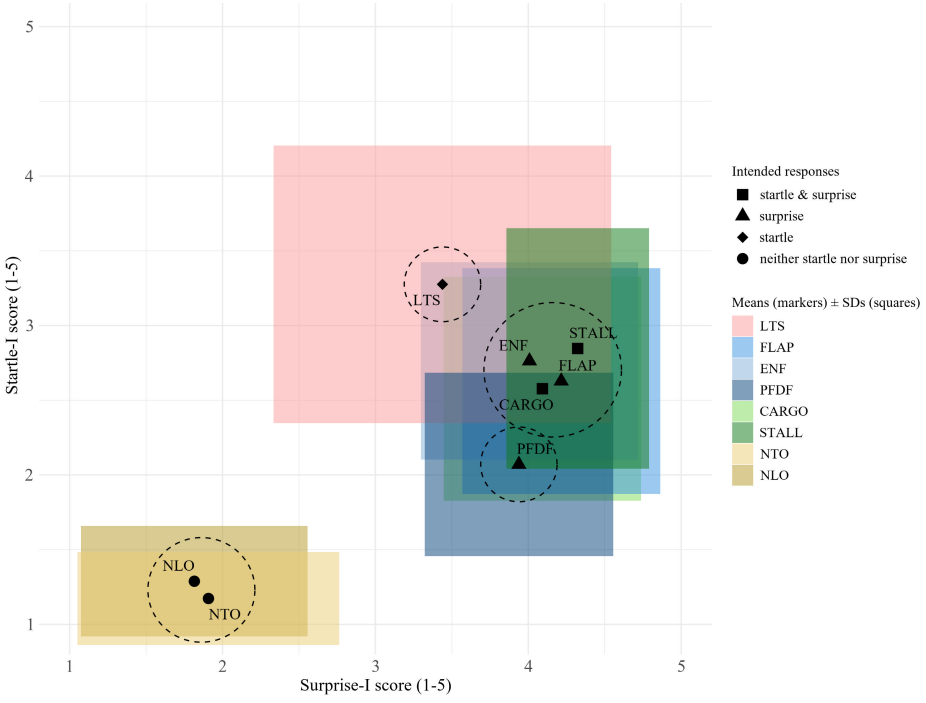


Figure 5.7: Mean values (markers) and SDs (squares) of Startle-I and Surprise-I for each scenario. Dashed circles denote scenarios groups with non-significant differences in mean startle and surprise levels.

Table 5.7: Pairwise comparisons of estimated marginal means between scenarios.

Comparison	Estimate	SE	*p*	Estimate	SE	*p*
	Startle level			**Surprise level**		
ENF - FLAP	0.13	0.16	0.99	-0.21	0.13	0.78
ENF - LTS	-0.51	0.19	0.12	0.57	0.25	0.31
ENF - CARGO	0.19	0.16	0.94	-0.08	0.14	1.00
ENF - NLO	1.47	0.12	< 0.0001	2.19	0.20	< 0.0001
ENF - NTO	1.59	0.11	< 0.0001	2.10	0.23	< 0.0001
ENF - PFDF	0.69	0.13	< 0.0001	0.07	0.14	1.00
ENF - STALL	-0.08	0.16	1.00	-0.32	0.12	0.14
FLAP - LTS	-0.65	0.20	0.02	0.78	0.24	0.03
FLAP - CARGO	0.05	0.17	1.00	0.12	0.12	0.97
FLAP - NLO	1.34	0.13	< 0.0001	2.40	0.19	< 0.0001
FLAP - NTO	1.46	0.12	< 0.0001	2.31	0.22	< 0.0001
FLAP - PFDF	0.56	0.14	< 0.0001	0.28	0.12	0.26
FLAP - STALL	-0.22	0.17	0.91	-0.11	0.10	0.96
LTS - CARGO	0.70	0.20	0.01	-0.65	0.24	0.13
LTS - NLO	1.99	0.17	< 0.0001	1.62	0.28	< 0.0001
LTS - NTO	2.10	0.16	< 0.0001	1.53	0.30	< 0.0001
LTS - PFDF	1.21	0.18	< 0.0001	-0.50	0.24	0.44
LTS - STALL	0.43	0.20	0.41	-0.88	0.23	< 0.0001
CARGO - NLO	1.29	0.13	< 0.0001	2.28	0.19	< 0.0001
CARGO - NTO	1.40	0.13	< 0.0001	2.18	0.22	< 0.0001
CARGO - PFDF	0.51	0.15	0.02	0.15	0.12	0.90
CARGO - STALL	-0.27	0.18	0.79	-0.23	0.10	0.28
NLO - NTO	0.12	0.07	0.70	-0.09	0.26	1.00
NLO - PFDF	-0.78	0.10	< 0.0001	-2.12	0.19	< 0.0001
NLO - STALL	-1.56	0.14	< 0.0001	-2.51	0.18	< 0.0001
NTO - PFDF	-0.90	0.10	< 0.0001	-2.03	0.22	< 0.0001
NTO - STALL	-1.67	0.14	< 0.0001	-2.42	0.21	< 0.0001
PFDF - STALL	-0.78	0.15	< 0.0001	-0.38	0.10	< 0.0001

5.4. Discussion

The construct validity of the Startle-I and Surprise-I was confirmed in a setting with highly ecological validity, including simulated in-flight events and a sample group of 26 professional pilots. To investigate the inventories' ability on measuring self-report startle and surprise, eight scenarios with varied startling and surprising stimuli were tested. MCFA was applied across two levels (i.e., within- and between-scenario) given the repeated-measures experimental design. The analysis was guided by the factor structure identified in Chapter 4, where a ML-EFA was performed. Results from the two-way ANOVA indicate that the averaged amount of variance in both inventories' items caused by differences between scenarios was generally larger than the variance caused by differences between participants. This outcome supports the intention of our experiment, to create significant differences between scenarios rather than between participants. This led to the clustering of data for the MCFA on Scenario instead of on Participant. The ICCs of all items emphasized the need

for applying a MCFA instead of a CFA to properly consider the between-scenario variance.

The goodness-of-fit for the total set of 11 items was compared between a one-factor model and a two-factor model (factors Startle and Surprise). The comparison revealed that the two-factor model provided a superior and acceptable fit to the data, whereas the one-factor model did not. Both models yield significant χ^2 values, which are typically indicative of poor model-data fit. However, the relatively small sample size likely inflated the χ^2 values, producing significant results even when the model-data discrepancies were minor [170]. In addition to the χ^2 test, other fit indices, CFI, TLI, RMSEA, $\text{SRMR}_{\text{within}}$, and $\text{SRMR}_{\text{between}}$ supported the adequacy of the two-factor model. Furthermore, the MCFA identified significant factor loadings at both within- and between-scenario levels, underlining the robustness of the two-factor structure across different levels.

The differentiation between the factors Startle and Surprise indicated by the two-factor model fit provides evidence for the construct validity of the Startle-I and Surprise-I. This finding supports that the Startle-I and Surprise-I can effectively capture pilot experiences of startle and surprise in an ecologically-valid setting. Additionally, this MCFA offers compelling evidence for the distinctiveness of the constructs of startle and surprise within the operational aviation context. The results support the hypothesis that the responses of startle and surprise are fundamentally distinct constructs with different causes and consequences [19, 35], even though **both** could impact pilots' performance to varying degrees.

Regarding the internal consistency of the Startle-I and the Surprise-I, the McDonald's ω across scenarios suggesting acceptable to excellent reliability. The observed variability can possibly be attributed to differences in scenarios' characteristics or to individual differences in interpretation of the items. Certain scenarios might evoke stronger/weaker (more consistent) responses than others, leading to differences in internal consistency.

From the comparison between ratings from inventories and VASs on measuring startle and surprise, the Startle-I and Surprise-I demonstrate a superior ability to distinguish between different levels of startle and surprise when compared to the Startle-VAS and Surprise-VAS (Table 5.6). This was unexpected, as the single-item VAS scores were shown to be highly correlated with inventory scores in Chapter 4, investigating startle and surprise stimuli that were unrelated to the aviation domain. Thus, although the visual analogue scales are efficient in capturing individuals' startle and surprise, and can be used quickly and immediately after target stimuli, the multi-item inventories were found to be more effective in an ecologically-valid context.

There are several limitations to this chapter. First, the findings are based on scenarios performed in a research simulator. Although this allows for high controllability and replicability, these may not simulate the level of stress, surprise, and high demand of real-world in-flight emergencies. Second, we employed a within-subjects experimental design involving 26 participants who were exposed to eight different test scenarios to explore the factor structure of Startle-I and Surprise-I. While this sample size enabled us to achieve meaningful insights into the model's goodness-of-fit, it limits the generalizability of our findings. The limited number of participants may reduce the statistical power necessary for detecting subtle nuances in the factor structures of Startle-I and Surprise-I. Future research should consider expanding the sample size and possibly incorporating a broader demographic (e.g., age, flight experience) to enhance the validity and robustness of these findings.

Third, the scenarios were designed to vary in induced startle and surprise, and the results indicate that the designed stimuli effectively elicited a wide range of variability in both startle and surprise. However, it remains challenging to independently elicit the responses of startle or surprise. Moreover, in intended surprising scenarios, stress from flight tasks could increase the level of startle response [23]. Fourth, all test scenarios were conducted in a single-pilot setting with a simplified twin-prop aircraft model that most pilots were not familiar with. Apart from the unfamiliarity, extra high workload was introduced by requiring pilots to fly manually, which could also affect their experienced startle or surprise, making them differ from the hypothesis. Fifth, the current chapter is limited in focus on the aviation domain, whereas the inventories have potential for broader applicability in other high-pressure environments involving human operators. Future research should aim to replicate these findings in different operational settings, as well as in different domains to enhance the generalizability.

5.5. Conclusion

This chapter provides strong and consistent evidence supporting the factor structure of the Startle and Surprise inventories, aligning with the prior chapter. The findings in this chapter, which used professional pilots (N = 26) as participants and settings with high ecological validity, were consistent with those reported in Chapter 4, in which video clips were applied to elicit startle and surprise. This demonstrates the reliability of Startle and Surprise inventories, across diverse contexts. These results underscore the inventories' applicability for assessing startle and surprise responses in practical contexts both at individual and scenario levels. The Startle and Surprise inventories open up opportunities for research into startle and surprise responses, their specific causal factors, and their specific effects on operator performance in various fields. This can lead to the development of evidence-based safety protocols and training interventions.

6

Criterion-related validity of the Startle and Surprise Inventories

To establish the criterion-related validity of the Startle and Surprise Inventories, this chapter discusses an investigation on the impact of startle and surprise on pilots' information-processing performance. Based on prior research, we hypothesize that both startle and surprise would have negative impact on information-processing performance, due to the cognitive disruption caused by the startling stimulus and the need to reframe and make sense of surprising situations. To test these hypotheses, 26 professional pilots are exposed to eight flight scenarios designed to elicit a wide range of startle and surprise. The information-processing performance is assessed using an auditory cognitive task during flight, with reaction time and accuracy serving as dependent measures. Linear mixed-effects models are used to analyse the relationships between self-report startle and surprise, and the auditory task performance. In addition, the statistical analyses included the examination of temporal patterns in reaction time and correlation analyses between age, flight hours, and the dependent measures.

The chapter is organized as follows: Section 6.1 reviews prior research on the influence of startle and surprise on performance, identifies key gaps in the literature, and outlines the hypotheses. Section 6.2 describes the methodology. Section 6.3 presents the results. Section 6.4 discusses the findings in relation to existing theory and operational relevance, and Section 6.5 concludes with key insights and recommendations for future research.

The contents of this chapter have been submitted for publication [180].

6.1. Introduction

Aviation operations are inherently complex and require the simultaneous integration of multiple sources, functions, and environmental configurations. Modern display technology has consolidated flight-critical information into fewer, more efficient displays to enhance situation awareness. However, the increasing complexity of displays and automation, and the risk of "information glut" have not necessarily alleviated pilots' mental workload during critical phases of flight, and instead could lead to attentional tunnelling and automation issues, such as mode confusion and automation surprise [11, 181, 182]. These cognitive challenges can significantly impact pilot's decision-making and performance, particularly in high-stakes situations [183, 184]. Unexpected failures or anomalies may further exacerbate these challenges, by triggering startle and surprise responses [5].

Research has shown that startle and the associated stress can temporarily disrupt cognitive processes, impair sustained attention, and reduce cognitive efficiency [98, 185, 186]. Neurobiological models suggest that startle activates survival circuits, prioritizing threat detection at the expense of cognitive processing, compromising the executive functions [187]. This shift in neural processing reallocates cognitive resources toward stimulus-driven responses, limiting working memory and attentional control necessary for task execution. The resulting cognitive overload may lead to narrowed attentional focus, reduced situation awareness, impaired decision-making, and decreased information-processing capacity [5, 97]. The severity and duration of this disruption vary based on individual differences [188], contexts, and characteristics of the eliciting stimuli [148]. This disruption is particularly concerning in high-risk operational contexts such as in aviation, where precise and timely decision-making is critical [19].

While several studies have demonstrated that startle can lead to transient cognitive disruptions, other studies have reported minimal or no detrimental effects, and in some cases even performance enhancements under high cognitive load conditions in the controlled laboratory environments [189]. These mixed findings underscore the importance of investigating the effects of startle within realistic and operationally relevant contexts.

Surprise has been shown to interrupt ongoing automatic cognitive processes by redirecting attention toward analysing the unexpected event and updating one's mental model of the situation [34]. This interruptive effect has been quantified in laboratory settings, for example, through increased latency in verbal and motor response tasks [190]. The subsequent sense-making and reframing processes required to interpret and respond to the unexpected event involve effortful, goal-directed cognitive activity [35, 77]. These processes are susceptible to disruption under high levels of stress, which can impair attention to task-relevant information and hinder the execution of appropriate actions [98].

Concluding, both startle and surprise may decrease pilot information-processing performance by disrupting ongoing cognitive processes. They may induce additional mental workload by distracting attention to the startling or surprising stimulus and requiring considerable working memory allocated to analyse the startling or surprising event.

In this chapter, we aim to investigate the effects of startle and surprise on pilots' information-processing performance in an ecologically-valid setting. Using a dual-task paradigm [191], information-processing performance was quantified using a secondary auditory cognitive task that was performed while pilots experienced the startling and surprising in-flight events. This auditory task was meant to represent a communication situation.

We hypothesized that both startle and surprise would impair pilots' information-processing performance, leading to reduced performance on the auditory task. To test these two hypotheses, pilots were exposed to eight in-flight scenarios in a flight simulator, which were designed to induce a wide range of startle and surprise responses. To measure startle and surprise, multi-item instruments developed and validated in Chapters 4 and 5, was employed. The insights obtained into the effects of startle and surprise on pilot information-processing performance in an operationally-relevant setting may contribute to the development of training programs aimed at enhancing pilot resilience during in-flight emergencies.

6.2. Method

6.2.1. Participants and apparatus

The same participants sample and apparatus described in Chapter 5 were used for experiment discussed in the current chapter. Please refer to Sections 5.2.1, 5.2.2 in Chapter 5 for detailed demographic information and apparatus setup, respectively.

6.2.2. Tasks and conditions

The experiment procedure and test scenarios were identical to that described in Chapter 5 (see Sections 5.2.3 and 5.2.4 for a detailed description).

6.2.3. Auditory task

The pilots were informed that the auditory task was designed to assess their capacity to process auditory information. In line with standard procedures, they were instructed to always prioritize aircraft control over the secondary auditory task. Figure 6.1 illustrates the auditory task as performed on the runway (baseline test) and during the flight test. Table 6.1 summaries the timing of the preset events and auditory task across test scenarios.

A "block" consisted of ten randomly generated numbers, ranging from 0 to 9, pronounced in the ICAO Phonetic Alphabet, where presented over the pilots' headset with 2.5 second intervals resulting in a total block duration of 28 s. Each block was preceded by an auditory warning: "The auditory task is coming". The target block in the flight test, where performance was collected, started at 5, 10 or 15 s before the preset startle and/or surprise event (i.e., lead time in Figure 6.1b) with 2, 4, or 6 additional numbers, respectively, and always continued for 28 s (10 numbers) after the event onset.

A non-target (distraction) block was also included in the ENF, FLAP, STALL, and LTS. These non-target blocks were presented without overlapping with target blocks or preset events, and always lasted 28 seconds. Their only purpose was to reduce participants' expectation of preset events, which were consistently paired with the target blocks.

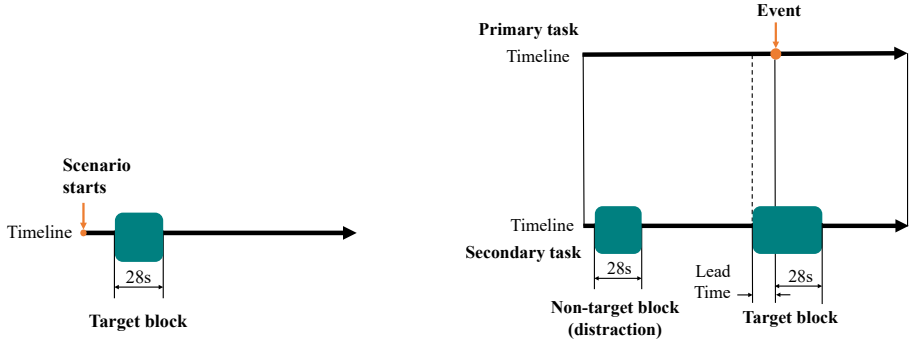
Participants were instructed to single-click the autopilot disconnect button with their thumb if the number was odd, and double-click if the number was even. For double-clicks, the interval between clicks had to be less than 500 milliseconds. Otherwise, the response would be recorded as one single-click and one invalid click.

Both aircraft control and auditory task were implemented in the Delft University Environment for Communication and Activation (DUECA). This ensured synchronisation between flight phase, auditory stimulus presentation, and responses acquisition.

Table 6.1: Timing of preset events and auditory task across scenarios.

ID	Event timing	Lead time[1]	Non-target (distraction) block timing
ENF	5 s after reaching 900 ft	10 s	Climbing through 100 ft
FLAP	At the selection of Flaps 25 in the base leg	5 s	Climbing through 800 ft
CARGO	10 s after reaching 200 ft	15 s	-
LTS	5 s after descending to 500 ft	5 s	Descending through 800 ft
PFDF	Descending to 600 ft	5 s	-
STALL	20 s after reaching 800 ft	5 s	Climbing through 150 ft
NTO	-	-	-
NLO	-	-	-

[1] Lead time as shown in Figure 6.1b.



(a) Auditory task performed on the runway (baseline test)

(b) Auditory task performed during the flight test

Figure 6.1: Auditory task in the test session.

6.2.4. Dependent measures

Auditory task reaction time

The main measure of the pilots' information-processing capacity was their reaction time to the secondary auditory task. The autopilot disconnect button was sampled at a frequency of 100 Hz, after which reaction times were calculated with a resolution of 0.01 s. The mean reaction time of the correct responses was obtained for the target block (after the event as in Figure 6.1b) and for the baseline (with the aircraft stationary on the runway as in Figure 6.1a) for each scenario. To clean the data, responses that were either extremely fast (i.e., quicker than the pilot's average response time in baseline) or missing were excluded as invalid. The mean reaction time during the flight test was then corrected by subtracting the mean baseline reaction time, resulting in the Delta Reaction Time (ΔRT).

Auditory task accuracy

As additional measure of pilots' information-processing performance, accuracy within the target block was assessed. This was calculated as the ratio of correct responses to the

total number of presented numbers in the target block (i.e., 10). To account for individual baseline performance, flight test accuracy was baseline-corrected by subtracting the baseline accuracy, yielding the Delta accuracy (ΔAC).

Self-report startle and surprise

Following each flight test scenario, participants completed the Startle Inventory (Startle-I) and Surprise Inventory (Surprise-I). The Startle-I and Surprise-I were developed and validated to measure an individual's startle or surprise response to a presented stimulus (as in Appendix B). There are six statements in the Startle-I and five statements in the Surprise-I, for which responses are scored on 1-5 Likert scales. The total scores of both inventories are the average of the items' scores, ranging from 1 to 5. McDonald's ω [177] of the current sample was $\omega = 0.88$ to $\omega = 0.96$ for the Startle-I, and $\omega = 0.77$ to $\omega = 0.96$ for the Surprise-I across test all scenarios, indicating acceptable to excellent internal consistency.

6.2.5. Statistical analysis

First, Pearson correlation analyses were conducted to preliminarily examine the relationships between Age, Flight hours, ΔRT, ΔAC, and self-report startle and surprise. To account for the repeated-measures design and the non-independence of observations both across scenarios and within participants, a modified "between and within formulation" [144] was applied. Specifically, for each dependent measure, a separate linear mixed-effects model was fitted, to remove the influence of different scenarios and individual differences. The between-participant correlations matrix was computed using the participant-level random effects extracted from the linear mixed-effects models for each dependent measure. The between-scenario correlations matrix was calculated by correlating the estimated marginal means for each dependent measure within scenarios, as derived from the linear mixed-effects models. The residual correlations matrix was computed using the residuals extracted from the linear mixed-effects models for each dependent measure.

To assess the effect of startle and surprise on ΔRT and ΔAC across different scenarios, linear mixed-effects models were applied using the `lme` function from the `nlme` package. The fixed effects were the ratings of startle (*Startle-I*), the ratings of surprise (*Surprise-I*), and scenario (*Scenario*), which was modelled as a categorical variable with eight levels. Additionally, if the sequence of scenarios (*Sequence*) was found to have a significant effect on ΔRT or ΔAC, it was also included in the linear mixed-effects model. To account for individual differences, participant number (*ID*) was included as a random effect.

Linear mixed-effects models were fitted using the `lme` function from the `nlme` package, while heteroscedasticity in the residuals was modelled using function `varIdent` to accommodate variance differences across scenarios in `R`.

Furthermore, the Intraclass correlation coefficient (ICC) [192] was calculated to assess the proportion of the total variance attributable to differences between participants. The ICC was derived from the random effect results, with a higher ICC indicating the notable between-participant variability, supporting the application of the linear mixed-effects model.

In addition to analysing the effects of startle and surprise on ΔRT, we conducted a descriptive analysis to explore the temporal pattern of ΔRT and valid responses during the auditory cognitive task in the target block. Specifically, for each scenario, the means and standard deviations of the ΔRT across the sequence of ten numbers were calculated, aiming

Table 6.2: Means and standard deviations of the dependent measures across eight scenarios.

Scenario	Δ *RT (s)*		Δ *AC (%)*		Startle-I (1-5)		Surprise-I (1-5)	
	Mean	***SD***	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***
ENF	0.46	0.21	-30.00	29.26	2.76	0.66	4.01	0.71
FLAP	0.33	0.25	-12.69	18.45	2.63	0.76	4.22	0.65
CARGO	0.33	0.25	-13.08	22.59	2.58	0.75	4.09	0.65
LTS	0.12	0.13	-5.38	8.11	3.28	0.93	3.44	1.10
PFDF	0.17	0.14	-2.31	9.08	2.07	0.61	3.94	0.62
STALL	0.29	0.23	-9.62	16.37	2.85	0.81	4.32	0.47
NTO	0.13	0.11	-1.92	5.67	1.17	0.31	1.91	0.86
NLO	0.09	0.09	-1.92	10.96	1.29	0.37	1.82	0.74

Note. ΔRT = Delta Reaction Time; ΔAC = Delta Accuracy; Startle-I = Startle Inventory; Surprise-I = Surprise Inventory.

to assess whether ΔRT stabilized or remained disrupted following the in-flight events.

6.3. Results

6.3.1. Overview of collected data

For two participants, the ΔRT data were missing in two scenarios (ENF and CARGO), due to an insufficient number of correct responses. Table 6.2 provides an overview of means and standard deviations of dependent measures across the eight test scenarios.

In general, Surprise-I scores were higher than Startle-I scores across scenarios. LTS was, as expected, the most startling scenario, and scored somewhat lower on Surprise-I than other surprising scenarios. CARGO and STALL scored, as expected, high on both startle and surprise. PFDF scored, as expected, relatively low on startle and high on surprise. ENF and FLAP exhibited high scores on both startle and surprise, which was unexpected.

Figures 6.2 and 6.3 show the pirate plots of the ΔRT and ΔAC. The plots represent the mean values (square markers with labels), interquartile range (IQR) in black lines, and distribution across scenarios. ENF, FLAP and CARGO were scenarios that required pilots to manually intervene, and these also exhibited the highest mean ΔRTs, lowest mean ΔACs, and highest variance on both. STALL did not require any manual intervention but appears to be similarly impactful as ENF, FLAP and CARGO. A low impact can be observed for LTS (startling, not surprising) and PFDF (surprising, not startling).

Scenarios associated with longer ΔRT, such as ENF, STALL, CARGO and FLAP, generally exhibited lower ΔAC, suggesting a potential trade-off between response speed and accuracy in the secondary task. However, scenarios with shorter ΔRT, such as NLO and NTO, demonstrated high ΔAC, indicating that faster responses did not always compromise the accuracy of the secondary task.

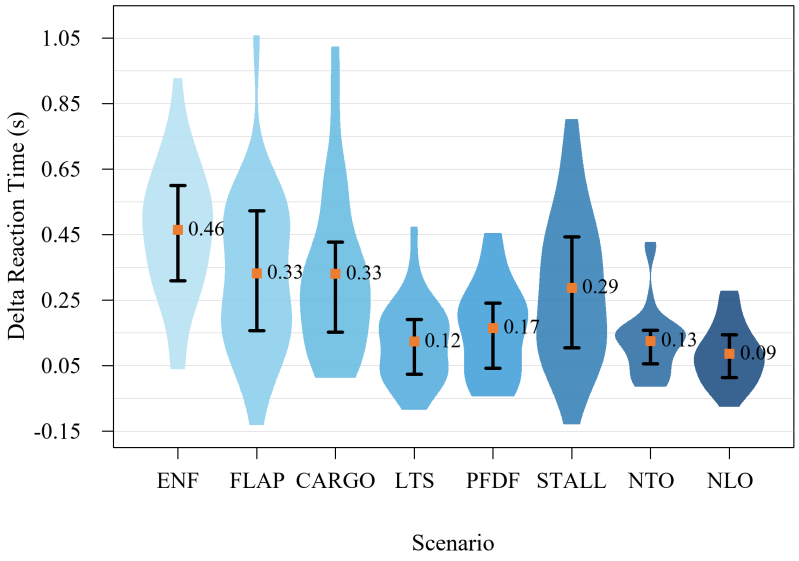


Figure 6.2: Delta Reaction Time of the auditory task across eight test scenarios (square markers indicate means, whiskers indicate interquartile range).

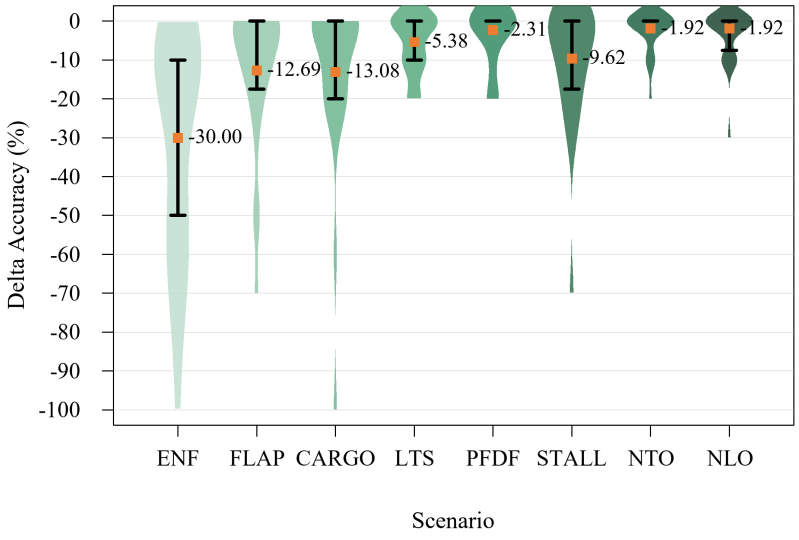


Figure 6.3: Delta Accuracy of the auditory task across eight test scenarios (square markers indicate means, whiskers indicate interquartile range).

Table 6.3: Summary of the linear mixed-effects model for Delta Reaction Time.

Effect	Estimate	*SE*	*t*	*p*	Variance[2]
Fixed effects					
(Intercept)	0.065	0.036	1.824	0.070	-
Startle-I (1-5)	0.049	0.017	2.817	0.005**	-
Surprise-I (1-5)	0.002	0.012	0.139	0.890	-
$\text{Scenario}_{\text{ENF}}$	0.257	0.053	4.849	< 0.001**	2.128
$\text{Scenario}_{\text{FLAP}}$	0.131	0.060	2.181	0.031*	2.756
$\text{Scenario}_{\text{CARGO}}$	0.137	0.058	2.379	0.019*	2.547
$\text{Scenario}_{\text{LTS}}$	-0.107	0.046	-2.315	0.022*	1.162
$\text{Scenario}_{\text{PFDF}}$	-0.008	0.039	-0.202	0.841	1.341
$\text{Scenario}_{\text{STALL}}$	0.076	0.055	1.388	0.167	2.155
$\text{Scenario}_{\text{NLO}}$	-0.045	0.023	-1.981	0.049*	0.895
Random effects (*SD*)					
Between-participant	0.073	-	-	-	
Within-participant	0.086	-	-	-	

Note. Number of observations = 206.
[2] Estimated multiplicative factor of the residual (i.e. within-participant) variance with respect to the reference scenario NTO. For example, the variance of the residuals for scenario ENF is 2.128 × 0.086 = 0.183.
** $p < 0.01$ (two-tailed). * $p < 0.05$ (two-tailed).

6.3.2. Effects of startle and surprise on ΔRT

No significant effect of the scenarios' sequence (*Sequence*) on ΔRT was found, hence it was not included in the model. The linear mixed-effects model of ΔRT was fitted as follows:

$$\Delta RT = 1 + Startle\text{-}I + Surprise\text{-}I + Scenario + (1|ID)$$

Results of the linear mixed-effects model for ΔRT are presented in Table 6.3. Higher Startle-I scores were significantly associated with an increase in ΔRT, $\beta = 0.049$, $SE = 0.017$, $t(171) = 2.817$, $p = 0.005$. This indicates that for every point scored higher on the Startle-I, ΔRT increased by 49 ms, after controlling for the effect of *Surprise-I*, *Scenario* and *ID*. In contrast, no significant effect of *Surprise-I* was found on ΔRT.

The estimated standard deviation of the between-participant level random effect was 0.073, corresponding to an estimated ICC of 0.422. This indicates that approximately 42.2% of the total variance in ΔRT is attributable to between-participant differences, after controlling for the fixed effects of *Surprise-I*, *Startle-I* and *Scenario*). These findings support the use of a mixed-effects model to account for the non-independence of observations.

Using NTO as the reference, ENF, FLAP, and CARGO resulted in significantly higher ΔRT. ΔRT was 257 ms higher in ENF, $\beta = 0.257$, $SE = 0.053$, $t(171) = 4.849$, $p < 0.001$. ΔRT was 131 ms higher in FLAP, $\beta = 0.131$, $SE = 0.060$, $t(171) = 2.181$, $p = 0.031$, and 137 ms higher in CARGO, $\beta = 0.137$, $SE = 0.058$, $t(171) = 2.379$, $p = 0.019$. Participants in the LTS and NLO showed significantly lower ΔRT compared to those in the NTO. ΔRT was 107 ms lower in LTS, $\beta = -0.107$, $SE = 0.046$, $t(171) = -2.315$, $p = 0.022$, and ΔRT was 45 ms lower in the NLO, $\beta = -0.045$, $SE = 0.023$, $t(171) = -1.981$, $p = 0.049$.

Table 6.4: Summary of the linear mixed-effects model for Delta Accuracy.

Effect	Estimate	*SE*	*t*	*p*	Variance[3]
Fixed effects					
(Intercept)	-0.594	2.845	-0.209	0.835	-
Startle-I (1-5)	-0.840	1.267	-0.663	0.508	-
Surprise-I (1-5)	-0.640	0.861	-0.744	0.458	-
Sequence	0.197	0.304	0.646	0.519	-
$\text{Scenario}_{\text{ENF}}$	-25.412	6.038	-4.208	< 0.001**	5.447
$\text{Scenario}_{\text{FLAP}}$	-8.115	4.123	-1.968	0.051	3.149
$\text{Scenario}_{\text{CARGO}}$	-8.629	4.734	-1.823	0.070	3.994
$\text{Scenario}_{\text{LTS}}$	-0.745	3.235	-0.230	0.818	1.324
$\text{Scenario}_{\text{PFDF}}$	1.707	2.627	0.650	0.517	1.494
$\text{Scenario}_{\text{STALL}}$	-4.710	4.114	-1.145	0.254	2.937
$\text{Scenario}_{\text{NLO}}$	0.053	2.623	0.020	0.984	2.437
Random effect (*SD*)					
Between-participant	4.498	-	-	-	
Within-participant	5.506	-	-	-	

Note. Number of observations = 208.
[3] Estimated multiplicative factor of the residual (i.e. within-participant) variance with respect to the reference scenario NTO. For example, the variance of the residuals for scenario ENF is 5.506 × 5.447 = 29.991.
** $p < 0.01$ (two-tailed).

6.3.3. Effects of startle and surprise on ΔAC

The sequence of test scenarios had a significant effect on ΔAC. *Sequence* was therefore included in the linear mixed-effects model of ΔAC:

$$\Delta AC = 1 + Startle\text{-}I + Surprise\text{-}I + Scenario + Sequence + (1|ID)$$

Results from the linear mixed-effects model for ΔAC are presented in Table 6.4. The model reveals that the fixed effects intercept was not statistically significant, indicating that the ΔAC in the reference scenario NTO did not differ significantly from zero. No significant effect of *Surprise-I* nor *Startle-I* was found on ΔAC.

The estimated standard deviation of the random intercept was 4.503, which corresponds to an ICC of approximately 0.475. This indicates that about 47.5% of the total variance in ΔAC stems from between-participant differences, further supporting the use of a mixed-effects model.

Using NTO as reference, ENF exhibited significantly lower ΔAC. ΔAC in ENF was -0.594 - 25.412 = -26.006, with β = -25.412, SE = 6.038, $t(172)$ = -4.208, $p < 0.001$.

6.3.4. Correlation analysis

Table 6.5 lists the Pearson correlations between Age, Flight hours, ΔRT, ΔAC, Startle-I and Surprise-I scores at between-participant, between-scenario and residual levels. At the between-participant level, both Age and Flight hours were significantly negatively associated

Table 6.5: Correlation matrices of the study variables.

	Age	FH	ΔRT	ΔAC	Startle-I	Surprise-I
Between-participant correlation matrix						
Age	1.000					
FH	0.917**	1.000				
ΔRT	-0.596**	-0.599**	1.000			
ΔAC	-0.107	-0.118	-0.401*	1.000		
Startle-I	-0.242	-0.301	0.237	-0.041	1.000	
Surprise-I	0.097	0.112	-0.038	-0.152	0.134	1.000
Between-scenario correlation matrix						
Age	-					
FH	-	-				
ΔRT	-	-	1.000			
ΔAC	-	-	-0.932**	1.000		
Startle-I	-	-	0.528	-0.508	1.000	
Surprise-I	-	-	0.718*	-0.543	0.817*	1.000
Residual correlation matrix						
Age	-					
FH	-	-				
ΔRT	-	-	1.000			
ΔAC	-	-	-0.161*	1.000		
Startle-I	-	-	0.162*	0.008	1.000	
Surprise-I	-	-	0.026	-0.044	0.161*	1.000

Note. FH = Flight hours; ΔRT = Delta Reaction Time; ΔAC = Delta Accuracy; Startle-I = Startle Inventory; Surprise-I = Surprise Inventory.
** $p < 0.01$ (two tailed). * $p < 0.05$ (two tailed).

with ΔRT, r = -0.596 and r = -0.599, respectively. These findings suggest that older and more experienced pilots demonstrated lower ΔRT in the secondary auditory task and experienced less startle to in-flight events.

At the between-scenario level, *Surprise-I* was significantly positively correlated with ΔRT, r = 0.718, and significantly positively correlated with the Startle-I, r = 0.817. Additionally, ΔAC showed a significant negative correlation with ΔRT at both between-scenario, r = -0.932, and between-participant levels r = -0.401. These findings indicate that both participants and scenarios with longer ΔRT tend to exhibit more errors in the cognitive task.

At the residual level (after removing participant- and scenario-level effects), *Startle-I* was significantly positively associated with ΔRT, r = 0.162 and with *Surprise-I*, r = 0.161. This would suggest that an unusually high *Startle-I* is associated with higher than usual *Surprise-I* and higher than usual ΔRT. In addition, ΔAC showed a significant negative correlation with ΔRT, r = -0.161. This means that if in one scenario the ΔRT is higher than what should be expected given the scenario and the participant, then more errors would occur in the cognitive task for this scenario.

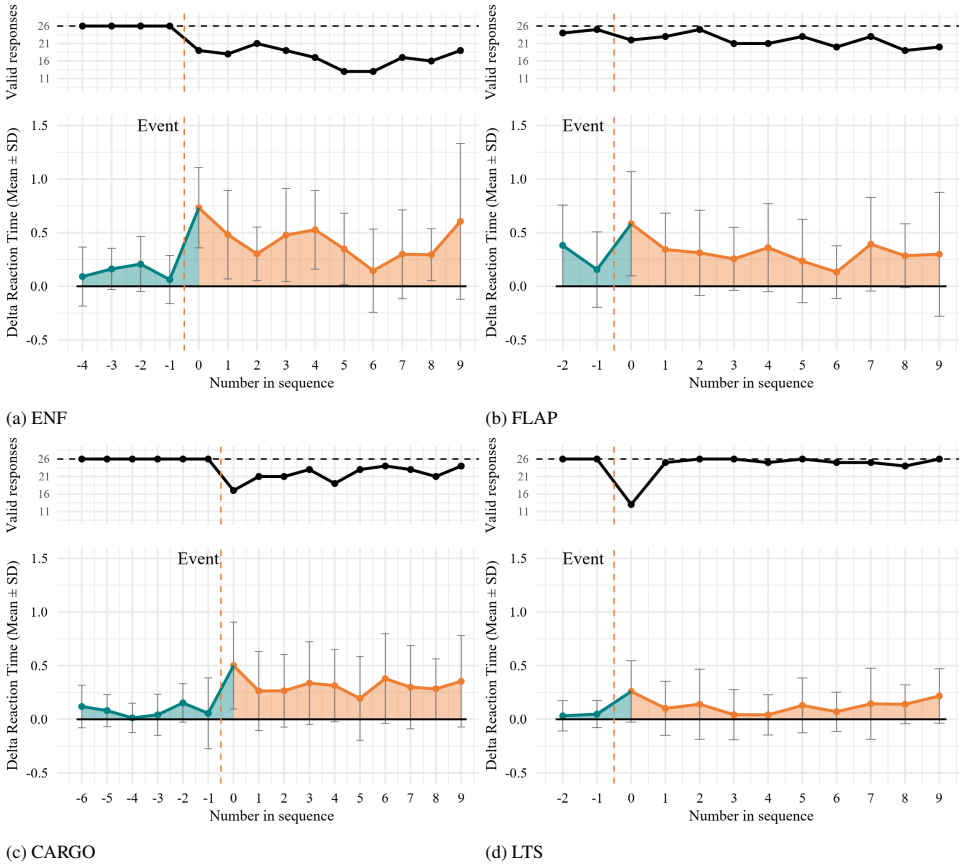


Figure 6.4: Mean Delta Reaction Time (± 1 SD) for target numbers across scenarios.

6.3.5. Temporal patterns of ΔRT

Figures 6.4 and 6.5 provide an overview of the ΔRT and the number of valid responses across the cognitive task duration in the target block. In the ΔRT plots, the black line represents a zero change in Reaction Time, while the shaded areas highlight deviations from this reference. In plots showing the number of valid responses, the gray dashed line indicates the total number of participants (N = 26). In all scenarios with preset events, a notable increase in ΔRT is observed immediately after the event (dashed vertical line), peaking at the first number in the sequence and coinciding with a dip in valid responses. For ENF, FLAP and CARGO, the scenarios where pilots had to intervene to control the flight path, the effect on ΔRT and valid responses can be seen to last throughout the measuring duration. Additionally, greater variability in ΔRT post-event is evident in these scenarios, as indicated by larger standard deviations. STALL can be seen to cause a similar pattern, possibly due to the persistent warning sound. The impact of LTS and PFDF can be seen to quickly subside, suggesting a more brief impairment in information-processing performance.

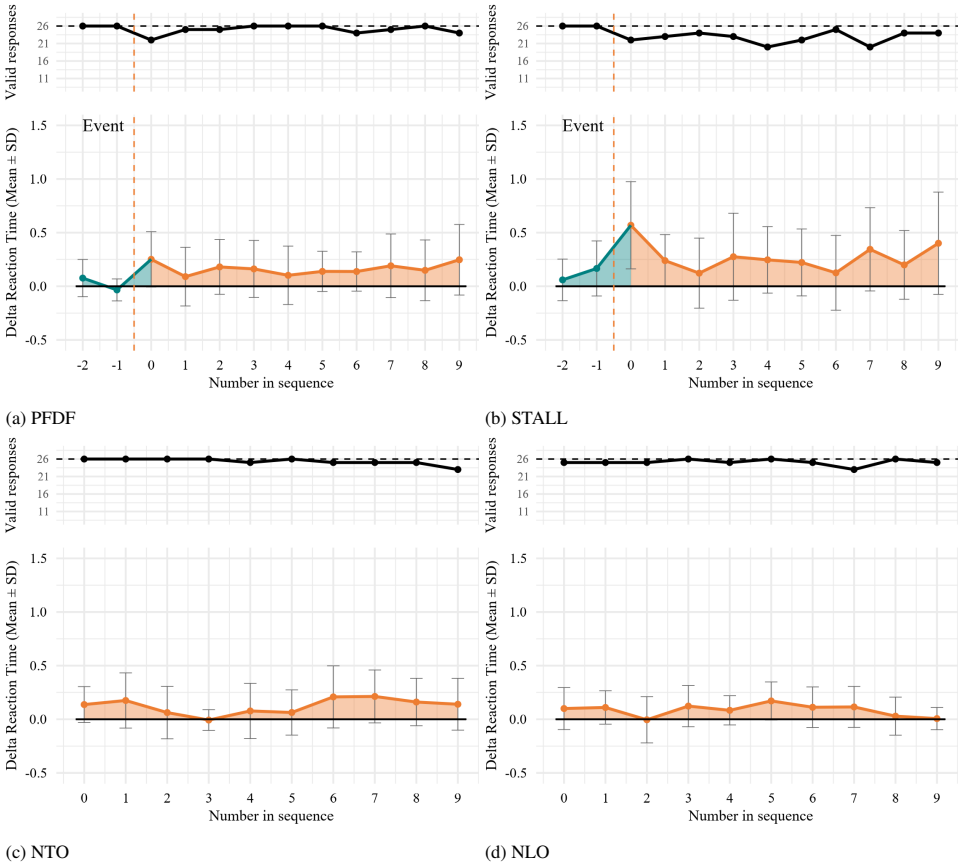


Figure 6.5: Mean Delta Reaction Time (± 1 SD) for target numbers across scenarios (continued).

6.4. Discussion

This study investigated the effects of startle and surprise on pilots' information-processing performance during simulated in-flight events. The results indicate that heightened startle responses significantly impaired information-processing speed, as evidenced by the increased ΔRT on the secondary auditory task. The finding aligns with prior research, suggesting that startle could induce temporary cognitive disruption [23], likely due to the rapid activation of survival-related neural circuits that prioritize threat detection over executive functioning [187]. The involuntary nature of the startle response appears to divert attentional resources from ongoing tasks, necessitating additional cognitive effort to reorient focus, which is supported by evidence of increased cerebral blood flow in prefrontal cortex regions [193].

The finding contrasts with recent research [189], which reported no performance impairment following startle and even noted slight performance improvements under high cognitive load conditions. One possible explanation for this discrepancy lies in the differences between laboratory environment and ecologically-valid setting. In controlled laboratory contexts,

the absence of perceived threat may enable participants to rapidly compensate for effect of startle, thereby minimizing its disruptive impact on ongoing tasks. However, real-world startle responses are more likely to occur in highly-demanding and potentially threatening situations, where the combined complexity and stress of the tasks environment may exceed compensatory mechanisms and exacerbate performance disruptions.

Contrary to our hypothesis, surprise did not have a distinct significant impact on secondary task performance. While surprise has been theorized to impose cognitive demands by prompting a need for sense-making and situation reframing [35, 77], our results show that its immediate effects on information processing were minimal in the in-flight events featured in our experiment. It could be that the surprise events in our study may not have been complex enough to elicit reframing efforts that would induce sufficient additional workload to impact secondary task performance.

Although no significant effect of surprise was found in the linear mixed-effects model, our study does not imply that surprise is less irrelevant for pilot training. In fact, significant correlations were observed between the ratings of surprise and ΔRT at the between-scenario level, indicating that higher levels of surprise are generally associated with greater information-processing performance disruptions (Section 6.3.4). The temporal pattern analysis in Section 6.3.5 also highlights the role of surprise in shaping responses' dynamics over time. One possible explanation is that the between-scenario correlations reflect general trends across different scenarios, whereas the linear mixed-effects model accounts for both within- and between-subject variations while controlling for additional factors. The mixed-effects model revealed that, after adjusting for the ratings of startle, scenarios and individual differences, the effect of surprise on ΔRT and ΔAC was non-significant. Nevertheless, previous studies have shown that unexpectedness in training is important to build cognitive flexibility and higher-level competences to deal with a wide variety of possible events [112, 194, 195]. Given its importance for training, more research is required into means to systematically introduce surprise in aviation scenarios in ecologically-valid settings.

We also observed a negative correlation between flight experience and ΔRT, suggesting that more experienced pilots were better at managing cognitive disruptions caused by in-flight events. This finding supports the idea that expertise and training can mitigate the impact of unexpected events [109], potentially by enabling pilots to rely on well-practiced procedural knowledge that requires fewer cognitive resources to execute under stress. This underscores the importance of tailored training programs designed to enhance resilience to startle responses, particularly for less experienced pilots.

Our findings highlight the necessity of incorporating startle management strategies into pilot training programs. Although current Upset Prevention and Recovery Training protocols acknowledge the role of startle and surprise in aviation safety, our results suggest that targeted interventions specifically aimed at mitigating startle-induced cognitive disruptions may be beneficial [196]. Stress management techniques, exposure-based training, and cognitive resilience exercises could help pilots develop strategies to minimize the disruptive effects of startle and maintain optimal performance under pressure [115, 197, 198].

Several limitations of the current study should be acknowledged, along with recommendations for future research. First, although the data revealed both within-scenario and between-scenario variation in Startle-I ratings, the intensity of the startle responses observed was likely lower than those typically experienced during real-world aviation emergencies.

This limitation may be attributed to the controlled nature of the simulation context, in which participants were aware that no actual threat was present. Future research could employ more immersive simulation techniques, introduce social stressors (e.g., performance evaluation and peer observation) or incorporate sudden-onset stimuli to elicit stronger startle responses, and further validate their effects under more ecologically intense conditions.

Second, the quantification of startle and surprise relied on self-report measures, which, although informative and psychometrically validated, are subject to potential biases such as individual differences in interpreting questionnaire items. Future research could enhance the assessment of startle and surprise by integrating self-report data with physiological indicators (e.g., reflex electromyogram [37], pupillometry [199] and heart rate [42]).

Third, the experimental setup involved single-pilot crew operating a twin-propeller model, which may not fully replicate the dynamics of real-world multi-crew operations. Investigating how startle and surprise interact in team-based settings could provide valuable insights into crew resource management strategies and training.

Fourth, while the application of the linear mixed-effects model on the effect of startle on ΔRT controlled for the surprise ratings, individual differences, and scenario differences, the relationships between startle ratings and secondary task performance may have been confounded by aircraft control demands in scenarios such as ENF, FLAP, and CARGO. However, our findings show that the LTS and STALL scenarios elicited the highest startle responses without imposing any (LTS) or similar (STALL) control demands as ENF, FLAP or CARGO, ruling out this potential confounder.

Fifth, the absence of significant findings regarding ΔAC might be due to the limited resolution of the secondary task measurement or to ceiling effects. Future studies might benefit from integrating additional metrics (such as physiological measures) or increasing the difficulty of the cognitive task, to better capture the fluctuation of information processing.

6.5. Conclusion

In summary, our study demonstrates that startle responses significantly impair pilots' in-flight cognitive performance, whereas surprise does not impose a comparable effect. These findings reinforce the importance of training interventions aimed at managing startle effects in high-risk environments. Additionally, the role of experience in mitigating these effects highlights the value of continued practice and exposure-based training. Future research should explore more nuanced measures of cognitive capacity and extend these findings to multi-crew settings to further enhance aviation safety protocols.

7

Discussion and conclusions

In this chapter, the main findings and contributions associated with each research objective are discussed. It reflects on how these address the research questions outlined in Chapter 1. The chatper also outlines the theoretical and practical implications of the findings, discusses the overall limitations of this research, and provides recommendations for future research on startle and surprise in the field of Human Factors and Applied Aviation Psychology.

Although startle and surprise have been widely acknowledged as critical human factors influencing pilot performance during unexpected situations, a significant gap remains in the systematic measurement of these responses and their integration into operational aviation contexts. This dissertation focused specifically on developing self-report measuring instruments to quantify startle and surprise, an essential step toward supporting the translation of theoretical constructs into operational application. To this end, three research objectives were formulated, including the conceptual analysis of startle and surprise (Chapter 2), the critical reviews of existing measures (Chapter 3), and the development, validation, and application of self-report measuring instruments across multiple samples and contexts (Chapters 4, 5 and 6). This concluding chapter provides an integrative reflection on how each research objective contributed to addressing the overarching aim of quantifying startle and surprise in an aviation operational context.

7.1. Research objective 1

The first objective of this dissertation was to use existing literature to conceptualize and describe the cognitive processes underlying pilot decision-making in unexpected situations. To this end, Chapter 2 intended to address the research question:

Key question 1

How do different cognitive models represent pilot decision-making and actions in unexpected situations?

To answer this question, three existing cognitive models, namely the Landman model of startle and surprise, the perceptual cycle model, and the three-level situation awareness model, were applied to analyse pilot decision-making and actions during two real-world flight incidents: Loganair Flight 6780 and US Airways Flight 1549. The comparative analysis showed that each model provides a distinct perspective on pilot decision-making in unexpected situations, highlighting different theoretical constructs and explanatory strengths.

The Landman model of startle and surprise emphasizes how discrepancies between perceived information and active mental frames trigger a reframing process, which is particularly challenging under acute stress. The model outlines how startle and surprise could influence perception, appraisal and actions guided by these active frames. However, the model does not describe how frames are initially selected, limiting its explanatory scope in the early stages of cognitive processing.

The perceptual cycle model focuses on the cyclical interactions between internal schemata, perceptual exploration and external environment, providing a structured framework for understanding how decision-making and actions emerge. However, the model treats the influence of schemata as unidirectional, and does not explicitly account for external factors, such as stress or fatigue, and internal factors, such as individual experience.

The three-level situation awareness model conceptualizes situation awareness as comprising three hierarchical stages, perception, comprehension, and projection. It integrates cognitive functions such as working memory, long-term memory, and attention into the decision-making process. Unlike the other two models, which emphasize dynamic interaction among cognitive processes, this model treats the three stages as sequential and

hierarchical. Additionally, as the model addresses schema selection, it does not explicitly account for schema modification or adaptation at higher levels of cognitive processing.

Recommendations

- No single cognitive model fully captures the dynamic and context-dependent nature of pilot decision-making and actions under unexpected conditions. Training and research should thus avoid overreliance on one single theoretical perspective.

- Reliable and validated measures of startle and surprise can complement cognitive models by empirically assessing the constructs of startle and surprise.

- Pilot training programs should be tailored to address two distinct but complementary cognitive processes in responding to unexpected events. First, intuitive and rapid pattern-matching responses can be strengthened through exposure to varied scenarios, repeated simulator training, and experience-based learning [198]. Second, more deliberate and reflective cognitive processes can be supported through meta-cognitive training, such as mnemonic strategies for managing stress, and targeted interventions to enhance situation awareness and decision-making under pressure [115, 196, 197].

7.2. Research objective 2

The second objective was to examine the relationships between existing self-report measures of startle and surprise, mental workload, stress, personality traits, and flight experience. In Chapter 3, the following research question was investigated:

Key question 2

How do personality traits and flight experience influence pilot cognitive and affective responses to simulated in-flight hazards?

The findings indicate that pilots with higher trait anxiety reported higher level of stress in response to startling and surprising events, highlighting a potential vulnerability in cognitive performance and emotional regulation under pressure. However, no significant effects were found between the trait of action orientation nor flight experience and pilot cognitive and affective responses. Flight experience not having a significant effect on startle or stress may indicate that even highly experienced pilots can possibly benefit from targeted training interventions to mitigate the impact of startle and surprise.

Additionally, strong within-individual correlations were found between perceived stress and mental workload, and between surprise and mental workload. The findings imply that the pilots who perceived a given scenario as more surprising or stressful, were also more likely to rate it more cognitively demanding, reinforcing the conceptual link between cognitive mismatch and sensemaking effort under stress conditions. The strongest between-individual correlation was found between startle and surprise, indicating potential overlap in individual susceptibility. We further investigated this in Chapter 6, in an attempt to disentangle startle and surprise.

Recommendations

- Pilot training can be individualized based on trait anxiety, with tailored stress exposure training [98, 197] and attentional control strategies [200] for those who need it most.
- Training interventions should target both novice and experienced pilots, as our results show that flight experience alone does not reduce cognitive or affective vulnerability to startle or surprise.
- Given the relationship between surprise and mental workload, pilot training programs should incorporate structured strategies that can be applied under conditions of uncertainty to mitigate cognitive overload. For instance, embedding targeted interventions within varied and unpredictable scenarios can support the development of sensemaking competencies, enabling pilots to manage surprise and stress without becoming cognitively overwhelmed.
- There is a critical need for reliable and validated tools to assess the distinct effects of startle, surprise, and related cognitive responses in dynamic and high-stress environments. To address this need, effective measurement approaches should integrate both subjective and objective indicators. For example, self-report instruments can be combined with physiological measures such as heart rate, electrodermal activity, and eye-tracking metrics that reflect arousal and cognitive load. Developing a comprehensive toolkit that combines these modalities would enhance the precision of individual difference assessments and inform the design of personalized training interventions aimed at improving performance under unexpected or high-pressure conditions.

7.3. Research objective 3

The third objective was to develop and validate psychometrically-sound self-report measures for startle and surprise. To achieve this, content validity, construct validity, criterion-related validity, and internal consistency of the developed multi-item scales were systematically evaluated across different samples and contexts.

7.3.1. Development and preliminary validation

Chapter 3 laid the foundation for the establishment of self-report measuring instruments to address the following research question:

Key question 3

How can startle and surprise be quantified with self-report measures?

This question was explored through a systematic three-phase process, which culminated in the development and validation of the multi-item Startle Inventory (Startle-I) and Surprise Inventory (Surprise-I), along with their more efficient counterparts, the Visual Analogue Scales for Startle and Surprise (Startle-VAS; Surprise-VAS). These instruments provide reliable and validated approaches for quantifying self-report startle and surprise.

The initial item set was formulated based on a comprehensive review of the literature and was subsequently refined through expert evaluation to ensure content validity. Construct

validity was conducted using participant ratings (N = 81) of nine video stimuli systematically varied to elicit startle and surprise responses. Multilevel exploratory factor analysis revealed strong psychometric properties for both the Startle and Surprise Inventories. Internal consistency was high for both measures, supporting their reliability as distinct constructs.

However, an unexpected third factor emerged, centred on cognitive comprehension, comprising items "I did not understand why it happened.", "Initially, it made no sense to me.", and "Initially, I was confused about it.". This suggests that initial incomprehension may constitute a distinct cognitive process, rather than being an inherent component of surprise response itself. This finding is particularly noteworthy given that previous conceptualizations of surprise often implicitly incorporate subsequent cognitive appraisal or sensemaking as integral components of the experience [35, 44].

One explanation for this distinct clustering could be that comprehension difficulties are not specific to surprise. Complexity, novelty or unfamiliarity of information may lead to initial incomprehension, without necessarily involving a mismatch with expectations. A second explanation could be that an initial incomprehension in response to surprising information may not always be consciously experienced, as the associated sensation can be brief. Surprise often reflects an immediate perception of expectation violation without necessarily involving prolonged incomprehension or difficulty in understanding the event. The identified distinction can be essential, as individuals may experience surprise due to unexpectedness yet not feel confused, particularly if the situation's meaning or context is familiar or can be rapidly understood.

7.3.2. Construct validity

Following the development and preliminary validation of the Startle and Surprise Inventories, Chapter 5 advanced the validation by examining the construct validity of the instruments in an ecologically-valid aviation setting, and addressed the following research question:

Key question 4

How valid are the Startle and Surprise Inventories as measures of startle and surprise in an ecologically-valid aviation context?

Using multilevel confirmatory factor analysis, the two-factor structure of the Startle and Surprise Inventories identified in Chapter 4 was confirmed in an operational aviation context. The dataset involved 26 professional pilots responding to eight simulated in-flight events designed to elicit startle and surprise responses varying in intensity. The analysis supported the distinct factors of startle and surprise and affirms the inventories' applicability to startle and surprise events as these occur in the operational context. Additionally, ratings from the inventories performed better than the corresponding visual analogue scales at differentiating levels of startle and surprise across scenarios.

These findings provide empirical support for the application of the Startle-I and Surprise-I in high-stakes domains such as aviation, where understanding the nature and impact of unexpected events is critical for ensuring safety and optimizing performance. The strong psychometric performance of the inventories in an ecologically-valid setting highlights their utility for investigating the operational impacts of startle and surprise, as well as evaluating

the effectiveness of mitigation strategies.

7.3.3. Criterion-related validity

Finally, the newly developed inventories were applied to investigate the relationships between experienced startle and surprise, and information-processing performance during in-flight events. Based on the prior research, both startle and surprise could impair information-processing performance. This study was also performed to obtain the first data regarding criterion-related validity of the Startle and Surprise Inventories. Thus, Chapter 6 addressed the research question:

Key question 5

How well do the Startle and Surprise Inventories predict pilot information-processing performance?

The Startle Inventory demonstrated strong criterion-related validity, as higher self-report startle scores significantly predicted impaired information-processing performance during in-flight events. This finding provides robust empirical support for the disruptive impact of startle on cognitive functioning, consistent with theoretical accounts that conceptualize it as a rapid activation of survival circuitry capable of interrupting ongoing cognitive processes.

The same relationship was not found for surprise. This discrepancy may be attributable to the complexity of modelling surprise effects and suggests that its impact may be more contextual, indirect, or delayed, in contrast to the more reflexive and acute nature of startle. Surprise could possibly interact with situation complexity to impact performance. In hindsight, applying the items that were removed from the inventory in Chapter 4, as they pertained to the third factor of initial incomprehension, would possibly have provided additional insight. Nevertheless, although surprise did not significantly affect immediate information-processing performance from the linear mixed-effects models, this does not necessarily imply that the Surprise Inventory lacks criterion-related validity. The between-scenario correlations still suggest that surprise contributes to the overall cognitive demand.

As part of the broader objective of assessing the Startle and Surprise Inventories, Chapters 4 and 5 also examined their reliability across multiple samples and contexts. Cronbach's α and McDonald's ω for both instruments demonstrated acceptable to excellent internal consistency, indicating their reliability in capturing pilots' subjective experiences of startle and surprise. The combination of demonstrated validity, reliability, and predictive utility for information-processing performance underscore the practical relevance of the Startle and Surprise Inventories, highlighting their potential as robust tools for assessing individual differences in affective and cognitive responses across various operational contexts.

Recommendations

- Test scenarios for pilots should be designed to explicitly distinguish between cognitive comprehension and surprise, thereby enhancing the understanding of cognitive processes underlying responses to unexpected events. To support this differentiation, future test scenarios could manipulate variables such as ambiguity, information availability, and event coherence. For example, a scenario may present a surprising but easily interpretable event (e.g., a system failure warning that is unexpected but well explained), versus one that is confusing but not necessarily surprising (e.g., an malfunction warning with unclear cause). Measuring responses to these types of events can help isolate the influence of surprise versus comprehension difficulty.

- The application of the inventories (including the previously-removed incomprehension items) can support the customization of interventions based on individual pilot profiles. By using the inventories longitudinally, it is possible to identify susceptibility to startle and surprise over time. This information can guide tailored training and interventions, such as exposure-based training, cognitive-behavioural techniques, and stress-management strategies, to help pilots develop strategies to minimize the disruptive effect of startle and maintain optimal performance under pressure.

- Although surprise did not significantly impact immediate information-processing performance, its role in building cognitive flexibility and preparedness is likely to be important. It could be that the surprise events applied were not complex or ambiguous enough to elicit significant reframing efforts that would induce sufficient mental workload to affect information processing. To more accurately reflect the real-world challenges pilots may encounter, future training scenarios could incorporate more cognitively demanding and contextually rich surprise elements.

- Additional validations are recommended to further strengthen the applicability of the newly developed instruments. In particular, criterion-related validation using physiological and behavioural data will help substantiate their robustness. Integrating physiological indicators [37, 42, 199] with validated self-report instruments may also enable a more comprehensive assessment of the real-time startle and surprise by capturing both subjective and objective responses.

- The construct validity of the Startle-I and Surprise-I have been examined in the single-pilot setting. Future research should further explore how startle and surprise manifest and interact within team-based environments. Applying the inventories in multi-operator contexts could provide valuable insights into how group dynamics and communication could influence collective performance to unexpected events or disruptions, which could support the development of team-based training interventions and systems aimed at enhancing system-wide robustness.

7.4. Final conclusions

Following the comprehensive investigations of the five key research questions under three research objectives, attention now turns to the main question:

Main question

How can startle and surprise be quantified in an aviation operational context?

The following final conclusions synthesize the key results and demonstrate how they collectively address the overarching research question:

- The Startle and Surprise Inventories represent the first systemically validated self-report measuring instruments designed to measure startle and surprise in response to specific stimuli or events, offering a robust and operationally-relevant foundation for quantifying startle and surprise in an aviation context.
- The assessment of startle and surprise can be effectively integrated into scenario-based simulation training. Application of the Startle and Surprise Inventories within structured simulations environments enables systematic tracking of operators' responses to unexpected events. This, in turn, can inform the refinement of training protocols or interventions to better support cognitive resilience and emotional regulation.
- The developed instruments can be applied to safety-critical domains, including but not limited to aviation, maritime navigation, healthcare, and military operations, to assess how startle and surprise influence operational performance, and training effectiveness in high-stakes situations involving unexpected events.

A

Content validity: expert open comments

In the phase of developing and validating the Startle and Surprise Inventories (Chapter 4), the initial set of 14 items for surprise, and 7 items for startle were formulated based on fundamental and applied literature on startle and surprise. The content validity of each item was reviewed by seven independent experts in the fields of Cognitive Science and Psychology. The experts were invited to assess the relevance of each item in capturing the experience of startle or surprise, and were also invited to provide reviews on the item formulations. This appendix presents their open comments in detail.

Expert 1: Items 3,11,14 are questionable because they might assume a native speaker level of familiarity with colloquialisms.

Expert 2: The ones marked not essential seemed to have other dimensions mixed with surprise (so they seemed less specific to surprise and/or startle. Not clear on the difference between surprise and startle).

Expert 3: I would personally not define a construct in terms of the construct (i.e., including "surprised" or "startled" as items in their respective scales). For me, several proposed "surprised" items overlap; same for "startle". I assume this was intentional, so I selected all. For any items that describe a state due to the reaction, they are relevant but not essential.

Expert 4: In rating these, I've tried to rate them as to their relevance to pilot performance, rather than the visceral response. For example, a person may feel angry but that may not necessarily affect their ability to perform. I believe that there are many occasions of surprise in one's aviation career. Some are quite significant and have serious impact on the flight, but some are just so unexpected that they result in confusion and can greatly disrupt a pilot's performance, both physically and cognitively. Hopefully there aren't many occasions of being startled. There's the old saying that flying is "hours and hours of boredom, punctuated by moments of stark terror." Regarding the term "stunned", I personally associate that with both startle & surprise, just in different degrees. I'm not sure how to differentiate that one.

Expert 5: Item 2 for startle: Better split this in two questions, one about angry and one about scared. Or leave angry out, because I think this is odd anyway in relation to startle. What are you angry about?

Item 7 for startle: Why is "burst" necessary, and can't you say "It made me quickly feel stressed". Also, I don't find frustration a logical emotion in relation to startle (maybe it is for surprise when you don't see the solution). An dit is certainly not the same as stress.

Expert 6: I ticked "not relevant" for "I predicted it beforehand" and "It was consistent with my expectations" because these two items are indications of NOT being surprised (i.e., negative wording). Is this a mistake or on purpose? Otherwise I would say that they are "essential". For the other items, I ticked essential when they are close to the definition of startle/surprise.

Expert 7: Overall, I believe the surprise portion of the scale focuses too much on "it" being something that DID happen, and currently does not adequately capture surprise when something DID NOT happen. As we (Kochan et al., 2004; 2005; Rivera et al., 2014; Talone et al., 2015) and others (including your group) have pointed out previously, accident and incident reports, as well as interviews with pilots, provide ample evidence that surprise can and does occur also frequently when something expected did not happen. Example: We expected holding and suddenly were cleared for a straight-in approach. Example: We expected to see the companion B737 to be at 11 o'clock same altitude, but there was only empty sky.

Surprise Items 6 and 9: If you want to capture the sensemaking portion of a surprise event, I would suggest that you change the wording to "At the time, ..." or "Initially, ..." as these are retrospective assessments, and what the individual saw or perceived at the time may later make perfect sense.

Surprise Items 3, 8, 10, 11, 14: I think I understand the premise of these items, which is to capture confusion associated with a surprising event, but I am not sure that using these terms is adequate, as they are (in my view) more narrative terms that are used when telling stories for emotional effect, but not what people would use to describe their reactions in a technical setting. There is a certain amount of hyperbole in these terms that, I believe, would say more about the personality of the respondent than the impact (or their assessment thereof) of the surprising event - these items may capture inter-individual variance more than inter-event variance. Also, you may get underreporting from pilots who may rarely saw "my jaw dropped" (except when the landing gear fell off the plane after take off).

Surprise Items 2 and 5: These two items are the only ones that are positively phrased but negatively scaled (i.e., a high endorsement would indicate a lack of surprise). In my experience of developing and testing scales (this is also captured in deVellis, 2003), reverse-coded items typically hang better together with one another than with the regularly coded items that are closer to them in terms of construct or content validity. Consequently, I have come to believe that the "pay attention" function of reverse-coded items is not worth the psychometric trouble, instead, adding an attention check item such as "For this, please mark the item at response 4" might be better.

Startle Items 2, 3, 7: I believe you are after the potential "fight or flight" part of what happens in response to startle here, but I would argue that these items do not adequately capture "startle" itself and are not unique to startle. In fact, if you would want to retain them, I would suggest you add an "Immediately following it" section that could relate to both surprise and startle. In fact, I would argue that we (as a scientific community) still have a long way to go in relaying to the aviation community and other stake holders that startle and surprise are not the same (Rivera et al., 2014), and breaking out commonalities across the two from their differences is thus important in scales as well.

I wonder whether any items might be useful that somehow capture the "what next?" portion of both startle and surprise. That is, are there items that could quantify the impact of the event/occurrence on the cognitive processes? The way I read them, the items currently are very focused on assessing the immediate event itself, e.g., "I was not mentally prepared for it.". What that may not capture is something along the lines of "I did not know what to do once it happened.", that is the follow-on consequence, how it interrupted on-going tasks, next steps, plans, response execution, etc. In this context, the work by Key Dismukes from NASA (and colleagues) on cockpit interruptions may hold some useful guidance on quantifying the consequence of a startle and/or surprise event on cognition, decision execution, etc.

The second thought I had was whether there might be some scale items under startle could focus more on measuring the perceived magnitude of the startling stimulus, rather than on the startle response itself. Something along the lines of "It was very

intense", "It was very loud/bright/strong"? The psychophysical literature may offer some guidance on how best to measure subjective impressions of stimulus magnitude.

B

Manual for the Startle and Surprise Inventories and Visual Analogue Scales

The contents of this appendix have been published as: J. Chen, A. Landman, O. Stroosma, M. M. van Paassen, and M. Mulder. *Manual for the Startle and Surprise Inventories and Visual Analogue Scales*. Delft, the Netherlands, 2025.

B.1. Introduction

Accurate and non-obtrusive measurement of startle and surprise is essential for advancing our understanding of human responses under stress, identifying causal factors, and assessing their distinct effects on operator performance. Insights gained from such research can inform the development of evidence-based safety protocols and targeted training interventions.

To support this aim, the Startle and Surprise Inventories (Startle-I; Surprise-I), and the Visual Analogue Scales for Startle and Surprise (Startle-VAS; Surprise-VAS) were developed and psychometrically validated as self-report measuring instruments for assessing startle and surprise in response to specific events, situations, or stimuli [167, 168]. Although initially developed within the context of human factors research in aviation, these instruments are designed to be broadly applicable across domains where unexpected events may impact human performance, including healthcare, maritime navigation, and military operations.

These instruments are intended for use in experimental, operational, and training environments, particularly those where acute stress and unexpected events are likely to occur, such as in-flight system failures. This manual provides guidance for human factors researchers and applied psychologists, on the standardized and scientifically rigorous use of the instruments. It is structured to include an overview of the instruments, administration guidelines, and a summary of their psychometric properties to support accurate application and interpretation in research and operational contexts.

B.2. Instruments overview

The Startle-I consists of six items (Appendix B.6), and the Surprise-I comprises five items (Appendix B.7). Simple and accessible language was used on the items to allow the inventories to be used in research with non-native English speakers. Response options for all items are presented on a 5-point Likert scale, which captures varying levels of agreement with each statement (1 = "*Strongly disagree*", 2 = "*Disagree*", 3 = "*Neutral*", 4 = "*Agree*", 5 = "*Strongly agree*"). For the Startle-I and Surprise-I, the score of each inventory is defined to be the average of all items' ratings, ranging from 1 to 5.

Additionally, item 2 ("I predicted it beforehand.") and item 4 ("It was consistent with my expectation.") in the Surprise-I are reverse-coded. For these items, a response of 5 should be recoded as 1, 4 as 2, 3 remains unchanged, 2 as 4, and 1 as 5.

The single-item Startle-VAS (Appendix B.8) and Surprise-VAS (Appendix B.9), each consists of a 100 mm horizontal line with tick marks at 10 mm intervals. The left endpoints are labelled with "*not startled at all*" and "*not surprised at all*", respectively. The right endpoints are labelled with "*extremely startled*" and "*extremely surprised*", respectively. For the Startle-VAS and Surprise-VAS, users are required to place a cross/mark on the horizontal line as answer to the question. The resulting score is the distance of the centre of the cross/mark to the left endpoint in centimetres, ranging from 0 to 10.

B.3. Administration guidelines

Consistent administration is essential to ensure the validity and comparability of data across studies. In the introductory text for both the multi-item inventories and visual analogue scales (VASs), clearly specify which particular stimulus or event the pronoun "it" refers to in each item. To ensure consistency across participants and conditions, standardized

instructions should be provided before application. Below is a suggested script:

> *"You are about to complete a questionnaire about your immediate reaction. Please respond honestly based on how you actually felt at the time of the specific event, not how you think you were supposed to feel or how you would normally respond. There are no right or wrong answers."*

The inventories and VASs should be completed as soon as possible after the onset of the event or stimulus of interest. Considering that startle is a rapid and transient response [20, 23] and surprise could involve cognitive appraisal [34, 35, 44], a delay in the measurement may affect the accuracy of self-reports, due to factors such as memory decay or reinterpretation of the event or stimulus.

B.4. Psychometric properties

B.4.1. Reliability

The internal consistencies of the Startle-I and Surprise-I have been evaluated across multiple samples and varied contexts. In a sample of 729 observations, 81 participants were exposed to nine video stimuli. Cronbach's α [201] indicated acceptable to excellent internal consistency, ranging from $\alpha = 0.714$ to $\alpha = 0.929$ for the Startle-I, and $\alpha = 0.843$ to $\alpha = 0.955$ for the Surprise-I [168].

In a separate study comprising 208 observations, 26 professional pilots experienced eight varied startling and surprising scenarios in an ecologically-valid aviation context. McDonald's ω [177] indicated similarly high internal consistency, with values ranging from $\omega = 0.88$ to $\omega = 0.96$ for the Startle-I, and $\omega = 0.77$ to $\omega = 0.96$ for the Surprise-I [167].

B.4.2. Validity

The Startle and Surprise Inventories

To evaluate the psychometric properties of the Startle-I and Surprise-I, multiple validity assessments were conducted. Content validity was established through expert evaluation by seven specialists in the fields of Cognitive Science and Psychology. These experts assessed the relevance of an initial set of items developed for measuring startle and surprise, derived from fundamental and applied literature [168]. An item was retained if at least 50% of the experts rated that item to be relevant for its construct [64].

Construct validity was initially examined using 729 observations from 81 participants, each of whom rated the retained items nine times following exposure to nine video clips [168]. Multilevel exploratory factor analysis with oblique, direct oblimin rotation was employed [152, 153]. Further evidence for construct validity in an ecologically-valid context was obtained from 208 observations involving 26 professional pilots, each exposed to eight simulated in-flight scenarios designed to elicit varied levels of startling and surprising responses [167]. Multilevel confirmatory factor analysis [176] was conducted to confirm the factor structures of the Startle-I and Surprise-I. These assessments collectively confirmed that the inventories can reliably and validly quantify self-report startle and surprise responses in operational contexts.

The Visual Analogue Scales for Startle and Surprise
Concurrent validity of the Startle-VAS and Surprise-VAS was assessed with Spearman correlations between the Startle-VAS and Startle-I, as well as the Surprise-VAS and Surprise-I, using 729 observations from 81 participants [168]. The ratings of Startle-VAS showed strong correlations with the Startle-I scores, $\rho = 0.778$ to $\rho = 0.877$. The ratings of Surprise-VAS highly correlated with the Surprise-I scores, $\rho = 0.681$ to $\rho = 0.903$. All correlations were statistically significant, supporting the concurrent validity of the Visual Analogue Scales for Startle and Surprise.

B.5. Contact and permissions

 When using or adapting these instruments, please cite:

doi: 10.4233/uuid:4aa39791-4d21-4427-b86a-628f52d17fbe

For questions regarding the administration, scoring, or interpretation of the Startle and Surprise Inventories and Visual Analogue Scales for Startle and Surprise, please contact the research team: startle-surprise-inventories@tudelft.nl.

B

B.6. The Startle Inventory (Startle-I)

The following statements refer to [the stimulus][1]. Please read each statement and circle the number that best represents your agreement with the statement.

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
1. It startled me.	1	2	3	4	5
2. It immediately made me feel scared or angry.	1	2	3	4	5
3. It made me physically flinch.	1	2	3	4	5
4. It caused my heart to suddenly beat harder or faster.	1	2	3	4	5
5. It shocked me.	1	2	3	4	5
6. It immediately caused stress or frustration to me.	1	2	3	4	5

[1] Specific stimulus description should be inserted here.

B.7. The Surprise Inventory (Surprise-I)

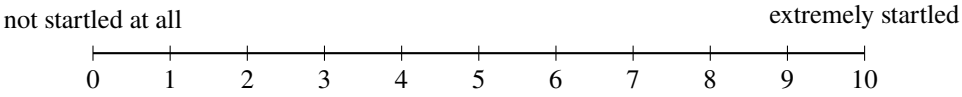
The following statements refer to [the stimulus][2]. Please read each statement and circle the number that best represents your agreement with the statement.

	Strongly disagree	Disagree	Neutral	Agree	Strongly agree
1. It surprised me.	1	2	3	4	5
2. I predicted it beforehand.	1	2	3	4	5
3. I did not see it coming.	1	2	3	4	5
4. It was consistent with my expectation.	1	2	3	4	5
5. It was unexpected.	1	2	3	4	5

[2]Specific stimulus description should be inserted here.

B.8. The Visual Analogue Scale for Startle (Startle-VAS)

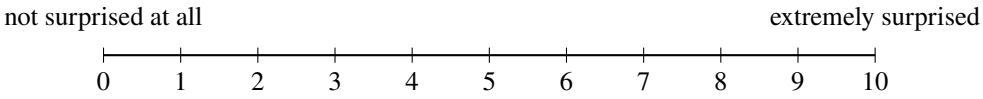
Please indicate, by placing a cross/mark on the line below[3], how **startled** you were by [the stimulus][4].



[3] When printed on A4 paper at 100% scale, the Visual Analogue Scale measures precisely 100 mm, as intended.
[4] Specific stimulus description should be inserted here.

B.9. The Visual Analogue Scale for Surprise (Surprise-VAS)

Please indicate, by placing a cross/mark on the line below[5], how **surprised** you were by [the stimulus][6].



[5] When printed on A4 paper at 100% scale, the Visual Analogue Scale measures precisely 100 mm, as intended.
[6] Specific stimulus description should be inserted here.

References

[1] Airbus. *A Statistical Analysis of Commercial Aviation Accidents 1958 - 2024*. Blagnac, France, 2025. URL: https://accidentstats.airbus.com/wp-content/uploads/2025/02/20241325_A-Statistical-analysis-of-commercial-aviation-accidents-2025-links.pdf.

[2] International Air Transport Association. *Loss of Control In-flight (LOC-I) Preventation: Beyond the Control of Pilots*. Montreal, Geneva, 2015. URL: https://www.iata.org/contentassets/b6eb2adc248c484192101edd1ed36015/loc-prevention-beyond-the-control-of-pilots.pdf.

[3] J. Wilborn and J. Foster. "Defining Commercial Transport Loss-of-control: A Quantitative Approach". In: *AIAA Atmospheric Flight Mechanics Conference and Exhibit*. AIAA 2004-4811. Providence, USA: American Institute of Aeronautics and Astronautics, 2004. DOI: 10.2514/6.2004-4811.

[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". In: *The International Journal of Aerospace Psychology* 27.1-2 (2017), pp. 2–14. DOI: 10.1080/10508414.2017.1365610.

[5] S. M. Casner, R. W. Geven, and K. T. Williams. "The Effectiveness of Airline Pilot Training for Abnormal Events". In: *Human Factors* 55.3 (2013), pp. 477–485. DOI: 10.1177/0018720812466893.

[6] V. Banks, C. K. Allison, K. Parnell, K. Plant, and N. A. Stanton. "Predicting and Mitigating Failures on the Flight Deck: An Aircraft Engine Bird Strike Scenario". In: *Ergonomics* 65.12 (2022), pp. 1672–1695. DOI: 10.1080/00140139.2022.2048897.

[7] International Air Transport Association. *Loss of Control In-Flight Accident Analysis Report Edition 2019 Guidance Material and Best Practices*. Montreal, Canada, 2019. URL: https://www.iata.org/contentassets/b6eb2adc248c484192101edd1ed36015/loc-i_2019.pdf.

[8] N. Sarter, D. Woods, and C. Billings. "Automation Surprises". In: *Handbook of Human Factors and Ergonomics*. Vol. 2. John Wiley and Sons, 1997, pp. 1926–1943.

[9] R. J. de Boer and K. Hurts. "Automation Surprise: Results of a Field Survey of Dutch Pilots". In: *Aviation Psychology and Applied Human Factors* 7.1 (2017), pp. 28–41. DOI: 10.1027/2192-0923/a000113.

[10] R. K. Dismukes, T. E. Goldsmith, and J. A. Kochan. *Effects of Acute Stress on Aircrew Performance: Literature Review and Analysis of Operational Aspects*. Washington, USA, 2015. URL: https://ntrs.nasa.gov/api/citations/20190002685/downloads/20190002685.pdf.

[11] R. J. de Boer and S. W. A. Dekker. “Models of Automation Surprise: Results of a Field Survey in Aviation”. In: *Safety* 3.20 (2017). DOI: 10.3390/safety3030020.

[12] Federal Aviation Administration. *Upset Prevention and Recovery Training*. Washington, USA, 2015. URL: https://www.faa.gov/documentLibrary/media/Advisory_Circular/AC_120-111_CHG_1_Ed_Upd_3.pdf.

[13] E. Salas, J. E. Driskell, and S. Hughes. “Introduction: The Study of Stress and Human Performance”. In: *Stress and Human Performance*. Ed. by J. E. Driskell and E. Salas. Psychology Press, 2013, pp. 1–45. DOI: 10.4324/9780203772904.

[14] European Aviation Safety Agency. *Loss-of-Control Prevention and Recovery Training*. Cologne, Germany, 2015. URL: https://www.easa.europa.eu/sites/default/files/dfu/NPA%202015-13.pdf.

[15] C. D. Spielberger. *Manual for the State-Trait Anxiety Inventory: STAI (Form Y)*. Consulting Psychologists Press, 1983.

[16] N. V. Ramanaiah, M. Franzen, and T. Schill. “A Psychometric Study of the State-Trait Anxiety Inventory”. In: *Journal of Personality Assessment* 47.5 (1983), pp. 531–535. DOI: 10.1207/s15327752jpa4705_14.

[17] F. R. H. Zijlstra. “Efficiency in Work Behavior: A Design Approach for Modern Tools”. PhD thesis. Delft, The Netherlands: Delft University of Technology, 1993. URL: https://resolver.tudelft.nl/uuid:d97a028b-c3dc-4930-b2ab-a7877993a17f.

[18] S. G. Hart and L. E. Staveland. “Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research”. In: *Advances in Psychology* 52 (1988), pp. 139–183. DOI: 10.1016/S0166-4115(08)62386-9.

[19] 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 58th Annual Meeting*. Vol. 58. 1. SAGE Publications, 2014, pp. 1047–1051. DOI: 10.1177/1541931214581218.

[20] M. Koch. “The Neurobiology of Startle”. In: *Progress in Neurobiology* 59.2 (1999), pp. 107–128. DOI: 10.1016/S0301-0082(98)00098-7.

[21] Y. E. M. Dreissen, M. J. Bakker, J. H. T. M. Koelman, and M. A. J. Tijssen. “Exaggerated Startle Reactions”. In: *Clinical Neurophysiology* 123.1 (2012), pp. 34–44. DOI: 10.1016/j.clinph.2011.09.022.

[22] T. D. Blumenthal. “Presidential Address 2014: The More-or-Less Interrupting Effects of the Startle Response”. In: *Psychophysiology* 52.11 (2015), pp. 1417–1431. DOI: 10.1111/psyp.12506.

[23] W. L. Martin, P. S. Murray, P. R. Bates, and P. S. Y. Lee. “Fear-Potentiated Startle: A Review from an Aviation Perspective”. In: *The International Journal of Aviation Psychology* 25.2 (2015), pp. 97–107. DOI: 10.1080/10508414.2015.1128293.

[24] A. S. P. Jansen, X. van Nguyen, V. Karpitskiy, T. C. Mettenleiter, and A. D. Loewy. “Central Command Neurons of the Sympathetic Nervous System: Basis of the Fight-or-Flight Response”. In: *Science* 270.5236 (1995), pp. 644–646. DOI: 10.1126/science.270.5236.644.

[25] A. Papadimitriou and K. N. Priftis. “Regulation of the Hypothalamic-Pituitary-Adrenal Axis”. In: *Neuroimmunomodulation* 16.5 (2009), pp. 265–271. DOI: 10.1159/000216184.

[26] S. Holand, A. Girard, D. Laude, C. Meyer-Bisch, and J.-L. Elghozi. “Effects of an Auditory Startle Stimulus on Blood Pressure and Heart Rate in Humans”. In: *Journal of Hypertension* 17.12 (1999), pp. 1893–1897.

[27] V. B. Nakagawara, R. W. Montgomery, A. Dillard, L. McLin, and C. W. Connor. *The Effects of Laser Illumination on Operational and Visual Performance of Pilots during Final Approach*. Washington, USA, 2004. URL: https://www.faa.gov/sites/faa.gov/files/data_research/research/med_humanfacs/oamtechreports/0312.pdf.

[28] R. I. Thackray and R. M. Touchstone. *Rate of Initial Recovery and Subsequent Radar Monitoring Performance Following a Simulated Emergency Involving Startle*. Washington, USA, 1983. URL: https://www.faa.gov/sites/faa.gov/files/data_research/research/med_humanfacs/oamtechreports/AM83-13.pdf.

[29] Civil Aviation Authority. *Flight-crew human factors handbook CAP 737*. Crawley, UK, 2016. URL: https://www.caa.co.uk/publication/download/14984.

[30] W.-U. Meyer, R. Reisenzein, and A. Schützwohl. “Toward a Process Analysis of Emotions: The Case of Surprise”. In: *Motivation and Emotion* 21.3 (1997), pp. 251–274. DOI: 10.1023/A:1024422330338.

[31] G. Horstmann. “Latency and Duration of the Action Interruption in Surprise”. In: *Cognition & Emotion* 20.2 (2006), pp. 242–273. DOI: 10.1080/02699930500262878.

[32] U. Neisser. *Cognitive Psychology: Classic Edition*. Psychology Press, 2014. DOI: 10.4324/9781315736174.

[33] J. Hansen and S. Topolinski. “An Exploratory Mindset Reduces Preference for Prototypes and Increases Preference for Novel Exemplars”. In: *Cognition and Emotion* 25.4 (2011), pp. 709–716. DOI: 10.1080/02699931.2010.496994.

[34] R. Reisenzein, G. Horstmann, and A. Schützwohl. “The Cognitive-Evolutionary Model of Surprise: A Review of the Evidence”. In: *Topics in Cognitive Science* 11.1 (2019), pp. 50–74. DOI: 10.1111/tops.12292.

[35] 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”. In: *Human Factors* 59.8 (2017), pp. 1161–1172. DOI: 10.1177/0018720817723428.

[36] M. W. Eysenck and M. G. Calvo. “Anxiety and Performance: The Processing Efficiency Theory”. In: *Cognition and Emotion* 6.6 (1992), pp. 409–434. DOI: 10.1080/02699939208409696.

[37] T. D. Blumenthal, B. N. Cuthbert, D. L. Filion, S. Hackley, O. V. Lipp, and A. Van Boxtel. "Committee Report: Guidelines for Human Startle Eyeblink Electromyographic Studies". In: *Psychophysiology* 42.1 (2005), pp. 1–15. DOI: 10.1111/j.1469-8986.2005.00271.x.

[38] H. S. Hoffman, R. R. Marsh, and C. L. Stitt. "Tests of a Principle of Reflex Modification: Modification of the Human Eyeblink-reflex is Independent of the Intensity of the Reflex-Eliciting Stimulus". In: *Animal Learning & Behavior* 8 (1980), pp. 81–84. DOI: 10.3758/BF03209733.

[39] M. A. Flaten and T. D. Blumenthal. "A Parametric Study of the Separate Contributions of the Tactile and Acoustic Components of Airpuffs to the Blink Reflex". In: *Biological Psychology* 48.3 (1998), pp. 227–234. DOI: 10.1016/S0301-0511(98)00018-0.

[40] J.-G. Gehricke, E. M. Ornitz, and P. Siddarth. "Differentiating between Reflex and Spontaneous Blinks Using Simultaneous Recording of the Orbicularis Oculi Electromyogram and the Electro-Oculogram in Startle Research". In: *International Journal of Psychophysiology* 44.3 (2002), pp. 261–268. DOI: 10.1016/s0167-8760(02)00008-9.

[41] C. Evinger and K. A. Manning. "Pattern of Extraocular Muscle Activation during Reflex Blinking". In: *Experimental Brain Research* 92 (1993), pp. 502–506. DOI: 10.1007/BF00229039.

[42] L. Kinney and D. O'Hare. "Responding to an Unexpected In-Flight Event: Physiological Arousal, Information Processing, and Performance". In: *Human Factors* 62.5 (2020), pp. 737–750. DOI: 10.1177/0018720819854830.

[43] L. A. Clark and D. Watson. "Constructing Validity: Basic Issues in Objective Scale Development". In: *Psychological Assessment* 7.3 (2016), pp. 309–319. DOI: 10.1037/1040-3590.7.3.309.

[44] M. K. Noordewier, S. Topolinski, and E. van Dijk. "The Temporal Dynamics of Surprise". In: *Social and Personality Psychology Compass* 10.3 (2016), pp. 136–149. DOI: 10.1111/spc3.12242.

[45] J. W. Antony, T. H. Hartshorne, K. Pomeroy, T. M. Gureckis, U. Hasson, S. D. McDougle, and K. A. Norman. "Behavioral, Physiological, and Neural Signatures of Surprise during Naturalistic Sports Viewing". In: *Neuron* 109.2 (2021), pp. 377–390. DOI: 10.1016/j.neuron.2020.10.029.

[46] G. Hajcak and D. Foti. "Errors Are Aversive: Defensive Motivation and the Error-related Negativity". In: *Psychological Science* 19.2 (2008), pp. 103–108. DOI: 10.1111/j.1467-9280.2008.02053.x.

[47] M. J. A. G. Henckens, E. J. Hermans, Z. Pu, M. Joëls, and G. Fernández. "Stressed Memories: How Acute Stress Affects Memory Formation in Humans". In: *Journal of Neuroscience* 29.32 (2009), pp. 10111–10119. DOI: 10.1523/JNEUROSCI.1184-09.2009.

[48] N. A. Kloosterman, T. Meindertsma, A. M. van Loon, V. A. F. Lamme, Y. S. Bonneh, and T. H. Donner. "Pupil Size Tracks Perceptual Content and Surprise". In: *European Journal of Neuroscience* 41.8 (2015), pp. 1068–1078. DOI: 10.1111/ejn.12859.

[49] C. E. Izard, D. Z. Libero, P. Putnam, and O. Haynes. "Stability of Emotion Experiences and Their Relations to Traits of Personality". In: *Journal of Personality and Social Psychology* 64.5 (1993), pp. 847–860. DOI: 10.1037//0022-3514.64.5.847.

[50] D. Watson and L. A. Clark. *The PANAS-X: Manual for the Positive and Negative Affect Schedule-Expanded Form*. Lowa, USA, 1994. DOI: 10.17077/48vt-m4t2.

[51] W. E. Kotsch, D. W. Gerbing, and L. E. Schwartz. "The Construct Validity of the Differential Emotions Scale as Adapted for Children and Adolescents". In: *Measuring Emotions in Infants and Children: Based on Seminars Sponsored by the Committee on Social and Affective Development During Childhood of the Social Science Research Council*. Ed. by C. E. Izard. Vol. 1. Cambridge, UK: Cambridge University Press, 1982, pp. 251–278.

[52] J.-S. Ricard-St-Aubin, F. L. Philippe, G. Beaulieu-Pelletier, and S. Lecours. "Validation Francophone de L'échelle des Émotions Différentielles IV (EED-IV) French Validation of the Differential Emotions Scale IV (DES-IV)". In: *European Review of Applied Psychology* 60.1 (2010), pp. 41–53. DOI: 10.1016/j.erap.2009.05.001.

[53] D. Watson and J. Vaidya. "Mood Measurement: Current Status and Future Directions". In: *Handbook of Psychology*. Ed. by I. B. Weiner. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2003, pp. 351–375. DOI: 10.1002/0471264385.wei0214.

[54] W.-U. Meyer, M. Niepel, U. Rudolph, and A. Schützwohl. "An Experimental Analysis of Surprise". In: *Cognition & Emotion* 5.4 (1991), pp. 295–311. DOI: 10.1080/02699939108411042.

[55] R. Reisenzein. "The Subjective Experience of Surprise". In: *The Message within: The Role of Subjective Experience in Social Cognition and Behavior*. Ed. by H. Bless and J. P. Forgas. Hove, UK: Psychology Press, 2013, pp. 262–282.

[56] T. Q. Tran, R. L. Boring, D. D. Dudenhoeffer, B. P. Hallbert, M. Keller, and T. M. Anderson. "Advantages and Disadvantages of Physiological Assessment for Next Generation Control Room Design". In: *2007 IEEE 8th Human Factors and Power Plants and HPRCT 13th Annual Meeting*. 2007, pp. 259–263. DOI: 10.1109/HFPP.2007.4413216.

[57] M. M. Bradley and P. J. Lang. "Affective Reactions to Acoustic Stimuli". In: *Psychophysiology* 37.2 (2000), pp. 204–215.

[58] C. Grillon and J. Baas. "A Review of the Modulation of the Startle Reflex by Affective States and Its Application in Psychiatry". In: *Clinical Neurophysiology* 114.9 (2003), pp. 1557–1579. DOI: 10.1016/s1388-2457(03)00202-5.

[59] J. C. McCroskey. "Self-Report Measurement". In: *Avoiding Communication: Shyness, Reticence, and Communication Apprehension*. Ed. by J. A. Daly and J. C. McCroskey. Los Angeles, CA, USA: SAGE Publications, 1984, pp. 81–94.

[60] C. J. L. Rossato, M. A. Uphill, J. Swain, and D. A. Coleman. "The Development and Preliminary Validation of the Challenge and Threat in Sport (CAT-Sport) Scale". In: *International Journal of Sport and Exercise Psychology* 16.2 (2018), pp. 164–177. DOI: 10.1080/1612197X.2016.1182571.

[61] I. B. Mauss and M. D. Robinson. "Measures of Emotion: A Review". In: *Cognition and Emotion* 23.2 (2009), pp. 209–237. DOI: 10.1080/02699930802204677.

[62] U. Neisser. *Cognition and Reality: Principles and Implications of Cognitive Psychology*. W. H. Freeman and Company, 1976.

[63] M. R. Endsley. "Toward a Theory of Situation Awareness in Dynamic Systems". In: *Human Factors* 37.1 (1995), pp. 32–64. DOI: 10.1518/001872095779049543.

[64] R. F. DeVellis. *Scale Development: Theory and Applications*. Third. SAGE Publications, 2012.

[65] A. Chang, G. A. Justin, P. M. Mathews, and J. D. Auran. "Unanticipated Rare Adverse Events and the Surgeon Startle Response in Ophthalmic Surgery". In: *Eye* 36.1 (2022), pp. 3–4. DOI: 10.1038/s41433-021-01703-x.

[66] D. M. Gaba, S. K. Howard, and S. D. Small. "Situation Awareness in Anesthesiology". In: *Human Factors* 37.1 (1995), pp. 20–31. DOI: 10.1518/001872095779049435.

[67] C. M. Janelle and B. D. Hatfield. "Visual Attention and Brain Processes That Underlie Expert Performance: Implications for Sport and Military Psychology". In: *Military Psychology* 20.sup1 (2008), pp. 39–69. DOI: 10.1080/08995600701804798.

[68] P. J. Lang, M. M. Bradley, and B. N. Cuthbert. "Motivated Attention: Affect, Activation, and Action". In: *Attention and Orienting: Sensory and Motivational Processes*. Ed. by P. J. Lang, R. F. Simons, and M. T. Balaban. Psychology Press, 1997, pp. 97–135.

[69] U. Dimberg, M. Thunberg, and K. Elmehed. "Unconscious Facial Reactions to Emotional Facial Expressions". In: *Psychological Science* 11.1 (2000), pp. 86–89. DOI: 10.1111/1467-9280.00221.

[70] M. Green. "How Long Does It Take to Stop? Methodological Analysis of Driver Perception-Brake Times". In: *Transportation Human Factors* 2.3 (2000), pp. 195–216. DOI: 10.1207/STHF0203_1.

[71] S. D. Kreibig. "Autonomic Nervous System Activity in Emotion: A Review". In: *Biological Psychology* 84 (2010), pp. 394–421. DOI: 10.1016/j.biopsycho.2010.03.010.

[72] I. B. Mauss, R. W. Levenson, L. McCarter, F. H. Wilhelm, and J. J. Gross. "The Tie That Binds? Coherence among Emotion Experience, Behavior, and Physiology". In: *Emotion* 5.2 (2005), pp. 175–190. DOI: 10.1037/1528-3542.5.2.175.

[73] M. M. Bradley and P. J. Lang. "Measuring Emotion: The Self-Assessment Manikin and the Semantic Differential". In: *Journal of Behavior Therapy and Experimental Psychiatry* 25.1 (1994), pp. 49–59. DOI: 10.1016/0005-7916(94)90063-9.

[74] J. T. Cacioppo and L. G. Tassinary. "Inferring Psychological Significance from Physiological Signals". In: *The American Psychologist* 45.1 (1990), pp. 16–28.

[75] K. L. Plant and N. A. Stanton. *Distributed Cognition and Reality: How Pilots and Crews Make Decisions*. CRC Press, 2016.

[76] M. R. Endsley. "A Taxonomy of Situation Awareness Errors". In: *Human Factors in Aviation Operations* 3.2 (1995), pp. 287–292.

[77] 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*. Ed. by R. R. Hoffman. New Jersey, USA: Lawrence Erlbaum Associates Publishers, 2007.

[78] F. C. Bartlett. *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press, 1932.

[79] K. E. Weick. *Sensemaking in Organizations*. SAGE Publications, 1995.

[80] G. Klein. *Sources of Power: How People Make Decisions*. MIT press, 2017. DOI: 10.7551/mitpress/11307.001.0001.

[81] A. Rankin, R. Woltjer, and J. Field. "Sensemaking Following Surprise in the Cockpit: A Re-framing Problem". In: *Cognition, Technology & Work* 18.4 (2016), pp. 623–642. DOI: 10.1007/s10111-016-0390-2.

[82] J. W. Senders. "The Human Operator As a Monitor and Controller of Multidegree of Freedom Systems". In: *IEEE Transactions on Human Factors in Electronics* HFE-5.1 (1964), pp. 2–5. DOI: 10.1109/THFE.1964.231647.

[83] J. Rasmussen. "Skills, Rules, and Knowledge; Signals, Signs, and Symbols, and Other Distinctions in Human Performance Models". In: *IEEE Transactions on Systems, Man, and Cybernetics* SMC-13.3 (1983), pp. 257–266. DOI: 10.1109/TSMC.1983.6313160.

[84] J. Globisch, A. O. Hamm, F. Esteves, and A. Öhman. "Fear Appears Fast: Temporal Course of Startle Reflex Potentiation in Animal Fearful Subjects". In: *Psychophysiology* 36.1 (1999), pp. 66–75. DOI: 10.1017/s0048577299970634.

[85] J. E. LeDoux. "Emotion, Memory and the Brain". In: *Scientific American* 270.6 (1994), pp. 50–57. DOI: 10.1038/scientificamerican0694-50.

[86] S. Dehaene and L. Naccache. "Towards a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework". In: *Cognition* 79.1-2 (2001), pp. 1–37. DOI: 10.1016/s0010-0277(00)00123-2.

[87] J. M. Mandler. *Stories, Scripts, and Scenes: Aspects of Schema Theory*. Psychology Press, 1984. DOI: https://doi.org/10.4324/9781315802459.

[88] M. J. Adams, Y. J. Tenney, and R. W. Pew. "Situation Awareness and the Cognitive Management of Complex Systems". In: *Human Factors* 37.1 (1995), pp. 85–104. DOI: 10.1518/001872095779049946.

[89] A. Treisman and R. Paterson. "Emergent Features, Attention, and Object Perception". In: *Journal of Experimental Psychology: Human Perception and Performance* 10.1 (1984), pp. 12–31. DOI: 10.1037//0096-1523.10.1.12.

[90] C. D. Wickens, W. S. Helton, J. G. Hollands, and S. Banbury. *Engineering Psychology and Human Performance*. Routledge, 2021. DOI: 10.4324/9781003177616.

[91] D. Kahneman. *Attention and Effort*. Prentice-Hall Inc., 1973.

[92] G. D. Logan, S. E. Taylor, and J. L. Etherton. "Attention and Automaticity: Toward a Theoretical Integration". In: *Psychological Research* 62.2 (1999), pp. 165–181. DOI: 10.1007/s004260050049.

[93] R. C. Schank and R. P. Abelson. *Scripts, Plans, Goals, and Understanding: An Inquiry Into Human Knowledge Structures*. Lawrence Erlbaum Associates, lnc., 1977, pp. 151–157.

[94] Air Accidents Investigation Branch. *Aircraft Accident Report : Report on the Serious Incident to Saab 2000, G-Lgno Approximately 7 Nm East of Sumburgh Airport, Shetland 15 December 2014*. London, UK, 2016.

[95] National Transportation Safety Board. *Aircraft Accident Report: Loss of Thrust in Both Engines After Encountering a Flock of Birds and Subsequent Ditching on the Hudson River, Us Airways Flight 1549, Airbus A320-214, N106us Weehawken, New Jersey January 15, 2009*. Washington, USA, 2009. URL: https://www.ntsb.gov/investigations/accidentreports/reports/aar1003.pdf.

[96] C. D. Wickens. "Attentional Tunneling and Task Management". In: *2005 International Symposium on Aviation Psychology*. 2005, pp. 812–817. URL: https://corescholar.libraries.wright.edu/isap_2005/121.

[97] F. Dehais, M. Causse, F. Vachon, N. Régis, E. Menant, and S. Tremblay. "Failure to Detect Critical Auditory Alerts in the Cockpit: Evidence for Inattentional Deafness". In: *Human Factors* 56.4 (2013), pp. 631–644. DOI: 10.1177/0018720813510735.

[98] M. W. Eysenck, N. Derakshan, R. Santos, and M. G. Calvo. "Anxiety and Cognitive Performance: Attentional Control Theory". In: *Emotion* 7.2 (2007), pp. 336–353. DOI: 10.1037/1528-3542.7.2.336.

[99] H. Kohn. "The Effect of Variations of Intensity of Experimentally Induced Stress Situations upon Certain Aspects of Perception and Performance". In: *The Journal of Genetic Psychology* 85.2 (1954), pp. 289–304. DOI: 10.1080/00221325.1954.10532884.

[100] J. A. Easterbrook. "The Effect of Emotion on Cue Utilization and the Organization of Behavior". In: *Psychological Review* 66.3 (1959), pp. 183–201. DOI: 10.1037/h0047707.

[101] M. A. Staal and A. E. Bolton. "Cognitive Performance and Resilience to Stress". In: *Biobehavioral Resilience to Stress*. Ed. by B. J. Lukey and V. Tepe. Routledge, 2008, pp. 259–299. DOI: 10.1201/9781420071788.

[102] J. Klayman and Y.-w. Ha. "Confirmation, Disconfirmation, and Information in Hypothesis Testing". In: *Psychological Review* 94.2 (1987), pp. 211–228. DOI: 10.1037/0033-295X.94.2.211.

[103] A. Tversky and D. Kahneman. "Judgment under Uncertainty: Heuristics and Biases". In: *Science* 185.4157 (1974), pp. 1124–1131. DOI: 10.1126/science.185.4157.1124.

[104] S. Walmsley and A. Gilbey. "Cognitive Biases in Visual Pilots' Weather-Related Decision Making". In: *Applied Cognitive Psychology* 30.4 (2016), pp. 532–543. DOI: 10.1002/acp.3225.

[105] S. Darke. "Anxiety and Working Memory Capacity". In: *Cognition and Emotion* 2.2 (1988), pp. 145–154. DOI: 10.1080/02699938808408071.

[106] S. Day. *Airline Passengers Wait to Be Rescued on the Wings of a Us Airways Airbus 320 Jetliner That Safely Ditched in the Hudson River in New York*. Associated Press. 2009. URL: https://newsroom.ap.org/editorial-photos-videos/detail?itemid=7410f942432d41da80c7afd58de0b8b4&mediatype=photo.

[107] M. W. Eysenck, S. Payne, and N. Derakshan. "Trait Anxiety, Visuospatial Processing, and Working Memory". In: *Cognition & Emotion* 19.8 (2005), pp. 1214–1228. DOI: 10.1080/02699930500260245.

[108] K. Couric. *Capt. Sully Worried About Airline Industry*. CBS NEWS. 2009. URL: https://www.cbsnews.com/news/capt-sully-worried-about-airline-industry/.

[109] M. Causse, Z. K. Chua, and F. Rémy. "Influences of Age, Mental Workload, and Flight Experience on Cognitive Performance and Prefrontal Activity in Private Pilots: A Fnirs Study". In: *Scientific Reports* 9.7688 (2019). DOI: 10.1038/s41598-019-44082-w.

[110] D. Morrow, L. S. Miller, H. Ridolfo, N. Kokayeff, D. Chang, U. Fischer, and E. Stine-Morrow. "Expertise and Aging in a Pilot Decision-Making Task". In: *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting*. Vol. 48. 2. SAGE Publications, 2004, pp. 228–232. DOI: 10.1177/154193120404800201.

[111] G. Klein, B. Moon, and R. R. Hoffman. "Making Sense of Sensemaking 2: A Macrocognitive Model". In: *IEEE Intelligent Systems* 21.5 (2006), pp. 88–92. DOI: 10.1109/MIS.2006.100.

[112] A. Landman, P. van Oorschot, M. 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". In: *Human Factors* 60.6 (2018), pp. 793–805. DOI: 10.1177/0018720818779928.

[113] D. Boud and G. Feletti, eds. *The Challenge of Problem-Based Learning*. Psychology Press, 1997.

[114] R. Clewley and J. Nixon. "Penguins, Birds, and Pilot Knowledge: Can an Overlooked Attribute of Human Cognition Explain Our Most Puzzling Aircraft Accidents?" In: *Human Factors* 64.4 (2020), pp. 662–674. DOI: 10.1177/0018720820960877.

[115] A. Landman, S. H. van Middelaar, E. L. Groen, M. M. van Paassen, A. W. Bronkhorst, and M. Mulder. "The Effectiveness of a Mnemonic-Type Startle and Surprise Management Procedure for Pilots". In: *The International Journal of Aerospace Psychology* 30.3-4 (2020), pp. 104–118. DOI: 10.1080/24721840.2020.1763798.

[116] J. H. Johnston and J. A. Cannon-Bowers. "Training for Stress Exposure". In: *Stress and Human Performance*. Ed. by J. E. Driskell and E. Salas. Psychology Press, 1996, pp. 223–256. DOI: https://doi.org/10.4324/9780203772904.

[117] J. E. Driskell, E. Salas, J. H. Johnston, and T. N. Wollert. "Stress Exposure Training: An Event-Based Approach". In: *Performance Under Stress*. Ed. by J. Szalma and P. A. Hancock. CRC Press, 2008, pp. 287–302. DOI: 10.1201/9781315599946.

[118] K. Ebner and N. Singewald. "Individual Differences in Stress Susceptibility and Stress Inhibitory Mechanisms". In: *Current Opinion in Behavioral Sciences* 14 (2017), pp. 54–64. DOI: 10.1016/j.cobeha.2016.11.016.

[119] R. R. McCrae, P. T. Costa Jr, and T. A. Martin. "The NEO–PI–3: A More Readable Revised NEO Personality Inventory". In: *Journal of Personality Assessment* 84.3 (2005), pp. 261–270. DOI: 10.1207/s15327752jpa8403_05.

[120] G. D. Wilson, V. Kumari, J. A. Gray, and P. J. Corr. "The Role of Neuroticism in Startle Reactions to Fearful and Disgusting Stimuli". In: *Personality and Individual Differences* 29.6 (2000), pp. 1077–1082. DOI: 10.1016/S0191-8869(99)00255-X.

[121] P. Jylhä and E. Isometsä. "The Relationship of Neuroticism and Extraversion to Symptoms of Anxiety and Depression in the General Population". In: *Depression and Anxiety* 23.5 (2006), pp. 281–289. DOI: 10.1002/da.20167.

[122] M. Vollrath and S. Torgersen. "Personality Types and Coping". In: *Personality and Individual Differences* 29.2 (2000), pp. 367–378. DOI: 10.1016/S0191-8869(99)00199-3.

[123] K. A. Byrne, C. D. Silasi-Mansat, and D. A. Worthy. "Who Chokes under Pressure? The Big Five Personality Traits and Decision-Making under Pressure". In: *Personality and Individual Differences* 74 (2015), pp. 22–28. DOI: 10.1016/j.paid.2014.10.009.

[124] A. R. Hidalgo-Muñoz, D. Mouratille, R. El-Yagoubi, Y. Rouillard, N. Matton, and M. Causse. "Conscientiousness in Pilots Correlates with Electrodermal Stability: Study on Simulated Flights under Social Stress". In: *Safety* 7.49 (2021). DOI: 10.3390/safety7020049.

[125] H. Afshar, H. R. Roohafza, A. H. Keshteli, M. Mazaheri, A. Feizi, and P. Adibi. "The Association of Personality Traits and Coping Styles According to Stress Level". In: *Journal of Research in Medical Sciences: The Official Journal of Isfahan University of Medical Sciences* 20.4 (2015), pp. 353–358.

[126] C. D. Spielberger. "Anxiety: State-Trait-Process". In: *Stress and Anxiety*. Ed. by C. D. Spielberger and I. G. Sarason. Vol. 1. Hemisphere Publishing Corporation, 1975, pp. 115–143.

[127] J. S. Campbell, M. Castaneda, and S. Pulos. "Meta-Analysis of Personality Assessments As Predictors of Military Aviation Training Success". In: *The International Journal of Aviation Psychology* 20.1 (2009), pp. 92–109. DOI: 10.1080/10508410903415872.

[128] R. F. Baumeister, T. F. Heatherton, and D. M. Tice. *Losing Control: How and Why People Fail at Self-Regulation*. Academic Press, 1994.

[129] C. Englert, A. Bertrams, and O. Dickhäuser. "Dispositional Self-Control Capacity and Trait Anxiety As Relates to Coping Styles". In: *Psychology* 2.6 (2011), pp. 598–604. DOI: 10.4236/psych.2011.26092.

[130] J. P. Tangney, R. F. Baumeister, and A. L. Boone. "High Self-Control Predicts Good Adjustment, Less Pathology, Better Grades, and Interpersonal Success". In: *Journal of Personality* 72.2 (2004), pp. 271–324. DOI: 10.1111/j.0022-3506.2004.00263.x.

[131] J. Kuhl. "A Theory of Self-Regulation: Action Versus State Orientation, Self-Discrimination, and Some Applications". In: *Applied Psychology* 41.2 (1992), pp. 97–129. DOI: 10.1111/j.1464-0597.1992.tb00688.x.

[132] J. Kuhl and J. Beckmann. *Volition and Personality: Action Versus State Orientation*. Hogrefe & Huber Publishers, 1994.

[133] N. B. Jostmann, S. L. Koole, N. Y. van der Wulp, and D. A. Fockenberg. "Subliminal Affect Regulation: The Moderating Role of Action vs. State Orientation". In: *European Psychologist* 10.3 (2005), pp. 209–217. DOI: 10.1027/1016-9040.10.3.209.

[134] A. Landman, A. Nieuwenhuys, and R. R. D. Oudejans. "Decision-Related Action Orientation Predicts Police Officers' Shooting Performance under Pressure". In: *Anxiety, Stress, & Coping* 29.5 (2016), pp. 570–579. DOI: 10.1080/10615806.2015.1070834.

[135] J. Kuhl. "Individual Differences in Self-Regulation". In: *Motivation and Action*. Ed. by J. Heckhausen and H. Heckhausen. Springer International Publishing, 2018, pp. 529–577. DOI: 10.1007/978-3-319-65094-4_13.

[136] M. Piras, A. Landman, M. M. van Paassen, O. Stroosma, E. Groen, and M. Mulder. "Easy as ABC: A Mnemonic Procedure for Managing Startle and Surprise". In: *22nd International Symposium on Aviation Psychology*. 39. 2023. URL: https://corescholar.libraries.wright.edu/isap_2023/31.

[137] J. M. Diefendorff, R. J. Hall, R. G. Lord, and M. L. Strean. "Action-State Orientation: Construct Validity of a Revised Measure and Its Relationship to Work-Related Variables". In: *Journal of Applied Psychology* 85.2 (2000), pp. 250–263. DOI: 10.1037/0021-9010.85.2.250.

[138] A. Blunt and T. A. Pychyl. "Volitional Action and Inaction in the Lives of Undergraduate Students: State Orientation, Procrastination and Proneness to Boredom". In: *Personality and Individual Differences* 24.6 (1998), pp. 837–846. DOI: 10.1016/S0191-8869(98)00018-X.

[139] C. D. Spielberger. *State-Trait Anxiety Inventory: Bibliography*. Second. Consulting Psychologists Press, 1989.

[140] C. D. Spielberger, F. Gonzalez-Reigosa, A. Martinez-Urrutia, L. F. S. Natalicio, and D. S. Natalicio. "The State-Trait Anxiety Inventory". In: *Revista Interamericana De Psicologia/interamerican Journal of Psychology* 5.3 & 4 (1971), pp. 145–158.

[141] A. G. Hedberg. "Review of State-Trait Anxiety Inventory". In: *Professional Psychology* 3.4 (1972), pp. 389–390. DOI: 10.1037/h0020743.

[142] I. L. D. Houtman and F. C. Bakker. "The Anxiety Thermometer: A Validation Study". In: *Journal of Personality Assessment* 53.3 (1989), pp. 575–582. DOI: 10.1207/s15327752jpa5303_14.

[143] S. G. Hill, H. P. Iavecchia, J. C. Byers, A. C. Bittner Jr, A. L. Zaklade, and R. E. Christ. "Comparison of Four Subjective Workload Rating Scales". In: *Human Factors* 34.4 (1992), pp. 429–439. DOI: 10.1177/001872089203400405.

[144] J. Hox, M. Moerbeek, and R. van de Schoot. *Multilevel Analysis: Techniques and Applications*. Second. Routledge, 2010. DOI: 10.4324/9780203852279.

[145] N. S. Endler. "Stress, Anxiety and Coping: The Multidimensional Interaction Model". In: *Canadian Psychology/psychologie Canadienne* 38.3 (1997), pp. 136–153. DOI: 10.1037/0708-5591.38.3.136.

[146] T. Saunders, J. E. Driskell, J. H. Johnston, and E. Salas. "The Effect of Stress Inoculation Training on Anxiety and Performance". In: *Journal of Occupational Health Psychology* 1.2 (1996), pp. 170–186. DOI: 10.1037//1076-8998.1.2.170.

[147] M. A. Staal. *Stress, Cognition, and Human Performance: A Literature Review and Conceptual Framework*. Washington, USA, 2004. URL: https://ntrs.nasa.gov/api/citations/20060017835/downloads/20060017835.pdf.

[148] P. J. Lang, M. M. Bradley, and B. N. Cuthbert. "Emotion, Attention, and the Startle Reflex". In: *Psychological Review* 97.3 (1990), pp. 377–395.

[149] T. D. Blumenthal. "The Startle Response to Acoustic Stimuli Near Startle Threshold: Effects of Stimulus Rise and Fall Time, Duration, and Intensity". In: *Psychophysiology* 25.5 (1988), pp. 607–611. DOI: 10.1111/j.1469-8986.1988.tb01897.x.

[150] M. M. Bradley, B. Moulder, and P. J. Lang. "When Good Things Go Bad: The Reflex Physiology of Defense". In: *Psychological Science* 16.6 (2005), pp. 468–473. DOI: 10.1111/j.0956-7976.2005.01558.x.

[151] K. Hinkelmann and O. Kempthorne. "Latin Square Type Designs". In: *Design and Analysis of Experiments: Introduction to Experimental Design*. Chichester, UK: John Wiley & Sons, Ltd, 2007, pp. 373–417. DOI: 10.1002/9780470191750.ch10.

[152] S. P. Reise, J. Ventura, K. H. Nuechterlein, and K. H. Kim. "An Illustration of Multilevel Factor Analysis". In: *Journal of Personality Assessment* 84.2 (2005), pp. 126–136. DOI: 10.1207/s15327752jpa8402_02.

[153] B. G. Tabachnick, L. S. Fidell, and J. B. Ullman. *Using Multivariate Statistics*. Seventh. Pearson, 2013.

[154] H. F. Kaiser. "The Varimax Criterion for Analytic Rotation in Factor Analysis". In: *Psychometrika* 23.3 (1958), pp. 187–200. DOI: 10.1007/BF02289233.

[155] D. L. Streiner. "Figuring Out Factors: The Use and Misuse of Factor Analysis". In: *The Canadian Journal of Psychiatry* 39.3 (1994), pp. 135–140. DOI: 10.1177/070674379403900303.

[156] A. Satorra and P. M. Bentler. "A Scaled Difference Chi-Square Test Statistic for Moment Structure Analysis". In: *Psychometrika* 66.4 (2001), pp. 507–514. DOI: 10.1007/BF02296192.

[157] P. M. Bentler. "Comparative Fit Indexes in Structural Models". In: *Psychological Bulletin* 107.2 (1990), pp. 238–246. DOI: 10.1037/0033-2909.107.2.238.

[158] H. W. Marsh, J. R. Balla, and R. P. McDonald. "Goodness-Of-Fit Indexes in Confirmatory Factor Analysis: The Effect of Sample Size". In: *Psychological Bulletin* 103.3 (1988), pp. 391–410. DOI: https://doi.org/10.1007/BF01102761.

[159] J. H. Steiger. "Structural Model Evaluation and Modification: An Interval Estimation Approach". In: *Multivariate Behavioral Research* 25.2 (1990), pp. 173–180. DOI: 10.1207/s15327906mbr2502_4.

[160] B. O. Muthén. "Multilevel Covariance Structure Analysis". In: *Sociological Methods & Research* 22.3 (1994), pp. 376–398. DOI: 10.1177/0049124194022003006.

[161] R. B. Kline. *Principles and Practice of Structural Equation Modeling*. Fifth. Guilford Publications, 2023.

[162] L. K. Muthén and B. O. Muthén. *Mplus Statistical Analysis with Latent Variables User's Guide*. Eighth. Los Angeles, CA, USA: Muthén & Muthén, 2017. URL: https://www.statmodel.com/download/usersguide/MplusUserGuideVer_8.pdf.

[163] M. L. Bolton, E. Biltekoff, J. Wei, and L. Humphrey. "On the Level of Measurement of Subjective Psychometric Ratings". In: *Proceedings of the Human Factors and Ergonomics Society 66th Annual Meeting*. Vol. 66. 1. SAGE Publications, 2022, pp. 80–84. DOI: 10.1177/10711813226612.

[164] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, 2013.

[165] W. D. Johnson and G. G. Koch. "Intraclass Correlation Coefficient". In: *International Encyclopedia of Statistical Science*. Ed. by M. Lovric. Springer Berlin Heidelberg, 2011, pp. 685–687. DOI: 10.1007/978-3-642-04898-2_309.

[166] C. P. Ryffel, C. M. Muehlethaler, S. M. Huber, and A. Elfering. "Eye Tracking As a Debriefing Tool in Upset Prevention and Recovery Training (UPRT) for General Aviation Pilots". In: *Ergonomics* 62.2 (2019), pp. 319–329. DOI: 10.1080/00140139.2018.1501093.

[167] J. Chen, A. Landman, A. Derumigny, O. Stroosma, M. M. van Paassen, and M. Mulder. "Preliminary Multilevel Confirmatory Factor Analysis of the Startle and Surprise Inventories in the Flightdeck Context". Manuscript accepted by Journal of Cognitive Engineering and Decision Making. 2025.

[168] J. Chen, A. Landman, A. Derumigny, O. Stroosma, M. M. van Paassen, and M. Mulder. "Development and Validation of the Startle and Surprise Inventories and Visual Analogue Scales". In: *Ergonomics* (2025), pp. 1–14. DOI: 10.1080/00140139.2025.2529317.

[169] C. O. Ladd, P. M. Plotsky, and M. Davis. "Startle Response". In: *Encyclopedia of Stress*. Ed. by G. Fink. Vol. 3. Academic Press, 2000.

[170] B. Byrne. *Structural Equation Modeling With EQS: Basic Concepts, Applications, and Programming*. Second. Routledge, 2006.

[171] L.-t. Hu and P. M. Bentler. “Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives”. In: *Structural Equation Modeling: A Multidisciplinary Journal* 6.1 (1999), pp. 1–55. DOI: 10.1080/10705519909540118.

[172] R. Heck and S. L. Thomas. *An Introduction to Multilevel Modeling Techniques: MLM and SEM Approaches*. Routledge, 2020.

[173] J. Candlish, M. D. Teare, M. Dimairo, L. Flight, L. Mandefield, and S. J. Walters. “Appropriate Statistical Methods for Analysing Partially Nested Randomised Controlled Trials with Continuous Outcomes: A Simulation Study”. In: *BMC Medical Research Methodology* 18.105 (2018).

[174] T. A. Brown. *Confirmatory Factor Analysis for Applied Research*. Guilford Publications, 2015.

[175] O. Stroosma, M. M. van Paassen, and M. Mulder. “Using the SIMONA Research Simulator for Human-Machine Interaction Research”. In: *AIAA Modeling and Simulation Technologies Conference and Exhibit*. AIAA 2003-5525. Austin, USA: American Institute of Aeronautics and Astronautics, 2003.

[176] P. D. Mehta and M. C. Neale. “People Are Variables Too: Multilevel Structural Equations Modeling”. In: *Psychological Methods* 10.3 (2005), pp. 259–284.

[177] R. P. McDonald. *Test Theory: A Unified Treatment*. Psychology Press, 2013.

[178] J. C. Nunnally. *Psychometric Theory 3E*. Tata McGraw-Hill Education, 1994.

[179] A. W. Meade, E. C. Johnson, and P. W. Braddy. “Power and Sensitivity of Alternative Fit Indices in Tests of Measurement Invariance”. In: *Journal of Applied Psychology* 93.3 (2008), pp. 568–592.

[180] J. Chen, A. Landman, A. Derumigny, O. Stroosma, M. M. van Paassen, and M. Mulder. “Relationships between Pilots’ Startle and Surprise Responses and Information-Processing Performance during Simulated In-Flight Events”. Manuscript submitted for publication. 2025.

[181] C. D. Wickens. “Multiple Resources and Mental Workload”. In: *Human Factors* 50.3 (2008), pp. 449–455. DOI: 10.1518/001872008X288394.

[182] N. B. Sarter and D. D. Woods. “Team Play with a Powerful and Independent Agent: Operational Experiences and Automation Surprises on the Airbus A-320”. In: *Human Factors* 39 (4 1997), pp. 553–569. DOI: 10.1518/001872097778667997.

[183] R. Parasuraman, T. B. Sheridan, and C. D. Wickens. “A Model for Types and Levels of Human Interaction with Automation”. In: *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 30.3 (2000), pp. 286–297. DOI: 10.1109/3468.844354.

[184] M. S. Young and N. A. Stanton. “Malleable Attentional Resources Theory: A New Explanation for the Effects of Mental Underload on Performance”. In: *Human Factors* 44.3 (2002), pp. 365–375. DOI: 10.1518/0018720024497709.

[185] J. Valls-Solé, H. Kumru, and M. Kofler. “Interaction between Startle and Voluntary Reactions in Humans”. In: *Experimental Brain Research* 187.4 (2008), pp. 497–507. DOI: 10.1007/s00221-008-1402-0.

[186] M. Sehlström, J. K. Ljungberg, A.-S. Claeson, and M. B. T. Nyström. "The Relation of Neuroticism to Physiological and Behavioral Stress Responses Induced by Auditory Startle". In: *Brain and Behavior* 12.e2554 (2022). DOI: 10.1002/brb3.2554.

[187] J. LeDoux. "Rethinking the Emotional Brain". In: *Neuron* 73.4 (2012), pp. 653–676. DOI: 10.1016/j.neuron.2012.02.004.

[188] D. E. Bradford, J. T. Kaye, and J. J. Curtin. "Not Just Noise: Individual Differences in General Startle Reactivity Predict Startle Response to Uncertain and Certain Threat". In: *Psychophysiology* 51.5 (2014), pp. 407–411. DOI: 10.1111/psyp.12193.

[189] F. Schwartz, J. Deniel, and M. Causse. "Effects of Startle on Cognitive Performance and Physiological Activity Revealed by Fnirs and Thermal Imaging". In: *Scientific Reports* 15.6878 (2025). DOI: 10.1038/s41598-025-90540-z.

[190] J. R. Wessel and A. R. Aron. "On the Globality of Motor Suppression: Unexpected Events and Their Influence on Behavior and Cognition". In: *Neuron* 93.2 (2017), pp. 259–280. DOI: 10.1016/j.neuron.2016.12.013.

[191] C. D. Wickens. "Multiple Resources and Performance Prediction". In: *Theoretical Issues in Ergonomics Science* 3.2 (2002), pp. 159–177. DOI: 10.1080/14639220210123806.

[192] S. W. Raudenbush and A. S. Bryk. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Second. SAGE Publications, 2002.

[193] R. Parasuraman and D. Caggiano. "Neural and Genetic Assays of Mental Workload". In: *Quantifying Human Information Processing*. Ed. by D. McBride and D. Schmorrow. Rowman & Littlefield Publishers, Inc., 2005, pp. 123–155.

[194] E. Salas, K. A. Wilson, C. S. Burke, and D. C. Wightman. "Does Crew Resource Management Training Work? An Update, an Extension, and Some Critical Needs". In: *Human Factors* 48.2 (2006), pp. 392–412. DOI: 10.1518/001872006777724444.

[195] R. L. Helmreich, A. C. Merritt, and J. A. Wilhelm. "The Evolution of Crew Resource Management Training in Commercial Aviation". In: *Human Error in Aviation*. Routledge, 2017, pp. 275–288.

[196] D. Vlaskamp, A. Landman, J. van Rooij, W.-C. Li, and J. Blundell. "Airline Pilots' Perceived Operational Benefit of a Startle and Surprise Management Method: A Qualitative Study". In: *Proceedings of the 2nd International Conference on Cognitive Aircraft Systems*. SciTePress, 2024, pp. 29–34. DOI: 10.5220/0012927800004562.

[197] J. E. Driskell and J. H. Johnston. "Stress Exposure Training". In: *Making Decisions under Stress: Implications for Individual and Team Training*. Ed. by J. A. Cannon-Bowers and E. Salas. American Psychological Association, 1998, pp. 191–217. DOI: 10.1037/10278-007.

[198] M. Causse, F. Dehais, and J. Pastor. "Executive Functions and Pilot Characteristics Predict Flight Simulator Performance in General Aviation Pilots". In: *The International Journal of Aviation Psychology* 21.3 (2011), pp. 217–234. DOI: 10.1080/10508414.2011.582441.

[199] L. Leuchs, M. Schneider, and V. I. Spoormaker. “Measuring the conditioned response: A comparison of pupillometry, skin conductance, and startle electromyography”. In: *Psychophysiology* 56.1 (2019), e13283.

[200] B. A. Sari, E. H. W. Koster, G. Pourtois, and N. Derakshan. “Training Working Memory to Improve Attentional Control in Anxiety: A Proof-of-Principle Study Using Behavioral and Electrophysiological Measures”. In: *Biological Psychology* 121 (2016), pp. 203–212.

[201] L. J. Cronbach. “Coefficient Alpha and the Internal Structure of Tests”. In: *Psychometrika* 16.3 (1951), pp. 297–334. DOI: 10.1007/BF02310555.

Acknowledgements

Completing this dissertation marks the culmination of a challenging yet deeply rewarding journey that has shaped me as a researcher and an individual. Over the past four years, I have been fortunate to work in a supportive and inspiring research environment, together with dedicated supervisors, caring colleagues and friends who made this work possible. This journey has been filled with moments of discovery, challenges, and personal growth, and I am deeply grateful to everyone who have contributed to it in any way.

I would like to express my heartfelt gratitude to my promotors, Max and René. Max, thank you for this amazing opportunity to work on such a fascinating and meaningful topic at TU Delft. Even though your schedule was "unbelievable", you always made time to read my work, answer my questions, and offer constructive feedback. I am especially grateful for your support throughout the review process, during which I learned a great deal from your way of thinking and writing. And thank you as well for your travel tips, although the wind in Texel and the heat in Sicily made those trips quite a challenge at times, they still turned out to be some of my most memorable experiences in Europe. René, thank you for your guidance and support along this PhD journey. Your feedback and our discussions not only shaped this dissertation but also helped me grow as a researcher. It has truly been a privilege to learn from you.

I would also like to thank Annemarie and Olaf, whose guidance, expertise and presence have meant so much to me. Annemarie, I am deeply grateful for the incredible amount of time and care you devoted to supporting me. You always answered my questions with patience and reviewed every manuscript thoroughly. Your encouragement gave me confidence, not only in my work, but also in myself. I will always remember the time we shared at the EAAP conference in Greece, from presenting my work with your support to the unforgettable evening we danced together. Olaf, thank you for joining me on the SIMONA experiment journey. I truly appreciated the times we tested scenarios together, and I'm especially grateful for the opportunity to experience what it feels like to be a captain, if only briefly.

This journey would not have been possible without the contributions of many others. Alexis, thank you for sharing your statistical expertise and supporting this project throughout. Dirk and Louis, thank you for your DUECA expertise and for making it possible to run the high-fidelity scenarios in SIMONA. Xander, Ferdinand, Andries, and Harold, thank you for your assistance during the SIMONA sessions. And to all participants, thank you for your time and commitment, as this research could not have been completed without you.

My appreciation goes out to all my C&S colleagues and friends. Clark, Daan, and Junzi, Xuerui, thank you for your valuable advice and support, and the many moments of laughter that made this journey so much more enjoyable. Rowenna and Gijs, thank you for your time and effort in giving Dutch lessons (even if I may not have been the best student) and for generously sharing so much about Dutch culture along the way. My appreciation goes to all my fellow PhD colleagues and friends, Aidana, Bo, Chaoxiang, Cheng, Daan, Dequan, Esther, Fazlur, Giulia, Hang, Isabelle, Jan, Jasper, Kexin, Liming, Maria, Matthijs, Moji,

Prashant, Renzhi, Shenqi, Shushuai, Simon, Sravrow, Tinghua, Tiago, Till, Tom, Wenying, Yifei, Yilun, Yingfu, Yiyuan, and Ziqing. Thank you all for the many conversations, encouragement, companionship, and shared laughter, which brightened long days and made even the cold Dutch winters feel much warmer.

And finally, thank you Zixiong, for your love, patience, and encouragement. Though there were many miles between Karlsruhe and Delft, your support never failed to bring me strength, calm, and motivation. To my mum, thank you for always believing in me and for being there whenever I needed you. I could not have come this far without you.

Jiayu Chen - October 2025

Curriculum vitae

Jiayu Chen

14-07-1996 Born in Xi'an, China.

Education

2021–2025 **PhD in Aerospace Engineering**
Delft University of Technology, Delft, the Netherlands
Dissertation: Quantifying Startle and Surprise: Development of Measuring Instruments and Validation in an Aviation Context
Promotors: prof. dr. ir. M. Mulder, dr. ir. M. M. van Paassen
Copromotor: dr. H. M. Landman

2018–2021 **Master in Aeronautical Engineering**
Northwestern Polytechnical University, Xi'an, China
Thesis: Pilot's Cognition Model in MAV/UAVs Cooperative Engagement Environment

2014–2018 **Bachelor in Flight Vehicle Environment and Life-Supporting Engineering**
Northwestern Polytechnical University, Xi'an, China

Awards

2024 The Best Paper Award of the European Association for Aviation Psychology (EAAP) 35th Conference

2021 Chinese Scholarship Council Scholarship

2021 Distinguished Master Student

List of publications

Journal publications

4. **J. Chen**, A. Landman, O. Stroosma, M. M. van Paassen, and M. Mulder. "The Effect of Personality Traits and Flight Experience on Pilots' Cognitive and Affective Responses to Simulated In-flight Hazards". In: *Aviation Psychology and Applied Human Factors* 14.2 (2024), pp. 104-113.

3. **J. Chen**, A. Landman, A. Derumigny, O. Stroosma, M. M. van Paassen, and M. Mulder. "Development and Validation of the Startle and Surprise Inventories and Visual Analogue Scales". In: *Ergonomics* (2025), pp. 1-14.

2. **J. Chen**, A. Landman, A. Derumigny, O. Stroosma, M. M. van Paassen, and M. Mulder. "Preliminary Multilevel Confirmatory Factor Analysis of the Startle and Surprise Inventories in the Flightdeck Context". Manuscript accepted by *Journal of Cognitive Engineering and Decision Making* (2025).

1. **J. Chen**, A. Landman, A. Derumigny, O. Stroosma, M. M. van Paassen, and M. Mulder. "Relationships between Pilots' Startle and Surprise Responses and Information-Processing Performance during Simulated In-Flight Events". Manuscript submitted for publication (2025).

Conference publications

2. H. Xue, **J. Chen**, Y. Yuan, and X. Zhang. " Ergonomics Evaluation of Cabin Human–Machine Interface Based on SHEL Model". In: *Man–Machine–Environment System Engineering: Proceedings of the 19th International Conference on MMESE*. Zhengzhou, China, 2020.

1. H. Xue, **J. Chen**, and X. Zhang. "A Task Simulation and Ergonomics Analysis Method Based on JACK". In: *HCI International 2020–Late Breaking Papers: Digital Human Modeling and Ergonomics, Mobility and Intelligent Environments: 22nd HCI International Conference*. Copenhagen, Denmark, 2020.

Others

1. **J. Chen**, A. Landman, O. Stroosma, M. M. van Paassen, and M. Mulder, *Manual for the Startle and Surprise Inventories and Visual Analogue Scales*. Delft, the Netherlands, 2025.

Unexpected in-flight events can trigger startle and surprise, which could impair pilots' performance but remain difficult to measure. This dissertation addresses this gap by developing and validating self-report instruments to quantify startle and surprise in an aviation context.

Grounded in cognitive models, real-world incident analyses, and robust psychometric methods, the Startle and Surprise Inventories (Startle-I; Surprise-I) and Visual Analogue Scales (Startle-VAS; Surprise-VAS) are introduced and evaluated. Results from multi-phase studies involving field experts and professional pilots, provide strong evidence of validity and reliability.

The findings offer a scientifically validated framework for assessing pilots' responses to unexpected events, with broad implications for human factors research, evidence-based training, and safety-critical operations.

ISBN 978-94-6518-166-0