Regular Paper

Journal of Cognitive Engineering and Decision Making
2025, Vol. 0(0) 1–17

DOI: 10.1177/15553434251401784
journals.sagepub.com/home/edm


Preliminary Multilevel Confirmatory Factor Analysis of the Startle and Surprise Inventories using Simulated Flight Scenarios

Jiayu Chen[1], Annemarie Landman[1,2], Alexis Derumigny[1], Olaf Stroosma[1], M. M. (René) van Paassen[1], and Max Mulder[1]

Abstract

This study was designed to validate the factor structure of the Startle and Surprise Inventories using multilevel confirmatory factor analysis in an ecologically valid flightdeck setting. The Startle and Surprise Inventories were developed to assess self-report startle and surprise to target stimuli. As their use expands in operational settings, construct validity should be further examined in contexts with ecological validity. 208 observations were collected from 26 professional pilots exposed to eight scenarios with varied levels of startle and surprise in a motion-based simulator. After each scenario, pilots completed the Startle and Surprise Inventories. A two-factor model, comprising the constructs Startle and Surprise, demonstrated superior and acceptable fit over a one-factor model. All items demonstrated significant factor loadings at both within- and between-scenario levels in the two-factor solution. McDonald's ω ranged from $\omega = 0.88$ to $\omega = 0.96$ for the Startle Inventory, and $\omega = 0.77$ to $\omega = 0.96$ for the Surprise Inventory, indicating acceptable to excellent internal consistency. The findings offer empirical support for the construct validity and reliability of the Startle and Surprise Inventories in a highly ecologically-valid setting. The validated and reliable measures can inform evidence-based safety training protocols and interventions in aviation and other safety-critical domains.

Keywords

aviation, cognition, evidence-based training, psychometrics measures, pilot performance

Introduction

Aviation operations are inherently demanding, requiring pilots to judge situations correctly and apply appropriate procedures to ensure safety and efficiency. When exposed to sudden and unexpected events, pilots sometimes need to make split-second decisions. In such emergency situations, pilots may experience startle, surprise, or both. These responses could impair pilot performance (Landman et al., 2017b; Martin et al., 2015), while the resulting stress may further exacerbate their effects by degrading teamwork coordination, narrowing attentional, prompting rushed and

[1]Delft University of Technology, The Netherlands
[2]Netherlands Organization for Applied Scientific Research (TNO), The Netherlands

Corresponding Author:
Jiayu Chen, Department of Control & Operations, Delft University of Technology, Kluyverweg 1, Delft 2629 HS, The Netherlands.
Email: jiayu.chen.jade@outlook.com

unsystematic decision-making, and leading to task shedding (Dismukes et al., 2015; Salas et al., 2013).

Startle is a response to abrupt, intense (auditory, somatosensory, vestibular or visual) stimuli perceived as potentially threatening, characterized by involuntary physiological startle reflexes and generalized stress responses (Blumenthal, 2015; Koch, 1999; Martin et al., 2015). The startle reflex, occurring typically within 100 milliseconds after a stimulus, may involve a range of rapid motor responses, including eyeblinks, head movements, facial grimacing, shoulder elevation, arm abduction, elbow bending, forearm pronation, finger flexion, and abdominal contraction (Ladd et al., 2000; Leuchs et al., 2019). Following the reflex, generalized stress responses could be triggered, characterized by increased skin conductance, elevated systolic blood pressure, accelerated heart rate and pupil dilation (Dreissen et al., 2012; Holand et al., 1999). While the initially startle could impair performance due to transient physiological disruptions (typically lasting 1 to 3 s), individuals often recover rapidly. The overall impact on performance varies depending on task complexity and the contextual factors (Duchevet et al., 2025; Landman et al., 2017b; May & Rice, 1971; Staal, 2004; Thackray, 1965; Thackray & Touchstone, 1983).

Surprise is a cognitive and affective response, primarily evoked by unexpected (schema-discrepant) stimuli or events (Horstmann, 2006; Landman et al., 2017a; Meyer et al., 1997). It consists of responses to the unexpectedness of events, and subsequent sensemaking process of the schema discrepancy (Noordewier et al., 2016). Unexpectedness causes an automatic interruption of ongoing mental processes and signals a failure to anticipate future events, which can be disruptive and distressing due to the innate need for predictability. The occurrence of surprise is thought to alert individuals of the discrepancy, and motivate efforts to deeper learning or schema revision, leading to improved cognitive flexibility and adaptability (Reisenzein et al., 2019). However, if the discrepancy remains unresolved, the situation may be perceived as poorly understood, hindering the ability to focus on relevant information, make accurate projections, and execute appropriate actions.

To accurately investigate the causes and consequences of startle and surprise, it is essential to quantify these responses with validated and reliable measures. Such quantification allows for a deeper understanding of their physiological correlates, recovery dynamics, and task vulnerabilities to unexpected stimuli. In aviation, this approach enables the identification of factors that impair pilot performance during unforeseen emergencies, thereby informing the development of evidence-based strategies to mitigate adverse effects and enhance operational effectiveness in high-stakes environments.

The previous research detailed the development and preliminary validation 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), which were designed to assess individuals' startle and surprise responses to specific stimuli/events (Chen et al., 2025a). There are six statements in the Startle-I and five statements in the Surprise-I (Chen et al., 2025). Individuals indicate to what extent they agree with the statements on 5-point Likert scales (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 validation of the self-report measures started with content validation, during which seven experts in the fields of Cognitive Science and Psychology assessed the relevance of 21 items, which were derived from fundamental and applied literature on startle (Blumenthal, 1988; Bradley et al., 2005; Koch, 1999; Ladd et al., 2000; Lang et al., 1990; Martin et al., 2015) and surprise (Izard et al., 1993; Klein et al., 2007; Landman et al., 2017a; Meyer et al., 1997; Rivera et al., 2014). Based on this assessment, 19 items was retained for the multilevel exploratory factor analysis (ML-EFA) to examine the construct validity. To capture variation in responses, a group of 81 participants were exposed to nine video stimuli, designed to

elicit varying levels of startle and surprise. The ML-EFA resulted in an 11-item two-factor structure.

Concurrent validity of the Startle-VAS and Surprise-VAS was supported by significant correlations between the Startle-VAS and Startle-I scores, and between the Surprise-VAS and Surprise-I scores, respectively. These correlations ranged from $\rho = 0.778$ to $\rho = 0.877$ for the Startle-VAS, and from $\rho = 0.681$ to $\rho = 0.903$ for the 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 (Byrne, 2006). In contrast, CFA adopts a hypothesis-driven approach to test hypothesized factor structures while assessing the model fit to the data. The analysis incorporates various goodness-of-fit indices, including the significance of χ^2 test (Satorra & Bentler, 2001), Comparative Fit Index (CFI; Bentler, 1990), Tucker–Lewis Index (TLI; Marsh et al., 1988), Root Mean Square Error of Approximation (RMSEA; Steiger, 1990), and Standardized Root Mean Square Residual (SRMR; Hu & Bentler, 1999). Construct validity can be 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 (Heck & Thomas, 2020; Kline, 2023; Muthén, 1994). 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 attributable to between-cluster variance (Reise et al., 2005). In intervention studies with nested designs, even an ICC as low as 0.05 can substantially impact statistical power (Candlish et al., 2018). By accounting for the clustered nature of data, MCFA provides more precise and reliable estimates of factor loadings and variances, minimizing biases introduced by data non-independence (Brown, 2015). 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 (Muthén, 1994).

In this study, we aim to systematically validate the factor structure of the Startle and Surprise Inventories in a highly ecologically-valid aviation setting through MCFA. Ecological validity is operationalized through the use of aviation manual control tasks, the inclusion of action-relevant in-flight startling and/or surprising events, and the simulation of performance consequences, such as loss of control. Based on the established two-factor model identified from the ML-EFA (Chen et al., 2025a), we hypothesize that the model demonstrates a good fit both at the within- and between-cluster levels.

A representative sample group of professional pilots was recruited, and simulated in-flight scenarios were used to elicit startle and surprise responses. Events in most of the scenarios were found to be effective to elicit startle and surprise responses in pilots (Chen et al., 2024), 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.

Method

Participants

26 currently employed professional pilots participated in the experiment, comprising 25 males and 1 female. The characteristics of the participants are listed in Table 1. All participants possessed either an Airline Transport Pilot License (ATPL) or a

frozen ATPL during 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 research 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.

Table 1. Characteristics of the Participants (N = 26).

	Mean	*SD*
Age (yrs)	43.77	12.95
Employed time (yrs)	17.73	13.26
Flight hours (large aircraft)	6566.12	6607.78
Flight hours (business jet)	1257.69	2803.71
Flight hours (small aircraft)	810.08	1258.38

Apparatus

The experiment was conducted in the SIMONA Research Simulator (Figure 1(a)) at the Delft University of Technology (Stroosma et al., 2003). This is a full motion simulator with a six degrees of freedom hydraulic hexapod motion system. The simulator has a collimated 180° horizontal by 40° 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 to establish a highly credible flightdeck operational environment. During the experiment, participants wore single-ear intercom headsets (ClearCom CC-110-X4).

A generic model of the Piper PA-34 Seneca III, a light multi-engine piston (MEP) aircraft, served as the aircraft model throughout the experiment. The flight controls (Figure 1(b)) included 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.

Tasks and Conditions

Experiment Procedure. A within-subjects experimental design was used, consisting of eight test scenarios with varied startling and surprising events presented to all 26 participants. The experiment procedure included briefing, familiarization, test session, and debriefing, and lasted approximately 2 hours per participant.

During the briefing, pilots received instructions on the simulator features, aerodynamic model, concepts of startle and surprise, and the "standard traffic pattern" (Figure 2). Startle was explained to participants as "a rapid, involuntary reaction to an abrupt and intense stimulus, typically perceived as a threat", and surprise was described as "a cognitive-affective response evoked by unexpected stimulus or events" with practical examples to clarify their mechanisms and implications.

This circuit was left-handed, started and ended on runway 18C, Schiphol Airport (EHAM). It was to be flown at 1000 ft with a speed of 115 kt. The flaps setting of 0° (UP) was required during take-off, 25° in base leg and 40° (LAND) in final leg. Rotate speed was 80 kt, minimum control speed was 80 kt, best climb speed was 92 kt, and landing approach speed was 90 kt. Pilots had these configurations also available on a kneepad. Pilots practised flying the circuit without crosswind in the familiarization session. By the end of the familiarization, all pilots confirmed that they could handle the aircraft model satisfactorily as single-pilot crew, with none requiring assistance in identifying the turn points of the circuit.

Pilots then proceeded with the test session containing eight test scenarios (Table 2), presented in a semi-counterbalanced order defined by a Latin square (Hinkelmann & Kempthorne, 2007). Before each scenario, participants were briefed on the weather condition (e.g. wind strength and direction, and weather code) through Meteorological Aerodrome Report (METAR) and the starting position. Participants started each scenario either from the take-off position on EHAM runway 18C, or in-flight position at an altitude of 800 ft in front of runway 18C, with an airspeed of 99 kt. Regardless of the starting position, participants were

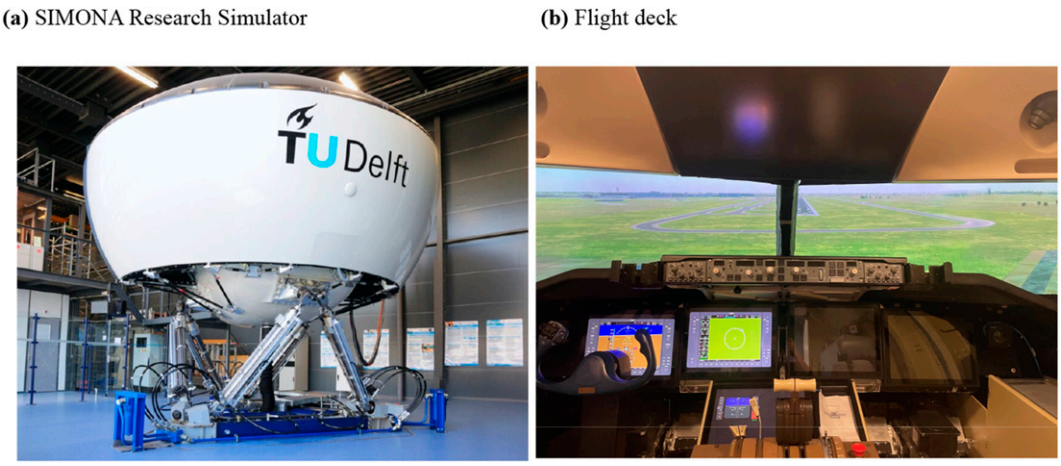

Figure 1. SIMONA Research Simulator and flight deck used in the experiment.

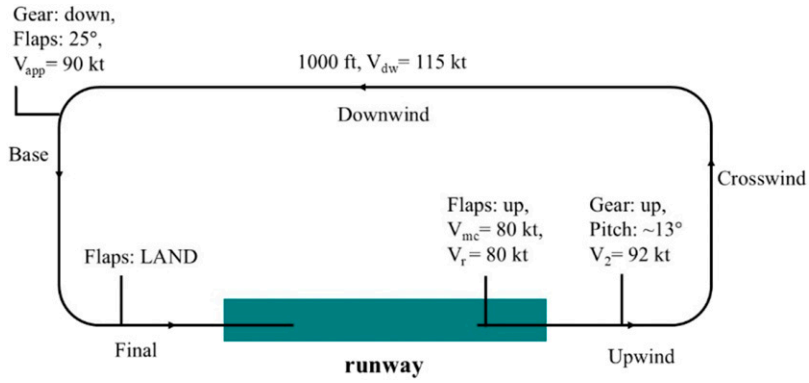

Figure 2. Standard traffic pattern used in the scenarios with required settings.

required to complete the circuit and perform a safe landing on EHAM runway 18C.

From 5 to 15 s before the startling or surprising event onset until 28 s after, pilots also performed a secondary auditory task that was used to measure information-processing performance in a different study (Chen et al., 2025b). Numbers ranging from 0 to 9, were presented through a headset at 2.5 s intervals in an auditory format, and pilots responded with a single autopilot disconnect button press for odd number and a double button press for even number. This button was located on the control column near the left thumb.

Immediately following each test scenario, pilots were instructed to complete the Startle and Surprise Inventories and Visual Analogue Scales assessing their startle and surprise levels, referring to the events that had just occurred in that scenario. After completing the eight scenarios, pilots were debriefed about the specifics of all simulated events in each scenario.

Test Scenarios. We aimed to include events that would elicit varied levels of startle and surprise. The characteristics of the scenarios are listed in Table 2, in which the third column indicates whether the preset event in each scenario was intended to induce startle, surprise, both, or neither. Startle was triggered by the sudden onset of intense auditory, visual, or vestibular stimuli. Surprise was intended to arise from situations that pilots did not expect. Scenarios designed to elicit

Table 2. Description of Preset Events and Intended Startle/Surprise Responses across Scenarios.

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

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

neither startle nor surprise did not contain unexpected nor sudden events.

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, with the sound played 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 out-the-window field of view. To mitigate surprise, the METAR for this scenario included "TSRA" (thunderstorm with rain), which signalled the possibility of thunderstorms.

The flap asymmetry (FLAP), engine failure (ENF), and primary flight display failure (PFDF) scenarios were designed to elicit surprise but no (or very limited) startle. In FLAP, participants encountered a malfunction when selecting Flaps 25° in base leg. The left flap remained in the UP position, causing an unexpected roll and yaw moment that could be counteracted using the column. In ENF, the right engine failed during take-off (5 s after reaching 900 ft), causing a roll and yaw moment that could be counteracted using the column and pedals. The PFDF started in the in-flight position. The PFD malfunctioned and went black during landing at 600 ft. Pilots could use the outside view or the backup display to continue landing.

The cargo shift (CARGO) and false stall warning (STALL) scenarios were designed to elicit both startle and surprise responses. In 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. In 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.

The normal take-off (NTO) and normal landing (NLO) scenarios featured no preset startling/surprising events during the circuit and were intended to induce low levels of startle and surprise. NTO and NLO started in the take-off and in-flight positions, respectively.

Table 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.

Measures of Startle and Surprise

Participants were instructed to indicate their experienced levels of startle and surprise, immediately after each scenario using the Startle-I, Surprise-I, Startle-VAS and Surprise-VAS (Chen et al., 2025). The six items of the Startle-I and the five items of the Surprise-I were randomized in order (Table 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 target stimulus within each scenario, as referred in the Startle and Surprise Inventories and Visual Analogue Scales, is stated in the rightmost column of Table 2.

Statistical Analysis

Two-way ANOVA and ICCs. The scores on reverse-coded items 4 and 6 were first 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 serve to evaluate whether a MCFA was necessary instead of a single-level CFA, as sufficiently high between-level variance (ICC > 0.05; Candlish et al., 2018) 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 and Surprise Inventories was analysed using MCFA for the clustered dataset (Mehta & Neale, 2005). 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 previous ML-EFA (Chen et al., 2025a), and an 11-item, one-factor model, in which all 11 items are considered as variables of a single factor.

The model goodness-of-fit was deemed acceptable if the χ^2 value was non-significant (Satorra & Bentler, 2001), CFI and TLI are greater than 0.90 (Bentler, 1990), and RMSEA (Steiger, 1990) and SRMR (both within- and between-level) are below 0.10 (Hu & Bentler, 1999). To further investigate the relationship between the factors Startle and Surprise in the two-factor model, standardized covariances between these two factors were computed at both the within- and between-level. The internal consistency was examined by obtaining McDonald's ω (McDonald, 2013) of the Startle-I and Surprise-I for each scenario, with values greater than 0.70 indicating acceptable internal consistency (Meade et al., 2008; Nunnally, 1994).

Manipulation Checks. To check whether designed scenarios induced the intended responses of startle and surprise, manipulation checks were conducted. Two linear mixed-effects models were applied to account for the repeated-measures design, with heteroscedasticity modelled in the residual structure. 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 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.

Results

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 4. Thus, data were clustered over scenarios for the MCFA.

The ICCs ranged from 0.34 to 0.74 (rightmost column in Table 4). Notably, all of the items had ICCs greater than 0.05, indicating considerable variance due to the between-scenario differences. Specifically, for nearly half of the items, more than 50% of the total variance was attributable to between-scenario differences, indicating the need for a MCFA to properly examine the factor structure.

Multilevel Confirmatory Factor Analysis

The 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 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 (Table 5), all items loaded significantly on their respective factors (i.e., absolute Z values were greater than 1.96 at the 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-scenario level, representing that pilots who tended to report higher levels of startle did not necessarily report higher levels of 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 $\omega = 0.88$ to $\omega = 0.96$ for the Startle-I, and $\omega = 0.77$ to $\omega = 0.96$ for the Surprise-I (Table 6).

Manipulation Checks

The ratings from the Startle-I, Surprise-I, Startle-VAS, and Surprise-VAS for each scenario are shown in pirate plots (Figure 3). These plots represent the mean values (square markers with labels), interquartile range in whiskers, and estimated ratings distribution. across different scenarios. 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 may suggest that the multi-item inventories were more effective than

Table 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		

the VASs in differentiating between subjective startle and surprise. 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. This high variation in responses also facilitated the application of MCFA.

Figure 4 presents the averaged levels of startle and surprise across scenarios. The dashed circles indicate groups of scenarios that do not have significantly different levels of startle and surprise as analysed using the linear mixed-effects models (see Table 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 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.

Most scenarios elicited the intended responses on self-report startle and surprise. However, ENF and LTS were found to have no significant difference between both their startle ratings, $ENF_{startle} = 2.76$, $LTS_{startle} = 3.28$, and surprise ratings, $ENF_{surprise} = 4.01$, $LTS_{surprise} = 3.44$. No significant differences were found on startle ratings between ENF and CARGO, $ENF_{startle} = 2.76$, $CARGO_{startle} = 2.58$, ENF and STALL, $ENF_{startle} = 2.76$, $STALL_{startle} = 2.85$, FLAP and CARGO, $FLAP_{startle} = 2.63$, $CARGO_{startle} = 2.58$, FLAP and STALL, $FLAP_{startle} = 2.76$, $STALL_{startle} = 2.85$. No significant differences were found on surprise ratings between LTS and CARGO, $LTS_{surprise} = 3.44$, $CARGO_{surprise} = 4.09$, LTS and PFDF, $LTS_{surprise} = 3.44$, $PFDF_{surprise} = 3.94$.

Table 5. Factor Loadings, Standard Errors and Z Values From the MCFA of the Two-Factor Model.

		Within (scenario)-level			Between (scenario)-level		
Factor	Item	Loading[b]	SE	Z	Loading[b]	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

[b]Standardized loading.

Discussion

The construct validity of the Startle and Surprise Inventories was confirmed in a highly ecologically valid flightdeck setting, with 208 observations comprising 26 professional pilots. To investigate the inventories' ability on measuring self-report startle and surprise, eight simulated in-flight 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 previous research where a ML-EFA was performed (Chen et al., 2025a). Results from the two-way ANOVA indicated 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 yielded significant χ^2 values, which are typically indicative of poor model-data fit. However, the relatively small sample size likely inflates the χ^2 values, producing significant results even when the model-data discrepancies are minor (Byrne, 2006). In addition to the χ^2 test, other fit indices, CFI, TLI, RMSEA, $SRMR_{within}$ and $SRMR_{between}$ all 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.

Table 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

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 the startle and surprise responses of pilots in a highly ecologically valid setting. Additionally,

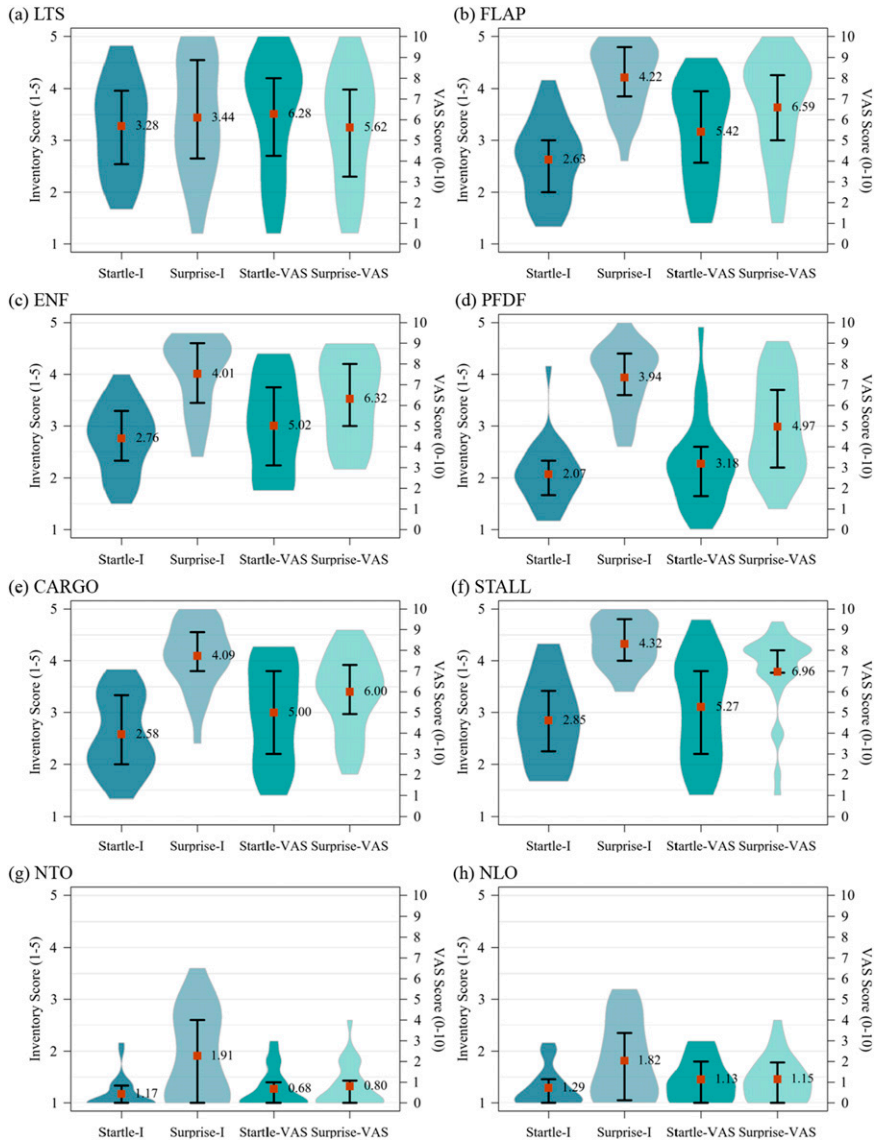

Figure 3. Ratings from Startle-I, Startle-I, Startle-VAS and Surprise-VAS across scenarios (square markers indicate means, whiskers indicate interquartile range).

the MCFA offers compelling evidence for the distinctiveness of the constructs of startle and surprise within the flightdeck operational context. The results support the hypothesis that the responses of startle and surprise are fundamentally and psychometrically distinct constructs with different causes and consequences (Landman et al., 2017a; Rivera et al., 2014), even though both could impact pilot (cognitive) performance to varying degrees.

Regarding the reliability of the Startle-I and Surprise-I, the McDonald's ω across scenarios indicated acceptable to excellent internal consistency. The variability across scenarios may be attributable to differences in scenario characteristics and individual differences in how to interpret and respond to the items. Specifically, scenarios that elicited more uniform and intense responses may have led to stronger inter-item correlations, thereby increasing McDonald's ω.

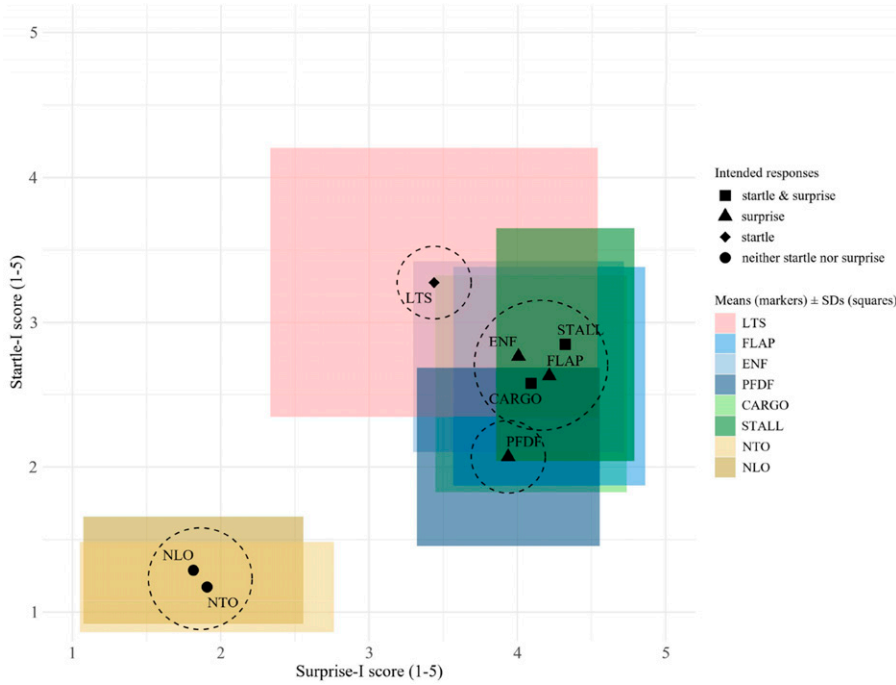

Figure 4. Mean values (markers) and SDs (squares) of Startle-I and Surprise-I for each scenario. Dashed circles denote scenarios groups with non-significant differences in mean startle and surprise levels.

From the comparison between ratings from inventories and VASs on measuring startle and surprise, the Startle-I and Surprise-I demonstrated a superior ability to distinguish between different levels of startle and surprise when compared to the Startle-VAS and Surprise-VAS (Figure 3). This was unexpected, as the single-item VAS scores were shown to be highly correlated with inventory scores in a previous research (Chen et al., 2025a), in which startling and surprising video stimuli were unrelated to the aviation domain. Thus, although the visual analogue scales are efficient in capturing self-report startle and surprise (can be used quickly and immediately after target stimuli), the multi-item inventories were found to be more effective in the ecologically valid flightdeck context.

There are several limitations should be acknowledged. First, the findings were based on the scenarios performed in a motion-based flight simulator. Although the setting 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, in which 26 participants were exposed to eight test scenarios. While the sample size was sufficient to yield meaningful insights into the model's goodness-of-fit, it limits the generalizability of the findings to broader populations. 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 more diverse demographic profile (e.g. age and flight experience) to enhance the generalizability and robustness of the findings.

Third, the results indicated that the designed stimuli effectively elicited a wide range of variability in both startle and surprise responses. However, it remains challenging to elicit startle and surprise independently, as these two responses often co-occur in high-stakes operational settings. This overlap may be evident in scenarios primarily designed to induce surprise, extra stress induced by workload to control the flight path may have heightened participants' startle responses (Martin

Table 7. Pairwise Comparison of the 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

et al., 2015). For example, in the FLAP and ENF scenarios, which were designed to evoke high levels of surprise, participants also reported that their startle levels was around the midpoint of the inventories/VASs.

Fourth, all test scenarios were conducted in a single-pilot setting with a simplified twin-propeller 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, this study provides empirical support for the construct validity of the Startle-I and Surprise-I in an ecologically valid context, extending previous findings from controlled laboratory setting (Chen et al., 2025a). The inventories showed acceptable to excellent internal consistency across various conditions. To examine their psychometric properties more comprehensively, future research could explore the criterion-related validity of the two measures by comparing subjective ratings with objective indicators, such as physiological responses (e.g. reflex electromyogram (EMG) and pupillometry; Blumenthal et al., 2005; Leuchs et al., 2019; Ryffel et al., 2019) or behavioural indicators (e.g. reaction time and micro-expressions).

Sixth, the current study is limited in focus on the aviation domain, whereas the inventories have potential for broader applicability in other high-

stakes environment involving human operators (Vlaskamp et al., 2025). Future research should aim to replicate these findings in different operational settings, as well as in different domains to enhance the generalizability.

Conclusion

This study provides strong and consistent evidence supporting the factor structure of the Startle and Surprise inventories, aligning with prior research. Using a sample of professional pilots (N = 26) and simulated in-flight scenarios, the findings support previous results obtained using video-based stimuli, further demonstrating the validity and reliability of Startle and Surprise Inventories across diverse contexts. The findings also highlight the inventories' applicability for assessing startle and surprise responses at both individual and scenario levels. Moreover, the Startle and Surprise Inventories enable further research into the antecedents and consequences of these responses, with potential implications for evidence-based safety protocols and training interventions in the flightdeck contexts.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Jiayu Chen https://orcid.org/0009-0006-0084-1661
Annemarie Landman https://orcid.org/0000-0003-3678-9210
Alexis Derumigny https://orcid.org/0000-0002-6163-8097
Olaf Stroosma https://orcid.org/0000-0002-4578-3171
M. M. (René) van Paassen https://orcid.org/0000-0003-4700-1222
Max Mulder https://orcid.org/0000-0002-0932-3979

References

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Blumenthal, T. D. (1988). The startle response to acoustic stimuli near startle threshold: Effects of stimulus rise and fall time, duration, and intensity. *Psychophysiology*, *25*(5), 607–611. https://doi.org/10.1111/j.1469-8986.1988.tb01897.x

Blumenthal, T. D. (2015). Presidential address 2014: The more-or-less interrupting effects of the startle response. *Psychophysiology*, *52*(11), 1417–1431. https://doi.org/10.1111/psyp.12506

Blumenthal, T. D., Cuthbert, B. N., Filion, D. L., Hackley, S., Lipp, O. V., & van Boxtel, A. (2005). Committee report: Guidelines for human startle eyeblink electromyographic studies. *Psychophysiology*, *42*(1), 1–15. https://doi.org/10.1111/j.1469-8986.2005.00271.x

Bradley, M. M., Moulder, B., & Lang, P. J. (2005). When good things go bad: The reflex physiology of defense. *Psychological Science*, *16*(6), 468–473. https://doi.org/10.1111/j.0956-7976.2005.01558.x

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford Publications.

Byrne, B. (2006). *Structural equation modeling with eqs: Basic concepts, applications, and programming* (2nd ed.). Routledge.

Candlish, J., Teare, M. D., Dimairo, M., Flight, L., Mandefield, L., & Walters, S. J. (2018). Appropriate statistical methods for analysing partially nested randomised controlled trials with continuous outcomes: A simulation study. *BMC Medical Research Methodology*, *18*(105), 105. https://doi.org/10.1186/s12874-018-0559-x

Chen, J., Landman, A., Derumigny, A., Stroosma, O., van Paassen, M. M., & Mulder, M. (2025a). Development and validation of the startle and Surprise inventories and visual analogue scales. *Ergonomics*, 1–14. https://doi.org/10.1080/00140139.2025.2529317

Chen, J., Landman, A., Derumigny, A., Stroosma, O., van Paassen, M. M., & Mulder, M. (2025b). Relationships between pilots' startle and Surprise responses and information-processing performance during simulated In-Flight events. (Manuscript submitted for publication).

Chen, J., Landman, A., Stroosma, O., van Paassen, M. M., & Mulder, M. (2024). The effect of personality traits and flight experience on pilots' cognitive and affective responses to simulated In-Flight hazards. *Aviation Psychology and Applied Human Factors*, *14*(2), 104–113. https://doi.org/10.1027/2192-0923/a000283

Chen, J., Landman, A., Stroosma, O., Van Paassen, M. M., & Mulder, M. (2025). *Manual for the startle and surprise inventories and visual analogue scales*. Delft University of Technology.

Dismukes, R. K., Goldsmith, T. E., & Kochan, J. A. (2015). *Effects of acute stress on aircrew performance: Literature review and analysis of operational aspects (No. NASA/TM-2015-218930)*. Washington, USA. Retrieved from. https://ntrs.nasa.gov/api/citations/20190002685/downloads/20190002685.pdf

Dreissen, Y. E. M., Bakker, M. J., Koelman, J. H. T. M., & Tijssen, M. A. J. (2012). Exaggerated startle reactions. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, *123*(1), 34–44. https://doi.org/10.1016/j.clinph.2011.09.022

Duchevet, A., Imbert, J.-P., Garcia, J., Lamirault, B., & Causse, M. (2025). Investigating the independent and combined effects of startle and Surprise in a simulated flight task. *Human Factors*, *67*(11), 1170–1187. https://doi.org/10.1177/00187208251342100

Heck, R., & Thomas, S. L. (2020). *An introduction to multilevel modeling techniques: MLM and SEM approaches*. Routledge.

Hinkelmann, K., & Kempthorne, O. (2007). Latin square type designs. In *Design and analysis of experiments: Introduction to experimental design* (pp. 373–417). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470191750.ch10

Holand, S., Girard, A., Laude, D., Meyer-Bisch, C., & Elghozi, J.-L. (1999). Effects of an auditory startle stimulus on blood pressure and heart rate in humans. *Journal of Hypertension*, *17*(12), 1893–1897. https://doi.org/10.1097/00004872-199917121-00018

Horstmann, G. (2006). Latency and duration of the action interruption in Surprise. *Cognition & Emotion*, *20*(2), 242–273. https://doi.org/10.1080/02699930500262878

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Izard, C. E., Libero, D. Z., Putnam, P., & Haynes, O. (1993). Stability of emotion experiences and their relations to traits of personality. *Journal of Personality and Social Psychology*, *64*(5), 847–860. https://doi.org/10.1037//0022-3514.64.5.847

Klein, G., Phillips, J. K., Rall, E. L., & Peluso, D. A. (2007). A data-frame theory of sensemaking. In R. R. Hoffman (Ed.), *Expertise out of context: Proceedings of the sixth international conference on naturalistic decision making*. Lawrence Erlbaum Associates Publishers.

Kline, R. B. (2023). *Principles and practice of structural equation modeling* (5th ed.). Guilford Publications.

Koch, M. (1999). The neurobiology of startle. *Progress in Neurobiology*, *59*(2), 107–128. https://doi.org/10.1016/S0301-0082(98)00098-7

Ladd, C. O., Plotsky, P. M., & Davis, M. (2000). Startle response. In G. Fink (Ed.), *Encyclopedia of stress* (Vol. *3*). Academic Press.

Landman, A., Groen, E. L., van Paassen, M. M., Bronkhorst, A. W., & Mulder, M. (2017a). Dealing with unexpected events on the flight deck: A conceptual model of startle and surprise. *Human Factors*, *59*(8), 1161–1172. https://doi.org/10.1177/0018720817723428

Landman, A., Groen, E. L., van Paassen, M. M., Bronkhorst, A. W., & Mulder, M. (2017b). The influence of Surprise on upset recovery performance in airline pilots. *The International Journal of Aerospace Psychology*, *27*(1–2), 2–14. https://doi.org/10.1080/10508414.2017.1365610

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1990). Emotion, attention, and the startle reflex. *Psychological Review*, *97*(3), 377–395.

Leuchs, L., Schneider, M., & Spoormaker, V. I. (2019). Measuring the conditioned response: A comparison of pupillometry, skin conductance, and startle electromyography. *Psychophysiology*, *56*(1), e13283. https://doi.org/10.1111/psyp.13283

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, *103*(3), 391–410. https://doi.org/10.1007/BF01102761

Martin, W. L., Murray, P. S., Bates, P. R., & Lee, P. S. Y. (2015). Fear-potentiated startle: A review from an aviation perspective. *The International Journal of*

Aviation Psychology, *25*(2), 97–107. https://doi.org/10.1080/10508414.2015.1128293

May, D. N., & Rice, C. G. (1971). Effects of startle due to pistol shots on control precision performance. *Journal of Sound and Vibration*, *15*(2), 197–202. https://doi.org/10.1016/0022-460x(71)90534-7

McDonald, R. P. (2013). *Test theory: A unified treatment*. Psychology Press.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*(3), 568–592. https://doi.org/10.1037/0021-9010.93.3.568

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, *10*(3), 259–284. https://doi.org/10.1037/1082-989X.10.3.259

Meyer, W.-U., Reisenzein, R., & Schützwohl, A. (1997). Toward a process analysis of emotions: The case of Surprise. *Motivation and Emotion*, *21*(3), 251–274. https://doi.org/10.1023/A:1024422330338

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*(3), 376–398. https://doi.org/10.1177/0049124194022003006

Noordewier, M. K., Topolinski, S., & van Dijk, E. (2016). The temporal dynamics of surprise. *Social and Personality Psychology Compass*, *10*(3), 136–149. https://doi.org/10.1111/spc3.12242

Nunnally, J. C. (1994). *Psychometric theory 3e*. Tata McGraw-Hill Education.

Reise, S. P., Ventura, J., Nuechterlein, K. H., & Kim, K. H. (2005). An illustration of multilevel factor analysis. *Journal of Personality Assessment*, *84*(2), 126–136. https://doi.org/10.1207/s15327752jpa8402_02

Reisenzein, R., Horstmann, G., & Schützwohl, A. (2019). The cognitive-evolutionary model of Surprise: A review of the evidence. *Topics in Cognitive Science*, *11*(1), 50–74. https://doi.org/10.1111/tops.12292

Rivera, J., Talone, A. B., Boesser, C. T., Jentsch, F., & Yeh, M. (2014). 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*, pp. 1047–1051). Sage Publications. https://doi.org/10.1177/15419312145812

Ryffel, C. P., Muehlethaler, C. M., Huber, S. M., & Elfering, A. (2019). Eye tracking as a debriefing tool in upset prevention and recovery training (UPRT) for general aviation pilots. *Ergonomics*, *62*(2), 319–329. https://doi.org/10.1080/00140139.2018.1501093

Salas, E., Driskell, J. E., & Hughes, S. (2013). Introduction: The study of stress and human performance. In J. E. Driskell & E. Salas (Eds.), *Stress and human performance* (pp. 1–45). Psychology Press. https://doi.org/10.4324/9780203772904

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*(4), 507–514. https://doi.org/10.1007/BF02296192

Staal, M. A. (2004). *Stress, cognition, and human performance: A literature review and conceptual framework (no. NASA/TM-2004-212824)*. Washington, USA. Retrieved from. https://ntrs.nasa.gov/api/citations/20060017835/downloads/20060017835.pdf

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*(2), 173–180. https://doi.org/10.1207/s15327906mbr25024

Stroosma, O., van Paassen, M. M., & Mulder, M. (2003). Using the SIMONA research simulator for human-machine interaction research. In *AIAA modeling and simulation technologies conference and exhibit*. Austin, USA: American Institute of Aeronautics and Astronautics.

Thackray, R. I. (1965). Correlates of reaction time to startle. *Human Factors*, *7*(1), 74–80. https://doi.org/10.1177/001872086500700109

Thackray, R. I., & Touchstone, R. M. (1983). *Rate of initial recovery and subsequent radar monitoring performance following a simulated emergency involving startle (No. FAA-AM-83-13)*. Washington, USA. Retrieved from. https://www.faa.gov/sites/faa.gov/files/dataresearch/research/medhumanfacs/oamtechreports/AM83-13.pdf

Vlaskamp, D., Pollitt, A., Blundell, J., Landman, A., Groen, E. L., van Paassen, M. M., Stroosma, O., & Mulder, M. (2025). Startle and Surprise in helicopter operations: Reported prevalence and application of mitigation strategies. *Cognition, Technology & Work*, *27*(3), 579–590. https://doi.org/10.1007/s10111-025-00811-y

Jiayu Chen earned her MSc in Aerospace Engineering from the Northwestern Polytechnical University in 2021. She is currently a PhD candidate in the section Control and Simulation, Delft University of Technology.

Annemarie Landman received her PhD in Aerospace Engineering at Delft University of Technology in 2019. She is currently working as a

scientist in the Training and Performance Innovations department at TNO Human Factors, and as assistant professor at the section Control and Simulation, Delft University of Technology.

Alexis Derumigny received his PhD in Applied Mathematics from the National School of Statistics and Economic Administration (ENSAE Paris) in 2019. He is currently assistant professor in the Department of Applied Mathematics, Delft University of Technology.

Olaf Stroosma earned his MSc in Aerospace Engineering from Delft University of Technology in 1998. He is currently a senior researcher at the section Control and Simulation, Delft University of Technology, where he manages the SIMONA Research Simulator facility.

M. M. (René) van Paassen is associate professor in the section Control and Simulation, Delft University of Technology, where he received his PhD in Aerospace Engineering in 1994.

Max Mulder is a full professor in the section Control and Simulation, Delft University of Technology, where he received his PhD (cum laude) in Aerospace Engineering in 1999.